McGill University, July 21-22, 1986

Proceedings of:

Montreal Symposium on Speech Recognition
Symposium sur la Reconnaissance de la Parole

PROCEEDINGS

MONTREAL SYMPOSIUM ON SPEECH RECOGNITION
SYMPOSIUM SUR LA RECONNAISSANCE DE LA PAROLE

July 21-22, 1986

McGill University, Montreal

UNITS AND THEIR REPRESENTATION
IN SPEECH RECOGNITION

UNITES ET LEUR REPRESENTATION POUR LA
RECONNAISSANCE DE LA PAROLE

Program Committee : P. Mermelstein, M. Lennig, D. O'Shaughnessy

Secretariat : Bell-Northern Research, 3 Place du Commerce, Verdun, Quebec H3E 1H6, Canada

SPONSORS

CANADIAN ACOUSTICAL ASSOCIATION

BELL-NORTHERN RESEARCH

BELL CANADA

INRS-TELECOMMUNICATIONS (UNIV. DU QUEBEC)

McGILL UNIVERSITY

4. SYLLABLE, DEMISYLLABLE, AND DIPHONE MODELS

Monday Afternoon II: 4:00–5:45 pm
Chairman: M. Lennig

Contributed Papers

5. PHONETIC MODELS

Tuesday Morning I: 8:30–10:00 am
Chairman: P. Mermelstein

Contributed Papers

6. ACOUSTIC-PHONETIC ANALYSIS

Tuesday Morning II: 10:15 am–12:00 pm
Chairman: D. O'Shaughnessy

Contributed Papers

7. WORD RECOGNITION SYSTEMS

Tuesday Afternoon I: 1:30–2:30 pm
Chairman: T. Crystal

Contributed Papers

8. ACOUSTIC-PHONETIC DECODING

Tuesday Afternoon II: 2:45–4:15 pm
Chairman: M.A. Bush

Contributed Papers

PERIPHERAL PREPROCESSING IN HEARING AND PSYCHOACOUSTICS AS GUIDELINES FOR SPEECH RECOGNITION

Eberhard Zwicker

Institute of Electroacoustics, Technical University München, Arcisstr. 21, D-8000 München 2, F R Germany

Introduction

Modern electronic equipment realizing network of system-theory as well as signal-theory strategies was a strong motor within the last 15 years pushing speech recognition systems to better and better results (for summaries see for example DeMori, 1979; Terhardt, 1978). Nevertheless, this progress is not comparable with the much larger progress of the data processing system like computers, memories, signal processors. Therefore we may ask for other and better guidelines to organize speech recognition systems. Since the human hearing system is still by far the best speech recognition system in every respect, it may be very helpful to simulate this system as much as we know about it. This idea is not new. Our research group seems to offer proposals in this direction each seventh year (Zwicker, 1971; Zwicker et al., 1979), this paper included. Other groups have accepted this approach in part by using critical band filtering (Klatt, 1982), by using loudness-time functions for segmentation (Mermelstein, 1975; Schotola, 1984), or more in general by using loudness-critical band rate-time patterns as preprocessed data base (Ruske, 1985 and this volume).

Hearing research made progress in the last seven years especially in the field of peripheral preprocessing in the cochlea. The Mössbauer technique was used in carefully performed animal experiments in order to measure basilar membrane displacement at lower levels (Patuzzi et al., 1984). For research in human cochlear preprocessing, the oto-acoustic emissions became a very effective non-invasive tool in order to get insight into this system (Zwicker, 1979; 1986a). The peripheral preprocessing system acts in advance of the neural data processing. The data to be processed are displacements, velocities or accelerations, i.e. AC-values, which are correlated to the sound pressure time function. This kind of preprocessing ends at the synapses of the inner hair cells in the organ of Corti. Then neural data processing starts. Its function can be studied in humans almost exclusively by psychoacoustical experiments. The neural processing with regard to speech recognition may be devided into two parts, the extraction of basic auditory parameters, such as loudness, pitch, roughness, timbre, fluctuation strength, duration together with the selection of the dominant parameters which form the input data to the second part, the subsequent segmentation, classification and recognition.

Although the general topic of our laboratory's research is "human hearing" and not specifically "speech recognition" we may be able to offer to the research area of speech recognition some usable tools which can help to solve some of actual problems by imitating the best speech recognizer, the human hearing system. A paper like this should deal with all three topics mentioned: (1) peripheral preprocessing up to the first synapses, (2) extraction of basic auditory parameters and selection of dominant ones, and (3) segmentation, classification and recognition. We are not active in topic (3). Therefore, I will concentrate on topics (1) and (2) in this paper.

1. Peripheral preprocessing

Based on a hypothesis (Zwicker, 1979) which was not very well founded on real facts and which did not fit into the trends at that time we completed a model of peripheral processing which looks like well founded on the measured facts known now. The model incorporates three assumptions: Only inner hair cells transfer information towards higher neural levels; the outer hair cells act as nonlinear saturating active AC-amplifiers; and form together with the hydromechanic system of the cochlea many feedback loops, which may even oscillate although at very low levels.

The physiological and anatomical view of the model was outlined formerly (Zwicker and Manley, 1983), and the simplified model realized in an analog version (Zwicker, 1984; 1986a) and in a computer version (Zwicker and Lumer, 1985). The behaviour of a combination of linear and nonlinear networks often is difficult to describe. In our case, with a strong frequency selectivity included, its behaviour can be outlined as a quasi linear system the nonlinearity of which is expressed in level dependencies. This way, the most prominent characteristics of the analog model simulating our hearing system's preprocessing are described in the following paragraphs.

A schematic diagram of two sections out of 90 in the analog model is shown in Fig. 1. The upper part represents the hydromechanics of the (passive) inner ear in the dual form in regard to the one normally plotted. This way, voltages can be used as values of interest instead of currents. The velocity-corresponding voltages are picked up through a transformer, amplifiedin an amplifier with symmetrically saturating nonlinear characteristic and feed back through a large resistor. This amplifying part with feedback represents the action of the outer hair cells. The inner hair cells are not shown explicitly but the output of each section of the model represents the input to the inner hair cells which is there transformed into neural spike activity and transmitted towards higher centers belonging to topic (2).

Before describing the behaviour of the peripheral preprocessing simulated in the model in some detail, it may be didactically helpful to compare the most important characteristics with those achieved in formerly used simple broadcasting receivers. Such receivers have a knob to choose the station we want to listen to: A resonant circuit produces the frequency selectivity needed. Otherwise we would hear many stations at the same time and the loudest one would disturb all the other softer ones we may be interested in. The sharper the tuning the better the separation of different stations. Normal passive frequency selective systems have been found not to be sharp enough and also not sensitive enough. There-

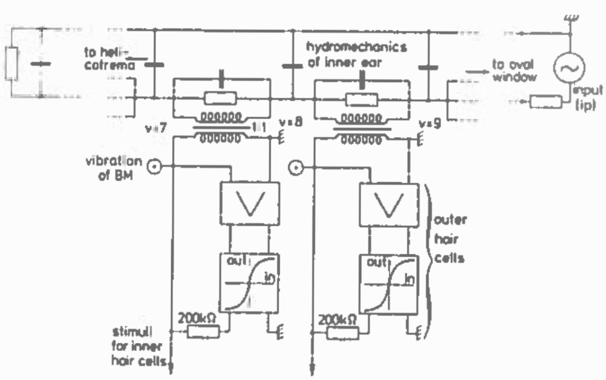


Fig. 1: Schematic diagram of a peripheral preprocessing model containing nonlinear active feedback.

fore, the simple broadcasting receivers of the 30's got - besides the tuning knob and the volume knob - a third, namely the feedback knob. Using active systems, the feedback could be controlled by this knob. Turning it to the right, the tuning was sharpened and the selectivity enhanced so that faint broadcasting stations could be received as well. This feedback knob, however, was a capricious tool: turning the knob a little bit too much to the right, feedback resulted in a very loud squeezing selfoscillation of the system. This was a strong handicap of those systems. Nevertheless, the most selective and most sensitive adjustment could be achieved by setting the knob just before the set where it starts to oscillate. Such feedback systems basically are not very stable and therefore are not used anymore.

Our inner ear, however, seems to make use of this strategy in a very interesting variation: it combines the feedback system with a saturating nonlinearity so that - for very faint sounds - the whole system can act near the oscillation point with large selectivity and large sensitivity. For loud sounds, however, the sensitivity is reduced automatically and the tuning widened. Such a behavior is very meaningful: the large sensitivity is needed for faint sounds only, not for loud sounds. But what about the annoying loud oscillations? The saturating nonlinearity acts at faint levels already, leading to the fact, that oscillations can be produced only with very small amplitude. Depending on the metabolism of the inner ear the system may oscillate a very little bit or not, an effect which was actually measured as sound pressure in the closed ear canal of more than 50% of normal hearing human subjects (Schloth, 1983; Dallmayr, 1985). The level of these spontaneous oto-acoustic emissions is mostly below threshold and therefore neither audible nor disturbing (no relation to tinnitus was found for these low-level emissions!).

This nonlinearity established in the outer hair cells creates an important characteristic: the large dynamic range of the sounds received is reduced strongly already at the level of basilar membrane vibration. Our inner ear acts in many parallel channels - and not in one channel only as the broadcasting receiver does - but all these channels act frequency selective so that the introduced nonlinearity does not disturb the information. This way, the ingenious and very effective construction of the inner ear uses all advantages of the above mentioned system and pushes its disadvantages in the background.

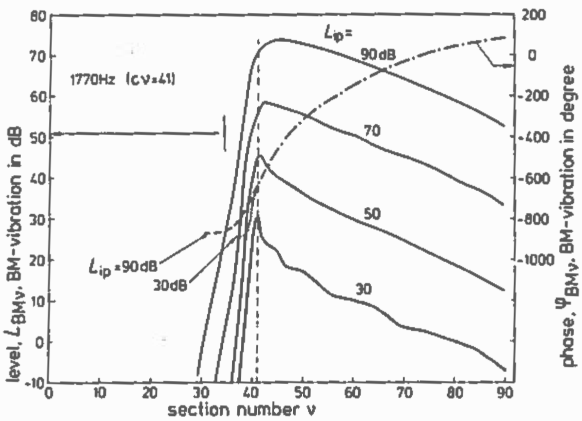


Fig. 2: Level $L_{BM\nu}$ of the voltage equivalent to basilar membrane vibration and its phase $\varphi_{BM\nu}$ as a function of the section number ν corresponding to place along the basilar membrane. Parameter is the input level L_{ip} of the 1770-Hz tone.

The so fare more generally described characteristics of the model are shown quantatively in Fig. 2. The level $L_{BM\nu}$ and the phase $\varphi_{BM\nu}$ corresponding to the level of the vibration of the basilar membrane and to its phase are plotted as a function of ν, the number of sections of the model corresponding to the place along the basilar membrane. The level-place patterns are plotted for an input frequency of 1770 Hz and input levels L_{ip} of 30, 50, 70, and 90 dB. The comparison of the four curves indicates the increasing place selectivity (corresponding to frequency selectivity) with decreasing input level. The peak strongly indicated for 30 dB at the characteristic place $c\nu=41$ disappears more and more for increasing input level. The increasing slopes of the curves are very steep but flatter for the decreasing part towards large numbers ν and level independent. The two phase-place patterns show an expected behaviour of strong phase lag with decreasing ν which depends near the characteristic place $c\nu$ on input level L_{ip}.

The effect of compressing the dynamic range is most clearly seen in the relation between level $L_{BMc\nu}$ at the characteristic place and the input level L_{ip} as indicated in Fig. 3. There, an input range of (100-40)dB=60dB is reduced to (80-39)dB=41dB. The slope of this output-input function amounts in a large range close to 0.5.

The model of peripheral preprocessing explains very well the existance and the behavior of oto-acoustic emissions (Zwicker, 1986b) and also the unusual frequency-difference and level dependence of the $(2f_1-f_2)$-difference tones (Zwicker, 1986c). More important for speech recognition seems to be the fact outlined in Fig. 2: the unsymmetric shape of the level-place patterns with the extremly steep rise, the level-dependent 3dB bandwidth which corresponds for normal speech level of 60dB to a $\Delta\nu$ of about 8 i.e. to the critical bandwidth, and the compression of the dynamic range especially at medium levels.

2. Extraction of basic auditory parameters

Following the peripheral nonlinear active preprocessing in the cochlea, the information picked up as vibration of the basilar membrane is transferred by 3500 inner hair cells into neural spike patterns. Since the tonotopic organization remains toward higher neural centers, it can be assumed that the information used for speech recognition is hidden in the neural spike rate-place-time pattern. This pattern is the basis of the extraction of basic auditory sensations such as loudness, pitch, roughness, timbre, fluctuation strength, or duration. Presuming that the temporal variations of these parameters bear the relevant speech information the processes leading to these parameters have to be outlined. Since neurophysiological methods can not be applied for this search, psychoacoustical ones are only usable. However, the models based on psychoacoustical experiments must be in line with the peripheral preprocessing. This means that the reduction of signal flow

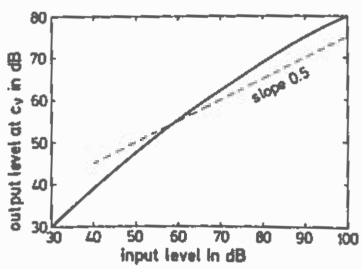


Fig. 3: Input-output relation expressed in 1770-Hz tone levels measured at the input and at $c\nu=41$.

produced stepwise from the sound pressure time function of speech to the final recognition by our hearing system can not be reversed: something lost in the first parts can not show up again at a later stage of processing.

The specific loudness-critical band rate-time pattern seems to be that fundamental psychoacoustical pattern, from which all basic auditory sensations are derived. It is approximated by the subdivision of the auditory frequency band into 24 adjacent critical bands. The amount of specific loudness in each channel is proportional to the square root of the sound pressure, and post-masking is already incorporated in its temporal structure.

To give an impression of such a specific loudness-critical band rate-time pattern, Fig. 4 shows it for the spoken word "ELECTROACOUSTICS" simplified in such a way that only the values of the even numbered bands between 2 and 22 are plotted. On top of the eleven time functions of the specific loudnesses N'_{ν}, the total loudness N is also indicated. Its time function changes much more slowly in relation to specific loudness but still contains important information useful for segmentation.

The extraction of the basic auditory sensation out of the specific-loudness pattern is described in a former paper (Zwicker et al., 1979). Meanwhile several pitch extractors have been discussed (Hess, 1983), some of them are also based on preprocessed auditory patterns (Terhardt, 1979; Terhardt etal., 1982a,b). Also pitch strength was studied in many details (Fastl, 1980) indicating that some kinds of pitch are much more impressive than others, additional data on roughness (Kemp, 1982; Aures, 1985) on timbre and sharpness (Aures, 1985.), and on subjective duration (Fastl, 1982b) have confirmed the effectivness of the use of specific loudness-critical band rate-time patterns.

An other basic auditory sensation, the fluctuation strength, added to the mentioned collection (Fast, 1982a, 1983, 1984). It is a sensation which seems to be useful for indicating the rhythm of speech (Köhlmann, 1982, 1985a,b) but may also produce hints for better and more meaningful segmentation (Köhlmann, 1985a,b). It is interesting to note that fluctuation strengh as a function of modulation frequency has its maximum near 4 Hz, a value for which the loudness-time function of speech shows its maximal spectral component as well (Fastl, 1982a).

The selection of dominant parameters is the last but in view of signal flow reduction still important step in using psychoacoustical results and models in speech recognition. The dominant changes of the basic

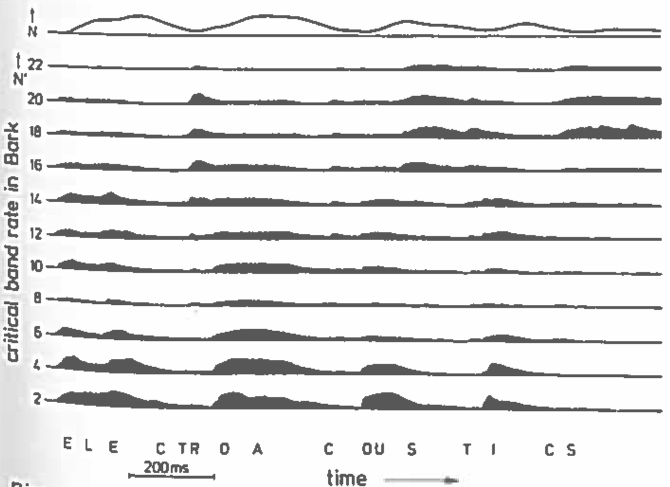


Fig. 4: Example of a specific loudness-critical band rate-time pattern. Total loudness N(t) on top.

auditory sensations are the features we listen to during speech recognition. In order to weight the different changes in a proper way, they should be expressed in just noticeable differences as units. Using this kind of psychoacoustical measure, the dominance shows up very clearly, so that for differences for a factor of two, the smaller one can be almost ignored, while for showing equal numbers of units, the changes of two auditory parameters are equivalent to each other so that both have to be taken into account (Suchowerskyj, 1977a,b).
For speech recognition, the size of the information flow to be handled by the recognition procedure is a very important value. Since normal speech in a quiet room offers an information flow of roughly 100.000 bit/s, this is too much to be processed and has to be reduced. In the specific loudness-critical band rate-time patterns, the flow is reduced to some 10.000 bit/s. Transferring these patterns into time functions of basic auditory sensations may reduce the flow for an additional factor of four. The extraction of only the dominant parameter changes decreases the flow for about a factor of two. This means that a signal flow closely to 1000 bit/s remains to be handled by the recognition procedure (see Fig. 5).

Two experiments produced results which are in line with these numbers, although very precise values can not be given. The first experiment made use of a single-board on-line system for speaker-independent isolated word recognition (Daxer and Zwicker, 1982). The influence of changes of (a) the number and frequency distance of channels, (b) the amplitude quantization, and (c) the dynamic range on recognition performance was explored. The results indicate that 10 to 20 filters based on critical band rate, 30 dB of dynamic range with only three or four bits per channel are sufficient. Using a sample frequency of 50 Hz, this leads to about 2500 bit/s. The second experiment used a vocoder system which was based on the specific loudness-critical band rate-time pattern (Knebel, 1980) and especially on sharpness (Fastl, 1982c) to devide speech into relevant features and to resynthesize it again. Speech intelligibility tests were used to check the effectivity. The results indicate that an information flow of about 1400 bit/s is sufficient to produce intelligibility scores of 90%. This means that a flow in the order of 1000 bit/s may be sufficient for speech recognition if an effective preprocessing system acts meaningfully, i.e. in our view, in a similar way than our hearing system.

3. Discussion and conclusion

Since computers and processors became so very popular in recent years, I have often been asked what is the difference between modern electronic systems and our hearing system in view of speech recognition. My reply was similar to the following sentences: (1) a very basic difference seems to be that electronics almost exclusively uses one very perfect, almost ideal line or processor or computer in order to solve a problem, while most of the biological sensory systems use very many, very poor lines or processes in parallel. This way, even with one or a few lines broken we are still able to hear although not as perfect as before. (2) Biological systems prefer nonlinear devises or at least combinations of linear and nonlinear devises, while we have learned through our education in mathematics and system theory to think more easily in linear systems. (3) Biological systems make much more use of adaptation and of feedback, often combined with each other, while we normally take care to avoid feedback in order to keep our electronic systems stable, and adaptive memories are coming in use only slowly.

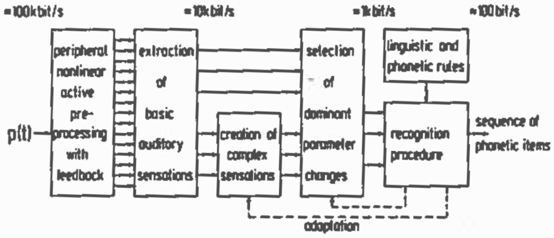


Fig. 5: Blockdiagram of a speech recognition system based on cochlear preprocessing and psychoacoustics.

Summarizing the strategies used by our hearing system which are discovered so far and which may be used in human speech recognition, a system as that shown in Fig. 5 can be offered. It contains the nonlinear peripheral preprocessing with active feedback, followed by the extraction of basic auditory sensations, out of which complex auditory sensations like virtual pitch or rhythm may be created. All these sensations are checked for dominant changes. The speech recognizing procedure makes also use of nonauditory information like linguistic rules and phonetic rules and finally produces a sequence of phonetic items.

It may be necessary to add to this simplified structure of a speech recognizing system based on auditory models other parts which take care of the many adaptive procedures available in hearing. We can adapt to reverberation, even to a strongly frequency-dependent one. We also adapt quickly to the characteristics of a speaker, however, to do so we need a larger information flow than in adapted situation. This can be given either by ideal, i.e. noiseless transmission of a new information or by a redundant information at the beginning of a speech, as for example "ladies and gentlemen". Adaptation is identical with strong feedback which is indicated in Fig. 5 by dashed lines and can be studied psychoacoustically in the same way as we have studied hearing sensations. Therefore and in contrary to ideas popular some 15 years ago (Pierce, 1969), we have seen and still see in the results of hearing research an effective help in order to find new or to improve realized ideas useful in speech recognition.

Acknowledgements and hints

The author is indebted to Dr.-Ing. habil. Hugo Fastl for several fruitful discussions. Most of the work described in this paper was carried out in the Sonderforschungsbereich 50, "Kybernetik" as well as 204 "Gehör", supported by the Deutsche Forschungsgemeinschaft.

Assuming that the literature in fields other than speech processing is not that well known to the readers of this article, the author has preferred to cite papers mainly on newer psychoacoustics of which reprints are still available in München.

References

Aures, W. (1985), Acustica 58, 268-281.
Dallmayr, C. (1985), Acustica 59, 67-75.
Daxer, W. and Zwicker, E. (1982), Speech Communication 1, 21-27.
DeMori, R. (1979), Signal Processing 1, 95-123.
Fastl, H. (1980), In: Psychophysical, Physiological and Behavioural Studies in Hearing, Delft, University Press, 334-339.
Fastl, H. (1982a), Hearing Research 8, 59-69.
Fastl, H. (1982b), Hochschul-Verlag, Freiburg.
Fastl, H. (1982c), Acustica 51, 99-102.
Fastl, H. (1983), In: Hearing - Physiological Bases and Psychophysics, Springer Verlag, 282-288.
Fastl, H. (1984), In: Fortschritte der Akustik, DAGA'84, Verl.: DPG-GmbH Bad Honnef, 739-742.
Hess, W. (1983), Pitch Determination of Speech Signals, Springer Verlag.
Klatt, D.H. (1982),J.Acoust.Soc.Am. 71 (S1),S111(A).
Kemp, S. (1982), Acustica 50, 126-133.
Knebel, H. (1980), In: Fortschritte der Akustik, DAGA'80, VDE-Verlag, Berlin, 671-674.
Köhlmann, M. (1982), In: Fortschritte der Akustik, FASE/DAGA'82, Verl.: DPG-GmbH, Bad Honnef, 903-906.
Köhlmann, M. (1985a), Acustica 56, 120-125.
Köhlmann, M. (1985b), Acustica 56, 193-204.
Mermelstein, P. (1975), J.Acoust.Soc.Am. 58, 880-883.
Patuzzi, R., Sellick, P.M. and Johnstone, B.M. (1984), Hearing Research 13, 19-27.
Pierce, J.R. (1969), J.Acoust.Soc.Am. 46, 1049-1051.
Ruske, G. (1985), In: New Systems and Architectures for Automatic Speech Recognition and Synthesis, Springer Verlag, 593-611.
Schloth, E. (1983), Acustica 53, 250-256.
Schotola, T. (1984), Speech Communication 3, 63-87.
Suchowerskyj, W. (1977a), Acustica 38, 140-147.
Suchowerskyj, W. (1977b), Biol.Cyb. 26, 169-174.
Terhardt, E. (1978), Elektr.Rechenanl. 20, 178-186.
Terhardt, E. (1979), In: Hearing Mechanisms and Speech, Springer Verlag, 281-291.
Terhardt, E., Stoll, G. and Seewann, M. (1982a), J.Acoust.Soc.Am. 71, 671-678.
Terhardt, E., Stoll, G. and Seewann, M. (1982b), J.Acoust.Soc.Am. 71, 679-688.
Zwicker, E. (1971), In: Pattern Recognition in Biological and Technical Systems. Springer Verlag, 350-356.
Zwicker, E. (1979), Biol.Cyb. 35, 243-250.
Zwicker, E. (1984), J.Acoust.Soc.Am. 76, p. 35.
Zwicker, E. (1986a), A hardware cochlear nonlinear preprocessing model with active feedback. J.Acoust. Soc.Am., in press.
Zwicker, E. (1986b), "Oto-acoustic" emissions in a nonlinear cochlear hardware model with feedback. J.Acoust.Soc.Am., in press.
Zwicker, E. (1986c), Suppression and $(2f_1-f_2)$-difference tones in a nonlinear cochlear preprocessing model with active feedback. J.Acoust.Soc.Am., in press.
Zwicker, E. and Lumer, G. (1985), In: Peripheral Auditory Mechanisms, Springer Verlag, 250-257.
Zwicker, E. and Manley, G. (1983), In: Biophysics, Springer Verlag, 671-682.
Zwicker, E., Terhardt, E. and Paulus, E. (1979), J.Acoust.Soc.Am. 65, 487-498.

REPRESENTATION OF THE FIRST FORMANT IN SPEECH RECOGNITION AND IN MODELS OF THE AUDITORY PERIPHERY

Dennis H. Klatt

Room 36-523, Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract. The frequency and amplitude of the first formant are not easy to measure as fundamental frequency (f0) varies in speech. Perceptual data indicate that the auditory system is not bothered by changes to f0, but processing strategies used in speech recognition, such as linear prediction, filterbank analysis, and the synchrony spectrum are seriously perturbed as f0 varies. The irrelevant variation makes it difficult/unreliable to perform phonetic comparisons between similar vowels based on simple ideas of pattern similarity. Of the possible solutions to this problem considered here, the one of greatest practical attraction is to implement a synchrony spectrum representation of vowel-like speech sounds, and a "learned pattern equivalence" approach to vowel phonetic-quality equivalence across different fundamental frequencies.

DFT magnitude spectra (25.6 ms Hamming window) of the lowest 1 kHz of a series of 5 kHz synthetic vowels are shown in Figure 1. All synthesis parameters have been held constant across stimuli except for the fundamental frequency of voicing (f0), which has been assigned a different constant value for each stimulus. The stimuli were devised to illustrate the problem of estimating the frequency (F1) and level (A1) of the first formant as fundamental frequency changes.

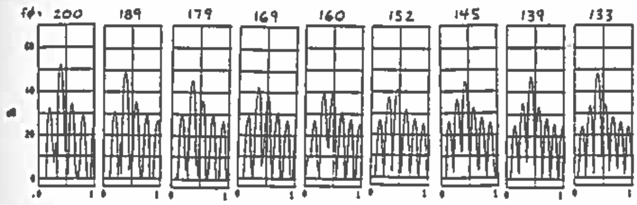


Figure 1. DFT magnitude spectra of 9 synthetic vowel stimuli varying only in f0.

The first formant frequency is 400 Hz in each synthetic waveform, and the first formant bandwidth is 50 Hz. These values, as well as the chosen frequencies and bandwidths of higher formants (F2=1800 Hz, B2=140, F3=2900, B3=240, F4=3800, B4=350), are typical for a vowel such as in the word "bit" (Klatt, 1980). Fundamental frequencies were selected in equal logarithmic steps from 133 Hz to 200 Hz. For the lowest fundamental, the third harmonic is exactly aligned with the 400 Hz first formant frequency; for the highest fundamental in the set of stimuli, the second harmonic is exactly aligned with the first formant frequency. For stimuli with intermediate values of fundamental frequency, no harmonic is exactly aligned with F1, and one has to interpolate by eye to determine the probable location of the first formant. This interpolation is not easy to perform automatically, as will become clear when we discuss the performance of various popular algorithms for formant estimation. There is a tendency for the first formant frequency estimate to be biased toward the frequency of the most intense harmonic, resulting in an error of up to plus-or-minus 8 percent for this stimulus set (Table 1).

Furthermore, the amplitudes of harmonics close to F1 are considerably less intense for intermediate stimuli of the stimulus set. The harmonic amplitudes are determined by the transfer function of the vocal tract, which peaks rather sharply at 400 Hz. If no harmonic is near F1, the strongest harmonic can be attenuated by up to 9 dB, resulting in a spectral peak that is attenuated by as much as 6 dB (filter banks) or 8 dB (linear prediction), which agrees with theory (Fant and Liljencrants, 1962) and measurements of real speech (Fintof, Lindblom and Martony, 1962). The formant amplitude misestimates of linear prediction are a result of misestimating formant bandwidths by a considerable factor (Atal and Schroeder, 1975).

STIM	f0	F1	HARMON	FB	LP
A	200	400	400	400	400
B	189	400	378	382	389
C	179	400	358	367	384
D	169	400	338	371	398
E	160	400	amb.	401	425
F	152	400	456	430	436
G	145	400	435	430	432
H	139	400	417	417	423
I	133	400	399	400	400
MAX ERROR:			+16% -15%	+7% -8%	+9% -4%

Table 1. First formant frequency predictions of nearest harmonic hypothesis (HARMON), peak location in wide-bandwidth filter bank (FB), and linear prediction spectrum (LP). Error increases if f0 is increased or BW1 is decreased.

According to one theory (HARMON in Table 1), the first formant is perceived to be the frequency of the strongest harmonic, at least for fundamental frequencies such that the ear can resolve individual harmonics (Chistovich, 1971).

According to a second theory, the formant peak is found by smoothing the spectrum in frequency such that individual harmonics are not seen (Chistovich et al., 1979). This proposal is similar in effect to earlier models which proposed to weight the importance of two strong harmonics according to the relative strength of their auditory representations (Carlson, Fant and Granstrom, 1975). In order to test the predictions of this theory, a particular smoothing algorithm was chosen -- the dft spectrum was smoothed by a 300-Hz wide Gaussian filter. As can be seen from Table 1, the energy smoothing model predicts that the perceived formant frequency will be somewhere between the "true" 400 Hz synthetic formant and the strongest harmonic. The amount of formant shift with changes to fundamental frequency is, however, quite large (see also Lindblom, 1962; Monsen, 19xx). Stimuli C and F differ by 63 Hz according to this model, which is 16 percent of F1. This difference would be easily audible because the JND for F1 is about 3% (Flanagan, 1955; Mermelstein, 1978). Thus Stimuli C and F should be heard as different vowels (/i/ and /I/) if this model were an accurate predictor of perceptual formant shifts with changes in formant/harmonic relationships. Apparently, the problem with the energy smoothing model is that a harmonic changes amplitude very rapidly as it slides down the skirt of a formant with a narrow (50 Hz) bandwidth. As soon as a harmonic is reduced by 4 to 6 dB below an adjacent harmonic, it hardly influences the location of the peak in the energy-smoothed spectrum.

According to a third theory, linear prediction spectra (autocorrelation form, 14-pole, 25.6 ms Hamming window) can extract F1 as the peak in the LP spectrum. Linear prediction fits an all-pole model to the waveform (Atal and Hanauer, 1971; Markel, 1972) or spectrum (Makhoul, 1975), thereby providing a method for effectively interpolating between harmonic locations to infer formant peaks. It is a particularly good model to apply to these stimuli since they were generated by an all-pole synthesizer and have virtually no noise or voicing source irregularities. The predictions of the linear prediction model are shown in the final column of Table 1. Linear prediction is not much better in performance than simple energy smoothing: there is a 52 Hz swing in the predicted F1 from stimulus C to F, which is a 13 percent change. Also, there is a slight bias toward overestimating F1 because the first harmonic amplitude is attenuated by the first difference analysis calculation. The reason that linear prediction does no better than the energy smoothing model is that the autocorrelation method uses a window of several pitch periods in duration, which means that the model must try to predict not only the damped vocal tract response to the first excitation at the beginning of the window, but also the time and magnitude of additional later glottal excitations and damped responses to them (Atal and Schroeder, 1975).

Perceptual Data. Does the human perceptual apparatus employ processing strategies which make all of these stimuli sound like exactly the same vowel (F1 the same) with the same loudness (vocal effort the same)? Naively, one might expect that if these stimuli are played in succession, one would hear not only a change in pitch, but also changes in loudness, spectral tilt, and vowel quality.

(1) First Formant Amplitude and Perceived Loudness. To see whether formant amplitude changes produce loudness differences across stimuli, Stimulus E was synthesized in its standard form and with 1,2,...6 dB added to the voicing sound source intensity. This set of stimuli was compared with both Stimuli A and I in unaltered form, using an "AX" randomized sequence in which subjects made a forced choice as to whether the first or second member of the pair was louder. Results from four listeners indicate a perceptual equal-loudness crossover at 2.0 dB. Thus when the pair of harmonics straddling F1 are 8 dB less intense (Stimulus E) than the single harmonic identical to F1 (Stimulus I), one must increase the level by only 2 dB to match subjective loudness.

Normally, it is said that loudness of a vowel depends primarily on the energy at F1, since this is usually the most intense part of the spectrum. We see that this is not the entire story because Stimuli E and I differ by 6 to 9 dB (depending on how energy near F1 is estimated), whereas an increase of only 2 dB makes these stimuli sound equally loud. Other possible determinants of vowel loudness are (1) the intensities of harmonics below F1, (2) energy in higher formants, (3) spectral tilt, and (4) the inferred shape of the vocal tract transfer function, i.e. the transfer function peak height instead of physical energy present at F1. Any one of these other potential cues could account for our loudness judgement results.

The variation in spectral amplitude of F1 as f0 is changed may be just as serious a deficiency of these spectral representations as mislocations of F1 in frequency. Any speech recognition device employing a distance metric that is sensitive to differences in relative formant amplitudes, such as the Itakura (1975) linear-prediction minimum prediction residual, or a filter-bank-based Euclidean metric (Plomp, 1970), will see considerable differences as f0 varies, even though the vowel is phonetically constant. This irrelevant variability can swamp out an ability to make fine phonetic distinctions in any current recognition device employing filter banks or linear prediction representations.

(2) First Formant Frequency and Perceived Vowel Quality. What kind of a perceptual effect on vowel quality is to be expected when f0 is changed? One possibility is that the auditory system somehow is able to extract the true F1, so vowel quality is unaffected. A second possibility is that the auditory system is fooled, or partially fooled, in exactly the same way as our processing schemes. A third possibility, one that somewhat confounds the choice between these alternatives, is that a change in f0 automatically invokes a kind of vowel-normalization process such that vowels spoken at higher f0 are assumed to come from shorter vocal tracts (Miller, 1953; Fujisaki and Kawashima, 1968; Carlson, Granstrom and Fant, 1970; Schwartz, 1971; Slawson, 1968; Traunmuller, 1982; Syrdal, 1985). A listening test was devised to distinguish among these alternatives (Klatt, 1985). Results showed convincingly that the auditory system is able to recover the true F1 with no bias toward the strongest harmonic, but there is also an automatic normalization process which makes it seem as if the vocal tract is shorter as f0 increases.

DISCUSSION

Our perceptual results are consistent with those of an excellent earlier paper that addressed the same issues (Carlson et al., 1975). They too found a regular shift in phonetic perception consistent with the view that f0 affects expectations of the vocal tract length of a talker. The authors examined their data to determine whether any phoneme boundary shifts could be attributed to perceptual biases toward the strongest harmonic, or toward a weighted mean of 2 or more harmonics. The weighting scheme that they employed was not the same as ours in that it did not weight harmonics according to their energy, and they did not examine an f0 range where harmonic biases go in an opposite direction from normalization biases, but the conclusions were the same -- there was no evidence of a bias toward the strongest harmonic as opposed to F1 (see also Florin, 1979; Assmann and Nearey, 1983; Darwin and Gardner, 1985).

So far this has been a largely negative paper: we have isolated defects in most speech processing algorithms that lead to unnecessary spectral confusions, but we have not provided any solutions. Three possible solutions are considered next.

Pitch-Synchronous Short-Window Analysis. If the analysis window is shorter than a single pitch period (e.g. windowed dft with a fixed 2 to 4 ms Hamming window, or covariance linear prediction during the inferred closed phase of glottal period) one can estimate the natural damped response of the vocal tract transfer function in the absence of excitations (Atal and Hanauer, 1971). This type of model is attractive, but is not easy to implement in a practical speech analysis system in such a way as to avoid occasional gross errors. If the window is misplaced, some very irregular spectra can be generated. The greatest problem with this kind of model is finding the time of glottal closure. Misplacements are particularly probable for high pitches and in noise. Until such time as analyses of this type can be made to mimic human perception consistently, we will have reason to doubt the validity of the technique as a speech analysis tool. An alternative might be to attempt to model the vocal tract transfer function using linear prediction, while simultaneously modeling the glottal waveform by some other appropriate representation (Milenkovic, 1986).

Auditory Modeling: Synchrony Detection. Sachs et al (1982) have shown that a measure of the tendency of neural firings to be synchronous with aspects of the basilar membrane displacement waveform has important advantages for speech processing. The synchrony measure is far less sensitive to changes in intensity of a vowel than are the average firing rate data. Synchrony data are also more immune to background noise and reverberation distortions (Allen, 1985), and they are not strongly affected by spectral tilt and formant amplitude variation (Srulovicz and Goldstein, 1983) which agrees with data on phonetic perception (Klatt, 1982). Processing schemes based on synchronous responses are reviewed in Carlson and Granstrom (1982), Delgutte (1984) and Seneff (1984). Thus it is of interest to determine whether any of these measures of synchronous response contains a representation of F1, and if so, is the estimate biased toward the strongest harmonic?

An answer comes directly from the Sachs et al. data, and from theoretical analysis of the waveforms observed at the outputs of the low-frequency critical band filters in this type of model. Physiological data and current models agree that the auditory system resolves individual harmonics near F1 for stimuli such as our family of synthetic vowels. Nowhere in the neural pattern are there time intervals between firings that are the inverse of F1. Only intervals related to harmonics are present. There is essentially only a sine wave at the outputs of these simulated mechanical filters because of a kind of FM capture effect that makes the strongest harmonic dominate the synchrony response in any channel (Allen, 1985). It will therefore be up to the central nervous system to figure out the first formant frequency from the relative proportions of fibers responding to each of the harmonics (and perhaps the relative phases of synchrony across channels). We can say little about the existance or details of such a calculation at this point.

Spectral Pattern Equivalence Sets. One interesting alternative that is not usually considered in speech recognition devices is that the harmonic pattern in the synchrony response is not processed centrally to recover an estimate of F1, but rather serves as a pattern vector in its raw form [Dick Lyon (personal communication) has expressed a similar viewpoint]. The CNS would then have to learn pattern equivalence sets across different fundamental frequencies, even though there may not be striking pattern similarity for equivalent vowel tokens. The total number of patterns in such a system would be much larger than the largest current vector quantization pattern set, but the approach, given sufficient labeled training data (see e.g. Kopek, 1985 for one of a number of possible implementation methods), could potentially overcome a number of other puzzling aspects of cross-speaker variability, as well as some of the distortions to a normal formant shape caused by (1) truncation effects (Fant and Ananthapadmanabha, 1982), (2) other source-tract interactions (Fant, 1985), (3) breathy-normal-creaky voice quality variations (Fant et al., 1985), and (4) vowel nasalization (Hawkins and Stevens, 1985). These four factors can introduce additional errors in algorithms designed to measure formant frequencies based on the detection of spectral peaks, and forcefully call into question the desirability of simple-minded approaches to the extraction of the frequency of F1 from speech waveforms (Bladon, 1982), although there can be no question of the importance of changes in F1 for vowel perception (Klatt, 1982). [This research was supported by ARPA.]

REFERENCES

Allen, J. (1985), "Cochlear Modeling," IEEE ASSP Magazine, Jan., 3-29.

Assmann, P.F. and Nearey, T.M. (1983), "Perception of Height Differences in Vowels", J. Acoust. Soc. Am. 74, S89 (A).

Atal, B.S. and Hanauer, S.L. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Am. 50, 637-655.

Atal, B.S. and Schroeder, M.R. (1975), "Recent Advances in Predictive Coding: Applications to Speech Synthesis," in G. Fant (Ed.) Speech Communication, Uppsala, Sweden: Almqvist and Wiksell, Vol. I, 27-31. [Reprinted in Markel, J.D. and Gray, A.H. (1976), Linear Prediction of Speech, New York: Springer-Verlag, 188-189.]

Bladon, A. (1982), "Arguments against Formants in the Auditory Representation of Speech", in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical Press, 95-102.

Carlson, R., Granstrom, B. and Fant, G. (1970), "Some Studies Concerning Perception of Isolated Vowels", Speech Transmission Laboratories Quarterly Progress and Status Report 2-3, Royal Institute of Technology, Stockholm, 19-35.

Carlson, R., Fant, G., and Granstrom, B. (1975), "Two-Formant Models, Pitch, and Vowel Perception", in G. Fant and M.A.A. Tatham (Eds.), Auditory Analysis and Perception of Speech, New York: Academic Press, 55-82.

Carlson, R. and Granstrom, B. (1982), "Towards an Auditory Spectrograph," in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical.

Chistovich, L.A. (1971), "Problems of Speech Perception," in L.L. Hammerich, R. Jakobson and E. Zwirner (Eds.), Form and Substance, Copenhagen: Akademisk Forlag, 83-93.

Chistovich, L.A., Sheikin, R.L., and Lublinskaja, V.V. (1979), "Centers of Gravity and Spectral Peaks as Determinants of Vowel Quality", in B. Lindblom and S. Ohman (Eds.), Frontiers of Speech Communication Research, London: Academic, 143-158.

Darwin, C.J. and Gardner, R.B. (1985), "Which Harmonics Contribute to the Estimation of First Formant Frequency?", Speech Communication 4, 231-235.

Delgutte, B. (1984), "Speech Coding in the Auditory Nerve II: Processing Schemes for Vowel-Like Sounds", J. Acoust. Soc. Am. 75, 879-886.

Fant, G. (1985), "The Voice Source: Theory and Acoustic Modeling", in I.R. Titze and R.C. Scherer (Eds.), Vocal Fold Physiology: Biomechanics, Acoustics and Phonatory Control, xx.

Fant, G. and Ananthapadmanabha, T.V. (1982), "Truncation and Superposition", Speech Transmission Labs QPSR 2-3, Royal Institute of Technology, Stockholm, 1-17.

Fant, G. and Liljencrants, J. (1962), "How to Define Formant Level: A Study of the Mathematical Model of Voiced Sounds," Speech Transmission Labs QPSR-2, Stockholm, Sweden: Royal Institute of Technology, 1-8.

Fant, G., Lin, Q.G. and Gobl, C. (1985), "Notes on Glottal Flow Interaction," Speech Transmission Labs QPSR 2-3, Royal Institute of Technology, Stockholm, 21-45.

Fintof, K., Lindblom, B. and Martony, J. (1962), "Measurements of Formant Level in Human Speech," Speech Transmission Labs QPSR-2, Stockholm, Sweden: Royal Institute of Technology, 9-17.

Flanagan, J.L. (1955), "A Difference Limen for Vowel Formant Frequency", J. Acoust. Soc. Am. 27, 613-617.

Floren, A. (1979), "Why Does [aa] Change to [ao] when F0 is Increased?", PERILUS I, Institute of Linguistics, Univ. Stockholm, 13-23.

Fujisaki, H. and Kawashima, T. (1968), "The Roles of Pitch and Higher Formants in the Perception of Vowels," IEEE Trans. AU-16, 73-77.

Itakura, F. (1975), "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans ASSP-23, 57-72.

Hawkins, S. and Stevens, K.N. (1985), "Acoustic and Perceptual Correlates of the Non-Nasal/Nasal Distinction for Vowels," J. Acoust. Soc. Am. 77, 1560-1575.

Klatt, D.H. (1980), "Software for a Cascade/Parallel Formant Synthesizer," J. Acoust. Soc. Am. 67, 971-995.

Klatt, D.H. (1982), "Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step", Proc. ICASSP-82, 1278-1281.

Klatt, D.H. (1985), "The Perceptual Reality of a Formant Frequency," J. Acoust. Soc. Am. 78, S81 (A).

Kopek, G.E. (1985), "Formant Tracking Using Hidden Markov Models," ICASSP-85, 1113-1116.

Lindblom, B. (1962), "Accuracy and Limitations of Sonagraph Measurements," Proc. 4th Int. Congr. Phonetic Sci., The Hague: Mouton, 188-202.

Makhoul, J. (1975), "Spectral Linear Prediction: Properties and Applications," IEEE Trans ASSP-23 283-296.

Markel, J.D. (1972), "Digital Inverse Filtering: A New Tool for Formant Trajectory Estimation," IEEE Trans AU-20, 129-137.

Mermelstein, P. (1978), "Difference Limens for Formant Frequencies of Steady State and Consonant-Bound Vowels", J. Acoust. Soc. Am. 63, 572-580.

Milenkovic, P., (1984), "Model Reference Glottal Inverse Filter of High F0 Voice", J. Acoust. Soc. Am. 76, S2 (A).

Miller, R.L. (1953), "Auditory Tests with Synthetic Vowels," J. Acoust. Soc. Am. 25, 114-121.

Plomp, R. (1970), "Timbre as a Multidimensional Attribute of Complex Tones," in R. Plomp and G. Smoorenburg (Eds.), Frequency Analysis and periodicitY Detection in Hearing, Leiden: Sijthoff, 397-411.

Sachs, M.B., Young, E.D. and Miller, M.I. (1982), "Encoding of Speech Features in the Auditory Nerve, in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical, 115-130.

Seneff, S. (1984), "A Synchrony Model for Auditory Processing of Speech", in J. Perkell and D.H. Klatt, (Eds.) Variability and Invariance of Speech Processes, Hillsdale, NJ: Erlbaum, xx-xx.

Schwartz, R.M. (1971), "Automatic Normalization for Recognition of Vowels of All Speakers", S.B. Thesis, MIT, Cambridge.

Slawson, A.W. (1968), "Vowel Quality and Musical Timbre as Functions of Spectrum Envelope and Fundamental Frequency", J. Acoust. Soc. Am. 43, 87-101.

Srulovicz, P. and Goldstein, J.L. (1983), "A Central Spectrum Model: A Synthesis of Auditory Nerve Timing and Place Cues in Monaural Communication of Frequency Spectrum", J. Acoust. Soc. Am. 73, 1266-1276.

Syrdal, A.K. (1985), "Aspects of a Model of the Auditory Representation of American English Vowels", Speech Communication 4, 121-135.

Traunmuller, H. (1982), "Perceptual Dimension of Openness in Vowels", J. Acoust. Soc. Am. 69, 1465-1475.

APPLICATION OF AN ADAPTIVE AUDITORY MODEL TO SPEECH RECOGNITION

Jordan R. Cohen

Continuous Speech Recognition Group, IBM T. J. Watson Research Center, Yorktown Heights, N. Y., U. S. A. (Current Address: Institute for Defense Analyses, Thanet Road, Princeton, N. J. 08540, U. S. A)

ABSTRACT

An adaptive model of the firing rates found in the auditory nervous system was configured as a signal processor for the IBM speech recognition system. The signal processor was tested on sentences drawn from office correspondence. Several experiments were done in low noise office environments using various microphones and different speakers. The system performance improved substantially compared to performance using a standard signal processor.

INTRODUCTION

Speech recognition systems sample speech signals with a signal-processing front end. One school of thought suggests that an auditory model is the 'ideal' signal processor for such applications, but performance figures available to date do not support the choice of auditory models over more standard signal analyses. This note reports the development and testing of a signal processing algorithm based on some aspects of the mammalian auditory system.

COMMENTS ON THE IBM SPEECH RECOGNITION SYSTEM

Information about the IBM speech recognition system is widely avaliable (Bahl, Jelinek and Mercer, 1983; Nadas, et. al., 1981). The 5000-word vocabulary isolated word dictation system developed at IBM was designed from a communications theory view of speech recognition. It is assumed that a talker formulates a complete English sentence and transforms it into a noisy acoustic signal. This acoustic signal is then captured by an acoustic processor which produces a series of (vector quantized) labels, discrete in both time and identity, from which a decision is made about the most probable sentence given the acoustic input. The probabilistic implementation of the system allows training of the linguistic decoder, but the system performance depends on the reliability of the acoustic processor.

The acoustic processor consists of two sub-systems. A signal processor transforms the high-bandwidth speech signal into a vectorized time signal sampled at a modest rate, and a labeller quantizes the resultant vectors once each centisecond. The standard system uses 30 filter-bank energies once each centisecond as its signal processor, and labels are assigned on a minimum Euclidian distance basis relative to prototypical vectors derived from training data. The signal processor reported here replaces the filter bank with an auditory model.

THE MODEL

The auditory model consists of a frequency analysis followed by perceptually motivated scaling and nonlinear adaptation. The frequency analysis is performed by a 20-band filter bank whose center frequencies and bandwidths correspond closely to those of auditory critical bands (Zwicker, Flottorp, and Stevens, 1957), roughly modeling the selectivity of the auditory system. A compressive power-law transformation is applied to the output from each filter, approximating loudness scaling (Stevens, 1955) and reducing the variability of the vector signal as compared with the original. The compressed signals form the inputs to a reservoir-type model of neural firings (Schroeder and Hall, 1974) which relates stimulus intensity to auditory-nerve firing rate, and which captures certain of the onset and offset characteristics of the neural response.

SIGNAL ACQUISITION AND FILTERING

Speech is captured using a far-field desk-mounted microphone (PZM-6). The speech signal is bandpass filtered (180 Hz to 8 kHz), and is digitized. Power spectra are computed with an FFT. A critical band filter bank is approximated by summing the squared Fourier coefficients (intensity) in each of 20 non-overlapping bands spaced one critical band apart.

The output of each filter is converted from intensity to loudness level by mapping each output power to its equivalent based on the Fletcher-Munson curves (Fletcher and Munson, 1937) and an estimate of the gain of the acoustic system. A conversion to loudness is performed by taking the third (in practice, the fourth) power of the output energy, and scaling such that 40 dB = 1 sone.

SHORT TERM ADAPTATION

Following the lead of Schroeder and Hall (1974), short term adaptation is modeled by assuming the existence of a reservoir holding some amount (n) of neurotransmitter. The change in the amount of neurotransmitter available at time t is described by

$$dn/dt = A - (S_0 + S_H + Dq)n(t).$$

A, D, S_0 and S_H are constants (estimated from psychophysical data), q is the square root of the loudness from each filter, and n is an internal state associated with each filter. This equation states that the change in neurotransmitter is equal to the replacement rate A minus the product of the amount of neurotransmitter available at that time with the sum of the spontaneous rate constant S_0, a decay constant S_H, and a scale D times the square root of the input loudness. The firing rate of that channel is expressed as

$$f = (S_0 + Dq)n(t).$$

These transformations were incorporated into the test system, and the output of the signal processor was substituted for the filter bank outputs of the previous standard process (Das, 1983).

RESULTS

Four talkers recorded the standard 100-sentence training corpus, and then recorded a 50-sentence test corpus at a later time. Signal processing was done twice, once using the filter bank and a second time using the auditory model front end. The system was trained for each speaker using the standard forward-backward algorithm. Results were as follows:

Table 1. Error rate and decoding times for four speakers using two separate front end processes. FB = Filter Bank, AM = Auditory Model.

	Error rate for 50 sentences (%)		*Decoding time (min)*	
Speaker	FB	AM	FB	AM
JRC	6.3	4.7	77	48
FRJ	7.9	4.4	75	38
LRB	4.2	2.3	43	32
PAF	6.6	4.0	99	61
Average	6.3	3.9	74	45

Error rates are expressed as the percentage of incorrect words in the entire test corpus, counting homophones of the correct word as incorrect. Decoding time is the time for the search through the possible sentences, and does not include signal processing time, labelling, clustering, training, and other overhead. Both error rates and decoding times are significantly lower using the auditory model than using the standard filter bank. The overall error rate is reduced by 40 percent. Informal experimentation using different speakers and microphones confirmed the efficacy of the new front end. Several of these experiments are summarized in Table 2.

Table 2. Decoding error rates for various speakers and two microphones. All experiments were trained on 100 sentences of training data, and tested on 20 sentences of test data (299 words). The test text was the same in each experiment. ER = Error Rate (%)

Speaker	Microphone	ER	ER
RLM	lip	3.6	3.3
RHR	lip	7.0	4.6
MAP	lip	6.0	3.3
RLM	lavalier	22.0	2.6
MAG	lavalier	9.3	6.0

The lip microphone was a Sure SMS-10, mounted near the corner of the talker's lips, and the lavalier microphone was a dynamic mike hung from a standard lavalier mount. The word error rates decreased for every speaker, although the decrease for RLM using a lip mike is quite small. (Some of the errors in this corpus are "language model" errors, in that the word strings are highly improbable given our particular 5000 word trigram model. Thus it is extremely difficult to demonstrate error rates below 2 percent for this corpus and language model.) The reduction from 22 percent error to 2 percent error for RLM's recordings using the lavalier microphone is quite striking, but in a different series of experiments using only long- term adaptation, the error rate on this corpus was decreased to 5 percent; much of the decrease is due to gain normalization. Decoding times were always less using the new front end than with the previous signal procesor.

Speakers MAG and PAF are both female, the rest of the speakers in the experiments reported here are male. No consistent difference has been noted in our recognition results between male and female speakers.

SUMMARY

A simple auditory model was developed and tested as a signal processing system for the IBM speech recognizer. It decreases the number of errors made by the system by approximately 40 percent in controlled tests.

ACKNOWLEDGEMENTS

I am indebted to Raimo Bakis, with whom this concept was originally developed, and to Michael Picheny, Robert Mercer, and Fred Jelinek for their support. Lou Braida aided me in debugging my explanations, and David Nahamoo's cooperative competition was of great aid.

REFERENCES

1. Bahl, L. R., Jelinek, F., and Mercer, R. L. A maximum likelihod approach to continuous speech recognition. *IEEE Trans. on pattern analysis and machine intelligence*. PAMI-5, 1983, 179–190.

2. Das, S. K. Some dimensionality reduction studies in continuous speech recognition. *IEEE Conf. on Acoust., Speech, and Signal, Proc.* 1983, 292–295.

3. Fletcher, H. F. and Munson, W. A. Loudness, its definition, measurement and calculation. *J. acoust. Soc. Am.* 5, 1937, 82–108.

4. Nadas, A., Mercer, R. L., Bahl, L. R., Bakis, R., Cohen P. S., Cole, A. G., Jelinek, F., and Lewis, B. L. Continuous speech recognition with automatically selected acoustic prototypes obtained by either bootstrapping or clustering. *IEEE Int. Conf. on Acoustics, Speech, and Sign. Proc.* 1981, 1153–1155.

5. Scharf, B. (1978). "Loudness," in Carterette and Friedman (eds.) Handbook of Perception, Vol. IV, Academic Press, p. 180-242.

6. Schroeder, M. R. and Hall, J. L. Model for mechanical to neural transduction in the auditory receptor. *J. acoust. Soc. Am.* 44, 1974, 1055–1060.

7. Stevens, S. S. The measurement of loudness *J. acoust. Soc. Am.* 27, 1955, 815–829.

8. Zwicker, E., Flottorp, G., and Stevens, S. S. Critical band width in loudness summation *J. acoust. Soc. Am.* 29, 1957, 548–557.

Speech Recognition Experiments with a Cochlear Model

Richard F. Lyon
Schlumberger Palo Alto Research
3340 Hillview Ave.
Palo Alto, CA 94304

Abstract

There are several ways that a computational model of auditory processing in the cochlea can be applied as the front end of a speech recognition system. For an initial round of experimentation, the fine time structure in the model's output has been used to do spectral sharpening, yielding a "cochleagram" representation analogous to a short-time spectral representation. In later experiments, fine time structure will be exploited for a more detailed characterization of sounds, and for sound separation.

So far, experiments have been done with only two words ("one" and "nine") spoken by 112 talkers, to limit the range of phonetic variation to simple voiced sounds, while providing a good sample of inter-speaker variation. The structure of the vector space of "auditory spectra" has been examined through vector quantization experiments, which yield a measure of information content and local dimensionality.

The inclusion of more dimensions of perceptual variation, such as pitch and loudness, in a speech front end representation is both an opportunity and a problem. Much larger vector quantization codebooks and more training data may be needed to take advantage of the extra information dimensions. A product-code approach and an improved algorithm for finding the nearest neighbor codeword are suggested to help cope with the problem and take advantage of the opportunity.

Preliminary recognition experiments using a single codebook per word and no time sequence information have shown a performance of about 97% correct one/nine discrimination for talkers outside the training set, and 100% correct for second repetitions from talkers in the training set. Further experiments are currently underway.

1 Introduction

Our experimental cochlear model has been most recently described in terms of its performance on simple "physiology" experiments [1]. Those experiments concentrated on the role of the AGC stages, which serve to partially normalize the output representation in the face of a wide dynamic range of overall amplitude and overall spectrum variations. The dynamics of the gain control process help to preserve perceptually relevant information about loudness and spectrum, emphasizing short-term changes.

The output of the model is regarded as a sequence of vectors in n-space, representing n-channel perceptual spectra. Silence maps to the zero vector, and perceptually louder sounds map to points further from zero. But detailed characterizations of this pattern space are difficult, due partly to its high dimensionality.

The number of important dimensions of variation due to phonetic and talker identity is an important issue in designing recognizers to work in this space, and is discussed in the next section. The following section discusses a set of recognition experiments, including comparisons with LPC. Finally, improved vector quantization techniques to work in this pattern space are suggested in the last section.

2 The Space of Cochlear Spectra

In the current version of the model, 92 bandpass channels are used to span a range of about 23 barks (about 100 Hz to 10 kHz). By modeling hearing, it is hoped that sounds will map into 92-space in such a way that a simple Euclidean distance in that space will correlate well with perceptual distinctions. Therefore, it is expected that a low-distortion vector quantizer designed to minimize mean squared Euclidean error will preserve most of the relevant information in a cochlear spectra. To explore this notion, codebooks of different sizes and distortions were constructed from various training corpora.

To make codebooks, a modified k-means algorithm was used. In each pass over the training data, new codewords were added to the codebook whenever the distortion to a training vector exceeded a desired distortion bound; at the end of a pass, each codeword was moved to the average of the vectors that were closest to it. Compared to a straight k-means with codebook size doubling, we found convergence to about the same rms distortion for a given codebook size, but in fewer iterations. Having maximum distortion as an independent variable is also useful.

The resulting data on codebook size *vs.* rms distortion and max distortion for a training corpus of 112 talkers saying "one" and "nine" are shown in Figure 1. The desired value of max distortion, such that reconstructed cochleagrams have clear and continuous formant and pitch tracks, is probably less than the lowest tried so far.

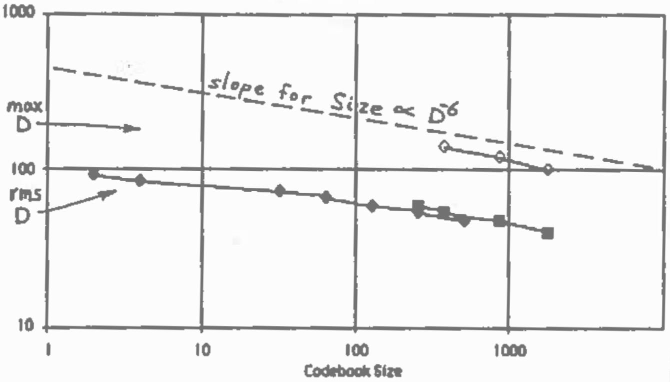


Figure 1: Codebook rms distortion (filled symbols) and maximum distortion (empty symbols) *vs.* codebook size.

The slope of the size *vs.* distortion curves (on a log-log plot) should reveal the dimensionality of the subspace that the codewords are packing into. Cutting the distortion by a factor of two will require a factor of sixteen in codebook size increase if there are four dimensions of variation to be covered.

The data show slopes corresponding to about 6 dimensions. Since the phonetic variation in the test corpus is quite small, much of this variation is probably due to talker differences. Since lower pitch harmonics are resolved in the spectrum, and loudness is not completely normalized out, these perceptually important dimensions contribute important dimensions of variation in the data that would not normally be seen in LPC and other common representations.

For the one/nine data, a codebook size of 1801 is barely adequate for high-fidelity coding of cochleagrams of the talkers in the training set. For the complete digit vocabulary, a codebook about five times larger would probably perform similarly. The distortion caused by using a codebook size of 383 is apparent in figure 2.

Based on these observations, it appears that representing a complete range of phonetic variation (eight or more dimensions), with reasonable fidelity would require a codebook size around 50,000 to 1,000,000. These sizes are far beyond normal practice in the speech recognition field, and require new techniques if they are to be useful.

3 Recognition Experiments with Cochleagrams and VQ Codebooks

Since training our existing recognizer [2] to use the cochlear spectrum pattern space will take considerable time, a much simpler test was undertaken first. Using the technique of Shore and Burton [3], a codebook was designed for "one" and another codebook was

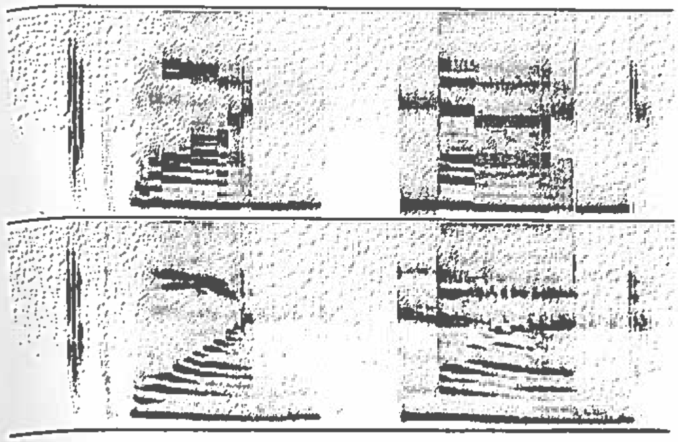


Figure 2: Cochleagram and vector quantized cochleagram of two digits by a talker outside the training set, with codebook size 383.

designed for "nine", using a single repetition of each word from each of the first 50 of the 112 talkers. Setting maximum distortion to 140 for both cases, the codebook for "one" reached a size of 261 and an rms distortion of 45.2, while the codebook for "nine" reached a size of 272 and a 5% higher rms distortion of 47.3.

Recognition proceeded by comparing quantization distortions (rms or total squared distortion) using the two codebooks, without compensation for the different codebook characteristics. No endpoint detection was done, so the generous amount of silence and noise at both ends of the words was included in the distortion measurements.

Testing on the second repetition of the same words from the training talkers led to no errors (in 100 trials). This result is encouraging, since this recognition technique has not previously been very successfully applied to speaker-independent or multi-speaker problems.

Testing on the other 62 talkers showed a serious bias: there were no misrecognitions of "one" as "nine", but ten misrecognitions of "nine" as "one" (5 on first repetition, 5 on second repetition, mostly from different talkers). Overall, on this speaker independent condition, there are 10 errors in 248 trials, or 96% correct. While this does not approach the performance of a good speaker independent isolated digit recognizer on the "one/nine" discrimination task, it is quite respectable for this simple algorithm.

Using order 11 LPC as a parameterization for comparison, with an Itakura distortion measure, we obtained at best 2 errors in 100 trials from talkers in the training set (98% correct), for various codebook sizes, and 14 errors in 248 trials on the other talkers (94.4% correct). Surprisingly, even very small codebooks (2 to 16 codewords) performed well with LPC, so it was decided to go back and try the cochleagrams with small codebooks.

With cochleagrams, it was found that for talkers in the training set, larger codebooks work best (sizes 32 and up gave no errors), but that smaller codebooks do a better job of generalizing to talkers outside the training set (size 32 was optimal with 7 errors in 248 (97.2% correct), while sizes 16 and 64 both were both slightly better than the initial large-codebook experiment, with 9 errors each. These differences may not be significant.

For every codebook size except size 2, the cochleagrams gave fewer errors than the LPC, usually by more than a factor of two.

4 VQ Algorithm Improvements

In spite of the encouraging results with small codebooks, it seems that to take full advantage of the information in cochleagrams with large talker populations will require very large codebooks. There are (at least) two alternative approaches to making very large vector codebooks practical. First, better fast quantization algorithms can be used to reduce the time cost. Second, codebooks can be constructed as product codes built from a small number of moderate-size codebooks.

Our present quantization algorithm takes advantage of the triangle inequality that applies to the Euclidean distance metric, so that codewords too far from a current best guess need not be examined; this unfortunately requires a table of N^2 inter-codeword distances, and so is impractical for much larger codebooks. The FN algorithm [4] uses a tree structure with a branch-and-bound search algorithm to take advantage of the same inequality with less stored information. Another approach which looks promising is to store the dual of the multi-dimensional Voronoi diagram [5] of the code vectors, so that each code vector is linked to its neighbors; in this case, when the current best guess is better than any of the neighbors, no further codewords need be examined. Using the last frame's quantization index as a first guess is very effective in these algorithms. In any case, the auxiliary data structures should be designed such that they are easy to modify when expanding or iterating the codebook.

The product code approach [6] is an alternative way to encode many bits of information per symbol with low distortion and small codebooks. The code space is the direct product of smaller codes, each of which encodes a separate part of the information in the original vector. In the simplest case, the original vector to be encoded is simply split up such that some components (*i.e.*, cochleagram channels) are used as a small vector in one codebook, and the other components are used with one or more other small codebooks. But other vector processing operations could also be used to try to separate the information more cleanly into feature vectors of lower dimensionality. For example, one process could attempt to capture pitch information, another could try to capture first formant information, etc. As long as these "feature extraction" processes don't lose information, the overall vector quantization distortion can be made as low as desired (even if quantizing sub-optimally by independently quantizing with each small codebook). If each feature detecting process captures only one or two important dimensions of variation, the resulting codebooks could be quite small. The structure imposed on the code space by the product code may also be useful in some kinds of recognition algorithms.

5 Conclusions

The cochlear model produces a spectral representation that captures important dimensions of speech signals. Preliminary experiments show that cochlear spectra lead to about 50% fewer errors in a very simple recognition technique, compared to LPC. Taking full advantage of the extra dimensions of information in cochlear spectra with a wide range of phonetic material and a wide range of talkers may yet require very large vector quantization codebooks or other techniques to extract the relevant features.

6 References

[1] Richard F. Lyon and Lounette Dyer, "Experiments with a Computational Model of the Cochlea," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Tokyo, Apr. 1986.

[2] Marcia A. Bush and Gary E. Kopec, "Evaluation of a Network-Based Isolated Digit Recognizer Using the TI Multi-Dialect Database," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Tampa, Mar. 1985.

[3] J. E. Shore and D. K. Burton, "Discrete Utterance Speech Recognition without Time Alignment," *IEEE Trans. Inform. Theory* **IT-29**, pp. 473–491, July, 1983.

[4] K. Fukunaga and M. M. Narendra, "A Branch and Bound Algorithm for Computing *k*-nearest Neighbors," IEEE Trans. Computers, **c-24**, pp. 750–753, 1975.

[5] D. T. Lee and Franco P. Preparata, "Computational Geometry—A Survey," IEEE Trans. Computers, **c-33**, pp. 1072–1101, 1984.

[6] John Makhoul, Salim Roucos, and Herbert Gish, "Vector Quantization in Speech Coding," Proc. IEEE, **73**, pp. 1551–1558, 1985.

Supported by DARPA contract N00039-85-C-0583.

A SPECTRAL-TEMPORAL SUPPRESSION MODEL FOR SPEECH RECOGNITION

P. L. Divenyi

Speech and Hearing Research Facility, Veterans Administration Medical Center, Martinez, California, 94553, and Department of Speech and Hearing, University of California, Santa Barbara, California, 93106, U.S.A.

INTRODUCTION

Speech recognition systems, however heterogeneous in their conceptions and schemes, share at least one basic feature: the inclusion of a vocoder-type front-end. While many of the early, and some of the contemporary, systems adopted a pragmatic design for their front-end filter bank, there were some efforts (e.g., Chistovich et al., 1975; Searle et al., 1979) toward providing the recognizer with an input stage that was modeled after the human ear. The motivation for such a design was the desire to optimize the recognition process from the very first stage on. However, work by auditory physiologists on auditory nerve responses to speech (Young and Sachs, 1979; Delgutte, 1980) signaled a welcome convergence of interests by two groups of scientists on the problem of speech processing in the auditory system. More recent work by several investigators, some of which is included in the present symposium, has been directed toward designing recognizer front-ends that resembled the ear more-and-more closely, and toward examining effects of model parameter modifications on recognition performance.

Computational models of the auditory system fall into two major classes, depending on whether the calculations are performed in the time or in the spectral domain. The advantage of time-domain algorithms lies mainly in their speed, whereas spectrally-based algorithms may more closely approximate the actual auditory processes because they are able to deal more directly with non-linear filtering operations. The present model is spectral in the sense that the filtering computations are executed in the frequency domain.

DESCRIPTION OF THE MODEL

The present model has been built around the physiologically-based and fine-tuned spectral model proposed by Shannon (1979). That work stands out in that it computes the magnitude of peripheral auditory activity across all frequency-specific channels, taking into account passive and active cochlear filtering, compressive nonlinearity, and suppression on both sides of a given channel. It is, however, restricted to spectral processing. The present modeling work was undertaken in an effort to see how time-varying signals can benefit from spectral suppression, i.e., an enhancement of the contrast between channels differing in their activity level, as offered by the Shannon model. The five stages of this model are connected in a strict sequential order, i.e., without feedback loops.

1. The Spectral Estimator Stage.
The physical continuum of frequency was mapped into 120 discrete channels between 50 and 10kHz using the frequency-to-basilar membrane distance transformation proposed by Greenwood (1961). The purpose of the spectral estimator was to provide the inner ear simulator (that operated in the spectral domain) with an estimate of the input magnitude that excited each channel. This input magnitude had to reflect the duration of the assumed equivalent impulse response of the corresponding inner-ear filter, i.e., it had to be gated using a window whose length was a function of the inner-ear filter width. Thus, a separate magnitude estimate had to be made for the narrow active- and the wider passive filters of each channel (see Stage 3). We adopted a Hamming window with a skew that emphasized more recent events. We arbitrarily assigned a 10-Hz maximum frequency resolution to our 50-Hz channel and calculated the window length for each channel assuming linear impulse response and applying the Greenwood mapping. We also limited the minimum window length to 2 ms, in order to account for an indelible neural refractoriness. The actual estimation was represented by Direct Fourier Transform coefficients of the windowed input at the frequency corresponding to a given channel.

2. The Outer- and Middle-Ear Response Simulator.
To account for ear canal resonance and middle ear attenuation, we included a spectral shaping algorithm gradually falling off below 2.5 and above 4 kHz. The attenuation (in dB) was a linear function of basilar membrane distance.

3. The Inner-Ear Spectral Response Simulator.
This stage, the actual Shannon model, is characterized by two concurrently working filter banks. One of the banks consists of passive, broadly-tuned, linear filters having a hhigh (30-dB SPL) threshold. Filters in the other bank are active, sharply tuned, low-threshold filters with a nonlinear compressive response that makes any activity increment beyond 40 dB SPL negligible. The active filters are followed by a sub-stage representing the suppression of high tones by low tones. The output of this sub-stage is linearly added, channel-by-channel, to that of the passive filter bank. The output of the mixer is followed by the sub-stage of suppression of low tones by high tones. In sum, the output of the inner-ear simulator represents the magnitude of the activity in the auditory nerve across tonotopically organized channels. This output compresses a 120-dB dynamic range in the input into a 20-to-25-dB range in the output.

4. The Auditory Nerve Temporal Response Simulator.
Single unit studies have demonstrated that there is a sizable temporal adaptation effect in the response of single auditory nerve fibers (Smith and Zwislocki, 1975). This effect is characterized by a strong burst of activity at the onset of the stimulus followed by a gradual decrease, and by a moment of sudden decrease of the activity at stimulus offset, followed by a gradual recovery. We used Smith's theoretical expression for this temporal process, noting that the effect is independent in each channel and that the adapted output is affected only by the magnitude of the present and the immediately preceding output epoch, rather than by the input. Thus, the effect is not unlike that of a high-pass filter with a floor (i.e., the spontaneous activity level). It was implemented in our model as simple exponential differentiators having different time constants for adaptation (18 ms) and recovery (36 ms). This stage enhances temporal contrasts in the input.

5. The Temporal Integrator Stage.
Auditory psychophysical data, however, depict the auditory system as one with memory: Detection of signals at threshold and ~~detection of~~ envelope

fluctuations, for example, clearly speak for the existence of a low-pass process, i.e., of a leaky integrator. We implemented this stage as an exponential integrator placed on each channel at the output of the temporal adaptation stage. The time constant we chose was short (1.5 ms) -- in agreement with other workers (Penner, 1978). We also noted that, because this integrator operates on the compressed output rather than on the input, a single, short time constant must be capable of accounting for both temporal integration at threshold and envelope discrimination at suprathreshold levels.

EXAMPLES

We have completed several tests with simple, easily definable input signals, in order to obtain an optimized set of model parameters. The output of two simple signals, a 100-dB SPL, 2-ms click and a 50-dB SPL 50-ms Gaussian white noise burst, are shown in Fig. 1. We have also examined the behavior of the model in response to natural speech sounds. One example, the beginning of the phonetically-balanced sentence "The goose was brought straight from the old market" is shown as a spectrogram in Fig. 2 and as a "neurogram", or time-frequency channel model output, in Fig. 3. In addition, we have also examined a large number of natural CV utterances, in an attempt to search for invariant cues (not shown here).

SPEECH RECOGNITION TESTS

In order to see whether the model could embody an improved front-end to a cepstrum-based recognizer, we conducted a series of experiments on a natural sentence data base. Recognition performance with the raw output of the model as input to the recognizer was significantly poorer than when the front-end was a simple vocoder. Much of the performance degradation could be attributed to the presence of individual low harmonics that dominated the model output. It seems, therefore, that some type of feature detection would be necessary before the model could become a useful tool in automatic speech recognition.

ACKNOWLEDGMENTS

This research was conducted when the author was visiting at the Institut National de la Recherche Scientifique, Université de Québec, and has been supported by the Veterans Administration.

REFERENCES

Chistovich, L., Fyodorova, N., Lissenko, D., & Zhukova, M. (1975). Auditory segmentation of acoustic flow and its possible role in speech processing. In G. F. a. M. Tatham (Ed.), Auditory analysis and perception of speech (pp. 221-232). London: Academic.

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. J. Acoust. Soc Amer., 68, 843-857.

Greenwood, D. D. (1961). Auditory masking and the combination band. J. Acoust. Soc. Amer., 33, 484-502.

Penner, M. (1978). A power-law transformation resulting in a class of short-term integrators that produce time-intensity trades for noise bursts. J. Acoust. Soc. Amer., 63, 193-201.

Sachs, M. B., & Young, E. D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. J. Acoust. Soc. Amer., 66, 470-479.

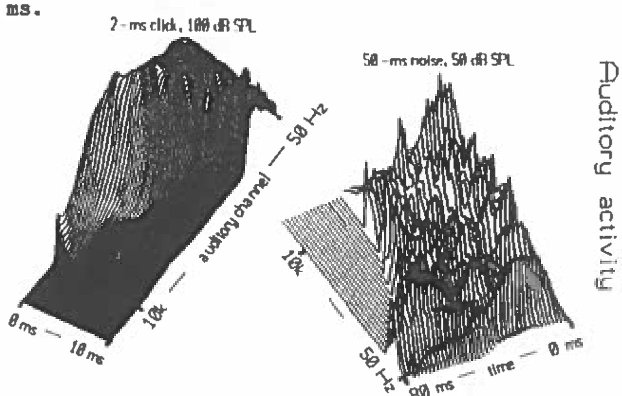
Searle, C. L., Jacobson, J. Z., & Rayment, S. G. (1979). Stop consonant discriminatin based on human audition. J. Acoust. Soc. Amer., 65, 799-809.

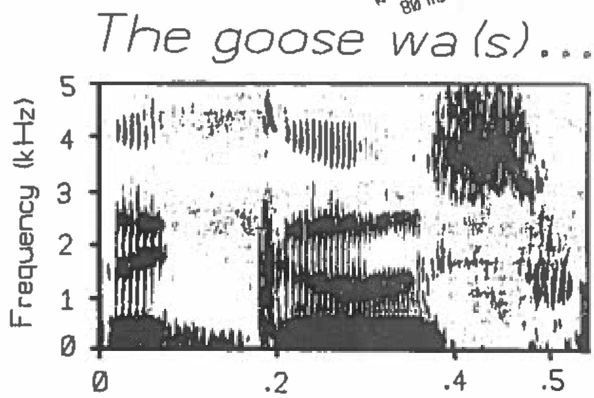
Shannon, R. (1979). A model for psychophysical suppression. J. Acoust. Soc. Amer., 65, S56.

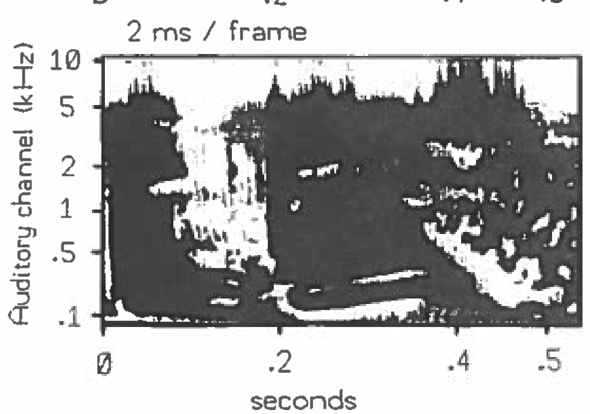
Smith, R. L., & Zwislocki, J. J. (1975). Short-term adaptation and incremental response of single auditory nerve fibers. Biol. Cybern., 17, 169-182.

FIGURE LEGENDS

1. a: 3-D picture of the model's response to a 2-ms click presented at 100 dB SPL. Frame size: .25 ms. Only the first 10 ms of the response are shown. b: 3-D picture of the model's response to a 50-ms burst of white noise presented at 50 dB SPL. Frame size: 2 ms. Only the first 80 ms of the response are shown.
2. Conventional spectrogram of the utterance "The goose wa(s)..." by a male talker.
3. Model output ("neurogram") of the same utterance. Difference between the darkest and the lightest parts of the output is 13 dB. Frame size: 2 ms.







THE AUDITORY PROCESSING OF SPEECH

SHIHAB A. SHAMMA

Electrical Engineering Dept & Systems Research Ctr
University of Maryland, College Park, MD. 20742.
Mathematical Research Branch, NIH. Bethesda, MD

abstract

The processing of speech in the mammalian auditory periphery is discussed in terms of the spatio-temporal nature of the distribution of the cochlear response and the novel encoding schemes this permits. Algorithms to detect specific morphological features of the response patterns are also considered for the extraction of stimulus spectral parameters.

The remarkable abilities of the human auditory system to detect, separate, and recognize speech and environmental sounds has been the subject of extensive physiological and psychological research for several decades. The results of this research have strongly influenced developments in various fields ranging from auditory prostheses to the encoding, analysis, and automatic recognition of speech. In recent years, improved experimental techniques have precipitated major advances in our understanding of sound processing in the auditory periphery. Most important among these is the introduction of nerve-fiber population recordings which made possible the reconstruction of both the temporal *and* spatial distribution of activity on the auditory-nerve in response to acoustic stimuli [1, 2]. Sachs et al. utilized such data to demonstrate the existence of a highly accurate temporal structure that is capable of providing a faithful and robust representation of speech spectra over a wide dynamic range and under relatively low signal-to-noise conditions [3, 4]. Their work has since motivated further research into the various algorithms that the central nervous system (CNS) might employ to detect and extract these and other response features, and the possible neural structures that underly them [5, 6].

In pursuit of these goals, we have constructed and analyzed the spatio-temporal response patterns of cat's auditory-nerve to synthesized speech sounds [4, 5]. These patterns are formed by spatially organizing the temporal response waveforms (or PST histograms) of the auditory-nerve-fibers according to their characteristic frequency (CF) [4]. The resulting display highlights the interplay of temporal and spatial cues across the fiber array and suggest novel ways of viewing cochlear processing and encoding of complex sounds [7, 5]. The availability of such experimental data, however, is at present limited by technical constraints and the massive amount of processing required to handle them. Thus, in order to analyze new speech tokens, and to facilitate the necessary manipulation of stimulus and/or processing conditions and parameters, we have developed detailed biophysical and computational models of the auditory periphery and used them to generate spatio-temporal response patterns to natural and synthesized speech stimuli. Various CNS schemes for the estimation of stimulus spectral parameters are then investigated based on these patterns.

The Cochlear Model :

Computational algorithms for the cochlear processing of speech are developed that are based on detailed biophysical formulations of linear basilar membrane mechanics and nonlinear hair cell transduction characteristics [8]. Basilar membrane analysis is based on detailed 3-D hydroelastic models that are quite efficient to compute [8, 9]. These models are used to generate the transfer functions at points along the cochlear length, which are then employed directly in all subsequent processing of speech sounds. The output (membrane displacement) at each point is transduced into hair cell intracellular potentials through two stages representing the velocity fluid-cilia coupling and the nonlinear hair cell. The latter stage can be approximated in most cases by a cascade of a compressive nonlinearity (of the form: $V = z.\exp(au)/(1+\exp(au))$ where (z,a,x) are constants with definite biophysical interpretations) followed by a low pass filter (time constant=0.1 ms). The final outputs then approximately represent the instantaneous probability of firing of the auditory-nerve fiber array. Many more detailed refinements have often been included in this model (e.g. synaptic adaptation mechanisms, middle and outer ear transfer functions, and some form of automatic gain control) to reproduce the finer details of the

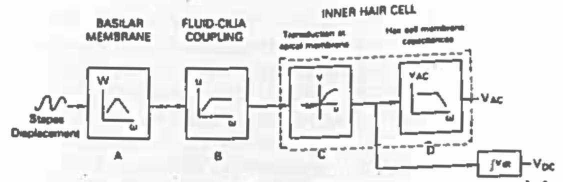


Fig.1: Schematic of the cochlear model stages [8].

responses. Nevertheless, the simpler model described above captures the major features of the experimental responses.

Examples of the model outputs are shown in Figs 2a,3 in response to a naturally spoken (female) /bɛt/ and a synthesized vowel /a/, respectively. In Fig.2a the response is to the onset of the vowel portion of the stimulus (whose spectrogram is shown in Fig.2b(right)). The periodic nature of the response is evident at regular intervals corresponding to the fundamental period of the stimulus. Strong harmonics, located near the formants of the vowel, dominate the response patterns over relatively broad segments of the channel array. Within each segment (e.g.$0.4 < CF < 1.6$ KHz) the travelling waves exhibit two important characteristics observed earlier in the experimental data: (1) Rapid apical decay due to the asymmetrical tuning of the basilar membrane amplitude. (2) phase shifts or delays in the response waveforms near the CF of the underlying harmonic, due to the rapid accumulation of phase-lag in the travelling wave near its point of resonance. The response to the plosive /t/ in /bɛt/ is also shown in Fig.2a, with its noisy character and high frequency content evident in the response patterns.

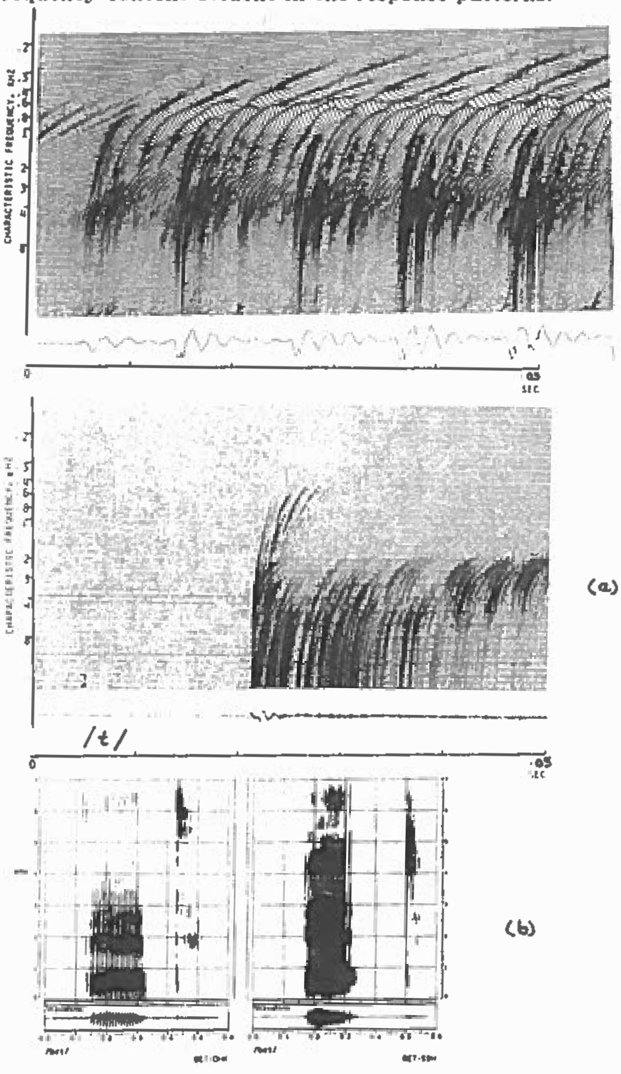


Fig.2: (a) Spatio-temporal responses of the cochlear model to selected portions of /bɛt/ spoken by a female. (b) Spectrograms of /bɛt/ spoken by a male (left) and a female (right) [12].

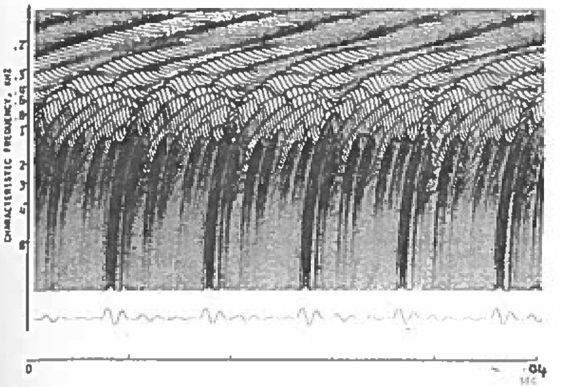


Fig.3: Spatio-temporal responses to synthesized vowel /a/. F0=130 Hz; F1=730 Hz; F2=1000 Hz; F3=2440 Hz.

The Central Processing of Auditory-Nerve Responses

This stage involves the extraction and utilization of the perceptually relevant cues from the response patterns of the cochlear nerve. Conceptually, it is a particularly difficult problem because the nerve patterns contain a rich variety of cues pertaining (in unknown ways) to a multitude of perceptual tasks. Thus, in studying a particular encoding scheme on the auditory nerve, or in implementing algorithms for automatic speech recognition applications, *a priori* decisions have to be made as to the appropriate response measures that need to be used and the ways these are to be combined. For instance, in the estimation of the spectral parameters of speech (e.g. formants) several measures have been proposed that range from purely spatial, i.e. discarding the fine temporal structure of the nerve responses (e.g. using the distribution of the *average rate* profiles across the tonotopically organized nerve-fiber array), to purely temporal, i.e. utilizing primarily the periodicities in the response as measures of the spectral content (e.g. the dominant frequency algorithm) [10]. Others in between include the Average Localized Synchronous Rate (ALSR) [3] and the Generalized Synchrony Detector [11].

An alternate approach is to view the response patterns essentially as 2-D spatio-temporal images with specific morphological features acting as spectral cues. One such feature, for instance, are the *edges* in the profiles of activity across the spatial axis created by one or both of the amplitude and phase changes eluded to earlier [5, 7]. The strength and position of the edges along the tonotopic axis are related to the signal spectral parameters through the dependence of the above two response characteristics on the frequency and amplitude of the stimulus (or its resolved harmonics in case of complex sounds). Edge detection algorithms, based on realistic biological lateral inhibitory network (LIN) topologies, can be used to extract these features and thus signify the spectrum of the underlying acoustic stimulus [5]. The LIN possesses several desirable properties which include: (1) A spatially distributed structure which is naturally suited for fast parallel processing implementations; (2) A robust performance in the presence of certain severe stimulus and/or channel distortions. The latter point is illustrated in the LIN outputs of Figs.4 under three conditions: (a) Moderate stimulus levels where few channels are saturated. (b) 40 dB higher stimulus levels where most channels are saturated: Despite channel saturation, the edges in the cochlear response patterns remain intact, and so do the LIN outputs near F1-F4 (These should be compared to the spectrograms of Fig.2.b). (c) Fig .4c simulates the case where the channel nonlinearity has a large slope [a], and the response waveforms become highly saturated. The outputs here are derived by a spatial first-difference operation evaluated *only* at the spatial zero crossings of the response pattern. The F1 and F2 are still extracted, though higher formants are now lost.

Acknowledgements

This work is supported in part by an Initiation grant from NSF, by the Mathematical Research Branch (NIH), and by a grant from the Minta Martin Foundation.

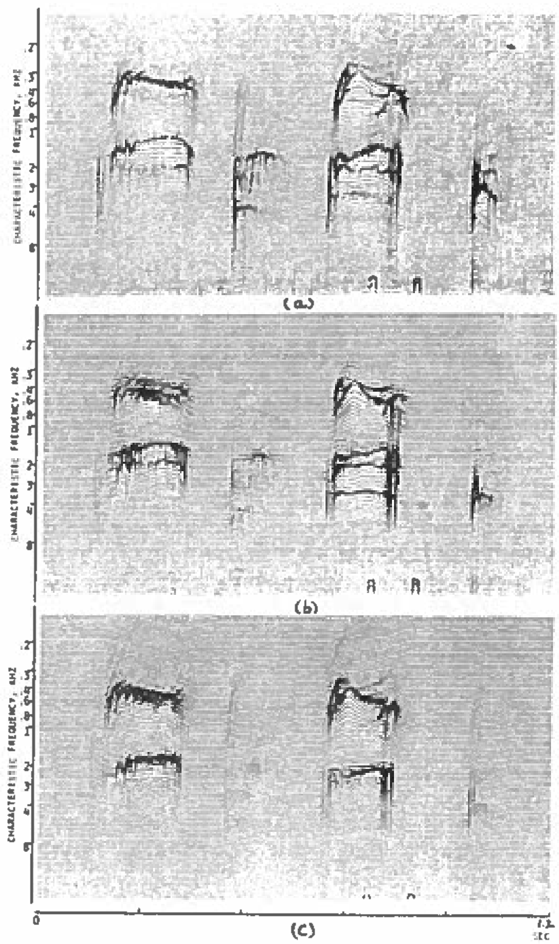


Fig.4: LIN estimates of spectral parameters of /b t/ whose spectrograms are shown in Fig.2. Parameters of the LIN network are published elsewhere [5]. (a) LIN outputs for moderate stimulus levels. (b) LIN outputs for high stimulus levels. (c) LIN outputs for severly saturated channels.

[1] M. B. Sachs and E. D. Young, "Encoding of steady state vowels in the auditory-nerve: representation in terms of discharge rate," *J. Acoust. Soc. Am.* vol. 66, pp. 470-479 (1979).

[2] R. R. Pfeifer and D. O. Kim, "Cochlear Nerve Fiber Responses: Distribution Along the Cochlear Partition," *J. Acoust. Soc. Am.* vol. 58, pp. 867-869 (1975).

[3] E. D. Young and M. B. Sachs, "Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Am.* vol. 66, pp. 1381-1403 (1979).

[4] M. I. Miller and M. B. Sachs, "Represtentation of Stop Consonants in the Discharge patterns of Auditory-Nerve Fibers.," *J. Acoust Soc. Am.* vol. 74, pp. 502-517 (1983).

[5] S. Shamma, "Speech processing in the auditory system. II: Lateral inhibition and the processing of speech evoked activity in the auditory-nerve ," *J. Acoust. Soc. Am.* vol. 78, pp. 1622-1632 (1985).

[6] B. Delgutte, "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds," *J. Acoust. Soc. Am.* vol. 75, no. 3, pp. 879-886 (1984).

[7] S. A. Shamma, "Speech Processing in the auditory System. I: Representation of speech Sounds in the responses of the auditory-nerve," *J. Acoust. Soc. Am.* vol. 78, pp. 1612-1621 (1985).

[8] S. A. Shamma, R. Chadwick, J. Wilbur, and J. Rinzel, "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," *submitted to the J. Acoust. Soc. Am.*, (1986).

[9] M. H. Holmes and J. D. Cole, "Cochlear mechanics: analysis for a pure tone," *J. Acoust. Soc. Am.* vol. 76, no. 3, pp. 767-778 (Sept. 1984).

[10] D. G. Sinex and C. D. Geisler, "Responses of Auditory-Nerve Fibers to Consonent-Vowel Syllables," *J. Acoust. Soc. Am.* vol. 73, pp. 602-615 (1983).

[11] S. Seneff, "Pitch and spectral estimation of speech based on auditory synchrony model," Working Papers on Linguistics, MIT (1984).

[12] V. Zue, "Speech Spectrogram Reading," Lecture Notes and Spectrograms, MIT (1985).

USING AUDITORY MODELS FOR SPEAKER NORMALIZATION IN SPEECH RECOGNITION

Anthony Bladon

Phonetics Laboratory, University of Oxford, 41 Wellington Square, Oxford OX1 2JF, U.K.

Auditorily-transformed versions of the speech spectrum may well be a useful way of reducing the apparently nonuniform physical differences between speakers. A speaker normalization technique of this kind is however justified to different degrees by different kinds of speech event. Does this presuppose a need for higher-level (phonetic class) information at the acoustic level in speaker-independent ASR?

> "It is obvious from our experiment that the unqualified assumption does not hold - auditory models used as speech recognition front ends will not consistently improve performance."

Blomberg et al.'s (1984) ominous words are ones which this symposium ought to take seriously to heart. They conflict with our initial theoretical expectations. This paper will not attempt to investigate what reasons lie behind the inconsistent results which some authors have found. Rather, we will focus on an aspect of the speech recognition task where the prognosis for auditory modelling promises to bear some fruit, namely, speaker differences (in speaker-independent speech recognition).

Speaker normalization for vowels

Normalizing the acoustic differences found between speakers - to take the best known example, differences of formant frequency in male and female vowels - used to be a formidable prospect. Fant showed how formant frequency differences were not just sex-specific, but also formant-specific and vowel-specific too. Methods of normalizing these data based upon reconstructing vocal tract shape fell foul of the problem that the solution to this exercise is nonunique. But if we apply auditory insights to the question, and compare not measured acoustic formants but auditorily transformed spectra, it can be shown that the nonlinearities which plagued Fant's data largely disappear. We argued (in Bladon et al., 1984) that the application of an auditory model which includes an auditory filter and a Bark scale, together with a displacement notion which has a simple physiological analogue, combine to generate a high degree of spectral match between male and female vowels. A large quantity of data, assembled by us and by others across a range of dialects and languages, has broadly supported this contention. Examples of vowels normalized in this way can be found in the above reference, and will be shown in the symposium.

How far is it worthwhile to extend an auditory model of speaker normalization beyond the vowel sounds? The theoretical answer seems to be: in part. At the present stage of research this answer has to be arrived at largely by inference from scattered pieces of the work of others, supported by some sporadic experimental confirmation of our own.

Voiceless vowels

Schwartz and Rine (1968) demonstrated that listeners could confidently identify a speaker's sex from individual steady-state vowels which were whispered. This is a finding of interest because it demonstrates that the role of voice pitch in speaker normalization is not a necessary one (though this does not exclude the possibility that pitch may have an ancillary role). As a result, the spectral characteristics of the whispered vowels are firmly implicated as a source for the listener's ability to identify sex.

Transforming the Schwartz and Rine whispered vowel spectra into auditory representations enables us to judge the effect of normalizing them by our method. It turns out that this procedure neutralizes much of the male/female difference. Whispered vowels, then, should be encompassed straightfowardly in an auditory model for speaker-independent ASR. It is not just voiced sounds which differ across speakers.

Plosives

This being so, what of plosives (voiced and voiceless)? The burst spectrum, widely believed to be of service as a differentiator of place in plosives, appears not to be a candidate for normalization. This statement derives from work in progress at Amsterdam by Weenink and remains to be fully confirmed. Weenink is finding that, while the plosive burst spectrum is sufficient to identify the plosive place in 85% of cases (thus corroborating the position long held by Stevens and others), listeners cannot identify the speaker's sex from the burst spectrum. When we recall the well-known templates for burst spectra, it is not difficult to guess why plosive bursts carry so little speaker information. The burst spectra are very variable, partly due to phonetic context; consequently the templates which fit each plosive are large, extensive both in frequency and in amplitude.

Even so, there is some evidence that normalization is appropriate for plosives, in respect not of their bursts but of their transitional spectra. This evidence comes from both production (O'Kane, 1984) and perception (Rand, 1971). Rand showed how, in a synthetic plosive-vowel sequence, the onset of formant transitions was at a frequency position which varied with speaker type. (His speaker types were "a large vocal tract" and "a small vocal tract".) He deduced that the same applies to the plosive locus frequency. In fact, unnoticed by Rand, the average [d] onsets needed to be 1.1 Bark different. It is striking, and unlikely to be coincidental, that this difference is reminiscent of an auditory displacement of the same magnitude which we have been discovering in vowel sounds.

The second piece of evidence is the measurements by O'Kane (1984) of locus frequencies, from the Australian English plosives spoken by 5 males and 5 females. She reported the overall locus ranges only in fairly gross terms: and, of course, ranges can give a misleading picture of the typical behaviour. Nevertheless, once again, when converted to a Bark scale, the female measured plosive loci can be seen to exceed the male values by a generally constant amount. One Bark would be a representative value. And so, while noting that plosive transitions have so far been only superficially investigated, it may be concluded that plosive transitions look like conforming to the normalization pattern.

Liquids and nasals

For many other classes of speech event there is at present no known evidence which would indicate how far, if at all, they are susceptible to variation with speaker-type, and hence, how far normalization is called for. This applies to laterals, nasals and trills, for instance. Prima facie, since these sounds have a prominent spectral content, they may possibly also carry the speaker-type information in a similar way to vowels. Alternatively, it may be that the spectral content in a nasal, with its large number of heavily damped formants, may be too elusive to have a clear auditory image which could be used in a normalization role. Pending further work, these matters have to be left open.

Fricatives

For fricatives, on the other hand, there is some well-documented evidence. Initially we will consider just the sibilant fricatives such as [s, ʃ, ç]. Schwartz (1968) published illustrations of speaker sex difference among voiceless English [s] and [ʃ]. Once again, we find that a conversion to auditory spectra leads to a greatly improved congruence of spectral shape.

Male and female [s] spectra were also investigated by us in British English. From a tightly controlled database and in an identical linguistic context, 55 male tokens (from five speakers) were compared with the same number of female tokens. Auditory spectra of these fricatives confirmed the tendency to congruence noted in the Schwartz data and further revealed that an especially constant feature of [s] was the (15 phons/Bark) low-frequency edge of the [s] peak. As with vowels and other sounds, this edge is so located as to suggest a constant male/female normalizing factor in auditory space.

Whether this behaviour extends to fricatives other than the sibilants mentioned is currently a matter of uncertainty: the basic work remains to be done. A fairly confident summary would be as follows. It is known from the study by Ingemann (1968) that speaker sex is identifiable from steady-state productions of the glottal fricative [h], with an accuracy comparable to that of the sibilants. Also identifiable at better than chance accuracy, according to the same study, are uvular [χ] and velar [x]. Spectra of these back fricatives show a somewhat vowel-like superimposition of vocal tract cavity resonances, and hence will be expected to behave in speaker normalization very much as vowels do. This is especially likely of [h] since the resonance patterns will not differ markedly from those of a whispered vowel. On the other hand the front fricatives [ɸ, f, θ] are not identifiable for sex. This is understandable, given that the front fricatives with little or no resonance cavity ahead of their friction source, do not have a very distinctive spectral shape. Intensity level is their prime cue. Speaker sex differences do not seem to exploit this.

Conclusion

Extrapolating somewhat beyond the rather superficial review above, it seems reasonable to say that, as a useful basis for speaker-independent ASR, an auditory model can in general be used to normalize the running-speech spectral shape. Fairly clear exceptions to this are the front fricatives (those which are more advanced than alveolar) and the plosive bursts, whose spectra appear not to be capable of signalling information on speaker type.

If this is so, then in an actual speech recognition system two empirically testable alternatives can be explored. One is the possibility that a decision on whether or not to normalize the currently incoming spectrum for speaker differences must be made, depending on a decision about its phonetic class. This alternative clearly implies the intervention of some higher-level expert. The other possibility is that no such decision needs to be made at all: the recognizer can safely normalize the whole signal, because those phonetic classes of event which do not show evidence of sex-based physical difference are anyway spectrally rather flat or heavily smeared.

In order to choose between these alternatives we propose to examine recognition test results to see whether (or how far) deterioration ensues, when the whole set of phonetic events in speech (as opposed to a partial set excluding front fricatives and plosive bursts) is first subjected to an auditorily-based normalization for speaker sex.

References

BLADON R.A.W., HENTON C.G. and PICKERING J.B. (1984a). Outline of an auditory theory of speaker normalization. In Van den Broecke M.P.R. and Cohen A. (eds.), Proceedings of the Tenth International Congress of Phonetic Sciences (Dordrecht, Foris), 313-317.

BLOMBERG M., CARLSON R., ELENIUS K. and GRANSTROM B. (1984). Auditory models in isolated word recognition. IEEE ICASSP 1984, 17.9.1-17.9.4.

INGEMANN F. (1968). Identification of the speaker's sex from voiceless fricatives. J. Acoust. Soc. Am. 44, 1142-1144.

O'KANE M. (1984). Extensions to the locus theory. In Van den Broecke M.P.R. and Cohen A. [see Bladon above], 331-337.

RAND T.C. (1971). Vocal tract size normalization in the perception of stop consonants. Haskins Labs. Stat. Rep. Sp. Res. 25/26, 141-146.

SCHWARTZ M.F. (1968). Identification of speaker sex from isolated voiceless fricatives. J. Acoust. Soc. Am. 43, 1178-1179.

SCHWARTZ M.F. and RINE H.E. (1968). Identification of speaker sex from isolated whispered vowels. J. Acoust. Soc. Am. 44, 1736-1737.

RECOGNITION OF WORDS WITH THE HELP OF THE SERAC-IROISE EXPERT SYSTEM

Xavier Marie, Martine Gérard, Guy Mercier

Centre National d'Etudes des Télécommunications, Centre Lannion A, Route de Trégastel, 22301 LANNION Cédex, FRANCE.

ABSTRACT

In order to test the performance of the acoustic-phonetic decoding module on the Serac-Iroise expert system, we have implemented a lexical analyzer, the function of which is to match each word of the task vocabulary against the phonetic hypotheses lattice. A one-stage dynamic comparison algorithm, initially designed for global recognition of connected words, has been adapted. Our knowledge-based approach makes it possible to improve performance significantly with the help of heuristics, e.g. concerning local constraints and the measure of similarity. Introducing phonological, syllabic and prosodic information into the lexicon allows refinement of the strategy by basing on islands of reliability. Such phonological phenomena as merging, spreading, insertion, deletion and confusion are dealt with in a rather flexible way : likelihood weights, penalty factors and thresholds of reliability are determined according to the most encountered recognition errors. The object - and rule-based representation gives advanced opportunities for system extension and modification.

1. Introduction

Conclusions after ARPA SUR and subsequent projects have led to reconsidering approaches to Automatic Speech Recognition (ASR). Separate contribution of the different knowledge sources are best modelized using Artificial Intelligence (AI) knowledge representation tools such as Production Systems (PS), that supply advanced features for formalizing expertise, comparing strategies and refining parameters and heuristics.

The KEAL system developed at CNET achieves multispeaker analytical recognition and interpretation of isolated sentences taken from a few hundred words'vocabulary. The SERAC system (Système Expert pour la Reconnaissance ACoustico-phonétique) has been designed to structure the knowledge acquired with Keal and to provide a flexible tool for maintaining, improving and extending it, eventually leading to a new ASR model.

The lexical analyzer MODEM (MOdule de DEtection de Mots) is dedicated to both validating the phonetic level and evaluating heuristics for connected word verification and techniques for representing lexical and phonological knowledge.

The Iroise system is a PS using an object-oriented problem-driven rule-based language of the OPS family, it consists in three functional components : the knowledge Base, the Inference Engine and the User's Interface.

An acoustic-phonetic module : feature extraction, sentence onset detection, centisecond labelling, segmentation into pseudo-syllables, segmentation into phones, hierarchical consonant and vowel recognition, and a prosodic module, which detects the extrema of the pitch and vocalic durations, and the type and main boundary of the sentence, have now been implemented in Serac : about 600 rules total.

2. Modem : a module for word detection

The principle of this detection is to match the phonetic lattice against phonetic frames taken from the set of possible words at a given instant, and finally keep the optimal sequence of words ; our heuristic comparison approach is based on a dynamic programming (DP) sequential algorithm and a measure of possible errors and similarity driven from lexical and phonological associated knowledge.

A main feature of MODEM is the intervention of phonetic, syllabic and phonological knowledge all along the process. The phonetic lattice shows as a sequence of phonetic segments (or frames) characterized with a set of weighed properties or attributes such as phoneme type (consonant, vowel, semivowel), cues like mode of articulation (fricative, plosive, nasal, liquid), place of articulation (labial, dental, velar, palatal, pharyngal), segment duration syllable number, and the best three phonetic hypotheses with associated confidence scores. In the IROISE representation language of structured objects, each phonetic frame is an instance of the frame class with fixed attributes.

3.1. The Lexicon

Each word in the lexicon also refers to contextual or phonemic information. For manageability's sake, the whole vocabulary is represented under the form of a unique list : the Lisp basic structure is thoroughly used. The functional notation makes access to the first element much easier, so special information is entered in the beginning of sublists referring to a particular object.

Elements of a 'phonemic frame' sublist are the possible realizations in decreasing likelihood order, preceeded with special makrs such as :

- obligatory : states the frame is an accentuated syllable's vocalic nucleus, and thus absolutely must be recognized within the best three hypotheses, and may not be omitted in any event ;
- optional : states the present frame can be dropped without penalty, as it is often in oral language ;
- non-optional : states the frames is to be fully penalized whenever omitted or ill-recognized.

The marks are very useful in determining the type of treatment to be applied during the similarity calculus. This highly modular representation makes automation easy when accessing very large lexicons, and enables the expert to introduce many other features of prosodic (pitch, duration), phonological (elision, nasalization) of phonotactic (phoneme combination rules) type.

3.2. Verification of words

3.2.1. The algorithm

It is originally a one-stage DP algorithm for connected word recognition (CWR), where words are considered as sequences of acoustic frames. The search space is a finite pattern of squares. The x-axis is the pronounced sentence, divided in N temporal frames (index i), while the y-axis is composed of M word templates (index k), each divided in J(k) frames (index j). A dissimilarity measure d(i,j,k) is used and a cumulative distance D(i,j,k) at point (i,j,k) is to be minimized to obtain the optimal sequence of templates (or optimal super-template).

1. Initialize D(1,j,k)= d(1,1,k)+...+d(1,J(k),k) ;
2a.for i:2..N do 2b-2e;
2b.for k:1..M do 2c-2e;
2c.D(i,1,k)= d(i,1,k)+min(D(i-1,1,k), D(i-1,J(k'), k'); k':1..M);
2d.for j:2..J(k) do 2e;
2e.D(i,j,k)= d(i,j,k)+min(D(i-1,j,k), D(i-1,j-1,k), D(i,j-1,k);
3.find k* such that D(N,J(k*),k*) be minimum, and trace back the path leading to (N,J(k*),k*).

Problems of i) practical implementation and ii) adaptation to recognition from phonemes are raised :

i) in order to reduce memory size, we use two column vectors Di(j,k) and Di-1(j,k) updated after every comparison, and two backpointer vectors Bi(j,k) and Bi-1(j,k) that state the instant when last template of the current supertemplate : T(i), beginning at frame F(i), terminates. The temporal complexity is M.N.J and the spatial complexity is 2(N+M.J) if J=moy(J(k)).

ii) modifications are to be introduced at the following levels :
- similarity measure (SM) between phonemes and dealing with phonological variations in a balanced way ;
- local constraints (LC) : choice of allowed transitions, introduction of penalty factors for phonological deformations ;
- normalization for keeping the SM homogene and optimal with the LC broadening ;
- heuristics dedicated to pruning the search and taking domain expertise into consideration.

3.2.2. Local constraints

They define the transition mode between two points of the search space. With Sakoe and Chiba's symmetrical LC, the path leading to (i,j) may come from :
(i-1,j): spreading ; (i,j-1): merging;
(i-1,j-1): normal ; (i-2,j-1): deletion;
(i-1,j-2): insertion,
segmentation errors and phonological phenomena being the main causes for abnormal cases. To normalize the cumulative similarity (CS) along a path, we use the following method: the length of a path always equals the sum L(i,j)=i+j of its ending coordinates. Penalizing abnormal transitions induces a strategy based on islands of reliability. Penalty factors depend on error type and template length; deletions and insertions are much more severely penalized as more likely to come from segmentation deficiencies.

3.2.3. Similarity

Given a point (i,j), the best path leading to it is determined using CS at points (i',j') from which transition is allowed, and similarity index (SI) s(i,j) between templates and the lattice :
S(i,j)= S(i',j') + (1-F) (L(i,j)-L(i',j')) s(i,j)
where s's factors are the penalty and normalization factors. The origin is the fictive point (-1,-1,-1) Points where a template terminates or almost terminates need a special treatment : the best among them are selected before being copied as points of -1 or -2 ordinate from where they will be reached without introducing any discontinuity in the DP process.

The SI is computed as the maximization of punctual similarity on every pair (lattice phoneme, template phoneme), the value of which is the phonetic score multiplied with the similitude between the two phonemes. The latter is computed once for all using an empirical measure : the C-V similitude is generally 0, while the V-V similitude is function of additive formantic distance, and the similitude between consonants is the weigthed mean of their hierarchized cues similitude : voicing, mode and place of articulation, with weights of (resp.) 3, 5 and 2, supposedly approximating the reliability of these cues detection.

3.3. Implementation

The objects used are the template and frame objects, representing information associated to the phonemic segment in the lexicon or in the phonetic lattice, and the path object, that characterizes the current search point : its attributes are : coordinates, type of transition leading to it, path length in the supertemplate, backpointer values, special marks, CS and SI. Loop control variables are also represented as objects.

The strategy of detection holds three stages :
- process control problems for computing general data (similitude matrices, similarity thresholds, penalty factors), loading files, commanding the DP loops, instantiating objects, pruning the search, displaying results and normalizing the distance ;
- search for adequate transitions with anterior path, according to the LC and existing marks ;
- path evaluation problems that compute the SI, select the best transition and validate the pat, with the help of heuristics for improving speed or deleting ill paths.

CONCLUSION

This makes 10 problems for about 60 rules total, calling a number of Lisp functions ; this skeleton is presently being extended to introduce new knowledge and efficient heuristics. A very useful development basis is supplied for evaluating phonetic decoding and adjusting heuristics able to improve lexical search in a general frame for ASR.

REFERENCES

(MER84) MERCIER G., GILLOUX M., TARRIDEC C., VAISSIERE J. "From Keal to Serac : A New Rule-Based Expert System for Speech Recognition" NATO/ASI, Bonas, Jul.1984.
(NEY85) NEY H. "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition" IEEE ASSP 32, no 2, Apr. 1982.
(SAK78) SAKOE H., CHIBA S. "Dynamic Programming Algorithm Optimization for Spoken Word Recognition" IEEE ASSP 26, n° 1, Feb. 1978.
(VIV85) VIVES R. "Mise en Correspondance Temporelle de Descriptions Phonologiques Syllabiques et Prosodiques de mots dans le Système de Reconnaissance de la Parole Keal" 14e JEP, Paris, 1985.

HIERARCHICAL RECOGNITION OF FRENCH VOWELS BY EXPERT SYSTEM IROISE-SERAC

Anne Bonneau, Mario Rossi

Institut de Phonétique, Aix-en-Provence (France)

Guy Mercier

CNET, Lannion (France)

ABSTRACT

We are presenting here an implementation of a French vowel recognition program under IROISE, an expert system for acoustic-phonetic decoding, used in CNET. The rules for recognition are based on polycontextual non-formantic cues ; the data are output from a 14-channel vocoder. The algorithm is represented by a binary tree with 37 hierarchized cues.

A rule under IROISE represents a branch of the tree. The first one follows the branch defined only by positive cues ; the second one puts the list of the first rule in its contextual part by eliminating the last cue. If the rule is applied we know that this cue is negative, because the preceeding rule was not set off, and we modify the cue's polarity. With this method, only the cues tested in the recognition phase will have the value "false".

Under IROISE, all cues are systematically tested even if they are not all used in any particular execution of the program. Then we call the algorithm in which every rule represents a branch.

We furnish the recognition results using this program on an initial corpus of 330 words pronounced by five male speakers and the results using rules under IROISE on digits pronounced by other speakers.

1. PRESENTATION DE SERAC-IROISE

SERAC est le module de reconnaissance acoustico-phonétique utilisant le langage du système-expert IROISE.

SERAC réécrit, en utilisant au mieux les possibilités de formalisation des connaissances offertes par IROISE, le module de reconnaissance acoustico-phonétique du système de reconnaissance de la parole KEAL, actuellement implanté en langage C sous UTS (IBM 3083).

En outre, il le complète en lui adjoignant de nouveaux modules de reconnaissance écrits par d'autres experts.

Le module de reconnaissance phonétique "KEAL-SERAC" commence par lire les échantillons spectraux ; les données acoustiques sont fournies par les analyses spectrales numériques effectuées toutes les 13,3 ms, par un vocodeur à 14 canaux ; il en extrait les paramètres acoustiques. Le paramètre le plus important, pour les programmes que nous implantons, est le vecteur des énergies, appelé "en", qui est constitué de la valeur de l'énergie dans chacun des 14 canaux du vocodeur pour un échantillon temporel donné.

Le module actuel de reconnaissance effectue l'étiquetage phonétique des échantillons, la segmentation en syllabes et en noyaux vocaliques, la reconnaissance des macro-classes... Nous intervenons après le module de détection des noyaux vocaliques.

2. RECONNAISSANCE DES VOYELLES

2.1. Le programme de reconnaissance des voyelles

La recherche des indices a été menée de deux façons : (1) par la méthode d'essai-erreur, (2) au moyen d'une analyse factorielle des correspondances. Le corpus comprenait 300 mots de forme CVCV enregistrés deux fois par 5 locuteurs masculins d'une moyenne d'âge de 25 ans.

La reconnaissance s'effectue selon un mode binaire dans une arborescence où tous les indices sont hiérarchisés : au total, 37 indices représentant essentiellement les traits Ouvert/Fermé, Aigu/Grave, Bémolisé/Non bémolisé, Nasal/Non nasal, Périphérique/Non périphérique. Ils sont fondés sur les variations d'énergie dans le spectre et calculés soit sur la partie centrale de la voyelle, soit sur plusieurs parties de celles-ci, dans le cas des voyelles nasales par exemple, caractérisées par la présence de deux segments distincts, un segment oral et un segmen nasal localisé dans les deux derniers tiers de la voyelle. En général, les voyelles sont tronquées de 20 % aux bornes afin d'éviter de prendre en compte les parties transitoires et de ne conserver que la partie stable de la voyelle.

Les indices portent, par pure commodité, l'étiquette d'un trait, mais il n'existe pas de relation biunivoque entre indices et traits. Les voyelles ne sont pas reconnues par traits mais par configurations d'indices.

2.2. Détermination des indices dans SERAC

Nous avons créé un nouvel objet, "phone-voy", qui représente la voyelle.

Chaque indice est un attribut de "phone-voy" et prend la valeur "vrai", "faux" ("inconnu" au lancement du programme). L'indice est détecté sur une partie déterminée de la voyelle ; les limites temporelles sont également des attributs de "phone-voy".

On teste si l'indice, ou plutôt son attribut, est "vrai" par un problème particulier qui porte le nom de l'indice testé. Nous avons regroupé dans un même fichier tous les problèmes qui testent la valeur d'indices d'un même trait.

Cinq fichiers sont créés pour : (1) les indices d'acuité, (2) les indices de bémolisation, (3) les indices d'ouverture, (4) les indices de nasalité, (5) les indices périphériques.

Dans la plupart des cas les indices sont détectés par des règles simples. Le langage Lisp peut coexister avec IROISE pour les règles plus complexes qui exigent en particulier des itérations.

2.3. Algorithme de reconnaissance des voyelles

Après l'évaluation de "n10", dernier indice testé, l'ensemble des attributs des indices positifs possèdent la valeur "vrai". Les autres sont implicitement "faux". On peut alors déclencher l'algorithme de reconnaissance.

Chaque règle de celui-ci correspond à un des chemins de l'arbre défini dans le par. 2.1. La première règle représente le chemin qui est défini uniquement par des indices positifs : la partie "contexte" de la règle est écrite de la manière suivante :

si (phone-voy ?pv (indic1 vrai) (indic2 vrai)...)

La deuxième règle reprend dans sa partie "contexte" la liste de la première en éliminant le dernier indice de celle-ci. On sait alors, si la règle est appliquée, que cet indice est négatif puisque la règle précédente n'a pas été déclenchée - les règles sont incompatibles entre elles - et on modifie la valeur de l'indice. Par cette méthode, seuls les indices effectivement testés lors de la reconnaissance d'une voyelle déterminée auront la valeur "faux", les autres resteront à "inconnu".

On réitère le processus jusqu'à ce que l'algorithme soit tout entier écrit sous forme de règles. On peut, de cette façon, créer un seul problème qui traite l'ensemble de l'algorithme.

2.4. Les indices fondamentaux d'antériorité (acuité) et d'ouverture (compacité)

Le premier axe de l'analyse factorielle des correspondances représente le trait Aigu/Grave, l'étude des corrélations entre les groupes de voyelles et les canaux permet de diviser le spectre en deux bandes fréquentielles : 650 à 1600 Hz, 1600 à 3400 Hz, à partir desquelles est calculé le principal indice d'acuité appelé AIGU1. L'indice recherche et compare les maxima spectraux dans chacune des bandes ainsi délimitées. En effet, l'énergie caractéristique des voyelles graves [u,o,o] est située dans la bande 650-1600 Hz tandis que celle des voyelles aiguës est située au-delà de cette bande. Il semble que l'énergie caractéristique de [a] soit située à la frontière mais cette voyelle peut apparaître dans l'une ou l'autre classe en fonction du contexte. [a] se comporte comme les voyelles vélaires sauf au contact des consonnes vélo-palatales [k-g] derrière lesquelles apparaît sa variante aiguë. La reconnaissance de la voyelle [a] dans la classe des /+ouvertes, +aiguës/ est accompagnée de la spécification du contexte vélo-palatal qui se trouve vérifiée dans 90 % des cas.

Le trait d'ouverture n'est pas représenté par le deuxième axe, plus complexe, mais l'examen du spectre des voyelles fermées met en évidence l'existence d'une zone d'anti-résonance dans les canaux 3 à 4 (650-1050 Hz) qui disparaît au fur et à mesure que F1 s'élève et que la voyelle devient plus ouverte. D'où l'indice d'ouverture "ouv1" :

si $EK1_2$ EK3 + EK4 alors -ouv1

Ki : ième canal. EKi : énergie dans le ième canal.

Les seules voyelles dont la classification pose un problème par cet indice sont [ ] et [u]. Pour [u], cet échec s'explique par la présence du F2 dans les canaux 3 et 4 qui réduit l'importance relative de F1.

Un second indice est alors proposé pour forcer [u] dans la classe des voyelles fermées. Cet indice, appelé "ouv4" permet de mettre en relief la proéminence du F1 :

si (EK1 - EK2) (EK3 + EK4) - EK1 alors -ouv4

Pour certains locuteurs, la voyelle [ ] est produite comme une voyelle fermée, en conséquence sa reconnaissance est prévue dans les deux classes vocaliques, ouvertes et fermées.

Le taux de reconnaissance des voyelles par l'ensemble des indices sur le corpus défini dans le parag. 2.1 est évalué à 86 % : 1 candidat est présenté dans 60 % des cas, 2 candidats dans 40 % des cas.

2.5. Reconnaissance des macroclasses

Le module de reconnaissance des traits vocaliques pour la reconnaissance des consonnes occlusives (A. Bonneau 1984), vise à regrouper les voyelles en quatre grandes classes vocaliques, selon les traits "ouvert-fermé" et "aigu-grave" :

a) Voyelles aiguës : /i,e, ,y/
Voyelles graves : /u,o, , , , /
[a, , oe, ø] peuvent, selon le contexte dans lequel ils apparaissent, appartenir à l'une ou l'autre classe.

b) Voyelles fermées : /i,u,y,(e),(ø),(o)/
Voyelles ouvertes : / , , ,a, ,( ),( ),(oe)
Il est difficile de déterminer a priori si, dans une syllabe donnée, en particulier les syllabes atones, le locuteur a prononcé la variante ouverte ou fermée de [e,] ; [o,] ; [ø,oe], c'est pourquoi ces phonèmes, mis entre parenthèses ci-dessus, ne sont pas pris en compte dans les % de reconnaissance selon l'indice d'ouverture.

3. RESULTATS ET CONCLUSIONS

SERAC-IROISE est écrit en Lisp dans le dialecte COMMON LISP, sur VAX 11/780, sous VMS. Le corpus choisi pour l'évaluation est constitué de 139 nombres (de 0 à 999) prononcés par 6 nouveaux locuteurs masculins. L'application des règles de reconnaissance à ces nouveaux locuteurs permet de tester dans quelle mesure les indices utilisés sont indépendants du locuteur :

Le pourcentage de reconnaissance pour les voyelles, pour ce nouveau corpus s'établit comme suit :

* 72 % de reconnaissance pour les nombres prononcés en parole continue. Le fort pourcentage des voyelles nasales [ ] dans les nombres est responsable de la chute du taux de reconnaissance ; cette voyelle, en effet, est la moins bien identifiée du système français. Sans la présence de cette voyelle, le taux de reconnaissance s'élèverait à 86 %. Une ou deux réponses (deux dans 37 % des cas) sont données pour l'identification de la voyelle à reconnaître. Le pourcentage de reconnaissance pour les traits aigu-grave et ouvert-fermé est de 97 %.

REFERENCES

1. Rossi, M., Nishinuma, Y., Mercier, G. (1983), "Indices acoustiques multilocuteurs et indépendants du contexte pour la reconnaissance automatique de la parole", Speech communication, North-Holland, pp. 215-217.
2. Bonneau, A. (1984), "Indices de reconnaissance des consonnes occlusives sourdes du français, en vue d'une application à la reconnaissance automatique de la parole", Thèse de doctorat de troisème cyle, Aix-en-Provence.
3. Mercier, M., Gilloux, M., Tarridec, C., Vaissière, J. (1984), "From Keal to Serac : a new rule-based expert system for speech recognition", Nato Advance Studies Institute, Bonas, Gers, France.

REPRESENTATION D'UN LEXIQUE POUR LA R.A.P.C A L'AIDE DE CONNAISSANCES PHONOLOGIQUES

J. Gispert, H. Meloni

G.I.A. Faculté de Luminy
70, route Léon Lachamp
13288 Marseille Cedex 09 France

I. INTRODUCTION

Le système présenté fait partie d'un ensemble complet de Reconnaissance Automatique de la Parole Continue. Il fait suite à l'étude présentée par (Méloni 1985).

Son but est de permettre la reconnaissance de mots à partir d'un treillis de phonèmes produit par le décodage acoustico-phonétique.

Le système porte sur la composante phonétique du lexique ; la composante graphique nécessaire à la sortie des résultats sera ajoutée ultérieurement sous forme d'un nouveau module.

Les règles de Phonologie Générative, énoncées déclarativement, permettent la dérivation des formes phonétiques des mots à partir de leur forme sous-jacente construite par combinaison d'éléments lexicaux. Parallèlement, des règles morpho-syntaxiques permettent de déduire les informations diverses provenant du regroupement des morphèmes.

Les accès aux mots représentés par ces structures sont codés séparément. Ils sont constitués au moyen d'informations syntaxiques et phonétiques, et des données fournies par le décodage Acoustico-Phonétique.

Le codage du lexique est automatique ; toutefois, certaines ambiguïtés doivent être levées manuellement.

Un ensemble de prédicats permet d'extraire toutes les informations codées : structure phonologique, morphèmes, caractéristiques syntaxiques, forme phonétique.

II. CONNAISSANCES UTILISEES

On ne peut envisager de coder explicitement chaque mot sous sa forme phonétique pour conserver au lexique un volume raisonnable. Il convient donc de se tourner vers un système linguistique tirant partie des parentés entre mots différents, construits à l'aide d'un radical et de préfixes et suffixes éventuels.

Dans le cas des conjugaisons, (Bescherelle 1980) fournit un catalogue des particularités verbales, selon une approche descriptive. J. Pinchon (Pinchon 1981) opère quelques regroupements et aboutit à un ensemble de formes un peu plus restreint.

La Phonologie Générative (Chomsky 1973) offre un cadre analytique dans lequel les parentés entre mots trouvent des explications par le biais de variations phonologiques. Nous avons donc choisi cette approche, pour laquelle on dispose, pour le Français, des travaux de (Schane 1968), (Plénat 1981,1985) et (Dell 1973,1984). On traite également des connaissances lexicales :

- découpage des mots en morphèmes,
- groupements de morphèmes possibles ou interdits,

et des connaissances grammaticales :

- catégories et attributs syntactico-sémantiques.

III. CODAGE DES REGLES

Les règles utilisent les traits articulatoires suivants : *consonantique vocalique haut bas avant rond nasal tendu*.

Chaque phonème dérivé ou sous-jacent est décrit par une clause associant son identificateur à son vecteur de traits, dans l'ordre ci-dessus.

Exemple :

<*aa,vocalique.bas.tendu*> → ;
<*Aa,vocalique.bas*> → ;

Aa représente le phonème /*a*/ sous-jacent lâche, et *aa* le phonème /*a*/ tendu, sous-jacent ou dérivé.

Des prédicats permettent :

- d'atteindre un trait dans la liste, par exemple *trait-cons(l,c)* donne la valeur positive ou negative du trait consonantique de *l*,
- de caractériser des classes de phonèmes (voyelle-orale, consonne ...).
- de modifier la liste de traits selon les besoins des règles (négation, affirmation, échange ...).

Les règles sont codées à l'aide de clauses Prolog sous une forme très proche de celle proposée par les linguistes. La tête de clause présente la séquence de phonèmes initiale et celle qui en dérive. La queue de la clause détermine les contraintes contextuelles d'application de la règle et les liens existant entre les deux formes.

Exemple :

nasalization(v.c.q,v'.q) →
  voyelle-orale(v)
  consonne-nasale(c)
  nasaliser(v,v')
  non-vocalique(q) ;

IV. INTERPRETATION DES REGLES

Les règles que nous utilisons sont (partiellement) ordonnées. Des méta-règles déclaratives décrivent cet ordre.

Chaque règle, à son tour, est appliquée de toutes les manières possibles sur la chaîne à transformer. La stratégie de synthèse force l'application d'une règle dès l'instant où son environnement est satisfait, puisque ceci correspond à la possibilité d'un phénomène phonologique.

Par contre, en analyse, il ne convient pas d'appliquer systématiquement toutes les règles possibles. En effet, certains phonèmes sont à la fois sous-jacents et dérivés ; ils peuvent figurer tels quels dans la forme phonologique ou bien être obtenus par dérivation. Il existe donc diverses formes sous-jacentes conduisant à la même forme phonétique. Les connaissances phonologiques ne peuvent résoudre cette ambiguïté, qui est levée par consultation du lexique.

Le système doit être capable de reconnaître une phrase, même prononcée en violation de certains phénomènes comme par exemple les liaisons. Les règles qui décrivent ces phénomènes pourront être facultatives. On admettra ainsi des formes phonétiques fausses sur le plan théorique, mais d'emploi assez fréquent.

V. CODAGE DU LEXIQUE

V. 1- Représentation des Morphèmes

Les éléments lexicaux sont des morphèmes. Ils sont représentés par des clauses indiquant la nature du morphème (prefixe, radical, marques de genre et de nombre, ...) et des données dépendant de cette information.

- L'identificateur du morphème est construit avec les phonèmes qui le constituent,
- les suffixes portent une information indiquant la catégorie syntaxique qu'ils produisent,
- les radicaux décrivent l'ensemble des mots construits autour d'eux, grâce à un terme qui indique les préfixes et suffixes possibles. Ce terme est constitué de doublets <préfixe,suffixe>, des opérateurs *et* et *ou* et de la fonction *event* (éventuellement),
- tous les morphèmes peuvent comporter des données particulières sur l'emploi des règles.

Exemple :
ddOoll(radical,ou(<vide,ou(vide,et(Oorr,event(Oozz)))>,
<Aann,et(Oorr,iirr)>))) → ;
Oorr(suffixe,nom) → ;
Oozz(suffixe,adjectif) → ;
Aann(prefixe,appris) → ;
iirr(suffixe,<verbe(3),appris>) → ;

V. 2- Codage des accès

Le radical permet d'accéder à tous les morphèmes constituant les mots qui lui correspondent. Par contre, les préfixes et suffixes n'offrent pas cette possibilité, le nombre de radicaux associés à chacun étant trop grand.

Pour la décomposition d'un mot, il est préférable d'avoir accès à chacun de ses morphèmes pour éviter de rechercher le radical n'importe où. L'analyse se fait de gauche à droite par identification de préfixes, puis du radical et enfin de suffixes. Cette recherche est bien entendu non déterministe.

Le phonème le plus à gauche dans un morphème donne un accès naturel pour une analyse de gauche à droite.

Exemple :
acces-morpheme(Oo,Oorr,Oo.rr.nil) → ;
acces-morpheme(dd,ddOoll,dd.Oo.ll.nil) → ;
acces-morpheme(Aa,Aann,Aa.nn.nil) → ;
acces-morpheme(ii,iirr,ii.rr.nil) → ;
acces-morpheme(Oo,Oozz,Oo.zz.nil) → ;

VI. TRANSFORMATION DES REGLES

Les résultats présentés par (Gispert 1986) montrent la nécessité de transformer les règles.

L'utilisation d'une règle choisie sur un contexte totalement indéterminé, produit les formes les plus générales transformées par cette règle. En appliquant à ces formes toutes les règles possibles de façon non déterministe, jusqu'à obtenir d'un côté une forme phonologique et de l'autre une forme phonétique. on définit tous les usages qu'il est possible de faire de cette règle. A chaque solution, on fait correspondre une macro-règle qui représente l'enchaînement des règles qui l'ont produite.

L'usage de ces macro-règles est possible grâce à des accès définis sur les phonèmes gauches des deux formes concernées. L'analyse d'un mot se fait maintenant ainsi :

- accès à une macro-règle par le premier phonème,
- unification de la forme phonétique donnée par la règle avec le mot à analyser,
- accès à une autre macro-règle par le premier des phonèmes restant à analyser, etc.

Avec ces méta-règles, les temps de calcul sur VAX 750 sont de l'ordre de la seconde.

Le remplacement des règles phonologiques par des macro-règles revient à déduire un catalogue des différents cas particuliers. Cependant, ce catalogue est obtenu automatiquement à partir des connaissances que les linguistes souhaitent manipuler. Il ressort donc que ce système peut convenir à la fois à la mise au point d'un jeu de règles et à son exploitation en situation de reconnaissance. Ceci justifie le détour par les règles de phonologie.

VII. CONSTRUCTION AUTOMATIQUE DU LEXIQUE

Un mot nouveau est proposé au système sous sa forme phonétique, avec ses attributs syntaxiques (catégorie, type d'emploi...). Le système en fait d'abord l'analyse phonologique, qui propose une forme sous-jacente dont il pourrait dériver.

L'analyse en morphèmes de cette forme ne peut être envisagée de toutes les manières possibles sans référence à des morphèmes connus. On imposera donc que les préfixes, suffixes et désinences soient tous répertoriés à priori dans le lexique. Seul le radical pourra donc être inconnu.

Ainsi ne peuvent subsister que certaines ambiguïtés d'analyse résultant de la confusion d'une partie du radical avec un préfixe ou un suffixe existant. Le choix de l'une ou l'autre forme est déterminé par d'autres mots de la même famille, qui possèdent le même radical. Ces ambiguïtés sont levées par l'utilisateur qui doit fournir au système un mot de même famille.

VIII. CONCLUSION

Cette étude met en évidence certains aspects de l'approche choisie par rapport au traitement automatique du problème :

- l'interprétation des connaissances proposées est délicate, certains aspects n'étant pas explicités,
- il est difficile de mélanger des règles provenant d'auteurs différents, celles-ci utilisant par exemple des jeux de traits différents,
- le système a permis de valider l'hypothèse de faisabilité sous Prolog (moyennant la compilation des règles),
- il constitue un outil de test pour de nouvelles théories phonologiques que l'on pourrait appliquer de la même manière.
- il est utilisable en reconnaissance

IX BIBLIOGRAPHIE

Bescherelle
Le Nouveau Bescherelle Hatier 1980

Chomsky N., Halle M.
The Sound Pattern of English Cambridge, Mass. MIT Press 1968

Dell F.
Les règles et les Sons, introduction à la Phonologie Générative Hermann Paris 1973

Dell F., Vergnaud J.-R.
Les développements récents en Phonologie : quelques idées centrales Forme Sonore du Langage Hermann 1984

Gispert J.
Représentation d'un Lexique pour la R.A.P.C. à l'aide de Connaissances Phonologiques
Séminaire Lexique GRECO GALF Toulouse Janvier 1986

Meloni H., Gispert J., Guizol J.
Un Système Expert pour l'Identification Analytique de Mots dans le Discours Continu. 5èmes Journées Internationnales Systèmes Experts Avignon 1985

Perennou G.
Base de Données Lexicale Rapport scientifique
GRECO communication parlée Juin 1984 CRIN Nancy 1983

Pinchon J., Coute B.
le Système Verbal du Français Nathan 1981

Plenat M.
L'autre conjugaison, ou de la régularité des verbes irréguliers
Cahiers de Grammaire n°3 Avril 81 Centre de Sociologie et de Dialectique Sociale Toulouse 1981

Plenat M.
Sur quelques aspects de la nasalisation en Français standard
Cahiers de Grammaire n°9, Toulouse 1985

Schane S.A.
French Phonology and Morphology Cambridge, Mass. MIT Press 1968

UN SYSTEME D'APPRENTISSAGE SYMBOLIQUE POUR LE DECODAGE ACOUSTICO-PHONETIQUE

J. Guizol

G.I.A., Faculté de Luminy, 70 Route Léon Lachamp
13288 Marseille Cedex 09 FRANCE

I - INTRODUCTION

Le système que nous présentons constitue une phase d'acquisition automatique de connaissances symboliques en vue du décodage acoustico-phonétique de la parole. A partir d'un ensemble d'exemples constitués par un codage de portions de signal issues des réalisations de phrases types, l'apprentissage a pour tâche de fournir une caractérisation du concept induit par le choix de ces exemples.

Les règles de stratégie, de réécriture et de généralisation sont définies en PROLOG [Colmerauer 83], de même que celles permettant de décrire les objets propres à l'application ou les contraintes de généralisation.

Les règles produites par la méthode comportent une information contextuelle et sont valuées en fonction du nombre d'exemples qu'elles vérifient et de la précision de détermination des objets qu'elles utilisent [Michalski 80a].

II - CODAGE DU SIGNAL

Notre système opérant un apprentissage par acquisition de concept (événement acoustico-phonétique, phonème, trait acoustique, etc.), les exemples sont constitués d'une représentation symbolique d'une portion de signal. Celle-ci est obtenue grâce à un ensemble de prédicats évaluables permettant, pour chacun des paramètres du signal, de déterminer maxima, minima, moyennes, pentes afin de modéliser les évolutions temporelles sous la forme de collines, vallées et portions monotones reliées entre elles par des relations situationnelles (coïncidence, succession, chevauchement, etc.) [Meloni 86].

A l'issue de ce traitement, nous disposons donc d'une part, de formes élémentaires issues de l'analyse du signal, représentant l'évolution dans le temps des divers paramètres, et d'autre part, de la chaîne phonémique associée. A partir de ces données, le module d'apprentissage détermine les configurations des diverses formes caractéristiques du phonème, de la classe de phonème ou du trait acoustique que l'on désire étudier.

III - PRESENTATION DE LA METHODE

L'apprentissage inductif que nous réalisons sur les données définies précédemment s'effectue sur des exemples positifs. Si nous avons choisi dans un premier temps de n'opérer que sur de tels exemples, c'est parce que nous doutions de la pertinence et de la validité de contre-exemples dans le domaine étudié. Toutefois, on note des ambiguïtés sur les règles produites caractérisant des classes proches. Afin de les réduire, nous utiliserons pour préciser une classe donnée, des contre-exemples constitués de représentants des classes concurrentes.

III . 1 - Structures des Exemples

Conçu dans l'optique d'être indépendant de l'utilisation qui en sera faite, le système admet des exemples dont la syntaxe est peu contraignante. Toutefois, dans un but d'efficacité, aucune interface n'a été prévue et ils doivent donc avoir une structure de termes PROLOG. Plus précisément, chacun d'eux est constitué d'une liste de n-uplets décrivant une conjonction de propriétés et de relations s'appliquant sur des objets.
Les propriétés sont des doublets formés d'une part d'un prédicat générique et d'une spécification, d'autre part d'un objet identifié par un terme. Les relations sont des triplets dont le premier élément est le symbole relationnel, le second un objet et le troisième une liste (de longueur quelconque) d'objets en relation avec le précédent.

Exemple :

<phoneme(ii),61>.<contexte-gauche(tt),56>.
<contexte-droit(tt),76>.<colline-r0(niveau1),62>.
<coincide,61,62>...

III . 2 - Organisation du Système

L'ensemble des règles constituant le système se sépare en deux parties totalement distinctes :

- une partie constituant le moteur d'induction proprement dit et dont les règles de production ou de stratégie présupposent uniquement les exemples mis sous la forme décrite précédemment ;
- une partie contenant la base de connaissance propre au domaine étudié (BCD) constituant un module facilement modifiable ou même interchangeable et qui permet de fournir des informations qui seront utilisées par le système (description arborescente des propriétés nécessaire à la généralisation par hiérarchie, modèle imposé à la forme généralisée, propriétés des relations, etc.).

III . 3 - Principe de Fonctionnement du Système

Le principe de fonctionnement s'inspire de la méthode proposée par R. S. Michalski [Michalski 80b]. La caractérisation d'un concept à partir des exemples s'opère pas-à-pas. A chaque étape, disposant d'une forme généralisée FG (issue de l'étape précédente) constituée d'une disjonction de règles, un nouvel exemple est proposé.

Dans un premier temps, le système s'assure que celui-ci n'est pas déjà inclus dans une des règles de FG. Ceci se fait très simplement sous PROLOG en transformant chaque terme de l'exemple en une clause unaire et en démontrant FG sur l'ensemble ainsi obtenu. Si l'exemple est vérifié, il est ignoré. Dans le cas contraire, le système va l'introduire dans chaque élément de la disjonction constituée par FG. En cas d'échec sur l'un d'entre eux, celui-ci demeure inchangé dans la nouvelle forme généralisée. En cas d'échec total, l'exemple est signalé à l'utilisateur.

Nous décrivons ci-dessous les étapes successives de l'introduction de l'exemple dans FG.

1) On procède tout d'abord à une factorisation de l'exemple en regroupant les arguments d'une propriété par conjonction interne :

$$P[A] \wedge P[B] \rightarrow P[A \wedge B]$$

2) On considère ensuite la disjonction de l'exemple factorisé avec chaque règle de FG transformée de la même manière. En utilisant la distributivité de $\wedge$ par rapport à

$\vee$ et la disjonction interne sur les arguments, on dégage les propriétés communes.

En supposant, par exemple que la règle considérée de FG est de la forme : $P/ A / \wedge R$ et l'exemple de la forme : $P/ B / \wedge E$, on aura la transformation suivante :

$$\{ P/ A / \wedge R \} \vee \{ P/ B / \wedge E \} \rightarrow P/ A \vee B / \wedge \{ R \vee E \}$$

A noter que A et/ou B peuvent être des conjonctions introduites par la factorisation préliminaire.

3) Des disjonctions d'arguments ainsi obtenues, on déduit des couples (formés d'un objet de FG et d'un objet de l'exemple considéré) qui sont ensuite valués en fonction du nombre de propriétés puis de relations vérifiées. Cette valuation correspond à un degré de pertinence du couple considéré, nécessaire lorsque l'on ne dispose que d'exemples positifs. Seuls ceux dont la valuation est supérieure à une valeur fixée par l'utilisateur dans la BCD seront retenus.

4) Les couples restants peuvent être regroupés en classes d'équivalence, deux éléments d'une même classe vérifiant des propriétés et des relations déductibles les unes des autres ou compatibles, voire identiques. Pour cela, le système utilise les connaissances de la BCD indiquant les propriétés des relations (symétrie, transitivité, inclusion, etc.).

Exemple :

Si le couple $<A,B>$ *vérifie* $P/ A / \wedge Q/ B / \wedge R(A,B)$
Si le couple $<C,D>$ *vérifie* $P'/ C / \wedge Q/ D / \wedge R(D,C)$
Si $P \Rightarrow P'$
Si R est symétrique

... alors, seul $<C,D>$ *sera retenu*
puisque P' est plus générale que P.

5) Il peut être nécessaire selon l'application réalisée d'imposer des hypothèses sur le contenu de la forme généralisée. Ce modèle minimum sera décrit dans la BCD. Dans le cas où ce dernier est non vide, et après regroupement des couples liés par une relation, seuls ceux verifiant les propriétés et/ou les relations contenues dans le modèle sont conservés.

6) A l'issue de cette série de filtres sur les couples, une forme généralisée est produite pour chaque regroupement obtenu. Elle est déterminée par la conjonction des propriétés et relations vérifiées par chacun des couples contenu dans le groupe.

A ce stade, chaque couple étant remplacé par une variable, si les propriétés de même prédicat générique ont une même spécification pour chaque élément du couple, le terme correspondant de la forme généralisée aura une structure identique. Dans le cas où les spécifications diffèrent, c'est leur disjonction qui spécifiera la propriété apparaissant dans la forme généralisée.

Ainsi, aucune information n'est perdue dans la généralisation, même si les propriétés diffèrent par leur spécification, chose que ne permettent pas certaines autres méthodes [Hayes-Roth 78, Guizol 85].

7) La généralisation par hiérarchie s'effectue en fin de traitement. On dispose pour cela dans la BCD d'autant de descriptions arborescentes des propriétés que de prédicats génériques, la spécificité des nœuds augmentant avec la profondeur. Les feuilles constituent en fait l'ensemble des valeurs possibles de la spécification d'une propriété dans les exemples de départ.

Par exemple, dans notre application, l'arbre décrivant les propriétés "*phoneme*", "*contexte-gauche*" ou "*contexte-droit*", est structuré selon la décomposition en traits acoustiques de Jakobson.

Cette généralisation va intervenir sur les propriétés dont la spécification est une disjonction. Après recherche du nœud de plus bas niveau, dont dépendent tous les éléments de la disjonction, elle s'opère de la façon suivante :

- Si ce nœud est la racine :
  - si tous les identificateurs de feuilles sont présents dans la disjonction, la propriété, devenue non significative, est alors supprimé.
  - dans le cas contraire la propriété demeure inchangée.
- Si ce nœud se situe en dessous de la racine, la disjonction est remplacée par l'identificateur affecté à ce nœud.

IV - CONCLUSION

Le système d'apprentissage que nous avons présenté constitue un outil très utile pour caractériser des concepts de façon automatique. En particulier, dans l'application que nous en faisons, il nous permet d'obtenir des règles décrivant des réalisations d'unités acoustiques ou phonétiques propres à un locuteur, nous dispensant ainsi de la laborieuse mise au point de règles ad-hoc.

Les temps de calcul sont assez conséquents, mais ce traitement devant être effectué une seule fois par locuteur, nous jugeons que cela ne constitue pas un réel problème et compense de toute manière le temps passé à déterminer les règles "à la main". D'autre part, le caractère systématique de la production des règles constitue un net progrès.

BIBLIOGRAPHIE

Colmerauer A., Kanoui H., Van Caneghem M.
PROLOG : Bases théoriques et Développements actuels ; TSI, vol. 2, N° 4, pp 271-311, juin-juillet 1983

Guizol J., Meloni H., Gispert J.
Inférence de règles d'adaptation au locuteur dans un système de R.A.P.C ; 14èmes J.E.P., pp 315-318, Paris 10-13 juin 1985

Hayes-Roth F., Mc Dermott J.
An Interference Matching Technique for Inducing Abstractions
Communications of the A.C.M, Vol. 21, N° 5, pp 401-410, 1978

Meloni H., Bulot R.
Un système de traitement des connaissances pour le décodage acoustico-phonétique ; Symposium on Speech Recognition, Montréal 21-22 juillet 1986

Michalski R. S.
Knowledge Acquisition Through Conceptual Clustering : A Theoretical Framework and an Algorithm for Partitioning Data into Conjunctive Concepts ; Policy Analysis and Information Systems, Vol. 4 N° 3, pp 219-244, 1980a

Michalski R. S.
Inductive Learning as Rule-Guided Generalization and Conceptual Simplification of Symbolic Descriptions ; Workshop on Current Developments in Machine Learning, CMU, Pittsburgh, July 16-18, 1980b

UN SYSTEME DE TRAITEMENT DE CONNAISSANCES POUR LE DECODAGE ACOUSTICO-PHONETIQUE

H. Meloni, R. Bulot

G.I.A., Faculté de Luminy, 70 route Léon Lachamp 13288 Marseille Cedex 09 France

I - INTRODUCTION

Les travaux que nous avons accomplis dans le cadre de la reconnaissance de la parole (Méloni 1982, 1984, 1985) séparaient assez nettement les traitements numériques, exécutés dans un langage algorithmique classique, des traitements de données symboliques effectués en PROLOG. Cette dichotomie artificielle interdisait l'interaction et l'optimisation contextuelles des deux processus. Afin de permettre une coopération simple entre toutes les sources de connaissances, nous proposons, sous la forme d'un ensemble de prédicats du langage PROLOG II (Colmérauer 1984), un environnement souple et efficace pour l'acquisition, la manipulation, l'évaluation, la représentation et le traitement d'informations acoustiques, phonétiques et linguistiques. Les outils développés permettent, de manière interactive, de produire diverses paramétrisations du signal, de décrire et reconnaître des formes simples, de définir et identifier des événements, des propriétés, des indices et des traits acoustico-phonétiques, de coder ces informations et les stratégies qui les utilisent et de structurer l'ensemble des résultats produits sous la forme d'un treillis d'unités valuées. Nous illustrons les possibilités nouvelles de cet environnement en présentant quelques particularités d'un système de décodage acoustico-phonétique réalisé entièrement sous PROLOG II.

II - PARAMETRISATION DU SIGNAL

Le but visé à ce stade du traitement est de caractériser de manière précise et peu coûteuse une portion de signal au moyen d'une suite de vecteurs de paramètres. Les limites de la zone codée, la nature des attributs retenus ainsi que leurs conditions d'évaluation sont déterminées en examinant des connaissances de niveau acoustique, phonétique ou phonologique.

II - 1 - Prédicats évaluables pour la paramétrisation

Nous disposons de 2 prédicats évaluables qui effectuent le calcul des paramètres et leur chargement dans une mémoire accessible à d'autres fonctions réalisant des opérations numériques complexes. Chaque vecteur, produit à intervalles réguliers de 10 ms, est constitué d'une vingtaine d'attributs temporels et spectraux (répartition spectrale de l'énergie, densité des passages par zéro, position, amplitude, émergence et largeur des pics, etc.). Les spectres lissés sont obtenus à partir des coefficients cepstraux ou de LPC dont un ensemble de variables définit les conditions de calcul (portion de signal traitée, méthode utilisée, nombre de coefficients, préemphase, rayon, etc.).

II - 2 - Utilisations des prédicats de paramétrisation

Les prédicats de paramétrisation du signal ont été employés, dans la phase d'acquisition des connaissances acoustiques, pour évaluer les conditions optimales du codage des sons correspondant aux diverses phases de phonèmes segmentés semi-automatiquement dans un ensemble de 130 phrases prononcées par 2 locuteurs.

La stratégie du système de décodage conduit à une paramétrisation globale d'un énoncé au moyen des 14 premiers coefficients de LPC, mais ces attributs sont localement recalculés lorsque certaines règles de niveau quelconque exigent d'autres conditions d'évaluation. C'est le cas notamment pour l'identification des traits des voyelles nasales, le traitement des explosions d'occlusives ou de portions sourdes du signal.

III - RECONNAISSANCE DES FORMES

Les outils proposés dans ce cadre ont pour objectif la modélisation et la symbolisation des évolutions temporelles de certains groupes de paramètres.

III - 1 - Prédicats évaluables de reconnaissance de formes

Sur un intervalle de temps, ils définissent des fonctions simples d'un paramètre telles que la mesure de ses extrema, le calcul de sa moyenne, la caractérisation de son instabilité. Ils déterminent également des fonctions complexes d'un ensemble d'attributs comme la recherche de formants, la valuation de la continuité ou de la monotonie d'un phénomène. Enfin, ils désignent et identifient des schémas de formes parmi lesquels les modèles de collines ou de vallées sont les plus utilisés. Des variables permettent de préciser les contours d'une forme ; ainsi, la définition d'un type de colline, pour un paramètre quelconque, sera donnée par sa largeur et son minimale, son émergence limite à gauche et à droite, le seuil maximum de déviation acceptable ainsi que le seuil de bruit au dessous duquel le paramètre n'est pas significatif.

III - 2 - Utilisations des prédicats de R.F.

A partir de schémas de formes simples, appliqués pour l'ensemble des paramètres sur les phrases de référence, nous avons sélectionné quelques dizaines de formes représentatives de portions de sons déterminées. Ces éléments (collines, vallées, zones monotones ou stables, etc.) concernent, souvent de plusieurs manières, la plupart des attributs temporellement continus.

Dans le système de décodage, les formes retenues définissent des événements acoustiques et phonétiques et constituent des repères pour guider le processus de reconnaissance vers les phénomènes les plus saillants.

IV - NIVEAUX SYMBOLIQUES DU SYSTEME

Les connaissances acoustico-phonétiques formelles d'un même type sont regroupées en niveaux pour leur présentation, mais chaque règle peut être sollicitée indépendamment de sa classe.

IV - 1 - Prédicats prédéfinis de l'environnement

Ces outils contribuent à rendre plus naturelle l'expression de la connaissance et définissent pour l'essentiel les fonctions suivantes :

- relations temporelles entre des unités du treillis de résultats (coïncidence, intersection, succession, union, adjacence, etc.),
- démonstrations particulières d'une liste de prédicats pour la gestion du contrôle et la visualisation des parcours (effacement déterministe ou complet, vérification de l'existence d'une ou de plusieurs solutions, saturation des effacements, impression de traces, etc.),
- opérations logiques sur des listes de prédicats (conjonction, disjonction, négation, implication, etc.),
- opérations arithmétiques diverses acceptant des fonctions en paramètres,
- manipulations complexes sur les arbres.

IV - 2 - Evénements acoustiques et phonétiques

Les événements acoustiques sont définis par regroupement de formes au moyen des prédicats qui décrivent des relations temporelles entre les éléments de base. Les unités engendrées ne reçoivent pas d'interprétation phonétique ; elles mettent en évidence la conjonction de propriétés acoustiques du signal et caractérisent généralement des segments infra-phonémiques.

Les événements phonétiques, identifiés à partir des événements acoustiques, des formes et des relations, constituent des unités que l'on peut associer directement à des phases spécifiques de phonèmes et de transitions (constriction, occlusion, explosion, etc.) ou à des regroupements de segments acoustiquement proches. Des règles contextuelles réunissent ensuite ces éléments pour désigner les limites des phonèmes ou décomposent certains d'entre eux à partir de critères plus fins pour séparer certaines voyelles des consonnes vocaliques qui les entourent.

Les quelques dizaines de clauses qui définissent ces connaissances opèrent dans des contextes souvent très différents suivant qu'il s'agisse d'événements "évidents" ou de segments tributaires de l'identification préalable de l'environnement. Ces règles sont indépendantes du locuteur, elles opèrent une partition peu ambiguë d'un énoncé en macro-classes pseudo-phonétiques. L'exemple ci-dessous décrit et évalue un type particulier d'événement vocalique :

```
evenement-voc(<voc(5),z>) ->
    forme(<colline1-er0,z>)
    inferieur(5,longueur(z))
    voise(z)
    ou (coincidence-sur(z,colline1-ebf) ,
        coincidence-sur(z,colline1-ap1) ) ;
```

IV - 3 - Traits pseudo-phonétiques

Chaque événement pseudo-phonétique est caractérisé par un faisceau de traits hiérarchisés dont chacun est défini par un ensemble de clauses. Les règles qui représentent ces connaissances utilisent de nombreuses informations contextuelles sur la nature, les paramètres, les propriétés, les indices ou les traits des sons adjacents. Cette étape de la reconnaissance fonctionne comme un filtre phonétique limitant le nombre des phonèmes candidats. Le choix des solutions les plus vraisemblables résulte de l'évaluation d'un score à partir de paramètres sélectionnés et ajustés en fonction de caractéristiques des segments contigus.

L'acquisition et l'évaluation de ces règles sont effectuées de manière interactive sur les phrases de référence. Une étude statistique des paramètres ou de certaines fonctions de plusieurs d'entre eux permet de désigner les attributs les plus discriminants pour la détermination d'un trait d'un phonème dans un environnement précis. Les règles qui en résultent sont immédiatement testées sur l'ensemble des situations où elles sont susceptibles de s'appliquer. La clause suivante décrit et évalue un indice du trait *grave* pour les occlusives sourdes :

```
acuite-occ-sourde(z,grave(2)) ->
    inferieur(cgh(z),3200)
    inferieur(afmedian(z),ebf(z))
    inferieur(afhaut(z),ebf(z))
    si-alors(inferieur(moins(fois(2,ebf(z)),10),
                       plus(afmedian(z),afhaut(z))),
             inferieur(afhaut(z),afmedian(z)))
    si-alors(inferieur(300,fbas(z)),
             inferieur(afbas(z),plus(6,ebf(z)))) ;
```

V - TREILLIS DES RESULTATS

Certains résultats, considérés comme définitivement acquis au cours du décodage d'un énoncé, sont conservés dans un treillis constitué de clauses PROLOG. Cette structure est rendue souple et efficace au moyen d'un ensemble de prédicats qui permet de réaliser des opérations telles que l'ajout et la suppression d'unités en un point quelconque, des parcours multiples, le repérage des zones libres, des accès diversifiés à une unité (par la position de ses bornes, par son type ou ses caractéristiques, etc.), la récupération des arguments et des limites d'un élément, etc.

Dans la phase de décodage acoustico-phonétique, la stratégie conduit à exécuter non séquentiellement les étapes suivantes :

- calcul des paramètres sur l'ensemble de l'énoncé,
- reconnaissances des formes constituant les noyaux nécessaires à la définition d'événements sûrs,
- identification et mémorisation dans le treillis des événements acoustiques évidents,
- recherche dans les zones libres des événements secondaires et transitoires,
- regroupement des segments étiquetés pour produire et ajouter des événements consonantiques,
- affinement de la segmentation des noyaux vocaliques pour déterminer, quelquefois de manière ambiguë, des événements vocaliques qui vont enrichir le treillis,
- identification des traits pseudo-phonétiques puis ajout des phonèmes les plus vraisemblables après filtrage et calcul du score,
- interprétation des zones non reconnues au moyen de l'ensemble des unités du treillis.

VI - CONCLUSION

La brièveté de l'exposé ne donne qu'une image imparfaite de la puissance potentielle de l'environnement proposé. Son utilisation pour les tâches d'apprentissage et d'acquisition des connaissances acoustico-phonétiques nous a permis de constituer très rapidement un important ensemble de règles dont certaines demeurent perfectibles. Le traitement de ces connaissances fournit des résultats bien supérieurs par certains aspects à ceux que nous obtenions au moyen de techniques classiques. La durée du processus de décodage demeure tout à fait raisonnable sur un mini-calculateur, et nous pouvons envisager de réaliser sous PROLOG un système complet de reconnaissance automatique de la parole.

BIBLIOGRAPHIE

Colmérauer A., Kanoui H., Van Caneghem M.
PROLOG : Bases théoriques et Développements actuels ; TSI, vol. 2, n° 4, juin-juillet 1983, pp 271-311

Méloni H.
Etude et réalisation d'un système de reconnaissance automatique de la parole continue ; Thèse de Doctorat d'Etat, Université d'Aix-Marseille II, Faculté de Luminy, février 1982

Méloni H.
Traitement des Contraintes Linguistiques en Reconnaissance de la Parole ; TSI, vol. 2, n° 5, septembre-octobre 1983, pp 349-363

Méloni H., Gispert J., Guizol J.
Traitement de connaissances pour l'identification analytique de Mots dans le discours continu ; congrès AFCET Informatique 5[e] Génération, Paris 5-7 mars 1985, pp 339-350

UTILIZATION OF MULTIPLE UNITS IN HUMAN AND MACHINE RECOGNITION OF CONTINUOUS SPEECH ---- PERCEPTUAL EVIDENCE AND A PROPOSAL FOR AN ASR SYSTEM

H. Fujisaki, H. Udagawa and N. Kanedera

Dept. of Electronic Eng., Faculty of Engineering University of Tokyo, Bunkyo-ku, Tokyo, 113 JAPAN

The ultimate goal of automatic speech recognition (ASR) is obviously to replicate the human capability of speech processing by machine. Research of ASR will thus profit very much from investigations into the human processes of speech perception/comprehension. Few studies, however, seem to have been made along this line. The present paper summarizes a series of psycholinguistic experiments conducted to elucidate certain aspects of human speech perception, especially in relation to the units of processing. On the bases of these experiments, we propose a new system for continuous speech recognition utilizing multiple units.

AN EXPERIMENT ON HUMAN SPEECH PERCEPTION

Objective and Method

While it is desirable to design a psychological experiment that would directly disclose the size of the unit of human speech perception, the difficulty of the problem led us to adopt an indirect approach. We first designed an experiment which would show that certain segments are not processed as independent perceptual unit in human speech recognition. In the following experiment, we investigated perception of connected speech in the presence of deleted syllables to find out whether such deletions are always noticed by the listener[1,2]. If they are not noticed by the subject, one would be able to infer that the subject is not treating the deleted syllables as independent perceptual units, but is recognizing the input speech as a sequence of larger units. The fact that the deletion of a certain syllable is not noticed would indicate that it does not impair perception of a larger unit containing the deleted syllable.

The original speech material was one minute of speech recorded by a male speaker reading a Japanese text at a normal speech rate of approximately 7 morae/sec. The speech signal was low-pass filtered at 4.8 kHz, sampled at 10 kHz with 12 bit accuracy for processing by a digital computer. A total of 25 CV syllables was deleted on the basis of visual inspection of the speech waveform on an X-Y plotter. In order to avoid artifacts, only CV syllables, each starting with an unvoiced consonant and being followed by an unvoiced stop consonant, were selected for deletion. Figure 1 illustrates an example of syllable deletion. In order to examine the effect of context on the noticeability of the deletion, the

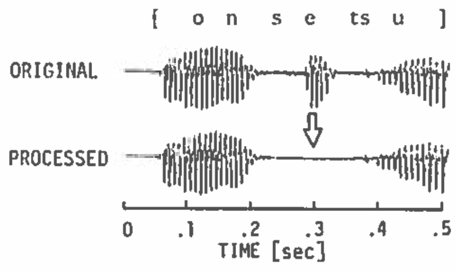


Fig. 1. An example of syllable deletion. The syllable [se] of the word "on'setsu" (meaning 'syllable') is deleted from the original signal.

following four types of test stimuli were prepared after deletion of the syllables.
(1) Segmented into lexical words and randomized.
(2) Segmented into prosodic words and randomized.
(3) Segmented at every pause and randomized.
(4) Without segmentation and randomization.
These stimuli were presented to each subject through a binaural headphone in four test sessions.

The subjects were three male adults with normal hearing. The subject's task was to count the total number of deleted syllables he could notice under each of the four test conditions. Each subject sat for the four test sessions at least five times.

Results and Interpretation

The results of the experiment is shown in Table 1 and the averaged results of the three subjects are shown in Fig. 2. The averaged probability of noticing the deleted syllables is approximately 70% under test condition (1), i.e., when the speech signal is segmented into lexical words and randomized, it drops only slightly under condition (2), but drops rather drastically below 40% under conditions (3) and (4), i.e., when the speech signal is either segmented at every pause or not segmented at all. The difference of results for condition (3) and for condition (4) is quite small.

Table 1. Probability(%) of detection of syllable deletion of each subject.

SUBJECT	LEXICAL WORD	PROSODIC WORD	CLAUSE	SENTENCE
A	77.3	77.3	40.8	40.8
B	69.3	54.7	34.4	32.8
C	69.3	64.0	40.8	37.6
AVERAGE	72.0	65.3	38.7	37.1

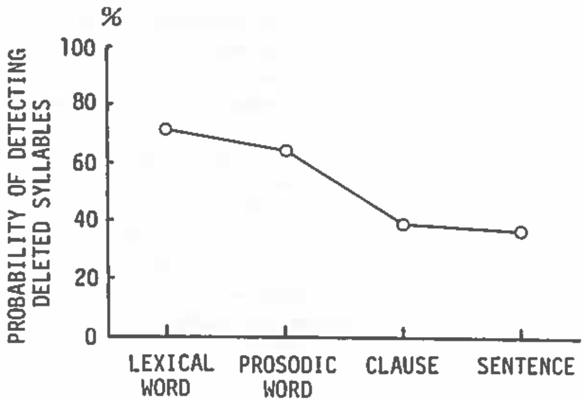


Fig. 2. Results of the perceptual experiment. Relation between size of given context and probability of detection of syllable deletion. Each circle expresses mean value of three subjects.

These results indicate that human listeners pay more attention to syllabic units in a word context, but pay much less attention when the context is as large as a clause or a sentence. In other words, the unit of speech perception is more likely to be syllable-sized when the available context is of the size of a word, but the unit is more likely to be word-sized when the context is as large as a clause or a sentence. Although further experimental studies are necessary, the result of the present experiment suggests that the unit of human speech perception is not unique, but is rather multiple.

FURTHER EXPERIMENTS

Although the above-mentioned experiment revealed the multiplicity of perceptual units, we still need to know the actual size of the units as well as the exact conditions at which one type of units is predominantly used. In this section, we describe some of further experiments being carried out or planned to investigate more deeply into the human processes of speech perception.

Size of Perceptual Units

Granting that the unit in perception of connected speech is larger than a syllable, we need to know whether it is a morpheme, a lexical word, or a prosodic word. The following experiment was designed to answer this question.

Since it has become clear that deletion of a syllable is more easily noticed at the initial position of a perceptual unit than elsewhere, the following three types of stimuli were prepared.

(1) Stimuli in which syllable deletions occur only at the morpheme-initial position which is not the word-initial position.

(2) Stimuli in which the same number of syllable deletions occur only at the word-initial position which is not the initial position of a prosodic word.

(3) Stimuli in which the same number of syllable deletions occur only at the initial position of a prosodic word.

The experimental procedure is the same as in the experiment described in the previous section. If there is no significant difference in the detection rate of syllable deletion among the three types of stimuli, we may infer that the perceptual unit in this case is most likely a morpheme. If the detection rate for the type (1) stimuli is significantly lower than for the type (2) stimuli, but the latter show no significant difference from the type (3), then we may infer that the perceptual unit is a lexical word. In the same vein, if the detection rate is significantly higher only for the type (3) stimuli, we may infer that the perceptual unit is a prosodic word or a still larger unit. Our preliminary results suggest that the latter case is most likely, although we still need more experimental data to confirm it.

Effect of Syntactic Roles on Detectability

Assuming that the unit in perception of connected speech is a prosodic word, one can naturally ask whether all the prosodic words in a sentence receive the same degree of attention and thus show approximately equal detection rate of deleted syllables, or they show different detection rate depending on the difference in their syntactic roles. This question can be answered by investigating the dependency/independence of the detection rate on the syntactic role of the prosodic word containing a deleted syllable. Preliminary results indicate that there are significant differences in the detection rate depending on the syntactic role.

Size of Context on Syllable Recognition

While it is true that most of the evidences and discussions in the foregoing sections are in favor of the use of units larger than the syllable, there are also cases where one has to rely on syllable recognition[3]. If, for example, we are to deal with a very large vocabulary, or even with an unlimited vocabulary, the system will occasionally have to recognize (or transcribe) unknown words syllable by syllable, just as a human listener will do when presented with an unknown word.

In order to design a recognition system whose performance is comparable to that of a human listener, it is thus necessary to know human perception of syllables in connected speech. It has been shown that a human listener needs a context of one syllable each immediately before and after the target syllable in order to be able to recognize with high accuracy the target syllable in connected utterances of one speaker[4]. Likewise, syllable recognition by machines will have to take into account the influences of the context of similar span.

OUTLINE OF AN ASR SYSTEM USING MULTIPLE UNITS

From the evidences and discussion in the foregoing sections, we have proposed a new system for continuous speech recognition based on template matching of multiplicity of liguistic units (idioms, prosodic words, and syllables)[2]. The system operates in the following four steps:

1) Extract acoustic parameters of input speech. (formant frequencies, fundamental frequency, band-limited power, etc.)
2) Detect syllable nuclei, prosodic word boundaries, and clause/sentence boundaries.
3) Detect and recognize frequently used idioms and prosodic words in the continuous speech signal by using their templates. For the portions of input speech where the template matching fails, syllables are detected and recognized by using context-dependent syllable templates.
4) Construct a lattice of (prosodic) word candidates based on the results of the preceding step. Syntactic and semantic coherence is evaluated for all combinations of candidates.

If real-time processing is not required, the system performance would be still more improved. When ambiguity remains, it can also be checked for global coherence to reduce the candidates and to obtain the most probable output. Global coherence is also utilized to re-examine and revise the results of recognition already obtained for a prior input. This is only possible when real-time processing is not required.

REFERENCES

[1] Fujisaki, H., K. Hirose, H. Udagawa, N. Kanedera and Y. Sato, "Considerations on units for continuous speech recognition based on human process of speech perception," Rec. Spring Meeting, Acoust. Soc. Japan, 3-1-14 (1985).

[2] Fujisaki, H., K. Hirose, H. Udagawa and N Kanedera, "A New Approach to Continuous Speech Recognition Based on Considerations on Human Speech Perception," Proc. 1986 IEEE Int. Conf. ASSP (1986).

[3] Fujisaki, H., K. Hirose, H. Udagawa, T. Inoue, T. Ohmori and Y. Sato, "Analysis of variability in the acoustic-phonetic characteristics of syllables for automatic recognition of connected speech" Trans. of the Committee on Speech Research, Acoust. Soc. Japan, S84-69 (1984).

[4] Kuwahara, H., H. Sakai, "Perception of vowels and CV syllables segmented from connected speech," J. Acoust. Soc. Japan 28, 225 (1972).

DISTORTION MEASURE EVALUATION USING SYNTHETIC SOUNDS AND HUMAN PERCEPTION

D. Tuffelli, H. Ye

Institut de la Communication Parlée (LA. CNRS. No.368)
46, Av. Félix-Viallet 38031 Grenoble
France

In this paper, we try to compare several distortion measures with the human's perception using synthetic sounds. Correlation and another measure of coherence is used. The goal of this research is to study the coherence between mathematical distortion measures and the human's perception. The results show there are some differences between them. But Itakura distortion measure is the best in the case of our isolated vowels.

I. INTRODUCTION

Distortion measures of speech is an important problem for speech processing: speech recognition; speaker identification; speech coding...etc.

Generally, there are 2 kinds of distortion measures. The first one is defined by means of a mathematical criterion, such as Itakura-Saito; cepstral; likelihood ratio and weighted Itakura-Saito[1,2] ... etc. The second is perceptually based measures, such as weighted slope metric(WSM)[3]; euclidean distance of critical-band spectra[5] and weighted likelihood ratio[6] ... etc.

The first approach is purely mathematical without any perceptual constraint. The second approach try to make use of perceptual properties with some model made from human's perception.

An early study has been done with difference limens of formants[7]. A recent study has been done on perceived phonetic distance[3].

Another more global type of comparison[8] was carried out between human performance (presented by confusion matrix) and an automatic recognition algorithm.

The work presented here tries to examine and to compare the previous 2 kinds of distortion measures with the data of a test of psychoacoustics which was especially designed for this goal.

II. EVALUATION OF DISTORTION MEASURES

Different Tested Distortion Measures

*Itakura distortion[10] is gain optimized Itakura-Saito measure which was originally introduced as an error matching function in maximum likelihood estimation of autoregressive spectral models.

$$d_{ita}(x,x') = \log(\alpha/\alpha_m)$$

where α is any residual energy and α_m is minimal residual energy.

*Cepstral distortion measure is an approximation of the L_2 norm of the log spectral distortion by first N terms.

$$d_{cep}(x,x') = \sum_{i=-N}^{N} (c_i - c'_i)2$$

*2 other kinds of distortion measure a priori bad are tested: euclidean distance of linear prediction coefficients and autocorrelation coefficients (from LPC preprocessing).

*Weighted slope metric proposed by Klatt is a perceptually based distortion measure[3].

$$d_{wsm}(x,x') = Ke|E-E'| + \sum_{i=1}^{Q} K(i)*[S(i)-S'(i)]^2$$

Ke and K(i) are coefficients. We take Ke=0,K(i)=1 (according to [9] error is minimal with these values). Here Q=18 (Some values differ from Klatt).

*Another perceptually based distortion measure was proposed by Plomp[5]. Late it was used by Carlson (1979) and Blomberg(1983).

$$d_{plm}(x,x') = (\sum_{i=1}^{Q} |L_i - L'_i|)^{1/p}$$

where L_i is critical band spectra in band i and p=1 or 2.

*Another simple slope distortion measure (called here D_{hsl}) is defined by a Hamming distance on a set of Fn parameters[4]. Where

F_n = 1 if X(n+1)>=X(n)and X(n+1)>=threshold
0 otherwise

and X(n) is smoothed spectrum either in linear or in Mel frequency scale.

A classic method of evaluation of different distortion measures is to test them in a recognition system. So one can judge their performance according to their error percentage of recognition. This is often expensive and time consuming.

Psychoacoustic Tests

A test of psychoacoustics has been designed to produce pertinent histograms which can be easily compared with the curves of distortion measures.

The test was carried out with steady state synthetic vowels. 12 pairs of french vowels have been chosen. Each pair vowel is close so that there is not a third vowel between the vowels of a pair. A series of 11 sounds has been synthesized for each vowelpair by linear interpolation of their formants. The data of formants are from Mrayati (1976).

During the test, an auditor had to listen to the previous series of sounds between 2 references (these 2 references are phonetic references, that is vowels labels and the sounds were not given in the test) and discriminate every presented sound to one of the 2 asked references with forced choice. 12 histograms have been built with 9 auditors from 132 sounds (12*11).

In fact this is a similarity measure of the tested sound to vowels. Auditor will discriminate a sound to one class if it seems more similar to its reference than another one.

Distortion Measure Curve

The same signals have been used for distortion measure calculation. For reason of comparison we calculate

$$D_s(x,V1,V2) = d(x,V2)-d(x,V1)$$

where V1, V2 are 2 references and x is any sound of the series of sounds synthesized by linear interpolation between V1, V2. d is a distortion measure.

The evaluation is made by correlation and percentage of errors which will be defined in next section.

III. EXPERIMENTAL RESULTS

Normalized Correlation Measure

It is often used to compare a distortion measure and human perception.

If x and y are regarded as Euclidean vector,
$r = \cos\theta = (x,y) / (||x||.||y||)$

Percentage of Error

*For human perception, there is a statistic frontier (an arrow below) between 2 vowels. Every auditor made some error with respect to this frontier. The mean of this error for all auditors is denoted by E_h. For example, a histogram is presented in Fig 1, E_h is the sum of shaded region.

*For distortion measure, a percentage of error E_{dis} is defined. This percentage is computed by the ratio of 2 lengths: the length of the interval between the distortion measure frontier (zero crossing) and human statistic frontier, and the length between 2 references in Fig 2.

Some Figures

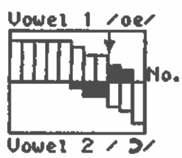


Fig 1.

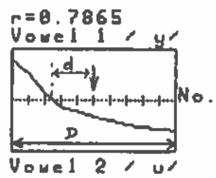


Fig 2. $E_{dis}=d/D$

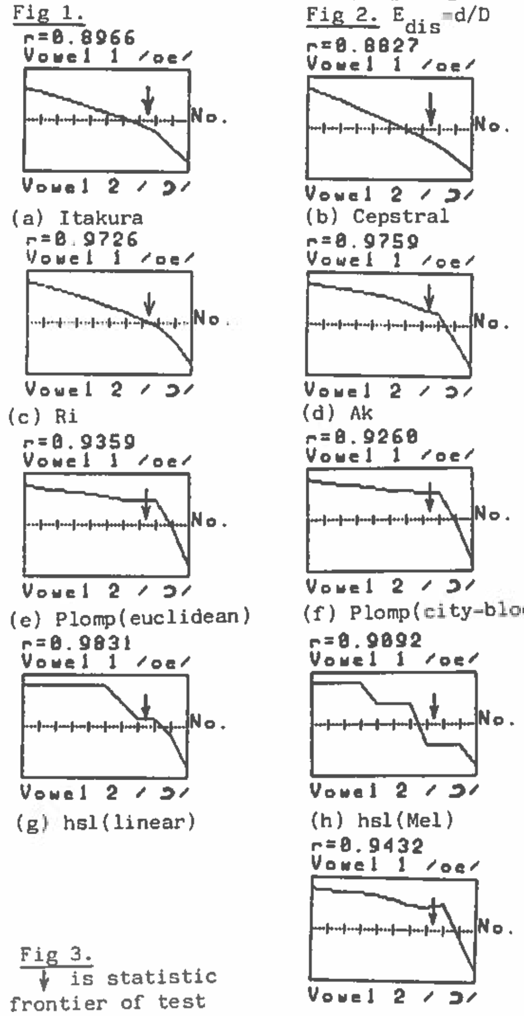


Fig 3.
↓ is statistic frontier of test

(i) WSM

Some Results

We present here a part of results about the correlations and the percentages of errors. All results are means of 12 tests. This correlation is between all points of 2 curves: D_s and histogram of human perception.

	Correlation	Percentage
Auditors		9.6%
Itakura	0.857	10.1%
Cepstral	0.832	13.7%
Plomp(city-block)	0.824	15.3%
Ri(euclidean)	0.77	17.0%
Plomp(euclidean)	0.80	17.8%
Hamm. slop(linear)	0.74	18.3%
Hamm. slop(Mel)	0.77	19.6%
WSM	0.70	21.0%
Ak(euclidean)	0.64	22.6%

Another type of correlation can be computed from the different frontiers. For example correlation between frontiers of Itakura and these of cepstral over 12 tests is 0.993, it corrresponds to an angle of 6.7°; and correlation between Itakura and Ri is 0.9, it corresponds to an angle of 25.8°.

IV. CONCLUSIONS

The main mathematical distortions are better than perceptually based distortions but the test we have done is favourable to mathematical distortions (the sounds vary by formants shifts only). As it was expected the Ak coefficients are not good ones. Sometimes very bad frontiers are obtained which are difficult to explain. A very high correlation between Itakura and Cepstral measures is observed.

The most difficult choice, in this work, is the set of formants of references. The chosen set is considered as representative of french vowels. Surprisingly it is very well adapted to Itakura distortion.

REFERENCES

1. R.M.GRAY, A.BUZO, A.H.GRAY, Y.MATSUYAMA, "Distortion Measure for Speech Processing" IEEE Trans. ASSP-28, No.4, pp367-376, 1980

2. P.L.CHU, D.G.MESSERSCHMITT, "A Frequency Weighted Itakura-Saito Spectral Distance" IEEE Trans. ASPP-30, No.4, pp545-560, 1982

3. D.H.KLATT, "Prediction of Perceived Phonetic Distance from Critical Band Spectra: a first step" ICASSP pp1278-1281, 1982

4.T.K.VINTSJUK,A.G.SHINKAJ,apco8,lvov,3,pp19-24,1974 (in russian)

5. R.PLOMP, "Timbre as a Multidimentional Attribute of Complex Tones" FREQUENCY ANALYSIS AND PERIODICITY DETECTION IN HEARING Ed. PLOMP, 1970

6.K.SHIKANO, M.SUGIYAMA, "Evaluation of LPC Spectral Matching Measure for Spoken Word Recognition" Trans. IECE, Vol. J65-D, No.5, pp535-541, 1982

7. R.VISWANATHAN, J.MAKOUL, W.RUSSELL, "Towards Perceptually Consistent Measure of Spectral Distance" ICASSP pp485-489, 1976

8. M.ESKENAZI, J.S.LIENARD, "Recognition of Static State French Sounds Pronounced by Several Speakers; Comparison of Human Performance and an Automatic Recognition Algorithm" Speech Communication 2(1983) pp173-177

9. N.NOCERINO, F.K.SOONG, L.R.RABINER, D.H.KLATT, "Comparison Study of Several Distortion Measures for Speech Recognition" Speech Communication 4(1985) pp317-331

10.F.ITAKURA, "Minimum Prediction Residual Principle Applied to Speech Recognition" IEEE ASSP-23, No.1, 67-72, 1975

11. M.MRAYATI, Contribution aux Etudes sur la Production de la Parole, Thése d'Etat, INPG 1976

THE SYLLABLE AND LANGUAGE PERCEPTION

Geoffrey S. Nathan

Department of Linguistics
Southern Illinois University at Carbondale
Carbondale, IL 62901.

This paper proposes that languages have an active process of syllabification that takes a string of phonemes as input and organizes those sounds into a hierarchical syllable structure. This process acts as a filter on perception of input, such that native speakers hear both their own and foreign languages as if the sounds had been organized to follow the syllabification processes of thir own language. The results of several psycholinguistic research programs can be offered as evidence for this claim.

This paper is an exercise in 'armchair phonetics', in which I will argue that syllabification in language is an active process, in the sense used by Natural Phonology (see, e.g. Stampe 1973, Donegan and Stampe 1978). That is, I will argue that the assembly of segments into syllabic units is an activity carried out by the speaker in real time as speech is produced--a process governed, among other things, by rate of speaking, degree of care in the production of the speech and the purpose to which the speech is being put. Furthermore, the setting of segments into suprasegmental organization, although governed by universal tendencies, allows certain options which speakers may, on occasion, choose to exercise.

Although the syllable has had a tenuous position in recent linguistic and phonetic theory, many have argued eloquently for its existence. The earliest modern discussion of the concept of syllable, including an extensive discussion of the basis of syllable division and the idea that the shape of the syllable is governed by the sonority hierarchy, can be found in Sievers (1885:179-183). Sievers is also the first to argue that syllabification is a heuristic rather than an algorithm: 'Equally, one can, to a certain extent, give arbitrarily different syllabifications to any one of several sounds of an assembled string like [aia].' (179, my translation).

Since that time, numerous scholars have argued that the sounds of language are assembled into larger units that appear to be actively used in the production and perception of speech. An extensive discussion appears in Stetson 1951, although some of his contentions have since been disproven (Ladefoged 1982). Kozhevnikova and Chistovich (1965) argue that instructions to the articulators are sent in syllable-sized chunks (122), while several researchers have recently presented functional arguments based on the nature of the speech-producing mechanism for the syllable as a unit of sound organization--Studdert-Kennedy 1975 and, particularly, Lindblom 1983 are two notable recent works on the subject.

I will begin by discussing a Hebrew prayer, known as the Shma. The prayer is sung to a traditional melody, and consists of two lines. For the first line, there is only one possible setting for the words, but for the second line, there are two possible ways in which the words and the music can be coordinated, and both are used, apparently interchangeably:

1 a) boru.uch shem kevod ma.alchuto leolam vaed

b) boru.uch shem kevod malchuto.o leolam vaed

Since there are more notes than syllables, additional syllables must be created, and as 1) shows, there are two possibilities for the creation of the extra syllables. My primary argument is that rules of syllabification mediate between storage of sounds and their production (i.e. they are used in 'derivations') and between perception of sounds and their storage (i.e. their 'underlying representations').

My primary sources of evidence for this claim involve investigations that have been done in examining the acquisition of second language, where there appears to be conflict between the processes of syllabification in the languages involved.

An early paper on this subject is Brière et al. (1983). The authors note that, although neither /ž/ nor /ŋ/ can begin words in English, only one, /ŋ/ appears to offer any problems for native speakers of English learning a second language. They therefore suggest that the correct restriction on distribution of these phonemes is stated in terms of syllable distribution--/ž/ may occur in syllable-initial position (although, by accident, not in word-initial position), while /ŋ/ only occurs in syllable-final position, and hence may never occur word-initially. To study this issue they had native speakers of English produce words one syllable at a time following the beat of a metronome. (The words were controlled for such things as spelling and stress placement). They then studied what their subjects did with various consonants at the induced pauses occasioned by the enforced divisions the metronome produced. As one might expect, they found that while speakers produced such forms as 'lei.sure', they always divided 'sing.ing'.

For our purposes, however, a much more interesting result occured with words like 'city'. Although this word is normally pronounced with a voiced alveolar flap in American English, it was always pronounced as a voiceless, aspirated stop in their experiment. Various researchers (Stampe 1973, Kahn 1976) have argued that the choice of flap versus stop is controlled by syllabification. Syllable initial stops are aspirated, while syllable final (or ambisyllabic) /t/'s are flapped. Syllabication itself is driven by stress, with a stressed syllable attracting single consonants leftward away from an adjacent unstressed syllable. Since the highly unnatural isochronic stress pattern induced by speaking with a metronome made all /t/'s initial, it is not surprising that they came out aspirated. But this is to be expected only if the sounds are stored as /t/'s, with syllabification, and consequently segmental processes dependent on syllabification, occurring at the time of speech production.

In a much more recent publication, Eckman (1981) argued that there are 'natural processes' that speakers use when attempting to acquire a second language, even though these processes do not occur in any known natural language or historical change--ordinarily two major sources for the naturalness of phonological processes. He studied how native speakers of Spanish and Mandarin dealt with sound sequences that do not occur in their native languages but do in English--final voiced stops. Spanish speakers appear to begin using the well-known process of final devoicing, a traditional candidate for a natural process, and one that does not, as far as we know, occur in Spanish. Mandarin speakers, however, frequently deal with final obstruents through the insertion of a final schwa. Since the theory that Eckman follows (a version of generative phonology) requires that any systematic difference between target language and output be attributed to the presence of a rule, he is forced to posit a rule of 'schwa paragoge', which he also argues must be a natural process, since it occurs neither in the source nor the target language, and consequently cannot have been learned.

There is, however, an alternative explanation for the frequently attested action of learners (and borrowers) of adding syllables to foreign words to avoid unacceptable consonant configurations. The Japanese borrowing of 'baseball' as /beisuboru/ is a parallel example.

Let us suppose that principles of syllabification, as well as other phonological processes mediate between mental storage and pronunciation, and between hearing and mental storage. Foreign words, especially at the beginning of the study of a foreign language, will be storable only in native language terms. If the sounds perceived are, when produced in the L1, subject to L1 processes, then they will be so pronounced--thus native speakers of French unaspirate initial English voiceless stops and native speakers of English do the reverse. If the perceived sounds occur in positions in which they do not in the native language, unsuppressed natural processes which have never come into play in the first language may well do so in attempts at the L2. This explains the final devoicing of native speakers of Spanish. However, Spanish does have some final obstruents. Mandarin has no final obstruents at all. This forces native speakers to attempt something that I propose to term second language restructuring. They increase the phonetic substance of the target so that segments (such as final obstruents) that their native language patterns forbid them from producing will be retained. This creative restructuring of the input is not the same as the application of a natural phonological process, but is rather the invention of input which will be sufficiently immune to the natural processes the speaker already possesses that the otherwise deleted consonant will remain intact. Since the syllable-structure processes of Mandarin do not permit final obstruents, the creation of an additional syllable, particularly when it is made up only of the threatened consonant and a schwa, is a natural strategy for keeping phonic information that Mandarin and other universal processes would threaten.

A similar claim is made by Broselow (1984), who argues that the syllabification processes in English and Egyptial Arabic differ with respect to whether word boundaries play any role, with the result that English speakers misperceive word boundaries in Arabic and vice versa.

In conclusion, I will argue that the <u>process</u> of syllabification--that is, of setting strings of consonants and vowels to syllables--occurs as an active, ongoing, mental event in the speech production process. It is partly controlled by universal factors (more sonorant sounds are more likely to be syllable nuclei than less sonorant sounds), but also subject to language particular constraints. English allows syllabic nasals under certain limited, unstressed, circumstances, while French does not. French allows syllable-final consonant clusters (for example in 'quatre') that English does not. These processes apply to whatever 'underlying' (that is, mentally stored) strings the speaker has, whether native or foreign, and act as input filters constraining the possible set of underlying strings in the first place. However, despite their filtering effects, they allow for some slippage, particularly in differences between careful and 'sloppy' speech.

Finally, for speakers of one language learning a second, when input is encountered that would lead to impossible syllabifications (from the point of view of the native language) the input can be adjusted, either through the deletion of segments, or through the addition of supplementary segments which will allow the retention of the offending segments (usually consonants) by permitting the consonants to act as syllable onsets rather than codes. The addition of such 'epenthetic' consonants is itself not a natural process (i.e., serving neither morphophonemic nor allophonic speech adjustment roles) but is rather a creative use of language perception, adapting the input to the constraints the native speaker brings to the language.

<u>Bibliography</u>

Aronoff, Mark and Richard T. Oehrle, eds. 1984. <u>Language Sound Structure</u> Cambridge: MIT Press

Bell, Alan and Joan B. Hooper, eds. 1978. <u>Syllables and Segments</u>. Amsterdam: North Holland.

Brière, Eugene, Russell N. Campbell and Soemarmo. 1983. A need for the syllable in contrastive analyses. In Robinett and Schachter.

Broselow, Ellen. 1984. An Investigation of transfer in second language phonology. <u>IRAL</u> 22:253-269.

Donegan, Patricia and David Stampe. 1978. The syllable in phonological and prosodic structure. In Bell and Hooper pp 25-34.

Eliasson, Stig. Ed. 1984. Theoretical Issues in Contrastive Phonology. Heidelberg: Groos.

Eckman, Fred. 1981. On the naturalness of interlanguage rules. <u>Language Learning</u> 31.1:195-216.

Fromkin, Victoria A. 1985. <u>Phonetic Linguistics: Essays in Honor of Peter Ladefoged</u>. Orlando: Academic Press.

Fujimura, Osamu and Julie B. Lovins. 1982. <u>Syllables as Concatenative Units</u>. Indiana University Linguistics Club.

Kahn, Dan. 1976. <u>Syllable-based generalizations in English phonology</u>. Indiana University Linguistics Club.

Kavanaugh, James F. and James E. Cutting, eds. 1975. <u>The Role of Speech in Language</u>. Cambridge: MIT Press.

Kozhevnikova, V. and L. Chistovich. 1965. <u>Speech: Articulation and Perception</u>. Translation: Washington, D.C.: Joint Public Research Service.

Ladefoged, Peter. 1982. A Course in Phonetics. New York: Prentice Hall.

Lindblom, Björn. 1983. Economy of Speech Gestures in MacNeilage, ed.

MacNeilage, Peter F., ed. 1983. <u>The Production of Speech</u>. New York: Springer Verlag.

Robinett, Betty Wallace and Jacqueline Schachter, eds. 1983. <u>Second Language Learning</u>. Ann Arbor: University of Michigan Press.

Sievers, Eduard. 1885. <u>Grundzüge der Phonetik</u> Leipzig: Breitkopf und Härtel.

Stampe, David. 1973. <u>A Dissertation on Natural Phonology</u>. Indiana University Linguistics Club.

Stetson, R. H. 1951. <u>Motor Phonetics</u>. Amsterdam: North Holland.

Studdert-Kennedy, Michael. 1983. <u>Psychobiology of Language</u>. Cambridge: MIT Press.

______ 1975. From Continuous Signal to Discrete Message: Syllable to Phoneme. In Kavanaugh and Cutting.

Tarone, Elaine. 1984. The role of the syllable in interlanguage phonology. In Eliasson.

PREPLOSIVE FØ IN THE PERCEPTION OF /d/-/t/ IN ENGLISH

K.J. Kohler

Institut für Phonetik, Universität Kiel, Olshausenstr. 40, 2300 Kiel, F.R.G.

ABSTRACT

The importance of level vs. falling FØ contours on prestop vowels for the voiced/voiceless categorization is discussed in the light of perception test data for English "widen/whiten" and compared with corresponding data from German. Over the same set of complementary vowel/stop closure durations, level FØ leads to a greater number of /t/ responses than falling FØ.

INTRODUCTION

Kohler (1982, 1985) presented data from German which support the following points:

1. A measurable FØ contour is related to two factors: a global utterance intonation and local perturbations due to articulatory constraints.
2. In utterance-final disyllabic words of the type "leiden/leiten" ['la^{e}dn/'la^{e}tn] a falling terminal FØ contour changes its global character and consequently its meaning when the FØ peak is located either before/at the initial consonant/vowel boundary or right inside the stressed vowel.
3. In the case of a central peak on the stressed vowel, the FØ fall is delayed by a following voiceless vs. voiced stop consonant.
4. In the case of an early peak on the stressed vowel, the local FØ differences before voiced/voiceless stops disappear.
5. In perception, level vs. level+falling FØ patterns on the stressed vowel favor /t/ and /d/ responses respectively, compared with a continuously falling FØ throughout the stressed vowel.

This paper discusses comparable perception data from English.

PROCEDURE

The sentences "I am telling you I said widen/whiten." ['wa^{e}dn/'wa^{e}tn] with focus stress on the final word were the point of departure for constructing a listening experiment according to the principles outlined in Kohler (1985). Fig.1 represents the speech wave and fundamental frequency of the original sentence "I am telling you I said widen.", which was used for deriving the test stimuli. The duration of [a^e] was reduced from its value of 265 ms in the original "widen" to the value in the original "whiten" by six 10-ms steps (=7 Stimuli). To these vowels closure silences were appended which were increased from 70 ms in six 15-ms steps complementary to the vowel shortening. Three FØ patterns were generated with each vowel duration. (a) Level+falling (119-123-85 Hz); the level section represents the naturally produced fluctuation over the first 100 ms of the original [a^e]; the proportion of level to slope sections stayed the same in all the 7 stimuli. (b) Level (119-123-122 Hz). (c) Linearly falling throughout the vowel (119-85 Hz).

The same ranges of vowel and closure silence durations and very similar FØ patterns (as regards absolute values and timing) were used in the English stimuli as in the German ones.

A group of 12 native Southern British speakers were given the task of classifying the stimulus utterances as "widen" or "whiten" sentences by ticking the appropriate boxes on prepared answer sheets.

RESULTS

Fig. 2 shows the identification functions, as well as the binomial confidence ranges at the 5% level. The response curves for falling and level+falling patterns are very close together, except for the duration ratio of .64, and they are significantly different from level FØ at the low and middle duration ratios: level FØ leads to a higher number of /t/ responses.

DISCUSSION

Basically, the same results as for German have been replicated for English. There are two differences, however:

(a) There are generally more /d/ responses in the English test: the functions are shifted towards shorter ratios.
(b) The curves are closer together, and they are no longer separate for falling and level+falling.

These differences could, of course, be attributed to the different languages, and it might even be objected that such a comparison across languages and test groups is not legitimate. But it is possible to explain the divergencies of the German and English data by reference to Raphael, Dorman and Liberman (1975), who showed that the status of the prevocalic consonant influences the voiced/voiceless perception of postvocalic stops. Their results indicate that the longer the initial voiced formant transitions, the greater the lengthening of perceived vowel duration. In the case of English "widen" vs. German "leiden" the same argument applies since the sequence [w]+[a^e] constitutes a vocalic continuum with extremely long transitions and fuzzy segment boundaries, whereas [l]+[a^e] has a much clearer division. Consequently, [w] increases the perceived vowel duration more than [l]. The general strengthening of [d] responses in the English test is in line with these considerations.

Furthermore, a section of about 40 ms before the segmentation point set between [w] and [a^e] in the original stimulus has a level FØ of 119 Hz. It was not affected in the stimulus construction and therefore stayed the same in all three FØ sets. Thus the linearly falling pattern is preceded by a short level FØ, which, together with the fuzzy segment boundary, prevents it from becoming a different global pattern: linearly falling and level+falling FØ across the segmented [a^e] lead to identical response functions.

In conclusion, we can say that the prosody of the entire stressed syllable, i.e. its total temporal structure as well as its pitch contour, determines segmental voiced/voiceless recognition.

REFERENCES

Kohler, K. J. (1982). "FO in the production of lenis and fortis plosives," Phonetica 39, 199-218.

Kohler, K. J. (1985). "FO in the perception of lenis and fortis plosives," J. Acoust. Soc. Am. 78, 21-32.

Raphael, L. J., Dorman, M. F., and Liberman, A. M. (1975). "The perception of vowel duration in VC and CVC syllables," Stat. Rep., Haskins Laboratories 42/43, 277-284.

ACKNOWLEDGEMENT

I am indebted to Dr Andrew Butcher, Reading, for running the listening test.

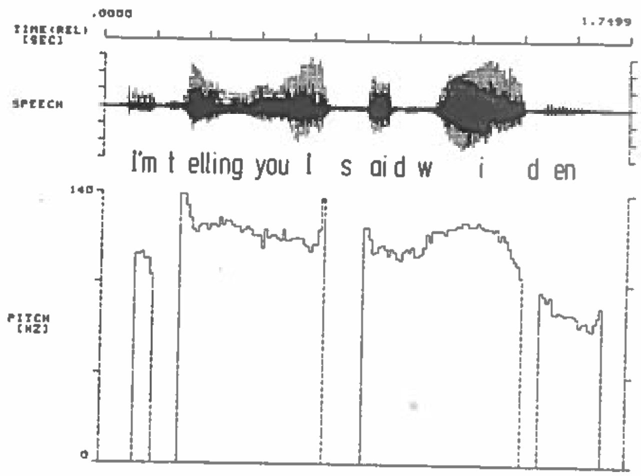


Fig. 1
Speech wave and fundamental frequency of the original sentence "I am telling you I said widen.", which was used for deriving the test stimuli.

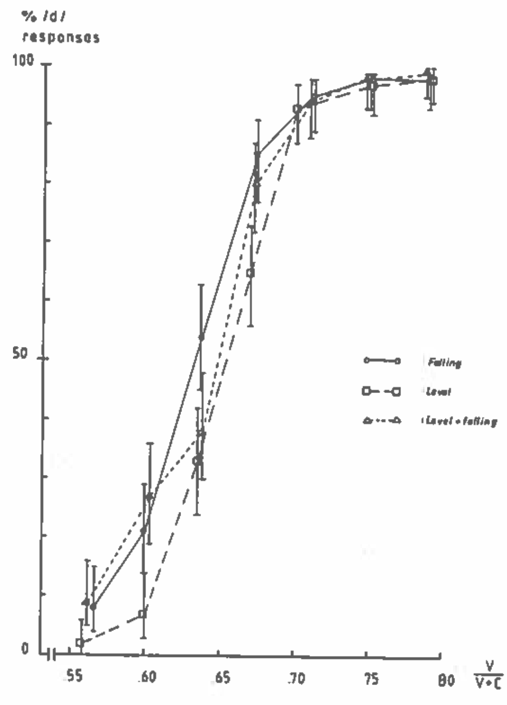


Fig. 2
Percentage /d/ responses as a function of vowel/(vowel + closure) duration ratios for three FØ conditions, and binomial confidence ranges at the 5% level; 12 listeners. At each data point N = 120.

THE EFFECT OF UNSTRESSED AFFIXES ON STRESS BEAT LOCATION IN ENGLISH

Robert Allen Fox and Ilse Lehiste

Speech & Hearing Science, Department of Communication, The Ohio State University, 324 Derby Hall; and Department of Linguistics, The Ohio State University, 204 Cunz Hall; Columbus, OH 43210 (This research was supported by Grant #1 R01 NS21121-01 from NINCDS to RAF, principal investigator).

Introduction

Although listeners commonly hear speech as rhythmical (Donovan & Darwin, 1979; Lehiste, 1972) it is not the case that the perception of rhythmicity arises from acoustic onset isochrony. For example, if sequences of monosyllables whose initial consonants differ in manner of articulation, are presented to listeners so that the acoustic onset-to-onset intervals are isochronous, the rhythm of the sequence will sound irregular. These sequences will sound regular to listeners only if systematic deviations from acoustic isochrony are introduced (Morton, Marcus, & Frankish, 1976; Fowler, 1979, 1983). Talkers behave in a similar manner in that when required to produce rhythmic sequences of monosyllables which contain different initial consonants the same kinds of deviations from isochrony are found (Allen, 1972a,b; Rapp, 1971; Fowler, 1979; Fowler & Tassinary, 1981). The term "stress beat" or "perceptual center" has been used in the literature to reference that point (or psychological event) in a stimulus upon which listeners/talkers base their rhythmic judgments.

In the past 15 years, a number of experimental studies have been directed at identifying the parameters which determine the location of this stress beat in both perception and production. Experimental results have supported the assertion that the stress-beat location is not universally linked to any particular articulatory or acoustic event, but rather can be shifted by the acoustic/articulatory characteristics of the entire syllable. For example, we have been engaged in research over the past two years examining the influence of several different phonetic parameters on the location of the stress beat in stressed CV or CVC monosyllables in both production and perception tasks. We have found that final consonant variations can shift the location of the stress beat for both talkers and listeners--an effect opposite in direction, but smaller in degree than, the shifts obtained by Fowler (1979; Fowler & Tassinary, 1981) when manipulating the initial consonant (Fox & Lehiste, 1985a). Similar results have been obtained when the medial vowel was modified (for both listeners and talkers); namely, that the stress-beat location shifts to a point later in the token as vowel duration increases.

The present study is a continuation of this line of inquiry and examines the effect of unstressed prefixes and suffixes upon the stress-beat location of stressed syllables in American English. Although the results to be presented today stem from a production task only (which we considered to be, necessarily, the first step in our research program), we anticipate that the listening tests will again show a similar effect. These data, then, should provide information about the relative timing of syllables in <u>both</u> production and perception and thus will provide relevant information about speech timing for speech recognition purposes.

Method

<u>Talkers</u>: There were three highly practiced American English talkers--two female, one male--naive to the purposes of the experiment.

<u>Stimuli</u>: The basic stimuli consisted of sets of seven-token sequences similar to those used by Fowler (1979; and Fox & Lehiste, 1985a,b). Each sequence was composed of 7 identical tokens, such as:

<u>peer</u> <u>peer</u> <u>peer</u> <u>peer</u> <u>peer</u> <u>peer</u> <u>peer</u>
<u>appear</u> <u>appear</u> <u>appear</u> <u>appear</u> <u>appear</u> <u>appear</u> <u>appear</u>

The tokens were either stressed monosyllables (the <u>basic form</u>) or 2-, 3-, or 4-syllable tokens. The latter were formed from the monosyllable by adding an unstressed prefix (e.g., <u>a</u>-, <u>con</u>-/<u>com</u>-, <u>de</u>-, <u>be</u>-) or an unstressed suffix (e.g., -<u>er</u>, -<u>ing</u>, -<u>able</u>), or both, to the basic form. Where possible, a prefix which, when appended to the basic form, produced a real word was chosen. The syllabic structure of the basic form included CV, CVC, CCVC, and CVCC. The initial and final consonants of the basic form included oral and nasal stops, fricatives, and liquids. In all cases the stressed syllable of the multisyllabic tokens corresponded to the basic form. Altogether 601 different tokens were constructed. The sequences were put into random order and presented to subjects on a CRT screen under the control of a PDP 11/23 computer, one at a time, in blocks of 52 (including distractor sequences).

<u>Procedure</u>: Talkers were instructed to read the 7-token sequence which appeared on the screen and to produce them in a rhythmic fashion "in time" with the timing pulse. If a talker was dissatisfied with his/her production on any trial, the talker was instructed to repeat the sequence. After successful completion of a trial, the talker hit the return key on the terminal which replaced the old sequence with the next sequence. Intersequence intervals were thus self-timed but averaged 2 sec in duration. The timing pulse was a 1000-Hz pulse, 100 msec in duration. The stimulus onset asychrony (SOA) between the timing pulses was 1000 msec. Talkers heard the timing pulses continuously throughout the experiment. Talkers were given a short break after every third block of stimuli.

<u>Measurements</u>: For each different basic form an acoustically defined point in the stressed syllable was selected which would, presumably, not change in its basic nature when prefixes and/or suffixes were added. These points included stop consonant release in those stressed syllables beginning with a stop (e.g., <u>do</u>, <u>bide</u>, <u>pose</u>, tone, broad), onset of medial vowel in those syllabes beginning with a fricative (e.g., <u>cede</u>, <u>seal</u>/-<u>ceal</u>), etc. The onset of this measurement point, relative to the onset of the timing pulse, was determined for each token. Of interest in this study is to determine whether or not the position (in time) of these measurement points shifted as a function of adding unstressed prefixes or suffixes. Although the stress beat does not seem to correspond to any particular acoustic event (cf. Fowler 1979; Marcus, 1981), we assume that if the position of these measurement points shifts in affixed tokens, relative to the unaffixed, basic form, it will indicate a concomitant shift in the stressed syllable's stress beat location.

Results and Discussion

Since the location of the acoustically defined measurement points differs across different stressed syllable types (e.g., those having syllable-initial stops vs. fricatives vs. liquids), it makes little sense to compare them directly across all affix conditions. However, if we take the location of the measurement point relative to the timing pulse in the <u>basic form</u> as a baseline location, we can calculate the <u>shift</u> of the measurement point, relative to the basic form, for all affixed versions of each basic form. To do this, the onset of the measurement point (relative to the timing pulse) of the basic, monosyllabic form of each token was subtracted from the onset of the measurement point in each of that form's variations. For example, the onset of the stop release (relative to the onset of the timing pulse) of the [p] in peer was subtracted from the stop release onsets (relative to the timing pulse) of the [p] in peer, peerer, peering, appear, appearer, and appearing. The resulting number indicates the shift of the measurement point relative to the basic, monosyllabic form. Shown in Table 1 are the mean shifts obtained for those tokens which were prefixed with a-, de-, and con-. Positive numbers indicate a shift of the measurement point to a position later than that of the basic form, negative numbers a shift to an earlier position.

Table 1. Mean shifts in "measurement points" in affixed conditions. Data are normalized relative to onset of measurement point of unprefixed, unaffixed basic form (with defined shift of 0.0). All data are in msec.

		Suffix			
		None	-er	-ing	-able
a- (N=137)	no prefix	0.0	-3.7	-2.9	-5.9
	prefixed	26.6	22.1	11.5	----
de- (N=35)	no prefix	0.0	-12.7	-4.3	-7.9
	prefixed	34.7	31.6	34.0	----
con- (N=22)	no prefix	0.0	-14.4	-18.5	-5.2
	prefixed	86.3	56.7	59.5	----
MEAN	no prefix	0.0	-6.5	-4.8	-6.2
	prefix	36.0	28.6	22.5	----

Although only of borderline significance in each case ($p<.08$, 1-tailed t-test and Wilcoxon), there is a small mean shift (to an earlier point) of the measurement points in the suffixed forms relative to the unsuffixed forms. This indicates that the location of the stress beat may occur later in the token when additional phonetic elements are appended and the overall duration of the token is increased. This result is consistent with the data obtained by Fox & Lehiste (1985a,b) who demonstrated that such shifts can be obtained by manipulating the medial vowel and final consonants of stressed monosyllables.

There is a much larger mean shift of the measurement point (relative to the basic form) in the opposite direction (i.e., later) when tokens have an unstressed prefix; all these shifts are significant at the .001 level (2-tailed t-tests). This indicates that the addition of a prefix shifts the location of the stress beat to a point earlier in the token. It is interesting to note that the a-, de- and con- prefixes produce a progressively greater shift of the location of the stress beat, respectively. This is most likely explained by the fact that although all three prefixes can be considered "unstressed," they are really not all unstressed to the same degree. The a- prefix usually has the least amount of stress and the con- prefix the most.

In order that these data can be examined globally in terms of the relative contribution of prefix, suffix, and prefix+suffix combinations, analysis of variance was done on a subset of these data--namely the a- data. The dependent variable used in this analysis was the time of the release of the initial stop consonant of the stressed syllable, relative to the onset of the timing pulse. These raw data were used instead of the normalized data because (1) one cell of the normalized data would have a variance of zero and (2) the stop release measure represents the same articulatory and acoustic event in all tokens. Shown in Table 2 are the relevant data averaged over basic forms and subjects.

Table 2. Mean onset of initial stop release of stressed syllable relative to onset of timing pulse in a- prefixed tokens. All measurements are in msec.

	Suffix			
	None	-er	-ing	MEAN
no prefix	-31.9	-39.6	-36.0	-35.8
prefixed	2.0	-1.0	-14.6	-4.2
MEAN	-15.0	-20.0	-25.3	-20.0

A two-way, repeated measures, analysis of variance (using basic form as the random variable) with the factors PREFIX and SUFFIX was done. The results showed a significant main effect due to PREFIX ($F(1,28)=34.2$, $p<.001$), a marginally significant main effect due to SUFFIX ($F(2,28)=2.28$, $p=.05$), but no significant PREFIX by SUFFIX interaction.

These results suggest that the location of the stress beat in stressed syllables in English can be affected by the addition of either an unstressed suffix or an unstressed prefix or both. The effects of such affixes on the stress beat are additive and independent of each other. In addition, the prefix seems to shift the stress beat differentially, as a function of its degree of stress. We are currently analyzing these data in terms of how well the durations of the prefix, affix, stressed syllable, etc. can predict the shifts in the measurement points (and, indirectly, the stress-beat location) and are conducting the appropriate, corresponding listening tests.

References

Allen, G.A. (1972a,b). The location of rhythmic stress beats in English: An experimental study. Parts I and II. Lang. Speech, 15, 72-100, 179-195.

Fowler, C.A. (1979). "Perceptual centers" in speech production and perception. Perc. & Psychophys., 25, 375-388.

Fowler, C.A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. J. Exp. Psych.: Gen., 112, 386-412.

Fowler, C.A. & Tassinary, L. (1981). Natural measurement criteria for speech: The anisochrony illusion. In J. Long & A. Baddeley (eds.), Attention and Performance, IX. Erlbaum.

Fox, R.A. & Lehiste, I. (1985a). The effect of final consonant structure on syllable onset location. J. Acous. Soc. Am., Suppl 1, 77, S54 (Abstract).

Fox, R.A. & Lehiste, I. (1985b). The effect of vowel quality variations on stress-beat location. J. Acous. Soc. Am., Suppl. 1, 78, S20 (Abstract).

Donovan, A. & Darwin, C. (1979). The perceived rhythm of speech. Proceedings of the Ninth International Congress of Phonetic Sciences, 2, 268-274.

Lehiste, I. (1972). Rhythmic units and syntactic units in production and perception. J. Acous. Soc. Am., 54, 1228-1234.

Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. Perc. Psychophys., 30, 247-256.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). Psych. Rev., 83, 405-408.

PHONOLOGICAL/PHONETIC OPPOSITIONS: BINARY OR GRADUAL? SOME EXPERIMENTAL CONTRIBUTIONS TO THE CURRENT ISSUE BASED ON THE ANALYSIS OF ITALIAN DATA FROM THE POINT OF VIEW OF SPEECH RECOGNITION.

P.Bonaventura(*),L.Prina Ricotti(*),J.Trumper(**)

(*) Fondazione "Ugo Bordoni"
Viale Trastevere, 108 - 00153 Roma, Italy

(**) Dept. of Linguistics
Universita' delle Calabrie, Cosenza, Italy

Abstract- In the present paper some evidence' is given for the existence of a gradual phonetic change in Italian stop consonants from the point of view of their defining distinctive features.

The four features of Mode 1 (Voicing), Mode 2 (Continuity), Place and Timing are assumed to be perceptually effective and are examined from the point of view of their significant correlation.

The experiment used synthetic stimuli for CV groups of phonemes obtained from an acoustic model which allows one to vary continuously the acoustic characteristics associated with the distinctive features that are being examined.

1. Foreword

This paper is based on the assumption that the acoustic or articulatory categories detected on the physical continuum are not homogeneous with the corresponding perceptual ones; in order to define such categories, it is essential to describe the relative variation of the significant parameters (see also [8]) on their own perceptual scales.

We have prepared a table (see table 1) of the relative values of the four most significant perceptual features for Italian consonants: in this system Timing is considered as a feature in itself which varies in combination with the others, but along a typical continuum.

It is thus obtained the "intrinsic" time of each perceptual entity (or "res percepta").

The table shows all the possible ratios between the relative values fixed for the experiment but it could be expanded (within given boudaries) to give a more suitable frame for the complex reality of a natural language.

2. Theoretical approach

Multivalue features scales commonly used (see [3], improvement on [7], etc.) are based on extrapolation from experimental observation on articulatory processes.

Such processes are effectively gradual: their graduality is implicit to the performance time of muscular commands.

The basic assumption of such approaches is that the sum of values (no longer conceived in binary terms) of a fixed number of features that are selected on the bases of their economicalness or "naturalness" defines each phoneme.

Scales are defined by giving fixed values to the beginning and end of possible continua which are segmented into different range groupings according to the language in question: e. g. possible cuts along a place of articulation continuum are:

α) 0 1 2 — /p/ /t/ /k/ — say English, Italian, Finnish.

β) 0 1 2 3 — /p/ /t/ /c/ /k/ — say Albanian.

γ) 0 1 2 3 4 — /p/ /t/ /k/ /q/ /?/ — say Classical Arabic.

We evince from (α) - (γ) that the possible continuum that defines the Front - Back, Frontedness or Place features is composed of the following positions that define phoneme ranges:

0	1	2	3	4	5
p	t	c	k	q	?

However, rather than claim that such levels or possible ranges of a continuum (cut into a graded plane or Gradatum) - and we can allow that levels may be only two (usual for Voicing) at the phonological level, as in English or Italian, or even only one as in Finnish, though n - valued at the phonetic level - are bound to the production level of our model (articulation), we would claim that when in fact we say that the Frontedness or Place features has three levels in Italian we are really referring to the perceptual interpretation of a position of the articulatory-acoustic space, to a process in which speech production categories are mapped on to corresponding perceptual categories.

The scheme we propose to represent the three phases of the whole process involved is as in Diagram I.

We can admit that the gradualness of features is effectively codified at the 2nd LEVEL, where perception takes places on the basis of a set of perceptual features that refer to the constant relations between physical values individuated along a perceptual scale formed of acoustic parameters given in the LEVEL ONE INPUT (i.e. capacity to select given parameters of the human ear). This has a filter function with respect to the acoustic signal and allows for the trasmission of only certain components of the complex signal.

It is at this level that binary choices (see [2]) operate and are observed effectively to operate, though uniquely on the acoustic form of a given feature.

The scope of the present paper is - by means of straightforward perceptual experiment of identification - to give a demostration of the non-categoricalness (or gradualness) of perception, that, as we shall see, operates on the basis of precise rules connected in the neural topology and functioning.

This renders discrete the continuum of acoustically selected parameters and combines segments obtained as a parameter of time (this parameter at the 2nd LEVEL corresponds to the intrinsic timing factor in [1].

The full set of these perceptual relation (particularly complex - we shall skip over details here, but the question is being studied) furnishes a definition of three phonemes belonging to the class of STOPS based on the reciprocal values of three essential perceptual features that we have so far identified.

Phonemes are organized on the basis of the values evinced from the perceptual scales for each features that describes a phoneme as a res percepta; this organization is schematized in table 1.

Numbers are not numbers in a set a natural numbers, but exist uniquely in a relational plane.

3.Method

We have varied a first parameter (F2) along a continuum composed of constant intervals of 100 Hz each obtaining 13 variable stimuli ranging across three levels identified respectively as the articulatory categories: LABIAL, ALVEO-DENTAL and VELAR.

The V.O.T. of the components of this sequence was also varied in 10 ms steps for negative V.O.T. and in 5 ms steps for positive V.O.T., affecting the first part of the transitions.

On such a sequence a simple labelling test was performed with a set of 18 unexperienced native italian listeners; a straightforward identification test was chosen as in our experience on sequences involving the variation of a single feature ranging within the classes of Stops, Fricatives and Affricates, any kind of discrimination tests, both ABX and 4IAX, gave poor results.

This is due to the procedure of the test in itself that affects perception in presenting too long sequences with respect to the temporal capacity of STM [4], in fact the initial part of the signal are lost and can't be processed in relation to the last parts.

Several other experimental procedures ([8], [10]) seem to give some evidence for a non categorial perception at the acoustic (our 1st) level, while categorical perception at the phonetic (our 2nd) level.

We are actually interested in verifying if the relative degrees attributed to the perceptual features in table 1 play an effective role in the perception process.

From our standpoint it is thus enough to check the labelling thresholds through an identification task.

4.Results and Discussion

The mean results of the identification tests are shown in Diagram 2. We assume that the correct proportion among the amounts of the cues for each perceptual feature involved must be maintained constant to obtain a definite categorization, otherwise the identification scores will not be significant, resulting in misunderstandings or even no labelling at all of the stimuli.

Timing is assumed to be fixed for all the consonantal parts of the syllables. It pertains to each phoneme and is calculated from the relative times of the parameters (see below).

Table 2a shows the values of the actual cues of the three features along the acoustic continuum; Table 2b shows the combinations of the actual values of the perceptual features as obtained from Table 1 with the corresponding labels attached to every stimulus according to their labelling rate in the test.

As shown in Table 2b the phoneme /g/ is characterized by the triple 0 1 0. While Mode 2 is unaltered, the cues of Mode 1 and Mode 3 are changed: Voicing moves, through regular intervals, from value 1 to value 0 and Place from 0 to 3.

The two extremes are characterized in Diagram 2 by the categorization as /ga/ for stimuli n. 1 to n. 5 and as /pa/ for n. 11 to n. 13. In the intermediate range (stimuli n.6 to n.10), the bias to perceive a /ka/ at the 5th and 6th steps can be explained considering that the variation of the cue of Voicing to the value 0 is not yet strong enough to interact with that of Place and to polarize identification in the Alveolar area. In the range of stimuli with a set of values similar to the initial one, only the variation of Voicing is relevant in perception (/ka/: 0 0 0).

Stimuli n.7 to n.10 are either perceived as /da/ or nonidentified at all as a speech sound. Perception of /da/ (triple 2 1 0) is explained as an interpretation of an ambigous Place value 1 (only values 0 to 2 are distinctive for Stops) as a level 2, coupled with a previous level 1 of Voicing

The possible evaluation of a level 1 of Place corresponds to the occurrence of nonexistent triples, either 1 0 0 or 1 1 0 , that determines the exit from the speech mode, given that this situation is not predicted in the subsystem of Italian we already described. The interval where there is a restructuring of unexpected values with a fictitious combination or quite none is considered a black hole in perception.

5.Conclusions

The results are consistent with the assumption that each single amount of any acoustic cue is not relevant in itself to select a feature, as the auditory system filters the signal transmitted by the receptors using a special code, activated when the listener is in the speech mode (see [9]). This code is a structure formed by the relation between the temporal amounts of every significant parameter in input and the whole temporal frame of the actual segment as it is realized. This means that a special rule binds the relative times of every single parameter among them, and with the global time that they share to form a " res percepta "; the correct proportions can be predicted from the relations of a complete model of perception and production (P.Bonaventura,"Preliminary studies for an MCC model of perception", to be published).

This could be a tentative explanation also for the insertion, erasure and reajustement of phonemes along the speech chain.

REFERENCES

[1] Fowler,C.A.: "Coarticulation and Theories of Extrinsic Timing". Report Speech Research SR-57

[2] JaKobson,R. Fant,G. Halle,M.: "Preliminaries to Speech Analysis". Massachussets Institute of Technology Press, Cambridge, Mass., 1963

[3] Ladefoged,P.: "Preliminaries to Linguistic Phonetics". Univ. of Chicago Press, 1971

[4] Massaro,D.W.: "Understanding Language". Academic Press, New York, 1975

[5] Massaro,D.W. Cohen,M.M.: "Categorical or Continuous Speech Perception: a New Test". Speech Commun.,2,1983

[6] Massaro,D.W. Cohen,M.M.: "The Contribution of Fundamental Frequency and Voice Onset Time to the /zi/-/si/ Distinction". JASA,60,1976

[7] Miller,G.A. Nicely,P.: "An Analysis of Perceptual Confusions among some English Consonants". JASA,27,1955

[8] Oden,G.: "Integration of Place and Voicing Information in the Identification of Synthetic Stop Consonants". Journal of Phonetics, 6, 1978

[9] Pisoni,D.B. Lazarus,J.H.: "Categorical and Noncategorical modes of Speech Perception along the Voicing Continuum". JASA,55,n.2,1974

[10] Sawusch,J.R. Pisoni,D.B.: "On the Identification of Place and Voicing Features in Synthetic Stop Consonants". Journal of Phonetics, 2, 1974

Diagram 1

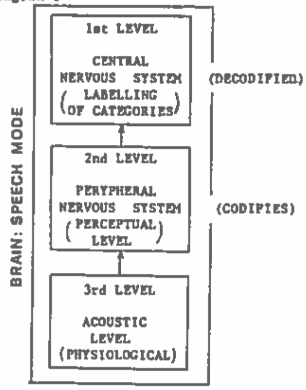


Table 1

Phoneme Symbols	m	n	ɲ	r	l	ʎ	f	s	ʃ	ts	tʃ	p	t	k	v	z	dz	dʒ	b	d	g
MODE 1 (Voicing)	2	2	2	2	2	2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
MODE 2 (Continuity)	0	0	0	3	1	1	3	3	3	1	1	0	0	0	3	3	1	1	0	0	0
MODE 3 (Place)	3	2	1	2	2	1	3	2	1	2	1	3	2	0	3	2	2	1	3	2	0
TIMING	t^{m}	t^{n}	$t^{ɲ}$	t^{r}	t^{l}	$t^{ʎ}$	t^{f}	t^{s}	$t^{ʃ}$	t^{ts}	$t^{tʃ}$	t^{p}	t^{t}	t^{k}	t^{v}	t^{z}	t^{dz}	$t^{dʒ}$	t^{b}	t^{d}	t^{g}

Diagram 2

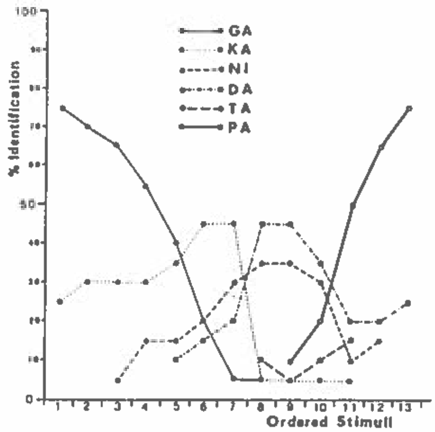


Table 2a

Stimulus Number	Parameter Values		
	F2	V.O.T.	Burst Duration
1	2200	+90	5
2	2100	+80	5
3	2000	+70	5
4	1900	+60	5
5	1800	+50	5
6	1700	+40	5
7	1600	+30	5
8	1500	+20	5
9	1400	+10	5
10	1300	0	5
11	1200	-5	5
12	1100	-10	5
13	1000	-15	5

Table 2b

Feature Values			Labels
Place	Voicing	Continuity	
0	1	0	ga
0	1	0	ga
0	1	0	ga
0	1	0	ga
0	1	0	ga
0	0	0	ka
0	0	0	ka
2/1	1/0	0	da/ni
2/1	1/0	0	da/ni
2/1	1/0	0	da/ni
3	0	0	pa
3	0	0	pa
3	0	0	pa

A MODEL OF THE PERCEPTIVE PHONETICS, ATTENDED BY THE HUMAN MEMORY

S.J.Mrchev

ul."Jordan Mishev" 21A,8600 Jambol,Bulgaria

The ifluence of the perceptive phonetics for systems with AI is actualized,in model describing: 1)HUMAN MEMORY (sensor and imagine-bringing instantaneous memory;short-term memory - direct,operative,buffer;long-term memory,super long-term and meta memory) 2)PERCEPTIVE PHONETICS (The zone perceptive basis of the natural language has been reserved in the long-term human memory like standarts and principles of these standarts,with the segmental and supersegmental common and complicated language units and their features).

Recent developments in perceptual phonetics - part of the study of human perception of speech are associated with advances in the fields of psycho-linguistics,knowledge engineering and applied AI,pattern recognition,etc.

We assume a zonal organization of templates in long-term memory (LTM) with the following structure :

- there are some primary (atomic?) phonetic units expressed by some domain of the space of values of certain parameters (each of them corresponding to a single measurable physical characteristics)the rest of the units being compound and corresponding to composite systems of domains of the parameters;
- to compound templates there correspond parameters of two types (type A and tape B). The units of type A are compound units for which the set of characteristics is the same for any two opposite units and these are distinguished only by the integral value of the compound parameters,e.g. the compound characteristics of accented and unaccented syllable: they both have the same set of parameters,such as duration of the vocal part, duration of the consonant part,intensity of the syllabic peak,frequency of the basic tone,etc.;they have non-intersecting regions of values of tho compond parameters (so that syllables with or without accent could be told appart).The units of type B are distinguished from opposite units by the existence of a parameter which is absent in the representation of the counterpart (any phoneme is an example of this type of unit);
- thus in contrast to units of type A the identification of units of type B may be based on a specific set of characteristics and not on an integral compound characteristics (as happens in case of units of type A).Based on experimental data,a hypothesis is put forward in 2 that the compound parameters of units of type B can themselves be composed by units of type A.In particular,distinct differential characteristics of the phonemes,occuring in different units,can be established by summing up the values of its components ;
- the templates in LTM of the phonetic units which correspond to sound images in EIM can be represented as zones of identical perception (ZIP).These ZIPs correspond to regions in the space of parameter values in which any two realizations are identified.So any change of the values of the parameters within the limits of the region leads to perceptually indistinguishable realizations. Such a view on the functioning of the templates is founded on ignoring in the perceptual basis of the language of the variations which are small.On the other hand identical reaction to phisical features that are "near" enough is phisiologically natural.In this respect it resembles the law of "all or nothing" ;
- an immediate neighbourhood of a ZIP is the zone of similarity to the template (ZST).The ZIPs of disctinct phonetic units do not intersect,moreover they have non-interecting closures in the topology,generated by the notion on nearness,while the ZST may well have non-empty common parts and this is one of the explanations for ambiguous perception ;
- for units that do not have a corresponding sound images the existence of a zone of identical reactions can also be conjectured as well as of zones of similarity ;
- the categorical character of speech sounds' perception is rejected,i.e. we do not need the notion of different speech sounds being comprehended in two completely different ways:"categorial" and "non-categorial";
- the boundaries of the zones (in particular of ZST) are quite unstable.This could explain the process of chage of the phonetical background of a language.The unstability of the boundaries have been established by experiments and it seems to be a result of different extralinguistic factors.A very substantial shift in the boundaries can be observed when a specific psychological attitude is adopted during the experiment - a fact that leads sometines to assimilative or contrastive perceptual illusions,and for this matter should be taken into account when determining templates' boundaries by phonetic experiments.

Under units of primary perception we understand templates for such segments and supra-segments of the speech flow that are operative in establishing the "sounding of an utterance.In experiments with uncommon combinations of consonants the stimili have been comprehended with big distortion.This fact shows that the units of primary perception are not the phonemes,i.e. in theperception of unusual combinations of consonants comparison is carried out not with the templates of some phonemes,but with templates of their combinations.If in the set of templates in the perceptual basis of the human mind there is no suitable tamplate (exactly fitting) the sound image is mapped to all the nearest such templates (in the topology) and to all combinations of them until a suitable combination is found and a satisfactory similarity is established.Of course,another possible explanation is that phonemic templates are indeed the templates of primary units and un the perception of a soundimg word a simultaneous correction is taking place.But data from 2 and 5 supports the view that this is not the case and that the units of primary perception are not the phonemes,but sertain their compounds,in particular - the syllables.One more reason for this is the fact that in experiments with perception of syllables the reaction time for single phonemes is much greater than the reaction time for syllables themselves.Thus one is bound to insist that the real formative units are the syllables.

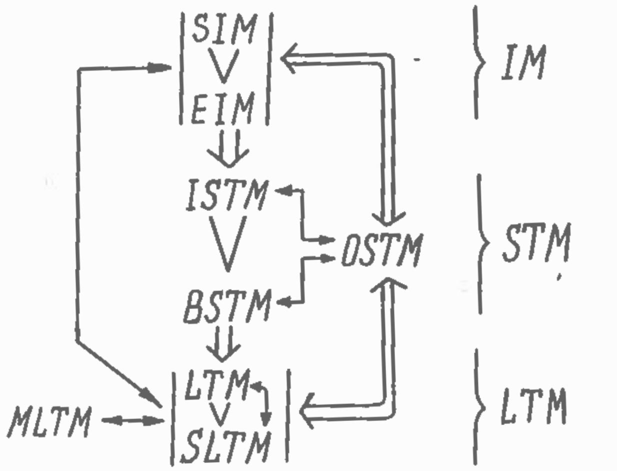
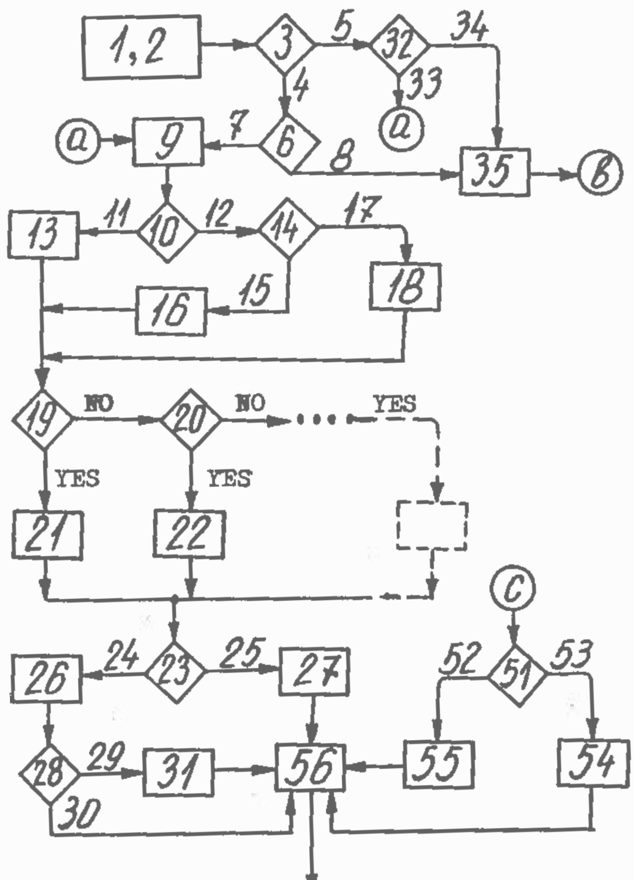


Fig.1.Internal structure of human memory(HM) (Instantaneous memory: SIM - sensory instantaneous memory; EIM - eidetical instantaneous memory. Short-term memory:OSTM - operative short-term memory;ISTM - immediate short-term memory;BSTM - buffer short-term memory. Long-term memory:LTM - long-term memory; SLTM - super long-term memory;MLTM - meta long-term memory. ⟹ , ⟹ - information flow, ⟵⟶ - control and feed back).



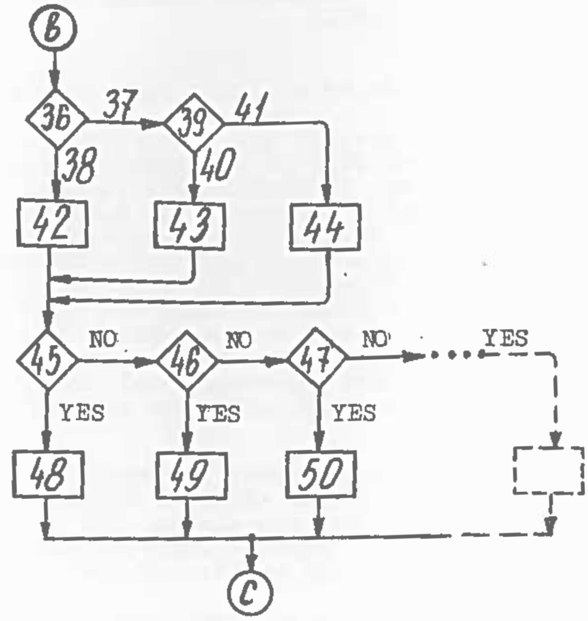


Fig. 2. A flow-chart of a part of the verbal phonetic perception based on the system of human memory (1.Phonetic units as speech chains in SIM and EIM;2.Templates in LTM;3.Is the phonetic unit a segment or a suprasegment? 4.Segment; 5.Suprasegment;6.Do the templates correspond to sound images or not?7. Corresponding to sound images templates;8. Otherwise;9.Comparison in OSTM of segments from EIM with templates from EIM with templates from LTM;10.Are the compared images simple or compound?11.Simple;12.Compound;13.Comparison of the simple segment image (coming from EIM) with primary templates;14.Is the compared compound image of type A or type B? 15.Type A;16.Type A comparison (using integral values);17.Type B;18.Comparison of type B(coincidence of all component parameters and nearness of their values);19.Is the image-unit from EIM an intonation model?20.Is it a rhythmic structure?21.Comparison with appropriate templates from LTM;22.The same; 23.Do we have a perfect fit (i.e. we are inside the ZIP)?24.In the ZIP;25.In ZST;26.Interacting OSTM and LTM recognize the unit of the semantic zonal space of LTM;27.A similarity is established;28.Are parameter compensating?29.Compensating parameters;30.Not the case;31.The compound unit is a "syllable/type A";32.The suprasegment unit has a template of sound image? 33.Sound image;34.Not a sound image;35.Comparison with templates which are not templates of sound images;36. Simple or compound? 37.Simple;38.Compound; 39.Type A or type B? 40.Type A; 41.Type B;42. Comparison in OSTM of the simple image, coming from EIM, with templates from LTM;43.Comparison of compound images of type A;44.Comparison of compound images of type B;45.Is this unit-image a feature of phonemes? 46.Is it a bemolity? 47.Is it a diffusion? 48.Establishing the corresponding property;49.Establishing a bemolity;50.Establishing a diffusion;51.Is the fit exact? 52.Yes,it is in ZIP;53.No,it is in ZST;54.Evaluation of the nearness;55.Interacting OSTM and LTM recognize the composition (identity reaction);56. Forming the recognized unit-percept; 57.Ready for a new round.).

SYLLABLE-BASED PHONOLOGICAL RULES AND THEIR IMPLICATIONS FOR SPEECH RECOGNITION

Daniel Kahn

Bell Communications Research ("Bellcore")
435 South Street
Morristown, NJ 07960 USA

Abstract

Rules can be written which describe with fair accuracy the perceived syllabic structure of English. Once syllabic structure is established, many important phonological rules find natural expression in terms of this structure. In particular, phonemes tend to be modified under the influence of conditions that exist within the syllable in which they reside or when they play a particular role within their syllable. These observations provide support for the syllable-based approach to speech recognition, but the explicit rules that arise from syllabic phonology are applicable to phoneme-based recognition as well.

1. Introduction

Phoneticians as well as workers in the field of automatic speech recognition (ASR) are well aware of the lack of anything close to a one-to-one correspondence between the phonemes of a language and acoustic events. While the complexity of the mapping from phoneme to sound does not preclude the creation of an effective ASR device whose basic unit of recognition is the phoneme, it is clear that the success of such an undertaking is dependent upon discovering the large set of relevant context-sensitive rules, making them explicit, and encoding them in the recognizer. Even proponents of such an approach recognize the enormity of the task (cf. Zue, 1985).

Whole-word template-matching (cf. Itakura, 1975; Rabiner & Levinson, 1981) is an approach to ASR which appears to obviate the need for the long and difficult program of discovery of the details of the phoneme-to-sound mapping. In this technique, no explicit decision is made regarding where in time each phoneme lies and what its identity might be. Rather, for each word in a vocabulary, a reference template is created consisting of a set of spectral representations computed at regular intervals in time, on the order of every 10 msec. The sequence of spectral representations of a word to be recognized is then compared to each of the templates (after time-normalization) and the unknown word is taken to have the same identity as the template to which it has the least total spectral "distance," appropriately computed.

Whole-word matching works very well for recognizing small vocabularies of words spoken in isolation. As vocabulary size increases, a disadvantage of this approach become apparent: a new template must be created, stored, and included in the distance calculation for each additional word in the vocabulary. In addition, much of the advantage of whole-word matching is lost in continuous speech, since word boundaries are not easily determinable and, in any case, cross-word-boundary phonology can greatly alter the isolated form of words.

It has occurred to several ASR researchers that most of the advantages of the phoneme-based approach (finite vocabulary size, straightforward extension to continuous speech in many cases) and of the whole-word template-matching method (no need for explicit representation of many complex contextual effects) can be combined in an approach to ASR in which the basic unit is the syllable or demisyllable. Inherent in the advocacy of syllable-based recognition is the assumption that most contextual variation on the part of phonemes is due to the influence of other phonemes within the same syllable, and that the effects of the environment outside the syllable in which a given phoneme lies can for the most part be considered second-order (cf. Fujimura, 1975; Mermelstein, 1975; Kahn et al, 1984).

In the last ten years several groups have taken important first steps toward the implementation of high-performance (demi)syllable-based recognition systems (e.g., De Mori et al, 1976; Ruske & Schotola, 1978; Zwicker et al, 1979; Hunt et al, 1980; Ruske, 1982; Rosenberg et al, 1983), and it is to be hoped that this work will continue.

I too have performed some (very preliminary) work in syllable-based (Kahn, 1982, 1983) and demisyllable-based (Rosenberg et al, 1983; Kahn et al, 1984) recognition, but the present paper is concerned with the linguistic motivation for the use of (demi)syllabic units in ASR. I believe, however, that not only does the phonological analysis discussed below argue for the wisdom of the (demi)syllable approach, but also that the explicit rule formulations that are an output of the syllable-based analysis can profitably be used in phoneme-based recognition.

2. The syllable in English phonology

In many languages it is obvious to native speakers how words of their language are to be syllabified, but English has both clear (*reply* = *re−ply*, not *rep−ly* or *repl−y*) and unclear (*pony* = *po−ny* or *pon−y*?) cases. This apparent indeterminateness has led the authors of many formal accounts of English phonology to deny the syllable a role in linguistic descriptions. This is unfortunate, because the concept of "syllable" is intuitively meaningful even to speakers of languages like English, and also because many phonological rules call out for descriptions in terms of the syllable, if only the concept could be formalized.

In Kahn (1980) I suggested an analysis of English syllable structure that I feel accounts well for both the clear and unclear cases of word syllabification, as well as for the syllabification of phrases in the case of continuous speech (where a syllable may extend across a word boundary). Most important, once syllabic structure is established in accordance with this analysis, many important phonological rules (sound modifications) can be expressed in a natural and compact way in terms of the syllable. In the limited space available here I will try to outline the analysis of English syllabification and discuss some examples of syllable-based rules. In all cases, I will have to omit details which may be significant but which do not, I believe, affect the correctness of the basic analysis.

2.1 Analysis of words and phrases into syllables

There is little controversy as to how many syllables a normally-spoken word contains. At the core of each syllable is exactly one vowel or other "syllabic" phoneme (like [n̩] of *button*). Each syllable will also contain zero or more non-syllabic phonemes (which I will imprecisely refer to as "consonants") before and after the vowel. Clearly any word-initial (-final) consonants must reside in the first (last) syllable of the word. Thus the question of interest is whether, in words of more than one syllable, to associate consonants that stand between two vowels with the preceding or following syllable.

In this regard, it is surely significant that any polysyllabic word of English can be broken down into syllables each one of which could stand alone as an English word without breaking the constraints on permissible *word*-initial and -final clusters. Thus English has words like *hamster*, corresponding to the permissibility of word-final /m/ and -initial /st/, but none like *hamkter* since there is no analysis of /mkt/ into permissible clusters. A natural conclusion from this observation is that English simply has a set of permissible *syllable*-initial and -final clusters, from which the facts about *word*-initial and -final clusters fall out as an immediate consequence.

The question remains how to correctly predict syllabifications in cases where more than one analysis is consistent with the cluster constraints (why *re−ply*, not *rep−ly*?). The answer appears to reside in the "maximal initial cluster" (MIC) principle: a syllable boundary is placed in a sequence of between-vowel consonants as far left as possible, consistent with the initial/final cluster constraints.

The MIC principle alone will, in general, predict correct syllabifications for what were referred to above as the "clear" cases. Even in the unclear cases, MIC appears to be correct, provided we look at overly precise, very-slow-speech pronunciations. In such speech we observe *po−ny*, not *pon−y*; *ci−ty*, not *cit−y*; *Pa−trick*, not *Pat−rick*.

Before returning to normal-rate syllabifications, it will be helpful to introduce a graphical representation of syllabification. Fig. 1 indicates that the word *reply* consists of two syllables, *re* and *ply*. Note that if we impose the natural constraint that the lines connecting syllables and phonemes may not cross, a whole class of syllabifications, like that in Fig. 2 in which the /r/ of *reply* is a member of the *second* syllable, become, quite appropriately, impossible to represent.

Now suppose that there are no further constraints on linking syllables and phonemes (aside from the one-syllable-one-vowel principle mentioned earlier). Then in addition to the syllabification of *pony* shown in Fig. 3, which, as noted above, is appropriate for the slow-speech pronunciation of this word, we might try to interpret the syllabification of Fig. 4. In Fig. 4, the /n/ of *pony* is shown as belonging simultaneously to both syllables, i.e., as being "ambisyllabic." I would suggest that this is the normal-rate syllabification of the word. The native speaker's inability to assign the /n/ of *pony* unambiguously to one or the other syllable in the normal-rate pronunciation of the word would then be attributed to the /n/ being ambisyllabic at normal rates (and in fact some phoneticians, in informal descriptions of English syllabification, have suggested that such consonants might be shared by two syllables). We can formalize the structural change in going from slow to normal speech as the addition of the line of association between /n/ and the first syllable.

The consequences of such an analysis go well beyond formalizing the intuition that certain consonants in English do, and others do not, reside fully in one syllable; there are phonological implications as well. For example, the simple rule "vowels become nasalized in English when followed by a nasal consonant in the same syllable" accounts for the /õ/ of *tone* and normal-rate *pony* alongside the /o/ of *poke* and slow-speech *pony*. French nasalized vowels are the result of a similar rule (*an* vs. *année*). Sect. 2.2 is concerned with examples of this type of rule.

We have not yet discussed under what conditions we observe ambisyllabicity; for ex., as opposed to *pony*, the syllabification of *reply* has the simple form given in Fig. 1 for both slow and normal speech. As discussed in more detail in Kahn (1980), it appears that the initial consonant of an *unstressed* syllable becomes ambisyllabic with a preceding vowel-final syllable. Thus it is the stress on the second syllable of *reply* that blocks ambisyllabification of the /p/.

To this point we have been discussing the syllabification of words in isolation. Turning to continuous speech, let us note first that it is always at least possible to pause between words, so a reasonable approach to continuous speech would be to postulate an initial level at which syllabification is in accordance with the "word-is-an-island" rules of the preceding paragraphs, with additional lines of syllabic association across word boundaries added by "continuous-speech rules." The most important of these rules appears to add a line of association (e.g., the dotted line in Fig. 5b) between the final consonant of a word and the initial syllable of a following *vowel-initial* word. This rule of "trans-word-boundary ambisyllabification" (TWA) can be understood when it is recalled that the clearly preferred syllable structure among the world's languages is ...CV-CV..., not ...VC-VC... Within words, this fact is reflected in the MIC principle. MIC is powerless, however, in the case of a word that happens to start with a vowel. In continuous speech, the unnatural situation of a vowel-initial syllable is remedied, where possible, by TWA. Thus *rocket* and *rock it*, syllabically distinct in slow speech (solid lines of association in Fig. 5), become homophonous at normal rates (addition of dashed lines).

2.2 Rules sensitive to syllabic structure

Many important phonological rules of English (and other languages) are best described in terms of syllabic structure. The outline of English syllabic structure given above is sufficient to illustrate several of these rules.

It is well known that the voiceless stops, and in particular /t/, take very different form as a function of environment. For example, /t/ is an aspirated stop in *tack*, an unaspirated stop in *stack*, a "flap" in *city* (Am. and Can. pronunciation) and is glottalized in *sit*. I would suggest that the rules responsible for these forms state that /t/, underlyingly an unaspirated stop, is aspirated when only syllable-initial, flapped when ambisyllabic, and glottalized when following a vowel and not syllable-initial. It is straightforward to confirm that these rules operate properly in simple cases like the words just cited, but the rules make other testable predictions. Thus in the phrase *Let Ann do it* we expect - and observe - glottalized /t/ in *let* if there happens to be a pause after the word but flapped /t/ in continuous speech, where TWA has applied. Similarly, in overprecise speech, where the (within-word) ambisyllabification rule fails to apply, the /t/ of *city*, normally ambisyllabic and flapped, has syllable-initial association only, and is aspirated. Of course, rules such as the ones that account for the various allophones of /t/ *could* be stated without reference to syllabic structure, but they would be grossly complicated, and would in fact be restating the independently-needed rules of English syllabification within the specific allophonic rules (cf. Kahn, 1981).

In standard British English and in parts of the Eastern U.S., /r/ is deleted in certain environments where spelling and the more "conservative" dialects would have it pronounced. The rule accounting for these facts, as it entered the language, is clearly syllable-conditioned and takes very much the form of the /t/-glottalization rule. Thus /r/ is lost when not syllable-initial, as in *form*, *for me*, *for(pause)Ann*, but is retained in *forest*, where /r/ is syllable-initial by MIC (and, irrelevantly, also syllable-final at normal rate by ambisyllabification), and *for(no-pause)Ann*, where /r/ is syllable-initial by TWA. French "liaison" is a more complex, though clearly related, phenomenon. If we regard a word like *vous* as consisting of the phonemes /vuz/ at an abstract level, and delete /z/ when not syllable-initial, then the TWA-like rule of French will account for *vous avez* [vuzave] vs. *vous l'avez* [vulave].

There is another very large class of rules which are clearly syllable-conditioned but differ in having been "frozen" at the lexical level. In most dialects of English, the vowel of *car*, through the influence of the following back phoneme /r/ (which until quite recently was pronounced in *all* dialects), has a distinctly more back quality than the vowel /ae/ of *cat*, *cap*, etc. (As suggested by the spelling, the vowels of *car*, *cat*, etc. were at one time identical.) The /ae/ of words like *carry*, however, was unaffected by the rule that modified *car*. We can account for these facts by imposing the natural condition that /r/ be fully in the syllable of the vowel it follows for it to have the backing effect. In accordance with this rule, words like *card* also have the backed vowel. The rule is "frozen" in the sense that words whose base form became subject to the rule now show the backed vowel even in non-base forms which should not be subject to the rule. Thus *starry* has the vowel of *star*, not of *carry*. Similar rules have affected other vowels: *her*, *herd* (vowel modified by /r/) vs. *hem*, *herring* (not).

A similar rule, but in the domain of consonants, accounts for the loss of /g/ in *long* [loŋ] vs. its retention in *longer* [ŋg]. Basically, /ŋg/ is simplified to /ŋ/ except when /g/ is syllable-initial. In the case of words of the form VngC...V, this rule correctly predicts [ŋ] without [g] (e.g., *angstrom* and *Yngve*) *except* when C is such that /gC/ forms a permissible initial cluster: *angry* (cf. grow), *linguist* ["ling-gwist"] (cf. Gwendolyn). Previous, non-syllabic analyses of *ng* did not properly account for these facts and could be made to only through explicit reference to the differential behavior of *gs* etc. vs. *gr* etc.; but clearly the correct course is to state the latter distinction once and for all in the (independently-required) permissible-cluster rules.

Additional examples of syllable-conditioned rules could easily be cited. At this point, however, let us note that a common feature of the rules that have been discussed is that they involve major changes, as viewed by the phonetician. That is, these rules delete segments or replace one well-defined phonetic element with another. Another class of rules, not generally considered to be in the realm of traditional phonology, deals with phenomena at a lower level. Thus, for ex., the phonetician (and the native speaker) hears the /i/'s of *bee* and *Dee* to be identical, even though the initial parts of the two vowels are spectrally quite distinct, due to the formant-transition phenomenon. Although the separation between a phoneme causing an acoustic modification and the modified phoneme is sometimes surprisingly large, it is probably fair to say that the strongest effects are found within the syllable and thus might be regarded as simply very-low-level syllable-based phonetic rules (cf. Malmberg, 1955; Fujimura, 1975, 1976).

3. Conclusion

This paper has been concerned with syllable-based phonetics and phonology and their relevance to ASR. Whether one is attempting to predict what phonemes are allowable in a particular environment or the precise acoustic shape of a given phoneme, local syllabic structure is most often found to be significant. In ASR systems based on syllabic units, such dependencies come "built-in." Even to the worker committed to phoneme-based ASR, however, syllable-based phonology is relevant because it offers compact and explicit formulations of many phoneme realization rules.

References

De Mori, R., Laface, P., & Piccolo, E., "Automatic detection and description of syllabic features in continuous speech," *IEEE Trans. on ASSP 24:5*, 365-79 (1976).

Hunt, M. J., Lennig, M., & Mermelstein, P., "Experiments in syllable-based recognition of continuous speech," *Int. Conf. on ASSP*, 880-3 (1980).

Itakura, F., "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on ASSP 23:1*, 67-72 (1975).

Fujimura, O., "Syllable as a unit of speech recognition," *IEEE Trans. on ASSP 23:1*, 82-87 (1975).

Fujimura, O., "Syllables as concatenated demisyllables and affixes," *J. Acoust. Soc. of America, 59:* S55(A) (1976).

Kahn, D., *Syllable-based Generalizations in English Phonology*, Garland Publishing, New York (1980).

Kahn, D., "Syllable-structure specifications in phonological rules," in *Juncture*, Aronoff, M. & Kean, M.-L., eds., Anima Libri (1981).

Kahn, D., "A syllable-parsing algorithm for telephone-quality speech," *J. Acoust. Soc. of America, 72:* S30 (1982).

Kahn, D., "A syllable-based connected-digit recognizer for continuous speech," *J. Acoust. Soc. of America, 74:* S31 (1983).

Kahn, D., Rabiner, L. R., & Rosenberg, A. E., "On duration and smoothing rules in a demisyllable-based isolated-word recognition system," *J. Acoust. Soc. of Am., 75:2*, 590-8 (1984).

Malmberg, B. (1955), "The phonetic basis for syllable division," in *Readings in Acoustic Phonetics*, Lehiste, I., ed., MIT Press.

Mermelstein, P., "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. of Am., 58:4*, 880-3 (1975).

Rabiner, L. & Levinson, S., "Isolated and connected word recognition - theory and selected applications," *IEEE Trans. on Comm., 29:5*, 621-659 (1981).

Rosenberg, A. E., Rabiner, L. R., Wilpon, J. G., & Kahn, D., "Demisyllable-based isolated word recognition system," *IEEE Trans. on ASSP 31:3*, 713-26 (1983).

Ruske, G., "Automatic recognition of syllabic speech segments using spectral and temporal features," *Int. Conf. on ASSP*, 550-3 (1982).

Ruske, G. & Schotola, T., "An approach to speech recognition using syllabic decision units," *Int. Conf. on ASSP*, 722-5 (1978).

Zue, V. recently estimated that another "30 to 40 years" of work in acoustic phonetics for ASR remained to be done and that what is currently known is just the "tip of the iceberg" (IEEE-ASSP Workshop on Speech Recognition, Dec., 1985).

Zwicker, E., Terhardt, E., & Paulus, E., "Automatic speech recognition using psychoacoustic models," *J. Acoust. Soc. of Am., 65:2*, 487-98 (1979). 2/18 09

Figure 1 Figure 2 Figure 3 Figure 4 Figure 5a Figure 5b

Syllable Network for Phonemic Decoding of Speech

V. Gupta, M. Lennig, J. Marcus, and P. Mermelstein**
Bell-Northern Research
3 Place du Commerce
Nuns' Island, Montreal, Quebec
Canada H3E 1H6

The decoding of speech into phonemes for large vocabulary speech recognition is made more reliable by restricting phoneme sequences to those which compose valid syllables. To apply this restriction when decoding a sequence of phonemes, we use a syllable network representing the valid syllables in Webster's 7th Collegiate dictionary.

Since major allophonic variants of a phoneme are determined by the phoneme's position within the syllable (e.g., prevocalic vs. postvocalic /r/), the syllable network can be used to represent allophonic variation by employing distinct allophone models of a phoneme in different positions within the network. A preliminary experiment using the syllable network in large vocabulary recognition to select appropriate Markov models for allophones shows promising results.

1.0 Introduction

In this paper, we describe the use of a syllable network when decoding speech as a sequence of phonemes in large vocabulary speech recognition. Phonemic decoding of speech without any restriction on valid phoneme sequences leads to a large number of hypotheses which do not obey the phonotactic constraints of the language. We have used a syllable network to restrict the possible phoneme sequences to correspond to sequences of valid syllables. The syllable network also serves to control the choice of positional allophones. Allophonic variation is represented by using different Markov sources (Bahl et al., 1983) for a given phoneme depending upon its position within the syllable network.

2.0 Syllable Network

A syllable network for English which generates all and only the 8157 English syllables is necessarily complex. Such a network can be obtained by first constructing a tree of all possible syllables and then merging the tree from both ends. Simpler networks overgenerate the English syllabary. We have constructed a syllable network of intermediate complexity to achieve a compromise between network complexity and overgeneration.

The syllabic onset, nucleus, and coda are the subunits of the syllable within which the tightest phonotactic constraints obtain (Selkirk, 1982). Thus, our syllable network includes separate subnetworks for each of these three subunits. The syllable network generates phoneme sequences of the form

$$(O_1(O_2(O_3)))N(C_1(C_2(C_3(C_4))))$$

where O_i stands for a consonant in the syllabic onset, N for the vowel in the syllabic nucleus, and C_j for a consonant in the syllabic coda. The parentheses imply that the segment is optional. Only the nucleus is compulsory in the syllable. The subnetwork for the onset allows a maximum of three consonants, while that for the coda allows a maximum of four.

The syllable network was created based on the 60,000 phonemic transcriptions contained in Webster's 7th Collegiate dictionary (henceforth, *the dictionary*). Starting with a rudimentary network, branches were added iteratively to account for syllables in the dictionary not generated by the network. The resulting network has 76 nodes and over 300 branches.

The phonotactic constraints can be tightened further by using a separate syllable network for each syllable position within the word. The maximum number of syllables for any word in the dictionary is 10 (except for one word which was excluded). The number of valid syllables decreases with increasing syllable position number within the word (Table 1). Note that the set of syllables which occur in the first position includes all syllables which can occur in any position.

Syllable position in word	Number of distinct syllables
1st	8157
2nd	6181
3rd	3931
4th	1718
5th	724
6th	306
7th	110
8th	36
9th	12
10th	2

Table 1. Number of distinct syllables possible at each position within the English word.

3.0 Use of the Syllable Network to Select Allophones

Allophonic variants of a phoneme are often determined by the phoneme's position within the syllable (e.g., prevocalic, postvocalic, intracluster). For example, the phonemes /l r w/ differ significantly in their prevocalic and postvocalic realizations. First and second formant trajectories move upward in most contexts when these phonemes appear in prevocalic position, while the formant trajectories move downward when these phonemes appear in postvocalic position. By using separate Markov sources for allophones which differ in position, we can account for such variation.

In some cases, allophones are conditioned by a more detailed positional specification. For example, the allophones of the nasal consonants which occur in the syllable-initial clusters /sm/ and /sn/ are realized as partially devoiced with a very short nasal murmur. Also, devoiced allophones of the phonemes /w j r l/ occur when preceded by a voiceless fricative as in *switch, few, three,* and *slide.* Allophones which are difficult to account for with the syllable network

* Also with INRS-Télécommunications, University of Quebec.

are those which depend on larger contexts than the syllable. For example, [ɾ], the flapped allophone of /t/, occurs ambisyllabically after a stressed and before an unstressed vowel, as in *butter*, pronounced [bʌɾɚ].

4.0 Preliminary Recognition Results

In a series of speaker-dependent, isolated word recognition experiments using the syllable network, the unknown word is decoded as a sequence of syllables, where each syllable corresponds to a path through the syllable network. Each of the syllable network's transitions is mapped to a Markov source allophone model. In the experiments we report, we vary this mapping. First, all occurrences of a phoneme are represented by a single Markov source. Then, separate Markov sources are used to represent a given phoneme occurring in the syllabic onset and in the syllabic coda. We use statistical decoding to compute between 200 and 600 most likely syllable sequences corresponding to words in the 60,000-word dictionary. Since our system does not employ a language model, all 60,000 words are assigned equal a priori probability. Thus, the perplexity of this task is 60,000.

The training set consists of 800 word tokens from arbitrary texts, 60 distinct words chosen to contain consonant clusters, and 100 distinct CVC words, where C stands for a stop or a liquid, i.e., one of the consonants /p t k b d g r l/.

Two test sets were used (see Appendices). The first, denoted *Chrysler*, is a 99-word automobile advertisement. The second is a 100-word list of CVC words where C is a stop or a liquid, having no words in common with the CVC training list. 59% of the words in the Chrysler test set and 6% of the words in the CVC test set are represented in the vocabulary of the training set. Training and test sets are disjunct.

Two experimental conditions are compared:

(1) One Markov source (one allophone) for each of the 39 phonemes in the syllable network.

(2) Stops and liquids are represented by two allophones each. One Markov source is used in the syllabic onset, the other in the syllabic coda. Other phonemes are represented by one allophone each.

The recognition results in Table 2 show the percent correct recognition in the top n phonetic transcriptions, where n is either 1, 5, 20, or 100. Use of distinct allophones for the stops and liquids as they occur in the syllabic onset and coda improves the performance only for the CVC test set.

test set	condition	$n=1$	$n=5$	$n=20$	$n=100$
Chrysler	(1)	60%	81%	91%	94%
	(2)	62%	81%	89%	94%
CVC	(1)	15%	36%	54%	67%
	(2)	21%	56%	77%	88%

Table 2. Percent correct recognition in top n choices.

5.0 Conclusions

The syllable network provides a convenient framework for the selection of different allophonic models depending upon a phoneme's position within the syllable. Separate allophones of stops and liquids for the syllabic onset and coda lead to a significant improvement in recognition of CVC words. The fact that no significant improvement is observed in recognition of arbitrary text suggests that a more general representation of allophonic variation in the multisyllabic environment and more complete training appropriate to that environment are necessary.

6.0 References

Bahl, L.R., Jelinek, F., and Mercer, R.L. (1983). "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-5(2), 179-190.

Selkirk, E.O. (1982). "The Syllable," *The Structure of Phonological Representations*, (Foris Publications, Dordrecht, Holland), 336-383.

Appendix: Chrysler Test Set

begin paragraph here is the confidence of front hyphen wheel drive comma the security of advanced electronics and the quiet comma smooth ride you expect in a fine luxury car period begin paragraph and here are the luxuries you demand period automatic transmission comma power windows comma power steering comma power brakes comma power remote mirrors and individual reclining seats standard period begin paragraph and finally comma here is the new technology of turbo-power period more power to move you period to accelerate period to pass period to cruise in serene comfort ellipsis yet with remarkable fuel efficiency period.

Appendix: CVC Test Set

but could back write put god book cut dead pull bed role top bad deal date doubt care look rock lip tool lack pair tear cup pale load pour dare dear kick tip leap cop lobe rob rub cab tub gale gag tag pig log bog rogue gab goat guile ball lower bit roll bird beat cool tall root coat rout luck core cat rare tale Paul coal pike beer pot peer tail cape robe lab goad dug gape tug dip rot rat cot cod tight tide tuck tack lull roar lure rope ripe reap rip pile tile curd pearl.

USING DIPHONES IN LARGE VOCABULARY WORD RECOGNITION

C. Vicenzi - D. Sciarra

Central Research Department
Elettronica San Giorgio, ELSAG S.p.A.
Via Puccini, 2 - 16154 Genova - ITALY

In this paper we present a large vocabulary, speaker dependent, isolated word recognition system with diphones as basic units, so that the training session is much faster and useful for any application. The system, tested on a vocabulary of 910 words on one speaker, gave a word recognition rate of 78%, slightly lower than an Itakura recognizer with whole word templates (WRR=85%).

INTRODUCTION

In a template-matching recognition system for large vocabulary applications, speaker dependence still seems to be an essential requirement for a satisfactory performance. On the other hand the classical and most common approach to isolated word recognition, the whole-word template matching, presents a serious drawback. In fact a training session where the whole vocabulary has to be uttered, even only once, becomes time consuming. Moreover, one or more repetitions of each word will be necessary for the extraction of reliable templates. The only practical solution to the problem is to use some kind of sub-word units to represent words. We chose the diphones, that, in our definition, include transitions between two phonemes, small portions of steady-state sounds and some longer transitional elements embracing three phonemes [1,2]. These units provided good performance in speaker-dependent connected speech recognition experiments with small and medium size vocabularies [3]. Moreover the diphones proved to be robust and economic units, as they are quite invariant with the context and a set of about 300 of them (corresponding to less than 400 templates) is sufficient to cover the whole Italian lexicon.

ISOLATED WORD RECOGNITION USING DIPHONES

The use of diphones is particularly appealing in speaker-dependent applications, as the training session for a new speaker, consisting in the utterance of a set of meaningful sentences, is only few minutes long. By means of an automatic technique [4], a diphone template inventory suitable for any application in the Italian language can then be derived from the collected speech material.

In the language model with diphones as basic units we assume that time warping may be allowed only during stationary diphones; templates for these units consist of a single spectral state, and appropriate lower and upper duration bounds ensure the time alignment capability. No warping is allowed on transitional diphones, whose templates consist of a sequence of spectral states of specified duration.
The model of a word consists of a lattice of diphones, where appropriate duration bounds are associated to each diphone. Alternative paths are present in order to deal with different possible pronunciations or phonetic variations [1]. Building up a word prototype as a lattice of diphone templates gives an accurate representation of the word, that is expected to work as well or better than the relevant whole-word template, as was shown in experiments on small-vocabulary connected word recognition [3]. As an example, similar words should be better discriminated as their representations coincide except for the actually phonetically different portions.
However, in the recognition of isolated words with no syntactic constraints, the use of a general lattice model becomes unsuitable, as the computational load and memory requirements of the decoding strategies may sensibly grow when the vocabulary size increases, making it hard to achieve a real-time performance. A compromise solution may be obtained if we consider that, in a classical isolated word matching, faster strategies can be implemented; in fact, as the speech model within a word consists of a regular lattice of spectral states, the same transition rules can be applied to any state.
Our approach then makes use of a diphone description of word templates in order to minimize the storage requirements, but, during the recognition phase, a spectral state description is recovered to speed up the matching. When building a word template, its lattice representation is translated into as many single path prototypes as needed, each one composed of a sequence of diphone labels and associated duration bounds. Each diphone label is then a pointer to the beginning of the spectral description of a diphone template in a common area containing the inventory. In the current implementation each spectral state description consists of 12 LPC Cepstral parameters computed every centisecond on a 25.6 msec portion of a 10 KHz sampled signal.
When a word prototype has to be matched in the recognition phase, its diphone label sequence is used to fetch the appropriate sequence of spectral states and to build in a work area a synthetic prototype according also to the duration bounds of each diphone. The input word, isolated by an end-point detection algorithm, can then be matched against each expanded prototype using an isolated word recognition approach and producing a cumulative distance score.
In a preliminary stage of our work, two Dynamic Programming algorithms were tested, obtaining essentially the same results. The former is derived from classical Itakura D.P. equations where weights are attached to skip and duplication transitions; the duration of stationary diphones is adjusted to the value that approximately gives the estimated duration bounds for that sound when the 2:1 warping of Itakura's equations is applied. In this way a sort of synthetic whole-word template is built, and the matching strategy loses any information

about the diphones that originated it. The latter algorithm (the one used to carry on the experiments) is more closely related to our diphone language model, as it allows time warping to be performed on stationary diphones only, giving a broader range of compression ratios than the usual 2:1 . In this matching strategy the transition portions of the reference pattern, as well as the mimimum duration portions of stationary diphones, are always completely traversed (no duplication, no skip), while skipping to the next diphone is only permitted on stationary diphones when their minimum allowed durations have already been reached.
The implementation of this technique has shown to be very efficient and less time consuming than the conventional ones; dynamic programming choices are not made at every frame of the reference pattern, but only on limited portions of it, corresponding to the variable length part of stationary diphones.

EXPERIMENTAL RESULTS

The complete approach was tested on the recognition of isolated words from the vocabulary of 910 names beginning with the same consonant "B". In a whole-word template training session made by a cooperative speaker it would take about 4 or 5 hours (with no breaks) to collect a single repetition of the entire vocabulary. Stress effects were not considered.
In our experimentation, a female speaker uttered a set of 36 meaningful sentences in a connected way, which constituted the training speech material for the extraction of the speaker-dependent diphone inventory. This session lasted ten minutes only.
An automatic bootstrapping procedure was then applied to extract the diphone templates: a forced recognition step was employed to determine the boundaries of each diphone occurrence in all the training sentences; the first occurrence of each transitional diphone was chosen as a template, while for each stationary diphone a clustering technique was applied to choose among all its occurrences one or more "representative" ones as templates. Generation of the templates for the words in the vocabulary was then automatically obtained by translating their orthographic forms into corresponding diphone sequences.
Two repetitions of the 910 words of this vocabulary were also collected from the same speaker, and an end-point detection procedure was applied to each word; we will refer to them as SET A and SET B.
In the first experiment a classical Itakura isolated word recognizer was run using in turn sets A and B as test or reference patterns (tests I1 and I2). Both of these experiments, as shown in Fig. 1, gave a Word Recognition Rate of 85%; in both cases, also, in 97% of the times the correct word was classified within the tenth position. These numbers were used as reference scores for the following experiment, where the diphone based isolated word recognizer was tested. Using SET A as test patterns, the diphone based templates gave a WRR of about 78% (see Fig. 1, test D) which is significantly lower; anyway, correct classification score within the N top candidates rapidly converges to that of I1 and I2 tests, indicating that the adopted approach should still be refined in order to achieve a better discrimination among similar words.
In fact, a qualitative inspection of the classification errors occurred, convinced us that, while the diphone language model seems to be adequate, in most cases misrecognition has to be ascribed to local confusion generated by diphone templates for some particular classes of sounds (such as liquids). We believe that a more accurate generation of the template inventory will yield more satisfactory WRR results.
This problem will be the focus of future work, together with the implementation of a sub-system that should restrict the number of word prototypes to be matched by means of a gross preclassification algorithm based on classes of diphones.

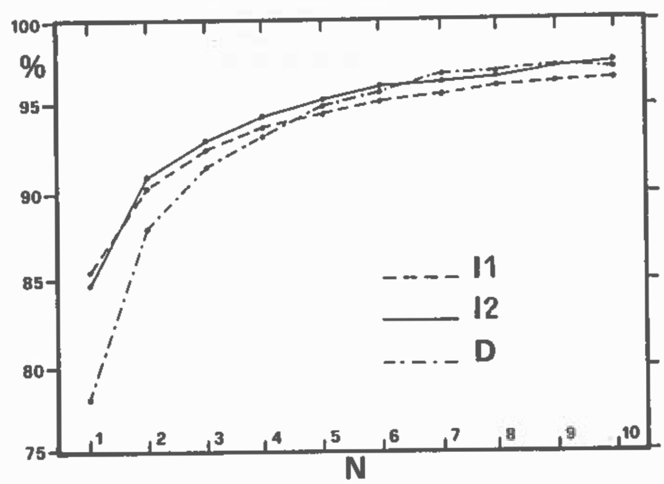


Fig. 1: Word recognition rates within the N (N= 1,..., 10) top candidates in the experiments I1, I2, D (see text).

REFERENCES

[1] A.M. Colla, C. Scagliola, D. Sciarra, "A Continuous Speech Recognition System using a Diphone-Based Language Model", Proc. ICASSP 1985 (31.9), Tampa (1985)
[2] A.M. Colla, "Some Considerations on the Definition of Sub-Word Units for a Template-Matching Speech Recognition System", Proc. Montreal Symp. on Speech Rec., Montreal (1986)
[3] C. Scagliola, "Language Models and Search Algorithms for Real-Time Speech Recognition", Int. J. Man-Machine Studies. Vol.22, n.5, pp. 523-547 (1985)
[4] A.M. Colla, D. Sciarra, "Automatic Diphone Bootstrapping for Speaker-Adaptive Continuous Speech Recognition", Proc. ICASSP 1984 (35.2), S. Diego (1984)

EXPERIMENTS ON THE USE OF DEMISYLLABLES FOR AUTOMATIC SPEECH RECOGNITION

G. Ruske

Lehrstuhl f. Datenverarbeitung, Techn. University of Munich, Franz-Joseph-Str. 38, D-8000 Muenchen 40, Fed. Rep. of Germany

Abstract: The paper describes methods for an explicit segmentation of the speech signal into demisyllable segments by evaluating the output of a loudness model. Syllable nuclei are indicated by maxima of a smoothed loudness function. Consonant clusters and vowels are introduced as decision units in order to reduce the inventory of classes. Two methods for classification of consonant clusters are compared: template matching and a feature extraction approach based on acoustic cues. Sentence recognition operates on phonetic word models adapted to the demisyllable structure.

1. INTRODUCTION

An important question in automatic speech recognition is the choice of basic units which have to be processed basically by the system. A segmentation procedure tries to divide the speech signal into individual parts (segments) in such a way that they can be processed as independently as possible. The segmentation can be performed implicitly when classification of the segments and determination of the segment boundaries are carried out in common. However, this usually requires an enormous expenditure of computing power. On the other hand, segmentation can be carried out explicitly by placing definite segment boundaries in the speech signal; classification now only has to treat the fixed segments. In this case, however, the system must be prepared for the fact that the segmentation step may cause errors, too. The subsequent stages of the system have to be able to correct these segmentation errors (see Sect. 5).

The speech recognition system described in this paper starts from an explicit segmentation into demisylables. These processing units have the advantage that the main coarticulation effects are contained within the segments. The number of classes can be drastically reduced when consonant clusters and vowels are used as decision units for classification.

Evaluation of the syllable structure in the speech signal is facilitated by using a loudness model of hearing /1/ for preprocessing. This model consists of a critical-band-rate filter bank with 24 band-pass filters; 22 channels are used in the system (50 Hz - 8.5 kHz). All channels are processed by a loudness model which simulates the masking effects in hearing. The outputs of the model are sampled every 10 ms; the 22 components constitute a so-called loudness spectrum, see fig. 1a. The total loudness N(t) is calculated as the sum of all 22 components; additionally a weighted sum of these components gives the so-called modified loudness $N_m(t)$ which is very useful for syllabic segmentation. Fig.1b displays the block diagram for the calculation of these functions.

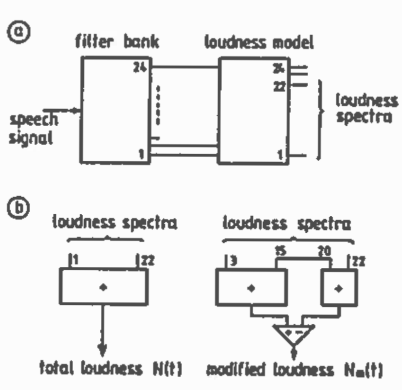


Fig. 1. a) Block diagram of preprocessing; b) calculation of N(t) and $N_m(t)$.

2. DEMISYLLABLE SEGMENTATION

The modified loudness $N_m(t)$ evaluates the frequency range which is dominated by the vowels. Therefore this function is especially suited to indicate the syllable nuclei (vowels and diphthongs). When this function is smoothed according to the average syllable rhythm in the speech signal, the local maxima of this function indicate the positions of the syllable nuclei. For this purpose a special smoothing filter (digital low-pass filter) has been applied having a Gauss-like impulse response h(t), see fig. 2; in the digital calculation this function corresponds to h(n) with $n = n \cdot \Delta t$ (10ms). This smoothing filter has been realized on the basis of an elementary filter with a rectangular impulse response; the output sample y(i) is calculated from the input signal x(n) as:

$$y(i) = 1/3\ (x(i-1) + x(i) + x(i+1))\ .$$

When this filter is placed k-times in series, the impulse responses of fig. 2 result. The repeating factor k now determines the time constant T of the filter, see fig. 2. This smoothing filter is applied to $N_m(t)$. The time constant T has been optimized using test material consisting of 23 sentences spoken six times; the speech material contained 2566 syllables altogether /2/. It is important to adjust the time constant T to the speaking rate: for a short time constant T many surplus syllable nuclei are marked (insertions), for long time constants T many nuclei are smoothed out resulting in omissions. Both effects contribute to the total segmentation error rate as depicted in fig. 3. It can be seen from the figure that an optimal value for T was reached for k=7 corresponding to a time constant T=55.7 ms (this is equivalent to a cut-off frequency of the filter f_g = 9 Hz). The minmum error rate was 3.66% (from 2566 syllables /2/). It has to be borne in mind that here only the maxima of the optimally smoothed function $N_m(t)$ were evaluated. A further reduction in the segmentation error rate is achieved by evaluating the spectral information at the positions of the maxima indicated by $N_m(t)$ /3,4/. As an extreme solution, a complete vowel classifier can be applied at each time instant in order to estimate the syllable nuclei /2/. In the realized recognition system a combination of both methods was implemented which has a segmentation error rate of about 4-8% in practical applications with continuous speech.

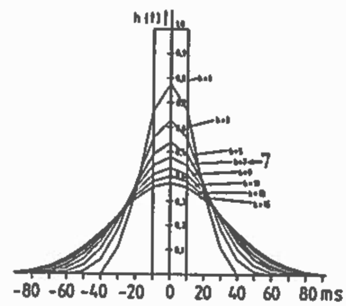


Fig. 2. Impulse response h(t) (from /2/).

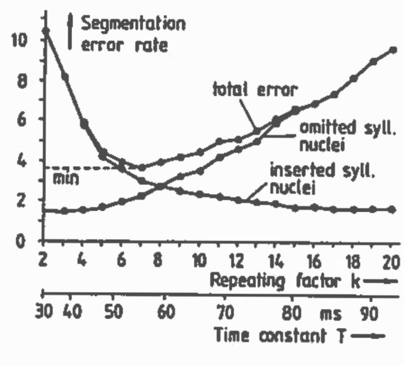


Fig. 3. Segmentation error rate for syllable nuclei as a function of T (from /2/).

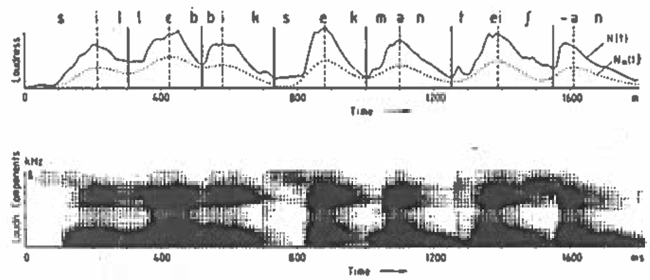


Fig. 4. Demisyllable segmentation of the utterance "syllabic segmentation".

Syllable boundaries are placed at local minima in the loudness N(t) between two consecutive syllable nuclei. When more than one minima are present, the lowest minimum is chosen /3/. This method yields in most cases a suitable boundary. The demisyllable segment now spans the range from the syllable boundary to the syllable nucleus, see fig. 4.

4. CLASSIFICATION OF CONSONANT CLUSTERS

Each demisyllable segment contains a consonant cluster and a part of the vowel from the syllable nucleus. The number of different units can be drastically reduced when consonant clusters and vowels are introduced as decision units for the classification. In the German language we only have to discriminate about:

- 50 initial consonant clusters,
- 20 vowels (inclusive diphthongs),
- 160 final consonant clusters.

That means that the demisyllable is seen as a segmentation and processing unit but not as a decision unit for recognition. In this way the huge inventory of different demisyllables can be avoided while preserving the advantages of demisyllable segmentation.

4.1 Classification by template matching

A first approach to recognition of the consonant clusters consists in using complete spectral-temporal templates of all consonant clusters. For this purpose a special time normalization procedure was developed called 'dynamic interpolation'. Details of this procedure have been described in /3,4/. After normalization of the demisyllable segment, a city-block metric can be applied for the calculation of similarity.

Experiments have been carried out with a test corpus of 368 initial and 384 final demisyllables which were automatically segmented in German words spoken by one male speaker /5/. This material contained 45 initial consonant clusters and 48 important final consonant clusters. The average recognition score using the template matching method amounted to 66% for initial and 75% for final consonants. These results can be seen as good as those typically obtained in automatic consonant recognition. Vowel recognition will not be discussed here.

4.2 Classification by feature extraction

A second approach starts from a description of those acoustic events within a demisyllable that are relevant for phonetic decoding. For this purpose the following features or 'cues' were measured: formants, formant transitions, formant-like links for nasals and liquids, turbulences (or bursts), pauses, and voice-bar within pauses or turbulences. These cues are characterized by spectral and temporal measurements. Since the number and order of consonants is restricted in syllable-initial and final position, initial consonant clusters could be completely described by 24 feature components and final consonant clusters by 31 components /5/.

The feature extraction methods are based on the evaluation of energy in several spectral bands and are described in /6/. The context dependencies are taken into account by collating all feature components derived from a demisyllable segment into a common feature vector. For comparison, this method was applied to the same speech material (see Sect. 4.1).

From the recognized consonant clusters the recognition scores of the single consonants were computed. The recognition scores were 4 and 7% lower as compared with the template matching approach /5/. However, it has to be borne in mind that the feature vectors consisted only of 24 or 31 components whereas the templates needed several hundred components for their representation. Thus the feature approach can indeed be seen as a suitable basis for the acoustic-phonetic analysis of demisyllables.

5. RECOGNITION OF SENTENCES

Demisyllable segmentation and recognition has been incorporated in a system which processes spoken sentences as a chain of connected words. This system is completely described in /7/ and will be summarized here only very briefly.

Each word of the vocabulary is represented by a phonetic word model containing the variations in pronunciation as well as possible segmentation errors. The models are constructed in such a way that they can be processed very efficiently by use of Dynamic Programming (DP) methods.

Sentence recognition is based on a 1-stage DP algorithm which determines the best match between a series of word models and the phonetic symbols (consonant clusters and vowels) provided by the classification stage. The word models and the DP transition rules take particular account of the syllabic structure of the utterance.

First experiments with a 75 word vocabulary resulted in recognition scores of 85% correct words in continuous speech without utilizing any grammatical or semantic information. These encouraging results demonstrate the efficient use of syllabic units in all stages of a speech recognition system.

References:

/1/ ZWICKER, E., Peripheral preprocessing in hearing and psychoacoustics as guidelines for speech recognition. Symp. Montreal 1986, Proc. of this conf.

/2/ GEYWITZ, H.-J., Automatische Erkennung fließender Sprache mit silbenorientierten Einheiten. Doct. Thesis, Techn. Universität München, 1984.

/3/ SCHOTOLA, T., On the use of demisyllables in automatic speech recognition. Speech Communication 3, Elsevier Science Publ., 1984, 63-87.

/4/ RUSKE, G., Demisyllables as processing units for automatic speech recognition and lexical access. In: "New Systems and Architectures for Automatic Speech Recognition and Synthesis" (R. DeMori and C.Y. Suen, eds.), Springer-Verlag, 1985, 593-611.

/5/ RUSKE, G., On the usage of demisyllables in automatic speech recognition. In: SIGNAL PROCESSING II: Theories and Applications, (H.W.Schüßler, Ed.), Elsvier Science Publ. (North-Holland), 1983, 419-422.

/6/ RUSKE, G., Automatic recognition of syllabic speech segments using spectral and temporal features. IEEE ICASSP, Paris, 1982, 550-553.

/7/ RUSKE, G. and WEIGEL, W., Automatic recognition of spoken sentences using a demisyllable-based dynamic programming algorithm. 12th ICA, Toronto, July 1986, in print.

HALF-SYLLABIC UNITS FOR SPEECH PROCESSING - AN AUTOMATIC SEGMENTATION

Mamoru NAKATSUI

Radio Research Laboratory, Ministry of Posts and Telecommunications, 4-2-1, Nukuikita-machi Koganei-shi, Tokyo, Japan 184

INTRODUCTION

The half-syllabic units proposed here are units each of which has segment boundaries at steady portions and preserves a transition between two phonetic units. Segment boundaries are basically determined by the minima (valleys) of gross spectral variation measure. The spectral variation measure is defined as the root-mean-square value of the slopes of the weighted regression lines calculated from LPC cepstrum parameters over several frames. The maxima (peaks) of the measure will serve as the reference points for further processing.

In speech synthesis by rule, it is primarily important to select synthetic units that have reasonably small size of inventory to represent spoken utterances and, at the same time, are easily concatenated. In speech analysis-synthesis system at very low-bit-rates such as phonetic vocoding, the units must, further, be automatically segmented and be suitable for interpreting into or matching with the reference units. These requirements on segmentation and matching or labelling are expected to be satisfied for speech recognition system in many cases and for providing useful tools for automatic generation of the inventory of concatenative units.

Syllables and Half-Syllables

One of the selections for the unit to be used in concatenation-based speech processing is the syllable. There have been several discussions and experiments on syllable as recognition unit (1-4). The syllable has been also used as a unit in synthesis by rule of Japanese (5). One of the disadvantages to using syllables as units is that the size of inventory representing spoken utterance is large. This problem can be solved by introducing smaller units such as the half-syllabic units proposed here, since much of the co-articulation among phonetic units is associated with transition regions and since boundaries at the steady portions outside transitions are easily definable.

There exist similar units known as dyads (6), diphones (7), or demisyllables (8) which have the common concept of incorporating the transition between phonemes. The context-dependent diphones have been utilized in constructing a phonetic vocoding system (9). The demisyllables originally proposed for use in a high-quality concatenative speech synthesis (8) have been successfully applied to constructing concatenative templates in the word recognition for large vocabularies (10).

Dynamic Spectral Feature

The gross spectral variation measure derived from a series of LPC cepstrum coefficients has been proposed as a dynamic measure investigating individuality of utterances (11). This dynamic measure has been used in the study on Japanese CV-syllable perception and it has been shown that dynamic spectral feature plays a primary role in phoneme perception (12). Usefulness of the dynamic measure in comparison with its static counterpart has also been shown in word recognition experiment (13). The dynamic measure has also been applied to the segmentation in a very low-rate speech coding where boundaries of the pattern are defined by the maxima of the measure (14).

The half-syllable-like unit has not yet been applied to processing Japanese utterance as far as we know. Our expectation for the units proposed is in the relatively small size of inventory in representing Japanese utterances, since Japanese has relatively simpler syllable organization than that of English. Our ultimate objective is to provide nearly universal units suitable for processing spoken Japanese. As the first step to that goal, our current interest is in confirming whether the proposed units meet the basic requirements, that they would be

1) automatically and reliably segmented,
2) closely related to certain linguistic units, and
3) suitable to acoustic phonetic observations

in the course of constructing the analysis-synthesis system like segment vocoder. This paper reports a preliminary experiment on segmentation of speech signal into the units proposed and some observations of the result with respect to the above requirements.

SEGMENTATION ALGORITHM

Speech sample is bandlimited to 4 kHz and digitized to 12 bits at sampling frequency of 10 kHz. Linear prediction (LP) analysis is carried out on a frame-by-frame basis (100 frames/s). Additional acoustic parameters currently used are a log power P, a zero-crossing count Z, a count for sign change of waveform X, and the first order PARCOR coefficients k1. The spectral variation measure D(j) for j-th frame is calculated by

$$D(j) = [\frac{1}{12}\sum_{i=0}^{11}\{u(i)\cdot a(i,j)\}^2]^{1/2}, \qquad (1)$$

where weight u(i) is currently one for all i and a(i,j) is the i-th coefficient of the weighted regression line of LPC cepstram parameter over several frames. A triangular weighting function is currently applied over seven frames.

With these acoustic parameters, signal processings on input speech are basically carried out in the following steps (descriptions in parentheses are associated with indications in Fig. 1):

1) appointing candidates for segment boundaries at local

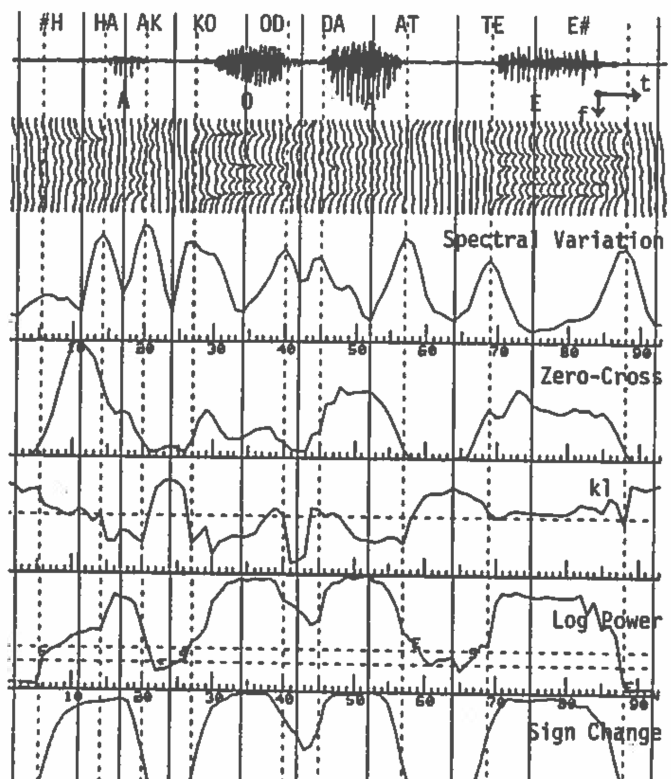


Fig. 1. An example of segmentation and acoustic parameters.

Table 1. Segmentation errors for 455 segments

Position	Initial	Middle	Final	Total
Deletion	7	7	0	14
Insertion	0	11	1	12
Total	7	18	1	26

maxima of spectral variation measure (vertical lines),

2) adjusting the segment boundaries by start and end points of speech interval (S and E),

3) classifying the boundaries into sub-groups of phonetic units and assigning candidates of vowel identity,

4) assigning the reference points at maxima of the variation measure for time arraignment in spectral matching with the reference patterns (dotted vertical line),

5) adopting weights for pattern matching inversely proportional to the normalized values of the spectral variation measure.

Among those steps, 3) to 5) are beyond the scope of this report. However some preliminary trials will be shown later. As for 2), a hysteresis characteristic is given to the decisions of speech interval (from S to E) providing two levels of thresholds for the log power P and the decisions for the non-speech interval associated with intervocalic unvoiced-stops are stabilized by referring the count of sign change X. The minimum (valley) just before S and that just after E were assigned as boundaries of the utterance.

RESULT OF PRELIMINARY EXPERIMENT

Sixty names of Japanese cities spoken by a male adult were used as the test material for segmentation process. It was estimated that the test material consisted of 455 half-syllabic units by our visual inspections.

Segmentation

Fig. 1 shows an example of segmentation where the segment boundaries are denoted by vertical lines and reference points for matching are denoted by the dotted vertical lines. Result of an automatic segmentation of the test material is summarized in Table 1. Correct rate of segmentation is more than 94 %. Most of the deletions of segment boundaries at word-middle are associated with intervocalic [r] and [g] sounds. These problems are going to be solved by the test material having wider spectral bandwidth. It is revealed that problems concerning deletions at word-initial and insertions at word-final are also due to inadequacy of the test material such as low signal-to-noise ratio and over-cuts at the beginning and the end of utterances. So, new test material suitable for our experiment is under preparation, because the current sample has been prepared for other experimental purpose.

Most of insertions of segment boundaries, extra boundaries than expected, are associated with nasal and unvoiced stop consonants. It is observed that extra segments correspond to nasalized vowels and aspirations after stop bursts. The detailed observation for much speech material from the point of view of acoustic phonetics should be made in order to give such solution and interpretation systematically.

Some Observation on Segments

Signal processings described below have not been fully automatized yet and, further, most of the observations have been based on a small set of test material. Alphabets at the top of Fig. 1 are our tentative labelling for the segments (units). Segment boundaries are first classified as either vowels or one of a consonantal group such as voiced-stops and unvoiced-fricatives using a set of acoustic parameters. Spectral distances between spectral frame of the boundary and single frame reference patterns including isolated five vowels and nasal murmurs were used as additional information in the classification.

Alphabets on the segment boundaries just below waveforms in Fig. 1 denote the first candidates of vowel identity showing minimum spectral distance. Ninety percent of vowel boundaries are identified as the first candidates and the remaining ten percent as the second for a sub-set of the test material having 40 vowels. Linear spectral matchings of the CV-type segments with the CV-syllable reference patterns were tried after pre-selections using those data on consonantal group and the first and second vowel candidates described above. In the matching, time arraignment between the segment and the reference pattern was adjusted in such a way that the reference points of both patterns coincide. It is observed that correct CV-syllable appears within the top three candidates for most cases in this arrangement.

CONCLUDING REMARKS

Although our experimental evidence is at quite a primitive stage, the half-syllabic units proposed seem to have potential to meet three basic requirements described above. Among many problems left to be solved, our current interests are in (1) preparation of speech material suitable for our objectives, including city-names at different speeds of utterence and conversational utterances, (2) improvement and tuning of the segmentation algorithm applicable for these speech data, and (3) the detailed observation of the units in acoustic phonetic aspect and systematic organization of classification algorithm.

REFERENCES

[1] O. Fujimura: "Syllable as a Unit of Speech Recognition," IEEE Trans. ASSP-23, 82-87, 1975.

[2] P. Mermelstein: "A Phonetic-Context Controlled Strategy for Segmentation and Phonetic Labeling of Speech," IEEE Trans. ASSP-23, 79-82, 1975.

[3] M. J. Hunt, M. Lennig and P. Mermelstein: " Experiments in Syllable-Based Recognition of Continuous Speech," Proc. ICASSP'80, Denver, 880-883 ,1980.

[4] H. Fujisaki, K. Hirose, T. Inoue and Y. Sato: "Automatic Recognition of Spoken Words from a Large Vocabulary Using Syllable Templates," Proc. ICASSP'82, Sandiego, #26.12 (Vol. 3), 1982.

[5] Y. Tohkura and Y. Sagisaka: "Synthesis by Rule using CV-syllable and Its Speech Quality," Trans. Commit. Speech Res. Acoust. Soc. Japan, #S80-47, Oct. 1980(Japanese Text).

[6] G. E. Peterson, W. S.-Y. Wang and E. Sivertsen: "Segmentation Techniques in Speech Synthesis," J. Acoust. Soc. Amer. 30, 739-942, 1953.

[7] N. R. Dixon and H. D. Maxey: "Terminal Analog Synthesis of Speech Using the Diphone Method of Segment Assembly," IEEE Trans. AU-16, 40-50, 1968.

[8] O. Fujimura, M. J. Macchi, and J. B. Lovins: "Demisyllables and Affixes for Speech Synthesis," Proc. 9th ICA, Madrid, #I-107, p. 513, 1977.

[9] R. Schwarts, J. Klovstad, J. Makhoul and J. Sorensen: A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model," Proc. ICASSP'80, Denver, 32-35, 1980.

[10] A. E. Rosenberg, L. R. Rabiner, J. G. Wilpon and D. Kahn: "Demisyllable-Based Isolated Word Recognition System," IEEE Trans. ASSP-31, 713-726, 1983.

[11] S. Sagayama and F. Itakura: "On Individuality in a Dynamic Measure of Speech," Spring Meeting of Acoust. Soc. Japan, #3-2-7, 589-590, 1979 (Japanese Text).

[12] S. Furui: "On the Role of Spectral Transition in Phoneme Perception and Its Modeling," 12th ICA, Tronto, 1986 (in print).

[13] S. Furui: "Speaker-Independent Word Recognition Using Dynamic Features," IEEE Trans. ASSP-34, No. 1, 1986 (in print).

[14] Y. Shiraki and M. Honda: "Very-Low-Rate Speech Coding Using Time Space Spectrum Patterns," Trans. Commit. Speech Res. Acoust. Soc. Japan, #S84-06, Apr. 1984(Japanese Text).

DEFINITION OF RECOGNITION UNITS THROUGH TWO LEVELS OF PHONEMIC DESCRIPTION

M.Cravero, R.Pieraccini, F.Raineri

CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.a. - Via G. Reiss Romoli 274 - TORINO (Italy)
Tel. + 39 11 21691 - Telex 220539 CSELT

ABSTRACT

In this paper a development system allowing the definition of different recognition unit sets is described. It takes into account acoustic, phonetic and phonologic knowledges. Such a system can be easily used to transcribe large lexicon into recognition units, starting from the ortographic form of the words. In the following a detailed description of the formalism used is given, along with some experimental results obtained by our unit set.

1. INTRODUCTION

A recognition unit set must include a certain number of informations belonging to different knowledge sources. Our recognition system, developed within a speech understanding project partially supported by ESPRIT Project No.26, takes into account the following:

a. Acoustic knowledge, i.e. the knowledge needed to hypothesize, recognize or verify an acoustic event by observing a set of features extracted from a speech segment.
b. Phonetic knowledge, that is the ability of dealing with the acoustic events and their relation to defined phenomenon classes (i.e. phonemes).
c. Phonological knowledge, namely the capability of transcribing each higher level segment (word, sentence) by means of the abstract categorization defined at the phonetic level.

In our system, the acoustic level is implemented by means of Hidden Markov Models (HMM); it means that each unit is described by an HMM in terms of number of states transition and emission probability matrices that are estimated with the Forward-Backward algorithm [1].

The other two knowledges are used to represent whatever Italian word in terms of basic units by means of a rule system that includes main phonetic and phonological variations. That interface between the acoustic knowledge (HMMs of units) and the lexical one is realized by a system based on two levels of description; the first one is the standard phonemic form of words along with additional forms accounting for inter-speaker variations. The second level is a description of each phoneme (the Underlying Phonemic Structure or UPS) by means of smaller units; they are mainly stationary segments and transitions [4]. Besides, a set of contextual rules handles the final transcription of a word in terms of stationary and transitional units.

This development system was designed to define an optimal unit set whose performance was experimentally evaluated within a recognition system. The optimal set proved to be a trade-off between phonemes and diphones; when the transition between two sounds is considered significant for the recognition of the two sounds themselves (i.e. plosive followed by sonorant), the corresponding diphone is included in the set, otherways the transition model is realized appending the two phonemic models.

2. PHONETIC TRANSCRIPTION

A module involved in the task of transcribing a lexicon into the corresponding defined elementary units must first translate an utterance from the ortographic form into the corresponding phonetic one. Italian language[2], as many others, has not an ortography faithful to the phonetics, in the sense that to each grapheme can correspond more than one phoneme, and some phonemes can be indicated by two graphemes (for istance the ortographic sequence "gl"can represent the unique phoneme ʎ of the IPA alphabet, or can be pronounced as the plosive "g" followed by the lateral "l").Besides that ambiguity inherent in the language, other problems arise: people coming from different Italian regions pronounce some words in different ways (i.e. the phoneme "s" of the word "casa" (house) is pronounced as a voiced phoneme by the northern people and as an unvoiced one by southern people).Moreover each speaker has their habits in the pronounciation of some words (for istance a schwa can be added or not to a word ending by consonant). These considerations suggested the idea of implementing a semi-automatic transcription: in the phase of lexicon creation, the operator introduces the new words one at a time; if an ambiguity is pointed out, all the possible trascriptions of the utterance are created and the manual intervention is required in order to decide if all these sequences are representative of the word (different pronounciations) or if some of them must be excluded being wrong.

3. UNDERLYING PHONETIC STRUCTURE

As said before the lower level of phonetic description consists in the so called Underlying Phonetic Structure (UPS); the idea is to transcribe each phoneme into a sequence of elements (Underlying Phonetic Elements or UPE) which show roughly uniform acoustic characteristics. Incidentally the alphabet used to describe UPS is the same as the phonetic one: while at the higher phonetic level each symbol represents a whole phoneme, at the lower UPS level a symbol represents a phoneme portion. To associate a UPS to each phoneme we use a set of rewriting rules as shown in Table 1, where the plus "+" symbol has the meaning of transition from the preceding or to the following phoneme; so the rule a=+a a a+ means that the phoneme a (on the left of the production) can be translated into a left transition (+a), a stationary portion (a) and a right transition (a+). In Table 1 a complete UPS for the Italian phonetic system is reported (the semicolon indicates geminate consonants). Notice that unvoiced plosives are translated as silence "-" plus transition to the following sound while voiced ones as stationary portion (the voicebar "b") plus transition.

ʃ = ʃ	l = +l l	b; = b; b;+
ɛ = +ɛ ɛ	m = m	dz; = dz; dz;+
ɔ = +ɔ ɔ	n = n	ts; = ts; ts;+
- = -	o = +o o	s; = +s; s;
ŋ = n	p = - p+	k; = - k;+
ʎ = ʎ ʎ+	s = +s s	t; = - t;+
ɱ = n	t = - t+	tʃ = tʃ tʃ+
ɲ = ɲ ɲ+	u = +u u	l; = +l; l;
r = +r r r r+	v = +v v v+	m; = m;
a = +a a	w = +u u u+	v; = +v; v; v;+
b = b b+	z = z	tʃ; = tʃ; tʃ;+
d = b d+	ʎ; = ʎ; ʎ;+	ʃ = ʃ
e = +e e	d; = b; d;+	dʒ; = dʒ; dʒ;+
f = f	n; = n;	r; = +r r; r; r+
g = b g+	ɲ = ɲ; ɲ;+	dʒ = dʒ dʒ+
i = +i i	f; = f;	dz = dz dz+
j = +i i i+	g; = b; g;+	ts = ts ts+
k = - k+	p; = - p;+	

Tab.1 - UPS for the Italian phonemes

Each phoneme is represented by means of a single UPS which is constituted by a sequence of UPE.In this way segments of different phonemes showing acoustic similarities can be treated by the same statisical model, as the voicebar of the voiced plosives.

The translation of a word from its phonetic form to its description in terms of recognition units starts with the translation of each phoneme into the corresponding string of UPE. For istance, according to table 1, the italian word APPARTIENE, rewritten by the ortographic to phonetic module in the sequence /ap;artjɛne/, can be translated into:

+a a - p;+ +a a +r r r r+ - t+ +i i i+ +ɛ ɛ n +e e

The second step detects where the transitions are possible; the rule to obtain a transition consists in merging two UPE's containing the symbol "+" in adjacent positions into one transitional unit. So, following the previous example, we obtain:

+a a - p;a a +r r r r+ - ti i iɛ ɛ n +e e
a - p;a a r r - ti i iɛ ɛ n e

It must be noticed that defining the UPS of the generic phoneme /x/ as x = +x x+ it comes out the classical diphone definition, while rewriting each phoneme by itself as x = x, we obtain the phoneme definition.

At this point the description of the word can be handled by a set of rules to take into account the possible effects of a particular phonetic context that cannot be catched by the generic UPS.

4. CONTEXTUAL RULES

Contextual rules can be expressed in the following general form:

U1_U2_....._Un=W1_W2_.....Wm

where Ui and Wj are generic recognition units and the production means that the sequence of units Ui(i=1,2...n) is translated into the sequence Wj(j=1,2...m). In our system rules are applied sequentially, in the given order, to the whole word. Table 2 gives an example of a rule set. From the third to the 18-th production, rules to obtain the stationary portion of /r/ only when it is in a non intervocalic context are described. The UPS of /r/ is made up of two consecutive stationary portions (+r r r r+); in fact, being impossible in the Italian language to utter an /r/ between two consonants, these rules make each vowel cutting away an /r/, so obtaining the desired transcription. The rules dealing with /v/ permit to define left transitions only for those /v/ inserted in a left vocalic context.

The rules 1 through 4 make the two vowels /o/ and /ɔ/ be represented by the same symbol /o/ as well as the two vowels /ɛ/ and /e/; this is done because of the acoustic similarity of the sounds and due to the fact that in Italian the use of the two o's and of the two e's depends on the speaker habits.

Finally the rule 17 transforms each geminate into the corresponding singleton as we demand the distinction between them to higher levels of knowledge.

1: #ɔ=#o	9: r_ru=ru	17: #;=#
2: ɔ#=o#	10: a_r=a_ar	18: a_v=a_av_v
3: #ɛ=#e	11: e_r=e_er	19: e_v=e_ev_v
4: ɛ#=e#	12: o_r=o_or	20: i_v=i_iv_v
5: r_ra=ra	13: i_r=i_ir	21: u_v=u_uv_v
6: r_re=re	14: u_r=u_ur	22: o_v=o_ov_v
7: r_ri=ri	15: r_rj=rj	
8: r_ro=ro	16: r_rw=rw	

Tab.2 - Contextual rules

Extending the rules to the previous example it can be easily obtained:

a - pa ar r - ti i ie e n e

This formalism, developed in order to easily transcribe large lexicons into recognition units given different unit definitions (included "phonemes" and "classical diphones"), was implemented by a program whose output is compatible both with the HMM training procedure and with a set of recognition and word verification systems.

5. PERFORMANCE EVALUATION

Recognition experiments [3] suggested that the best set of units is made up of 123 elements, precisely 22 stationary units and 101 transitional units. Hidden Markov models were trained by means of a 989 words vocabulary obtaining an average recognition rate of about 83% in isolated words belonging to vocabularies of monosyllables differing only for one phoneme (e.g. /aba/, /ata/, /aka/, etc.). Table 3 shows the correct recognition rate per phoneme.

b	87	z	93	dʒ	41
d	76	l	96	tʃ	67
g	90	r	77	j	63
t	70	ʎ	83	w	100
k	96	m	25	e	93
p	96	n	67	i	100
ʃ	96	ɲ	70	o	90
f	83	dz	96	u	100
v	41	ts	100	a	100
s	100				

Tab.3 - Correct recognition rate per phoneme.

6. CONCLUSIONS

A formalism was introduced to write a flexible system that permits the definition of a recognition unit set and the corresponding transcription of words and sentences from their orthographic description to a form that directly relates to the acoustic models of the units themselves. That is obtained using two levels of definition; the first one specifies the phonemes that constitute an utterance, while the second one splits each phoneme into stationary and transitional portions. A suitable set of units that relies on that concept was defined and tested obtaining encouraging results.

7. REFERENCES

[1] Baum, L. E., Petrie, T., Soules, G. and Weiss, N., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains", Ann. Math. Stat., 41, 164-171, 1970

[2] C. Tagliavini, A. Mioni "Cenni di Trascrizione Fonetica dell' Italiano", Patron Ed., Bologna 1983. (In italian)

[3] M. Cravero, R. Pieraccini, F. Raineri "Definition and Evaluation of Phonetic Units for Speech Recognition by Hidden Markov Models", Proc. of International Conference of Acoustic Speech and Signal Processing 1986, April 8-11, Tokyo, Japan.

[4] A.M. Colla, C. Scagliola, D. Sciarra, 'A Connected Speech Recognition System Using a Diphone-based Language Model', Proc. of International Conference of Acoustics Speech and Signal Processing 1985, March 26-29, Tampa, Florida

SOME CONSIDERATIONS ON THE DEFINITION OF SUB-WORD UNITS FOR A TEMPLATE-MATCHING SPEECH RECOGNITION SYSTEM

Anna Maria Colla

Elettronica S.Giorgio - ELSAG S.p.A..
Via Puccini 2, 16154 Genova ITALY

Some considerations on the definition of sub-word units suitable for speech recognition are exposed. An example of a kind of units particularly well-suited to syllable-timed languages is presented, together with some hints for the definition of similar units for different languages. Some experimental results are supplied.

FORMAL DEFINITION OF A SPEECH RECOGNIZER

A Speech Recognition System can be considered as a formal system $\mathcal{S}(\mathcal{U},\mathcal{R}, d)$, where $\mathcal{U}$ is a set of PHONETIC UNITS, $\mathcal{R}$ is a set of RULES for representing each utterance of a given task language $\mathcal{L}$ by means of elements of $\mathcal{U}$, and d is a SIMILARITY OR DISSIMILARITY MEASURE between any "segment" of an utterance and any element of $\mathcal{U}$. More formally we have:

$$d : T \times \mathcal{U} \longrightarrow [0,\infty)$$

$$d \ (t , u) = x$$

where $t \in T$ is a segment of an utterance of approximately equal size as the units, $u \in \mathcal{U}$ is a phonetic unit and x is a non-negative real number. The recognition system can also be considered as an operator acting on a given set S of utterances and yielding for each $s \in S$ an interpretation i(s) in the set $\mathcal{D}$ of all the permitted sentences of the task language $\mathcal{L}$:

$$\mathcal{S}(\mathcal{U},\mathcal{R}, d) : S \longrightarrow \mathcal{D}$$

$$(\mathcal{S}(\mathcal{U},\mathcal{R}, d))\ (s) = i(s) .$$

Actually $\mathcal{R}$ is a function $\mathcal{R}(\mathcal{L},\mathcal{U})$ of the language $\mathcal{L}$ on which the system operates and of the set $\mathcal{U}$ of phonetic units. $\mathcal{U}$ also can be regarded as $\mathcal{U}(\mathcal{L})$.

In a Template-Matching system each unit $u \in \mathcal{U}$ is represented by one or more templates expressed in a convenient form (e.g. as vectors or matrices of appropriate acoustic parameters).

DEFINITION OF THE SET $\mathcal{U}$ OF PHONETIC UNITS

Needless to say, the correct choice of the set $\mathcal{U}$ of the phonetic units (hence of $\mathcal{R}$) is of paramount importance for the efficacy of the whole recognition system.

The elements of $\mathcal{U}$ can be words: in this case the definition of the templates is quite natural and $\mathcal{R}$ simply is the set of grammatical rules apt to represent the permitted word sequences in sentences belonging to the task language $\mathcal{L}$; d can be any distance measure between the vectors of parameters (Mel-Based Cepstrum, LPC, and so on) chosen to acoustically represent input and templates. The calculation of d is made more complex by the need of achieving some time alignment between the input sentence and the templates.

In the Speech Recognition System described in [1] the language representation is an HTN [2]; $\mathcal{U}$ is a set of diphone-like units, which we simply have named "diphones"; $\mathcal{R}$ has also to comprise a set of rules to "translate" each word into a net of diphones and to specify the durations of the related events, and d is an Euclidean distance measure between the LPC-Cepstrum vectors respectively representing each time interval of an utterance and of the diphone templates.

The good results obtained in our diphone-based S.R. system are mostly due to the properties of the adopted set $\mathcal{U}$. Quite naturally, the dictionaries of diphone-like units have been designed taking the characteristics of the Italian language into account. The rhythm of Italian is syllable-timed, that is, syllables are pronounced in approximately the same space of time. Therefore units related to syllabic rhythm are particularly well suited to represent the Italian language.

Basic Hypotheses

The complete set $\mathcal{U}$ of the phonetic units we propose for the Italian language has been derived according to the following hypotheses:

- the transitory parts of speech must be as adequately represented as the stationary ones (whilst generally more emphasis is given to steady-state parts, which are longer);
- the units must be short in order to be fairly insensible to coarticulation (hence economical);
- the units must be related to syllabic rhythm (as Italian is a syllable-timed language);
- the duration of "transitory" units must to some extent be related to articulatory time constants.

The Diphones and Their Properties

According to the above hypotheses the diphones [1,2] are very short units: each stationary sound consists of one spectrum, while each transition is represented by a sequence of very few spectra (5-9). This indeed implies a fair insensibility to coarticulation between adjacent units. Therefore each diphone is in principle represented by one template per speaker (notable exceptions are the sounds affected by their position within a word, that is, vowels and sonorant consonants). The set $\mathcal{R}$ of rules is simply deducible from the phonetic strings corresponding to words, by means of a standardized procedure [3] consisting of 4 steps: orthographic-to-phonetic transcription, generation of the diphone sequences, context study for the choice of multiple templates for sonorants, definition of the duration rules.

The acoustic representation of the diphones is obtained by bootstrapping [3] the template(s) for each unit from a rather small training set by means of a forced recognition. The templates have to be well representative of the lexicon, and moreover should not become inadequate because of intra-speaker variability, which can be serious especially for steady-state sounds. The latter problem can be tackled by a definition of the diphone templates in accordance with a sort of probabilistic approach, where the prototypes are regarded as "average" or "modal" values of a distribution. One "average" template is derived for each steady-state sound and for each different prosodical context of each sonorant. These average templates are used in the same way as normal ones, regardless of the implicit variance.

The diphones have proved to be very effectual for the Italian language. In fact the representation they supply is:

- economical (at most 307 units, with about 350 templates per speaker);
- flexible (that is, apt to deal with pronunciation and duration variability both inter- and intra-speaker);
- automatically deducible for any word from its orthography, including template bootstrapping [3] (this makes the system easily trainable);
- Connected-Speech oriented (straightforward treatment of word coarticulation);
- yielding high scores of correct interpretation (ranging from 82% on the top candidate in a medium-large vocabulary I.W. recognition task, up to 99.5% in a Connected Digit recognition task [1]).

DIPHONE-LIKE UNITS FOR FOREIGN LANGUAGES

It can be hypothesized that, by rules similar to the above ones, adequate sub-word units can be defined also for languages other than Italian, and, in particular, that the units representing similar acoustic events in different languages, being only phoneme-dependent and not context-dependent, can be represented by means of the same templates.

The extension of the recognition system to languages other than Italian can be performed by adapting the different steps in the generation of the diphone representation to the peculiarities of the new language.

In particular the orthographic-to-phonetic transcription must be redesigned for any language, as the phonetic systems of various languages, although partially overlapping, are quite different from one another for a number of reasons: for instance an higher number of phonemes is generally required than for Italian, and above all the orthography is generally much more complex than the Italian one.

On the other hand the rules for the diphone lattice generation need only to be slightly modified, provided that the rhythm of the new language is syllable-based, that is, syllables are entirely pronounced or only their final vowels are not uttered (such as for instance in Spanish, French or German). For languages that are not syllable-timed, that is, languages whose rhythm is governed not by the syllable sequence(s), but by the sequence(s) of strong stresses (such as English or Swedish), the rules for deriving units like the diphones according to our definition are not so straightforward. The use of longer units, spanning the more complex phonetic events pertaining to these languages, is likely to be more appropriate.

EXPERIMENTAL RESULTS AND CONCLUSIONS

The correctness of the above hypothesis has been tested by trying to extend our definition of "diphones" to a language other than Italian and quite dissimilar from it, that is, German. A set of experiments on Connected German Digit recognition has been performed by the same recognition system used for Italian [1]. The test set is made up of 130 1-to-12 digit sentences generated at random, 662 words as a whole. The experimental results are shown in the Table below by the Word and Sentence Recognition Rates (WRR and SRR). Almost as satisfactory results as in the tests on Connected Italian Digits have been obtained both by using entirely new diphone templates (N), and partly re-using "old" diphone templates (O) previously derived for corresponding Italian events (e.g. "AI", "NO", and part of the steady-state sounds). The performance has been improved by submitting the rules of the German diphone lattice generation to some slight refinement, especially about duration of steady-state sounds. By the use of "average" templates for the stationary diphones a further better performance has been achieved (A).

EXPERIMENT	N	O	A
W.R.R.	97.3	96.2	98.2
S.R.R.	85.4	83.1	90.0

Summarizing, satisfactory results have been obtained in a Connected German Digit recognition task by a Template-Matching S.R. System based on diphone-like units as the ones which had proved to be so effectual for Italian [1]. These results show that the extension of such units to languages other than Italian is feasible.

Two are the crucial problems in the definition of diphone-like units for languages other than Italian: 1) the phonetic transcription; 2) the possible need of longer units for stress-timed languages. Moreover a context study is likely to be necessary in order to decide if the same rules for the selection of multiple templates are valid as with Italian.

REFERENCES

[1] A.M. Colla, C. Scagliola & D. Sciarra, "A C.S.R. System using a Diphone-Based Language Model", Proc. ICASSP 1985 (31.9), Tampa (1985)

[2] C. Scagliola, "C.S.R. Without Segmentation: Two Ways of Using Diphones as Basic Speech Units", Speech Communication, 2 (2-3), p. 199 (1983)

[3] A.M. Colla & D. Sciarra, "Automatic Generation of Linguistic, Phonetic and Acoustic Knowledge for a Diphone-Based C.S.R. System", in: R. DeMori & C.Y. Suen (Ed.) "NEW SYSTEMS AND ARCHITECTURES FOR AUTOMATIC SPEECH RECOGNITION AND SYNTHESIS", NATO ASI Series n. F16, Springer-Verlag (1985)

THE ROLE OF STRUCTURAL CONSTRAINTS IN AUDITORY WORD RECOGNITION

H. C. Nusbaum and D. B. Pisoni

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405, USA.

In the past, much of the research on human speech perception has focused on the recognition of acoustic-phonetic properties of isolated CV and CVC syllables. The tacit assumption of this research has been that our understanding of auditory word recognition is contingent upon solving the problems inherent in phoneme perception. By this assumption, auditory word recognition is equivalent to visual word recognition carried out one letter at a time. Indeed, most current theories of auditory word recognition directly reflect this sequential pattern matching approach to word recognition. However, a different perspective is that word perception may be approached as a problem of "weak" constraint satisfaction, in which the structural properties of words in the lexicon interact to specify the identity of an utterance. We will present the results of several analyses of the phonotactic constraints of word patterns that suggest the type of constraints that may be used by human listeners to mediate spoken word recognition.

RECOGNITION IN THE CONTEXT OF THE LEXICON

Context exerts an undeniably strong influence on perceptual processes. However, it is interesting to note that "context" is defined in almost all speech research by whatever stimulus information is presented immediately prior to or subsequent to a target stimulus. Thus, a phoneme is perceived in the context of a syllable, a syllable is perceived in the context of a word, and a word is perceived in the context of a sentence. In all cases, there are objectively definable physical dimensions to the context that is typically investigated. But there is another context that affects word perception as well: the implicit context of the mental lexicon. Beyond the listener's explicit knowledge about words, the structure and organization of the sound patterns of lexical entries may serve as an implicit context within which recognition occurs.

Marslen-Wilson and Welsh (1978) called attention to the potential importance of the structural properties of words with the cohort theory of word recognition. According to this theory, the initial sounds in a stimulus word activate all the words in the lexicon beginning with those sounds. Inappropriate candidates in the cohort are then deactivated when a mismatch occurs in comparing the left-to-right order of subsequent segments in the stimulus with the structures of activated candidates. The word that is ultimately recognized is the candidate that remains after all the other incompatible candidates have been deactivated.

According to cohort theory, the activated cohort of word candidates in the lexicon forms the mental context for spoken word recognition. However, unlike the sentential context that may precede a spoken word, this context has no physical dimensions that can be directly measured or analyzed. In the past, this has posed a problem for investigating the role of the lexicon in word recognition. However, several computer-readable databases of orthographic and phonetic representations of words have recently become available for analyzing the structural properties of words in the lexicon. The database used for all the analyses we will describe contains orthographic, phonetic, and syntactic information for 243,000 words (see Crystal, Hoffman, & House, 1977). Proper names and possessives were excluded from the analyses, leaving about 126,000 words that were examined in the database.

PHONOTACTIC PATTERNS IN THE LEXICON

Although the listener may be presented with spoken words as a temporally distributed sequence of segments, a recognition process need not compare these segments to lexical representations in a strict left-to-right order as claimed by some theories. Indeed, it is unclear how serial pattern matching strategies can recognize a word if the initial segment of the input is obscured, degraded or ambiguous. Since this initial segment is treated as the index into the lexicon, recognition could not proceed without a well-defined access point. An alternative approach is to view auditory word recognition as a constraint satisfaction process, in which the propagation of a number of weak constraints is used to specify the recognized word. When viewed as a constraint satisfaction process, a number of constraints may simultaneously be applied to the lexicon to refine the set of word candidates. Even if one constraint is inappropriate or uninformative, the intersection of the other constraints may still specify the correct word. Given this view, it is important to determine precisely which constraints are actually used during word perception.

The approach that we have taken to investigate structural constraints on human auditory word recognition was motivated by several recent studies that investigated the relative heuristic power of various classification schemes for large vocabulary word recognition by computers. Zue and his colleagues (Huttenlocher & Zue, 1984; Shipman & Zue, 1982) have shown that a partial phonetic specification of every phoneme in a word results in an average candidate set size of about 2 words for a vocabulary of 20,000 words. The partial phonetic specification consisted of six broad phonetic manner classes. Thus, with this approach, a recognition system need not accurately identify the phonemes in spoken words. Instead, only the most robust manner information must be coded. Using a slightly different approach, Crystal et al. (1977) demonstrated that increasing the phonetic refinement of every phoneme in a word from four broad phonetic categories to ten more refined categories produces large improvements in the number of unique words identified in a large corpus of text.

It is important to note that these computational studies examined the consequences of partially classifying every segment in a word. Thus, they actually employed two constraints: the partial classification of each segment and the broad phonotactic shape of each word resulting from the combination of word length with patterned phonetic information.

The analyses that we have carried out used a large lexical database of 126,000 words to study different constraints that might be appropriate for describing human auditory word recognition. This work extends the previous research of Zue and his colleagues to a much larger set of words. In addition, since human listeners are capable of recognizing much more phonetic information than just six manner categories, we have carried out analyses based on the assumption that human listeners will be able to identify some segments completely, while other segments will be unanalyzed.

The results of these analyses are quite revealing about the recognition constraints provided by the structural properties of spoken words. For the coarsest level of segmental analysis, that is knowing only the length of a word in number of phonemes, the search space is reduced from 126,000 words to 6,342 words. Clearly, word length is a very powerful constraint for reducing the candidate set in the lexicon by about two orders of magnitude, even without any detailed segmental phonetic information. Furthermore, the length constraint is strongest for relatively long words. If the length of a word is 21 segments, there are only two candidates out of 126,000 words. Thus, as word length becomes extreme, less detailed segmental information is needed to identify a word.

By simply classifying each segment as either a consonant or vowel (i.e., two categories), without providing any more detailed phonetic description, the reduction in the search space beyond the the length constraint phonotactic constraint is enormous. The number of candidates is reduced by an order of magnitude to 109 words averaged across different word lengths. Furthermore, it is interesting to note that much of this reduction in the candidate set is due to the specific phonotactic constraints provided by the ordering of consonants and vowels. If the segments in a word are classified with just two categories, as consonants or vowels, but the order information is removed, there are 1196 words in the average candidate set. This means that the phonotactic order information in the pattern structure of a spoken word accounts for an order of magnitude reduction in the candidate set size compared to just knowing the number of consonants and vowels, but not their arrangement.

Increasing the amount of phonetic detail for each segment to the six manner classes used by Zue and his colleagues reduces the search space by another two orders of magnitude from the CV classification scheme that maintains phonotactic order information. Using six categories for classifying every segment in each word reduces the average candidate set size to about 5.5 words from 126,000 words in the lexicon. This result agrees very well with the results reported by Shipman and Zue (1982) for a 20,000 word lexicon, indicating that this broad classification scheme is very powerful in reducing the number of word candidates in the search space. Increasing the lexicon by an order of magnitude from 20,000 words to 126,000 words only results in a tripling of the number of candidates from 2 to about 6 words. By any metric, partial information about every segment is an extremely effective constraint on the candidate set.

However, human listeners are capable of resolving much more phonetic detail than just six broad categories. One issue that can be raised then, concerns the constraint provided by complete phonetic information about some of the segments in a word compared to partial information about every segment in a word. Classifying every segment in a word provides two types of information: (1) partial phonetic information about every segment, and (2) the phonotactic "shape" of the entire word. By comparison, complete classification of some of the segments provides: (1) detailed phonetic information about a few segments, and (2) partial information about the phonotactic shape of a word. Based on the previous demonstration of the power of phonotactic shape with just two categories (i.e., consonant or vowel), it seems reasonable to predict that partial classification of every segment in a word should be more effective than complete classification of some of the segments in a word.

To test this prediction the following analyses were carried out: (1) the phonetic information in first half of every word was classified completely leaving the remaining segments unclassified, (2) the phonetic information in the last half of each word was classified completely leaving the first half unclassified, (3) only the consonants were phonetically classified leaving the vowels unlabeled, and (4) the vowels were phonetically classified leaving the consonants unlabeled. The results demonstrate that complete information about some of the segments in a word provides a more powerful constraint on the candidate set than partial classification of every segment. Classifying the beginning of words completely reduces the search space from 126,000 words to 1.7 words and classifying the last half of words reduces the candidate set to 1.9 words. By comparison, classifying only the consonants exactly and leaving the vowels unclassified yields a set size of 1.4 words, while classifying the vowels only yields a set size of 3.2 words. In each analyses, complete phonetic information about some of the segments in a word constrains the search space much more than partial classification of every segment. These results demonstrate that detailed phonetic information about some of the segments in a word provides enough constraint, in general, that other segments can be completely obscured or ambiguous without significantly impairing recognition. Moreover, to the extent that some phonetic information is available about other segments, the candidate set will be reduced further, probably to the extent of uniquely specifying the correct word.

CONCLUSIONS

The view of word recognition that emerges from these analyses differs substantially from serial pattern matching approaches. As more of a stimulus word is heard, the listener progressively narrows the candidate set based on the development of a phonotactic specification for the input. Over time, acoustic information in the stimulus is successively refined into more detailed phonetic representations. In some cases, only a broad phonetic description of segments may be computable and the phonotactic structure is used to further narrow the candidate set. This approach, called Phonetic Refinement Theory, is currently being implemented as a model of the recognition process. Although further research is needed, it is clear that computational analyses of the sound patterns of words can provide new information about the processes that mediate speech perception.

REFERENCES

Crystal, T. H. , Hoffman, M. K., & House, A. S. (1977) Statistics of phonetic category representation of speech for application to word recognition. Princeton, NJ: Institute for Defense Analysis.

Huttenlocher, D. P. , & Zue, V. W. (1984) A model of lexical access based on partial phonetic information. Proceedings of ICASSP-84, New York: IEEE Press, Volume 2.

Marslen-Wilson, W. D., & Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 10, 29-63.

Shipman, D. W., & Zue, V. W. (1982) Properties of large lexicons: Implications for advanced isolated word recognition systems. Proceedings of ICASSP-82, New York: IEEE Press.

SYLLABLE STRUCTURE OF ENGLISH WORDS: IMPLICATIONS FOR LEXICAL ACCESS

Michiko Kosaka and Hisashi Wakita

Speech Technology Laboratory, 3888 State St., Santa Barbara, CA 93105.

We parsed a large corpus of English words into syllables and into their constituents to determine the difference between high and low frequency words with respect to these structural properties. There are obvious applications of the results to the lexical access problem in large-vocabulary isolated-word speech recognition systems.

INTRODUCTION

One of the problems in the theories of word recognition involves the relationship between the frequency of usage of words and the structural properties of them. This question is interesting because (1) the differences in word frequency effects might be due to factors other than the frequency of usage, and (2) we might be able to clarify the nature of lexical access, i.e. whether words are accessed on the basis of an acoustic, phonetic or phonological representation. This question is also interesting for isolated-word large-vocabulary machine recognition systems because (3) certain structural constraints in lexical access have been shown to be very powerful in reducing the search space for candidate words. The precise form of the lexical entries is very important for continuous speech recognition systems.

MATERIAL AND METHODS

Brown Corpus words were used as the data. Following Pisoni, et.al., we defined high frequency words as those equal to or greater than 1000 words per 1 million (e.g. *the, of, many*), and low frequency words as those between 10 and 30 words per 1 million inclusively (e.g. *acceleration, bronchial, conjugate*). In addition, we defined mid frequency words to be 30 to 1000 words per 1 million exclusively (e.g. *able, measurement, strike*). These words were matched against the phonetic transcriptions of the SCRL dictionary, which resulted in a data base of a total of 7443 words. There were 91 high frequency words, 3072 mid frequency words and 4280 low frequency words.

Brown Corpus words might not be an ideal sample because the material is approximately 20 years old and because it is based on printed texts as opposed to a transcription of the spoken language. Nevertheless, because of a lack of other computer-readable data bases, we took the the Brown Corpus words to be our sample. It might be argued that word information from the spoken language is not an appropriate alternative, since we do not expect people to speak to the machines in the same way that they would speak to other people.

The phonetic transcriptions (ARPAbet) of these words were parsed by a syllable parser developed at STL. The syllabication of the parser is based on the maximum onset principle. Stress resyllabication was not included in this parser, since stress information was not available in a convenient form. Therefore, the onset count should be slightly over-represented for syllable-initial consonant clusters and slightly under-represented for syllable-final coda consonant count. The quantitative effect of this ommission is not clear, but we do not expect it to be significant.

This study focuses on the frequency of usage vs. syllable length and sub-syllabic constituents. A motivation for this is that previous studies on the phonological structural properties of words dealt exclusively with the identity of phonemes and their length in terms of phonemes [1, 2, 5, 6, 7].

WORD FREQUENCY AND LENGTH

Table 1 below shows the relationship between the word length (in syllables) and the frequency ranges of high, mid and low. Table 2 shows the relative frequency of occurrence within each frequency class. The results indicate that the high frequency words are different from mid and low frequency words and that they are from two separate populations. The Pearson correlation of mid and low frequency was 0.9. Thus the mid and low frequency words can be considered to be from the same population. That the two populations are independent can be seen from the proportion of one-syllable words. They are 0.88 0.35 and 0.23 for high, mid and low frequency words, respectively. The mean length for each group was 1.12, 2.01 and 2.33 for high, mid and low frequency words, respectively. One syllable is the median of high frequency words; whereas the median of mid and low frequency words are two syllables.

Table 1: Word Frequency and Length (Syllable)

length	high	mid	low	total
1	80	1073	978	2131
2	11	1199	1681	2891
3	0	541	1033	1576
4	0	211	429	640
5+	0	48	159	207
total	91	3072	4280	7443

Table 2: Word Frequency and Length (%)

length	high	mid	low	total
1	87.91	34.93	22.85	28.63
2	12.09	39.03	39.28	38.84
3	-	17.61	24.14	21.17
4	-	6.87	10.02	8.60
5+	-	1.56	3.71	2.78

WORD FREQUENCY AND SYLLABLE CONSTITUENTS

Difficulties in intelligibility of certain words have often been, in part, attributed to the lexical distance based on the frequency [1] and to the particular phonemes, or phoneme/grapheme ratios [2]. We investigated two factors that might account for such difficulties.

Word Frequency and Onset

The onsets were classified as nil (no consonant at the beginning of a syllale), cluster (two or more consonants at the beginning of a syllale) or simple (exactly one consonant at the beginning of a syllable). These three classes cover all the possible onsets. We hypothesized that high frequency words are simpler in the sense that it is low in consonant clusters and that simple and null onsets prevail. Table 3 summarizes the ratio of these occurrences.

These results show that the characteristics of high frequency words vs. mid or low frequency words is not in the composition of simple onsets. Simple onsets are by far the greatest proportion of all words in all frequencies. High frequency words are characterized by a relatively large proportion of null onsets and a very low proportion of consonant clusters with respect to low frequency words.

The results might be interpreted as the following. Null and simple onsets are simpler in that they are perceived and produced much more easily than the clusters. Clusters are complex components. They are more difficult to perceive and to produce. Another interpretation is to say that high frequency words are much more constrained phonotactically. In other words, fewer grammar rules are necessary to process high frequency words.

Table 3 also shows that within a population, the cluster onset decreases as the length increases, and in general, the nil onset increases (with the exception of mid frequency words). An instance of simplification seems to occur as the complexity, in terms of length, increases.

Table 3: Word Frequency and Onset: Composition Ratio within Frequency Class and Length (%)					
length	type	high	mid	low	total
1	nil	23.75	4.85	3.68	5.02
	cluster	1.25	22.09	30.16	25.01
	simple	75.00	73.07	66.16	69.97
2	nil	40.91	12.43	9.67	10.93
	cluster	0	13.22	16.21	14.91
	simple	59.09	74.35	74.12	74.16
3	nil	-	15.53	13.62	14.27
	cluster	-	10.41	13.39	12.37
	simple	-	74.06	72.99	73.36
4	nil	-	13.39	13.73	13.62
	cluster	-	9.00	11.80	10.88
	simple	-	77.61	74.47	75.51
5+	nil	-	13.11	14.74	14.37
	cluster	-	6.15	7.74	7.37
	simple	-	80.74	77.52	78.26
all	nil	27.45	12.08	11.42	11.77
	cluster	0.98	13.17	15.25	14.37
	simple	71.57	74.75	73.33	73.86

Table 4: Word Frequency and Coda: Composition Ratio within Frequency Class and Length (%)					
length	type	high	mid	low	total
1	nil	28.75	5.96	5.62	6.66
	cluster	6.25	10.16	49.80	28.20
	simple	65.00	83.88	44.58	65.13
2	nil	72.73	46.91	43.71	45.15
	cluster	0.00	4.34	12.73	9.20
	simple	27.27	48.75	43.56	45.67
3	nil	-	55.08	54.34	54.60
	cluster	-	4.68	8.23	7.01
	simple	-	40.23	37.43	38.39
4	nil	-	68.96	68.41	68.55
	cluster	-	2.25	3.79	3.24
	simple	-	28.79	27.80	28.05
5+	nil	-	76.23	76.90	76.75
	cluster	-	0.00	1.97	1.51
	simple	-	23.77	21.13	21.74
all	nil	38.24	46.12	50.24	48.53
	cluster	4.90	4.98	12.55	9.59
	simple	56.86	48.90	37.21	41.87

Word Frequency and Coda

The codas (syllable-final consonants) were classified in the same way as above into three classes: nil, cluster and simple. Our hypothesis was similar to the one for the onsets: that the high frequency words over represent nil and simple codas. Table 4 shows the relative distribution by frequency classes. The results indicate that while the hypothesis is true, the pattern of distribution is very different from the onset. The proportion of the clusters among the low frequency words ranges from 50% to 2%, while the comparable statistics for the onsets ranged from 30% to 8%. At the same time, the nil coda ranged from 6% to 77% for the same population, while the onsets ranged from 4% to 15%. Another striking fact is that the simple codas decrease in proportion to length in all frequency classes, in addition to the fact that their proportion for one-syllable length is lower than those for the onsets (except for the mid frequency words). The data on one-syllable length is important because there is no chance for stress resyllabication.

We demonstrated that there are structural differences among words of different frequencies along three dimensions: onset types, coda types, and syllable lengths. We have been able to show that there is a correlation between these properties and word frequencies.

LEXICAL INFORMATION AND LEXICAL ACCESS

There are several ways in which such lexical information can contribute to the lexical access problem in a speech recognition system. For example, syllable length of a word is potentially a very powerful device especially when a word is long. The length constraint was proposed and demonstrated to be effective [1, 3]. However, these proposals centered around phoneme length. The advantage of syllable over phoneme length is that the phoneme insertion and deletion errors can be avoided altogether. The disadvantage is that the cohort size is much larger.

Another possible constraint that can be used is the information on the type of onset. We have been able to identify 68 unique onsets over all the syllables of the complete set of sample words. We saw that the majority of English words favors the CV type of syllables. One might, for example, assign a probability associated with the types of onset prior to identifying the onset itself. It remains to be seen how powerful this constraint might be when this information is used even partially, e.g. at the beginnning of a word.

CONCLUSION

What is the relationship between word frequency and the phonological structure? We examined some of the phonological properties of English words which were not discussed before. We proposed a metric of simplicity to account in part for the structural differences between high and low frequency words. We also suggested that syllabic structural information might be used to organize the lexicon into equivalence classes in a speech recognition system.

REFERENCES

[1] Pisoni, D.B., Nusbaum, H.C., Luce, P.A., and Slowiaczek, L.M. Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Communication*, 1985, 4, 75-95.

[2] Landauer, T.K., Streeter, L.A. Structural Differences Between Common and Rare Words: Failure of Equivalence Assumptions for Theories of Word Recognition. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 119-131.

[3] Huttenlocher, D.P. and Zue, V.W. A Model of Lexical Access Based on Partial Phonetic Information. *Proceedings of ICASSP-84*, 1984, 2.

[4] Kucera, F. and Francis, W. *Computational Analysis of Present Day American English*. Brown University Press, 1967.

[5] Makino, S., Wakita, H. and Applebaum, T.H. Lexical Analysis for Word Recognition Based on Phoneme-Pair Differences. Talk delivered at ASA meeting, Minneapolis. October 1980.

[6] Denes, P.B. On the Statistics of Spoken English. *J. Acoust. Soc. America*, 1963, 35, 6 892-904.

[7] Greenberg, J.H. and Jenkins, J.J. Studies in the Psychological Correlates of the Sound System of American English. *Word*, 1964, 20, 157-177.

ON ACOUSTIC VERSUS ABSTRACT UNITS OF REPRESENTATION

Daniel Huttenlocher and Meg Withgott

Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, Massachusetts 02139 USA

Stanford University, Center for the Study of Language and Information, Stanford, California 94305 USA

Abstract: Postulating the existence of abstract representational units appears useful in speech research. For instance, such units can be used to partition a large lexicon for word-candidate hypothesization [8] [4], or to specify phonetic deletion and modification sites. However, since such linguistic representations have at best an indirect realization in the physical signal, it has proven difficult to build classifiers for these units. Therefore, recognition systems generally use less abstract units such as spectral templates. We argue that the difficulty of classifying abstract units does not preclude using these units in recognition. In particular, constraint-based systems provide a mechanism for exploiting abstract linguistic knowledge at the acoustic level.

Introduction

Work on lexical and phonological representation assumes the existence of abstract units such as phonemes or allophones. Powerful general principles have been identified operating under this assumption. However, attempts at developing recognizers which use similar units have met with difficulty (cf. [6]). Thus, systems for classifying the acoustic signal generally use representations which are far less abstract (e.g., templates, vector quantized spectra, etc.).

We consider some of the reasons that it is difficult to recognize abstract units such as phonemes from the speech signal. Then we turn to the limitations of current recognition systems. Finally we suggest how some of these limitations may be overcome by formulating lexical and phonological knowledge as constraints on acoustic data.

These constraint-based models can be used to specify that certain acoustic patterns are consistent with a given word. They may also specify that certain acoustic information is inconsistent with the presence of a given word. The critical idea is that of viewing recognition as consistency checking. This idea contrasts strongly with the use of abstract units in transformational systems.

Recognizing Abstract Units is Hard

The difficulty of recognizing abstract units such as phones or diphones from the speech signal is attributable to several factors. First is the problem of segmenting the speech signal into phonetic-sized units. Certain regions of an utterance do not clearly correspond to any particular phoneme or other abstract unit. Furthermore, segmentation errors cause the insertion and deletion of phonetic units.

Second is the difficulty of classifying the segments that have been identified. Variation across talkers causes a given abstract unit to have different realizations for different talkers. These may even overlap, as in the case of /s/ and /š/. Phonetic sized units can also be difficult to classify because they are distorted due to contextual effects (e.g., the /t/ in a /tr/ cluster). Third, certain regions of an utterance are often difficult to classify, such as unstressed syllables.

Thus, a given classifier will perform very well only in certain regions of an utterance, or for certain talkers. This suggests letting the classifier do "only as much as can be done reliably." However, this means that no single abstract level of representation is sufficient.

On top of all this, having identified a sequence of abstract units it is still difficult to do word recognition. Part of the problem is the phonological variation in the production of individual words. Deletion, epenthesis, and other phonological modifications can cause extreme departures from the canonical form.

The problem of mapping from a sound sequence to words is even harder in the case of continuous speech because the limit of the match is not generally known. For instance, it is well known that in fluent speech the phrase "did you go to the.." (/dɪd#yuʷ#goʷ#tʰuʷ#ðə/) can be produced as [dɪǰəgəɾəðə].

Considerable attention has been paid to the problem of recognizing words from phonetic sequences. The most common approach is to formulate transformational rules which characterize phonological variation. Such rules map lexical baseforms to surface phonetic strings. This mapping is then either used to expand each lexical entry into all possible surface forms, or to transform an input sequence into its possible underlying forms [7]. However, this assumes that all pronunciations can be anticipated and captured by the rules. Furthermore, since these rules are based on phonetic transcriptions, it is assumed that the output of the classifier is adequately detailed and relatively error free. These assumptions have not been borne out in actual speech.

Current Recognition Systems are Limited

Using acoustic representations for recognition seemingly bypasses the problems of classification and retrieving the underlying phonemic form. However, such systems only work for restricted tasks. While the IBM recognizer [1] is perhaps the most successful system to date, it appears to be reaching the limit of the approach.

The IBM recognizer searches the entire lexicon in recognizing each word. The most obvious consequence of this is the large amount of computation required. A more serious problem is that the distance between an unknown word and each lexical entry does not provide very strong discrimination among the possibilities. This is partly due to the fact that distance metrics are sensitive to acoustic differences, whereas phonological processes can cause large acoustic differences between pronunciations of the same word. These differences can be as large as those between different words, as when "balloon" is pronounced "b'loon", which is acoustically similar to "bloom".

As a result, the IBM system relies heavily on word tri-gram probabilities for its performance. These probabilities are obtained by observing word triples in a large training corpus. However, the use of tri-gram models makes it difficult to add new words because their probabilites must be estimated. Furthermore, tri-grams are not good models of novel sentences even from the same vocabulary. For a 1.8 million word corpus of text, the tri-grams found in one 1.5 million word subset covered only 77% of the tri-grams observed in the remaining 300,000 words [5].

Thus while tri-grams provide substantial constraint, they are too specific in that they don't capture general properties of English. However, a more general characterization of allowable word sequences is unlikely to provide nearly as much constraint. For example, attempts at using syntactic constraints in speech recognition have required using artificially simple grammars to appreciably limit the possible word candidates [6]. Therefore, some other source of constraint will be needed in order to develop the next generation of recognition systems.

A Look at Using Abstract Units in Recognition

There are three potential advantages of using abstract representational units in recognition. First, exploiting phonological information as a source of constraint in recognition requires using an abstract representation. Second, training a system (or adapting it to new speakers) can be greatly simplified by the use of abstract units. Third, abstract representations enable the use of non-exhaustive matching techniques in lexical access.

With respect to the problem of training, abstract sound units can be used to bootstrap the training process by representing each word in terms of component parts. Training then operates over this smaller set of units rather than over words. In a very large vocabulary system, such a bootstrapping process appears necessary. For example, the IBM system uses phonetic-sized units for training.

With respect to the problem of matching and lexical access, there are two ways in which abstract units can be used. The first is to search only part of the lexicon, rather than matching against every entry and picking the best match. The second is to match against only some of the information in each lexical entry being considered, depending for example on the certainty of the classifier.

While the use of abstract units can theoretically address such issues, the fact of the matter is that systems have been relatively unsuccessful at using abstract units. We claim that this can be traced to the framework within which abstract properties have been formulated, rather than to the use of abstract units per se.

For instance, if phonological rules captured the variability in speech, then lexical access could simply be done by table lookup. Yet as we noted above, there is substantial variability which cannot be accounted for by rules, and this causes classification errors. Thus, the transformational formulation does not get around the problem of exhaustive search of the lexicon.

Another approach which uses abstract units is to characterize what is stable or reliable about a given lexical entry, rather than trying to capture variability. This approach has been taken by Shipman, Zue and Huttenlocher in their work on partitioning the lexicon into equivalence classes of words sharing the same features. For example, manner of articulation features can be used to partition a 20,000 word lexicon into classes of only about 30 words on average.

Using this approach, ideally only that subset of the lexicon corresponding to a given feature sequence must be searched in lexical access. However, this assumes that each word has a small number of partial representations as output by the classifier. While the proposed partial representations are less sensitive to variability than phonetic representations, this still may not be a reasonable assumption.

Conclusion

In the previous section we have seen that systems which use abstract phonetic units have been developed based on the assumption that these units have reliable acoustic correlates. One example of this was transformational systems which view recognition as mapping between sequences of abstract units. In order to apply these transformations, the abstract units must first be reliably classifiable from the acoustic signal. Abstract units often do not have reliable acoustic manifestations, however. The absence of these correlates has led to the development of acoustically-based systems which do not use linguistic constraints at all.

While abstract units do not have reliable acoustic correlates, a given abstract unit is only consistent with certain acoustic patterns. Since constraint-based models can be used to specify what acoustic information is consistent with a given abstract unit, they are a convenient formalism for expressing such knowledge. In particular these models provide a means for expressing partial and redundant information [9] [2] [3]. This ability to exploit multiple levels of specificity means the classifier can be allowed to do as much as it can, while still using a lexical partitioning based on abstract representational properties.

Acknowledgments

This work was supported in part by the Defense Advanced Research Projects Agency under Office of Naval Research Contracts N0014-82-K-0727 and N0014-80-C-0505 to M.I.T., in part by Schlumberger Computer Aided Systems, and in part through an award from the System Development Foundation to the Center for the Study of Language and Information at Stanford.

References

1. Bahl, L.R., A.G. Cole, F. Jelinek, R.L. Mercer, A. Nadas, D. Nahamo, and M.A. Picheny "Recognition of Isolated Word Sentences from a 5000-Word Vocabulary Office Correspondence Task", *Proc. IEEE ICASSP*, 1983.

2. Fenstad, J. E., P-Kr. Halvorsen, T. Langholm and J. van Benthem. (To appear) *Equations, Schemata and Situations: A Framework for Linguistic Semantics* Dordrecht: Reidel. Also in CSLI-TR-29, Stanford Univ., 1985.

3. Grimson, W.E.L. and T. Lozano-Perez "Recognition and Localization of Overlapping Parts from Sparse Data", MIT Artificial Intelligence Laboratory Memo No. 841, 1985.

4. Huttenlocher, D.P. "Exploiting Sequential Phonetic Constraints in Recognizing Words", MIT Artificial Intelligence Laboratory Memo No. 867, 1985.

5. Jelinek, F. "Problems of Language Modeling for Speech Recognition", in preparation, IBM T.J. Watson Res. Ctr.

6. Klatt, D. Review of the ARPA Speech Understanding Project, *J. Acoust. Soc. Am.*, Vol. 62, No. 6, December 1977.

7. Oshika, B. T., V. W. Zue, R. V. Weeks, H. Neu, and J. Aurbach. The Role of Phonological Rules in Speech Understanding Research, *IEEE Trans. on Acoustics, Speech, and Signal Processing.* Vol. ASSP-23, 1, February 1975.

8. Shipman, D. and V.W. Zue "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems" *Proc. IEEE ICASSP*, 1982.

9. Sussman, G.J. and G.L. Steele "CONSTRAINTS: A Language for Expressing Almost Hierarchical Descriptions" *Artificial Intelligence*, Vol. 14, No. 1.

MODELS OF PHONETIC RECOGNITION I: ISSUES THAT ARISE IN ATTEMPTING TO SPECIFY A FEATURE-BASED STRATEGY FOR SPEECH RECOGNITION

Dennis H. Klatt

Room 36-523, Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract. This is the first of a set of papers from the MIT Speech Communication Group expressing conflicting viewpoints as to the nature of the speech perception process and the best way to approach the problem of speech recognition by machine. In this paper, it is argued that all models employing phonetic feature detectors (whose purpose is to make phonetic decisions so as to reduce the information content of the input representation prior to lexical search) are suboptimal in a performance sense. Such models are usually incompletely specified, and they do not confront certain theoretical problems that are discussed here. It is suggested that the LAFS model of precompiled acoustic expectations for familiar words (Klatt, 1979) has theoretically superior characteristics. However, aspects of the Stevens model described in the next paper (in particular, relational invariance at the acoustic feature detector level) are an attractive candidate for the front-end processor of a next-generation LAFS strategy.

What does it mean when someone says "I believe that phonetic features play an essential role in speech perception?" Can this philosophical position be translated into a practical strategy for speech recognition? The purpose of the present paper is to specify what must be present if a theory claims to be an instance of a phonetic feature based perceptual strategy. Along the way, we will point out some of the problems facing anyone wishing to build a speech recognition device having these characteristics. The paper is, in part, a challenge to those who embrace the phonetic feature basis of perception.

A literal translation (by me) of the phonetic feature concepts implicit in Jakobson, Fant and Halle (1963) or Chomsky and Halle (1968) to the domain of perception results in the procedure outlined in the block diagram of Figure 1. Similar models have been discussed by Studdert-Kennedy (1974) and Pisoni and Luce (1986).

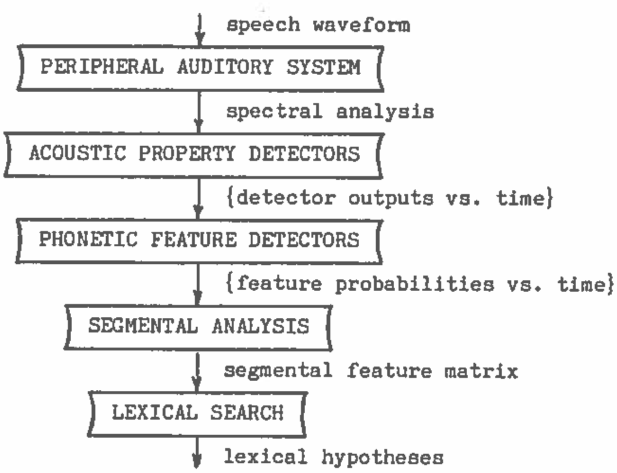


Figure 1. Block diagram of a "literal" phonetic feature detector model of speech perception.

Peripheral Processing. I assume that the peripheral processing stage provides at least two representations of input speech waveforms: (1) an average-firing-rate representation of the short-time spectrum (Goldhor, 1986), and (2) some sort of synchrony spectrum (Sachs et al., 1982; Allen, 1985). Details are not important to the issues at hand, although there is some hope that a properly designed simulation of peripheral processing, including critical bands, masking, adaptation, synchrony to formant frequencies, etc., will make the task of later modules easier by enhancing invariant acoustic characteristics of phonetic features and suppressing irrelevant variability.

Acoustic Property Detectors. A set of acoustic property detectors transform this spectral input representation into time functions that characterize the degree to which certain properties are present in the input at a given instant of time. These property detectors are assumed to differ from the raw input spectra in that they compute relational attributes of the signal which tend to be more invariant and "quantal" (Stevens, 1972) across phonetic contexts and across speakers than are the raw spectra. The acoustic property detectors are further assumed to differ from phonetic feature detectors (the next stage) in that they compute relatively simple general auditory properties which are useful for processing other signals as well as speech. Examples of possible auditory features are onset detectors, spectral change detectors, spectral peak detectors, formant frequency detectors, formant motion detectors, presence-of-voicing detectors, fundamental frequency detectors, nasal-formant detectors, etc.

Phonetic Feature Detectors. A phonetic feature detector has the task of examining an input set of auditory property values over a chunk of time, and making linguistic decisions that are language-specific. Of course aspects of the speech production/perception process constrain these decisions to be similar across languages (Stevens, 1972). A phonetic feature detector may make a relatively simple decision based on input from a single acoustic property detector, or, more typically, a feature detector combines information from several different auditory property detectors.

The decision of a phonetic feature detector is, in principle, binary -- reflecting the presence or absence of the feature at that instant of time. However, in a speech recognition context, it may be better to think of the detector output as expressing the probability of the presence of a particular feature at that time, given the acoustic evidence to date. In this way, one can represent real ambiguity and possibly recover later from inevitable errors. The output probability values may spend most of the time around zero and one, as a linguist would expect when the acoustic data are clear, but this is certainly not possible in the presence of background noise and other factors that influence articulatory performance. Experience with speech understanding systems has shown the undesirability of forcing an early decision when, in fact, representations incorporating uncertainty often permit correct resolution in later decision stages (Klatt, 1977). Even if phonetic feature outputs are probabilities, there is still a considerable reduction of information taking place at this stage; only about 20 or so feature "time functions" are available to represent phonetic events.

Segmental Analysis. Up to this point, the object of the computations has been to describe via phonetic features what is actually present in the acoustic signal, or equivalently, what articulatory gestures were used to generate the observed acoustic data. The segmental analysis stage must temporally "align the columns" of the set of parallel feature detector outputs so as to produce what can be interpreted as a sequence of discrete segments (the presumed form of the lexical entries). In the spirit of creating as much parsimony with current linguistic formalism as possible, I have assumed that the segmental representation is basically a feature matrix (Chomsky and Halle, 1968), but it can become a lattice of alternative matrices where necessary to describe segmentation ambiguity. One might also argue for additional levels of phonological representation to delimit syllables, onsets and rhymes, etc. (Halle and Vergnaud, 1980), or to group features into tiers that need not be temporally perfectly aligned (Clements, 1985; Stevens, these proceedings).

Entries in the matrix are, again, probabilities, but this time they indicate the likely presence/absence of more abstract "phonological" features -- reflecting the speaker's underlying intentions (to the extent that it is possible to infer such intentions from the acoustic data). For example, given evidence for a nasalized vowel followed by a [t], but with little or no evidence for a nasal murmur before or after the vowel, this stage of the analysis would postulate a nasal segment between the vowel and the [t], assign the nasality to it, and deduce the probable phonetic quality of the preceding vowel if it had not been nasalized.

Lexical Access. The lexical access module accepts as input the segment matrix (and perhaps prosodic information and syntactic/semantic expectations) in order to seek candidate lexical items. The mechanics of the matching process requires the development of sophisticated scoring strategies to penalize mismatches and deal with missing and extra segments. In general, word boundary locations are not known for certain, so that lexical probes may be required at many different potential starting points in an unknown sentence.

EXAMPLE

A schematic spectrogram of the utterance [ada] is shown in Figure 2. The spectrogram illustrates several cues that interact to indicate whether the plosive is voiced or voiceless.

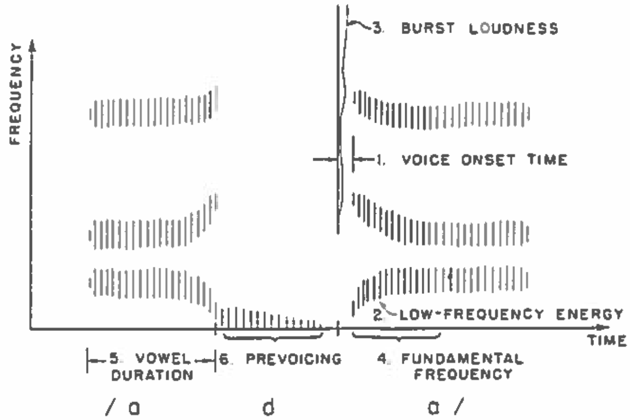


Figure 2. Six acoustic cues to voicing for plosives.

While six cues are identified in the figure (and Lisker, 1978, has catalogued 16 potential cues), it is by no means clear that the cues correspond to the outputs of six quasi-independent acoustic feature detectors. Proper analysis of this and other phonetic situations may reveal the existence of integrated detectors that combine at an auditory level some of the cues to voicing listed in the figure. Even so, the task of the voicing feature detector is a complex one, due to the difficulties enumerated below:

(1) When to Activate a Detector? Acoustic property detectors produce output time functions to indicate e.g. the location in time of an onset or the location in frequency of an energy concentration. However, these detectors do not make any decisions -- it is up to the phonetic feature detector to find the onset corresponding to the burst of a plosive, and the onset corresponding to voicing onset time so as to measure VOT. While these events are usually clear to the eye when inspecting a spectrogram, the viewer employs a great deal of speech-specific knowledge to reject visual onsets that don't look globally like plosive-vowel sequences. Programming a computer to behave reliably in this way has proven to be extremely difficult (see e.g. Delgutte, 1986). How much general speech knowledge must be employed by the voicing feature detector when trying to decide whether it is confronted by a plosive release?

(2) Feature Independence. If one task is to measure voice onset time by determining burst onset followed by voicing onset, the detector should probably be willing to accept a weaker burst as an onset if the plosive were labial than if it were not. Similarly, the VOT boundary between voiced and voiceless is probably somewhat shorter for labials. Is the voicing feature detector (a) permitted to know the place decision, (b) permitted to compute information required for an optimum voicing decision, or (c) forced to make an independent judgement of degree of voicing which will be corrected by the next level that has available all feature outputs?

(3) Time Functions vs. Event Sequences. The voicing decision involves multiple cues that occur at different times. The temporal location of release relative to closure can vary, making it hard to use fixed measurement points in combining information over time. Are each of the cues to voicing best thought of as time functions, as assumed thus far, or as events that occur in sequence and must be interpreted by a second decision level (what is the representation of knowledge and decision flow in a feature detector)?

(4) Cue Combination Rules. Ultimately, the voicing feature must combine all the available evidence into a single voicing decision (probability) that is the best decision possible at that given instant of time. Is the decision framework basically articulatory and Bayesian (compute the conditional probability of obtaining the observed data assuming the canonical articulatory pattern for a voiced plosive, and compare this with the conditional probability of obtaining the observed data assuming the articulatory pattern for a voiceless plosive)? How can the extremely rich set of alternative patterns of acoustic cues signalling voicing be programmed/learned in any practical model?

(5) Intended vs. Actual Articulations. Do the vowel feature detector outputs represent vowel qualities/articulations actually observed, or do they try to estimate underlying targets by discounting coarticulatory influences of adjacent segments?

(6) Phonetic Features or Segments. Are phonetic features identical in acoustic attributes for different segments? If not, would it be better to view perception as the problem of identifying segments from the temporal variations in acoustic property detector outputs? For example, [t,d,n] share a common place of articulation, and may share a single unifying integrated property, but it is unlikely that they share identical manifestations of place of articulation. Is there an inherent advantage to features, or is the advantage philosophical/genetic?

An alternative to the feature matrix as a segmental representation might be a column in which all possible phonetic segments are listed with an associated probability. Suppose we observe a voice onset time that is more compatible with [p,g] than with either [b] or [k]. It would be easy to specify highest probability for [p] and [g] within a segmental representation -- and some perceptual data suggests that this is appropriate (Oden and Massaro, 1978) -- but it is impossible to selectively favor this pair using only feature probabilities.

(7) Broad vs. Narrow Phonetic Representations. An intervocalic poststressed [p] is weakly aspirated, and so is somewhat ambiguous in voicing. The phonetic feature system, as described, does not permit specifying gradations of VOT, so this plosive will only be represented as having a slightly greater than chance probability of being voiceless. A word-initial highly aspirated [p] will generate more confident [p]-ness probabilities, and thus will better fit all lexical [p]'s, including those in poststressed position. This, and many other examples suggest that it is not a good idea to try to recover phonological segments (phonemes) prior to probing the lexicon because narrow phonetic information is useful in determining likely word-boundary locations, syllable structure and stress patterns (Church, 1986). To the extent that the segmental feature matrix produced by this model is somewhat inaccurate, or underspecified, or broadly phonetic, it is sub-optimal for lexical search.

DISCUSSION

We have identified a number of unsolved design issues which help to explain why phonetic feature extraction is not currently a popular method of automatic speech recognition. Phonetic features are hard to extract from acoustic data, and hard to convert to a representation suitable for probing the lexicon. A compelling list of theoretical and experimental reasons for believing that segments are perceptually real has been compiled by Pisoni and Luce

(1986); perhaps new methods of segment recognition and/or phonetic feature extraction can be devised to overcome the problems we have listed. Alternatively, the view that phonetic features are an essential aspect of language need not imply a belief in phonetic feature detectors for perception.

The Jakobson, Fant and Halle (1963) view of phonetics is that a very small number of universal binary distinctive features serves to describe language, both at the phonological and phonetic levels. Such a view, if adopted as a perceptual model, implies that the output of the phonetic feature detector stage is a rather broad phonetic characterization. The undesirability of a broad transcription became evident when we considered lexical search. A more narrow phonetic representation must be devised, perhaps by adding to the feature inventory. Also, feature outputs might take on continuous values representing strength of a cue rather than probability, in which case lexical representations can quantify expected position along a continuum of feature strength for each segment. However, in our view, phonetic feature detectors must make decisions and reduce the information content of the representation, or they become continuous recodings of the input which are no different in kind from those proposed for other non-featural non-phonetic models.

Relation to Perceptrons and Spreading Activation Models. There has long been an interest in simulating the presumed computational capabilities of neurons and neural assemblies (Hebb, 1949; Rosenblatt, 1962). One such model that captures the spirit of the phonetic feature detector model described in this paper has been proposed by Elman and McClelland (1986). Much is now known about the learning/generalization capabilities of this class of models (Minsky and Papert, 1969), and the implications are not entirely encouraging. I have described elsewhere specific problems with the Elman/McClelland implementation (Klatt, 1986b).

Relation to the Motor Theory. The motor theory of speech perception (Liberman et al., 1967; Liberman and Mattingly, 1986) advocates a transformation from acoustic data to articulatory representations. The claim is that segmental encodedness due to coarticulation, complex cue trading relationships, and other mysteries of perception can be better explained in articulatory terms. However, even if we grant that the motor theory proponents are correct and the outputs of the acoustic feature detector stage should be transformed into a model of the current hypothesized shape of an ideal vocal tract (Atal, 1975), such a transformation does not really solve most of the practical problems inherent in a phonetic feature model. Even ignoring the difficulty of determining a unique articulatory shape or trajectory from acoustic data (Atal et al., 1978), practical problems still center on making feature decisions and aligning features in order to represent the speaker's intended phonological segments, and then matching this highly reduced representation to lexical expectations. Furthermore, the rules needed to infer underlying features from articulatory shapes and dynamics may not be significantly easier to state algorithmically given present computer programming languages and pattern matching concepts.

Relation to Analysis by Synthesis. The model we have discussed might be considered as simply the initial stage of a more elaborate model of speech perception in which an important second module verifies lexical hypotheses by returning to the raw acoustic data to seek detailed confirmation/rejection. This "analysis-by-synthesis" model (see Halle and Stevens, 1962, the appendix in Klatt, 1979, Zue, 1985, or the companion Zue paper in these proceedings for a more detailed description) is in principle capable of overcoming errors and ambiguity in the initial hypothesization of words, and thus might tolerate imperfections and some featural indecisions.

Thus one way to simplify the task of the phonetic feature detector stage might be to suppose that these detectors only compute functions reflecting invariant attributes of features. More complex cue-trading relationships and context dependencies would then be handled at a later "analysis-by-synthesis" stage. The idea is that invariance-based features can be made to perform with an accuracy of perhaps 85% correct (Stevens and Blumstein, 1978; Kewley-Port, 1983), and this may be sufficient to access the lexicon. Shipman and Zue (1982) have shown that a broad-class acoustic classifier which avoids difficult decisions, such as place of articulation, can nevertheless significantly narrow the search among a large set of candidate isolated words. However, simulations of the continuous speech situation (Klatt and Stevens, 1973) suggest that the analysis-by-synthesis model is rapidly overwhelmed with lexical candidates when the phonetic matrix is underspecified, especially when the beginning time of a word is uncertain or there is an error such that no word matches perfectly.

The synthesis part of analysis by synthesis is intended to take advantage of the observation that synthesis rules are easier to state and less subject to ambiguous interpretation than corresponding (inverse) speech analysis rules. But synthesis is a fairly costly computational strategy, and is not a particularly plausible model of human perception (Klatt, 1979). An alternative, described next, is to precompute a knowledge representation equivalent to the synthesis stage of analysis by synthesis, and use it in direct analysis.

Relation to LAFS: Precompiled Acoustic Expectations. An alternative model of perception, "Lexical Access From Spectra" (Klatt, 1979; 1986a) proposes that the expected spectral patterns for words and for cross-word-boundary recodings are stored in a very large decoding network. Perception consists of finding the best match between the input spectral representation and paths through the network. No phonetic feature or segmental decisions are made as long as the system is dealing with familiar words.

For purposes of speech recognition, the advantage of a phonetic feature detector model over LAFS is in the possibility that relational invariants computed by acoustic detectors may go a long way toward combatting cross-speaker variability and discovering invariance. The disadvantages of a feature-based strategy are that it makes decisions too early (before lexical access), it has difficulty defining a representation that is appropriate for lexical access, and it requires expert specification of extremely complex decoding strategies in order perform well.

The advantages of the LAFS model are: (1) there is no assumption of phonetic feature invariance across segment types and across phonetic environment, so all phonetic sequence possibilities can be effectively treated as separate patterns if desired, (2) phonetics expertise is required only to set up the structure of the network, not to train/optimize it, and (3) no decisions are made too early since the first decision is a lexical one. The practical disadvantages of LAFS are that there may simply be too many cases to enumerate if all possible phonetic and lexical contexts are treated separately, and there is no well-motivated way to handle variability within and across speakers, except by defining alternative templates.

CONCLUSION

The initial stages of the phonetic feature detector model described in Figure 1 have the attraction of potentially taking advantage of (1) improved spectral representations of speech and (2) relational invariances that appear in the outputs of acoustic feature detectors. Succeeding stages of the model are far less attractive because it is unclear how to overcome the seven specific problems listed in the Example section. In preparing this review paper, I have come to the conclusion that there could be advantages to combining the attractive aspects of the initial stages of Figure 1 with the power of the LAFS model of lexical hypothesis formation. The result may be a LAFS model more capable of dealing with within-speaker and cross-speaker variability. Unfortunately, much basic research remains before an optimal acoustic-feature-based front end can be specified and interfaced with LAFS. [Research supported by NIH.]

REFERENCES

Allen, J. (1985), "Cochlear Modeling," IEEE ASSP Magazine, Jan., 3-29.

Atal, B. (1975), "Towards Determining Articulator Positions from the Speech Signal," in G. Fant (Ed.), Speech Communication, Vol. 1, Uppsala Sweden: Almqvist and Wiksell, 1-9.

Atal, B., Chang, J.J., Mathews, M.V. and Tukey, J. W. (1978), "Inversion of Articulatory-to-Acoustic transformation in the Vocal Tract by a Computer Sorting technique", J. Acoust. Soc. Am. 63, 1535-1556.

Chomsky, N. and Halle, M. (1968), The Sound Pattern of English, New York: Harper and Row.

Church, K.W. (1986), "Phonological Parsing and Lexical Retrieval," Cognition xx, xx-xx.

Clements, G.N. (1985), "The Geometry of Phonological Features," Phonology Yearbook, Vol. 2, Cambridge: Cambridge Univ. Press.

Delgutte, B. (1986), "xxxx," in J. Perkell and D. Klatt (Eds.), Variability and Invariance in Speech Processes, Erlbaum, xx-xx.

Elman, J. and McClelland, J. (1986), "xxxx", in J. Perkell and D. Klatt (Eds.), Variability and Invariance in Speech Processes, Erlbaum, xx-xx.

Goldhor, R. (1986), "A Model of Peripheral Auditory Transduction using a Phase Vocoder with Modified Channel Signals," ICASSP-86, 17.10. [See also ICASSP-83 1368-1371.]

Halle, M. and Stevens, K.N. (1962), "Speech Recognition: A Model and a Program for Research", IRE Transactions on Information Theory IT-8, 155-159.

Halle, M. and Vergnaud, J.R. (1980), "Three Dimensional Phonology," J. Linguistic Research 1 83-105.

Hebb, D.O. (1949), The Organization of Behavior, New York: Wiley.

Jakobson, R., Fant, G., and Halle, M. (1963), Preliminaries to Speech Analysis: the Distinctive Features and Their Correlates, Cambridge, MA: MIT Press.

Kewley-Port, D. (1983), "Time-Varying Features as Correlates of Place of Articulation in Stop Consonants", J. Acoust. Soc. Am. 73, 322-335.

Klatt, D.H. (1977), "Review of the ARPA Speech Understanding Project", J. Acoust. Soc. Am. 62, 1345-1366.

Klatt, D.H. (1979), "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access", in Perception and Production of Fluent Speech, R.A. Cole (Ed.), Lawrence Erlbaum Assoc. [See also J. Phonetics 7, 1979, 279-312.]

Klatt, D.H., (1986a), "The Problem of Variability in Speech Recognition and in Models of Speech Perception", in J. Perkell and D. Klatt (Eds.), Variability and Invariance in Speech Processes, Erlbaum, xx-xx.

Klatt, D.H., (1986b), "Response to Elman," in J. Perkell and D. Klatt (Eds.), Variability and Invariance in Speech Processes, Erlbaum, xx-xx.

Klatt, D.H. and Stevens, K.N. (1973), "On the Automatic Recognition of Continuous Speech: Implications of a Spectrogram-Reading Experiment", IEEE Transactions on Audio and Electroacoustics AU-21, 210-217.

Liberman, A.M., F.S. Cooper, D.S. Shankweiler, and M. Studdert-Kennedy (1967), "Perception of the Speech Code", Psychological Review 74, 431-461.

Liberman, A.M. and Mattingly, I.G. (1986), "The Motor Theory of Speech Perception Revised," Cognition xx, xx-xx.

Lisker, L. (1978), "Rapid vs. Rabid: A Catalogue of Acoustic Features that may Cue the Distinction,", Status Report on Speech Research SR-65, New Haven: Haskins Labs, 127-132.

Minsky, M.L. and Papert, S. (1969), Perceptrons: An Introduction to Computational Geometry, Cambridge MA: M.I.T. Press.

Oden, G.C. and Massaro, D.W. (1978), "Integration of Featural Information in Speech Perception", Psychological Review 85, 172-191.

Pisoni, D.B. and Luce, P.A. (1986), "Acoustic-Phonetic Representations in Word Recognition," Cognition xx, xx-xx.

Rosenblatt, F. (1962), Principles of Neurodynamics, New York: Spartan Books.

Sachs, M.B., Young, E.D. and Miller, M.I. (1982), "Encoding of Speech Features in the Auditory Nerve, in R. Carlson and B. Granstrom (Eds.), The Representation of Speech in the Peripheral Auditory System, Amsterdam: Elsevier Biomedical, 115-130.

Shipman, D.W. and Zue, V.W. (1982), "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," ICASSP-82, 546-549.

Stevens, K. N. (1972), "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data", in E.E. David and P.B. Denes (Eds.), Human Communication: A Unified View, New York: McGraw-Hill.

Stevens, K.N. and Blumstein, S.E. (1978), "Invariant Cues for Place of Articulation in Stop Consonants", J. Acoust. Soc. Am. 64, 1358-1368.

Stevens, K.N. and Halle, M. (1964), "Remarks on Analysis by Synthesis and Distinctive Features", Proc. of the AFCRL Symposium on Models for the Perception of Speech and Visual Form, in W. Wathen-Dunn (Ed.), Cambridge, MA: MIT Press.

Studdert-Kennedy, M. (1974), "The Perception of Speech," T.A. Sebeok (Ed.), Current Trends in Linguistics, The Hague: Mouton.

Zue, V.W. (1985), "The Use of Speech Knowledge in Automatic Speech Recognition," Proceedings IEEE 73, 1602-1615.

MODELS OF PHONETIC RECOGNITION II: AN APPROACH TO FEATURE-BASED RECOGNITION

K.N. Stevens

Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

Abstract An approach to speech recognition is proposed in which phonetic features are identified as acoustic properties in the speech signal, and lexical items are accessed directly without explicitly labeling phonetic segments. A possible advantage of such an approach is that a feature representation shows minimal modification as a consequence of the deletions and assimilation phenomena that occur in natural speech. Problems of determining acoustic correlates of features and of representing lexical items in terms of features are discussed.

In this paper I would like to argue that there are advantages to be gained by using phonetic features as primary units for identifying words. I hope to show that variability that occurs from speaker to speaker and from context to context can be taken into account in a natural way if features are used for representing utterances and if they form the building blocks for larger units by means of which utterances are identified.

Before discussing some of the advantages of features, and the structure of a speech recognition procedure based on features, let me first review some of the basic ideas underlying the concept of features.

Features and their Acoustic Correlates

A feature is a minimum unit in terms of which lexical items are represented (Jakobson, Fant, and Halle, 1963; Chomsky and Halle, 1968). Words that have different meaning (except for homonyms) have a different representation in terms of binary features. Thus, for example, the words mill and bill are differentiated on the basis of one of the features that characterize the initial segment -- in this case the feature sonorant. (Other features, such as nasal, may also play a role in this distinction. This concept of redundancy in the feature representation is discussed below.) It appears that about 20 features are needed to perform this function in language. Each lexical item is assumed to be represented in the mind of a speaker/listener in terms of patterns of features (with some further structure to this pattern).

Associated with each feature there is an acoustic correlate. This acoustic correlate, or property, is assumed to give rise to a pattern of response in the auditory system that is qualitatively different or distinct from the response pattern associated with other features. The property associated with each feature can be present in the sound with different degrees of strength. Features have articulatory correlates as well as acoustic or perceptual correlates, but in this paper our principal concern is with the acoustic correlates.

The acoustic properties that qualify as correlates of phonetic features tend to be relational and not absolute. Thus, for example, acoustic parameters such as the overall intensity of a component of the signal or the frequency of a particular spectral prominence, divided arbitrarily into two classes by a fixed intensity or frequency, would not qualify as the bases for the acoustic correlates of phonetic features. Parameters such as these show large interspeaker differences for the same utterance. Furthermore, there is no evidence to indicate a natural perceptual boundary or qualitative shift in the pattern of auditory response at an absolute intensity or an absolute frequency. On the other hand a property such as the frequency of one formant in relation to another could lead to qualitatively different auditory response pattern depending on whether the spacing between the two formants was greater or less than a critical value. (See, for example, Chistovich, Sheikin, and Lublinskaja, 1979.) Through proper selection of properties that describe spectral relationships, these properties can be speaker independent, since they do not depend on the speaker's vocal tract length or average fundamental frequency. Properties defining features can also be relational in the time domain. Thus, for example, a qualitatively different auditory pattern could result from an abrupt rise in spectrum amplitude in a broad frequency region as opposed to an abrupt fall in amplitude. In this case the relevant property is relational in the same sense that the amplitudes of spectral components at one time are interpreted in relation to the amplitudes of these components at an adjacent time.

There is a tendency for groups of features to be implemented more or less simultaneously, and consequently these features are naturally organized into segments. For example, within 10-20 msec of the release of a stop consonant, the sound contains properties identifying the features continuant and sonorant as well as the features related to place of articulation. In general, however, each feature is not specified for every segment. (For a discussion, see Halle, 1985, and references cited therein.) Sometimes just one feature might show a change at a point in time at which no other feature shows evidence for a change (e.g., the feature continuant in the initial consonant in /ča/, or the feature high in the vowel in /se/). On the other hand, some features may be defined for some segments, with no specification of these features for intervening segments. Thus, for example, in the word banana, the features indicating backness and high pitch are specified only on the second vowel and not in the other vowels, which are unstressed and reduced.

An important characteristic of the representation of an utterance in terms of features is that the representation usually has more features than the minimum number that are needed to distinguish the utterance from possible competitors. That is, there is redundancy in the feature representation. A consequence of this redundancy is that there is room for variability in the acoustic representation of an utterance. Not all features need to be marked in the signal, and the acoustic properties associated with these features can be present with different degrees of strength (Stevens, Keyser, and Kawasaki, 1986).

Situations often arise in which one or more features of one segment spread to a nearby segment, resulting in a change of some features of the segment, a specification of features that were previously unspecified, or even a deletion of the segment. Examples are: in miss you /s/ becomes [š], taking the palatal feature of the adjacent [j]; in at the, the sequence /t#ð/ can become [t̪], i.e., a dental t; in sit close in rapid speech, /t/ can lose its place features but retain the stop feature; in tree, the initial /t/ takes on the retroflex feature of the next segment. In many cases the spreading of features is allowed because there is redundancy in the feature description of a segment, and changing one or more features does not lead to misidentification of a lexical item. These assimilation phenomena often occur when there are two or more adjacent consonants, and they can occur within words or at the boundaries between morphemes or words. They appear to follow certain general principles, and linguists are working on models of feature organization that capture these principles in a natural way. (See, for example, Clements, 1985 and Halle, 1985.) The point is, however, that if the feature is used as a basic unit of representation these sources of variability in the speech signal can be accounted for in a rather natural manner.

Features, Variability, and Invariance

From the above discussion we can identify two principal sources of variability when an utterance such as a word is produced by different speakers with different speaking styles and in various contexts. One kind of variability arises mainly because different speakers have different vocal-tract sizes and shapes, and because talkers may use various speaking rates. This source of variability can be accounted for by proper specification of the acoustic correlates of the features. In particular, the acoustic properties should be relational so that they are insensitive to vocal tract size and speaking rate. Considerable progress has been made in specifying these acoustic properties, but much work remains to be done in this area. This research can be guided by an understanding of the psychophysics and physiology of hearing, and of theories of speech production.

The second source of variability arises because a speaker may modify the feature description that underlies an utterance or may make adjustments in the strength with which a feature is implemented. In some situations this modification is dictated by rules specific to the language, and in other cases the changes are optional and are influenced by speaking style. These modifications in the feature description appear to be capable of specification in terms of spreading of features across segments, such that features in one segment are changed as a consequence of particular feature values in an adjacent segment. The spreading can lead to changes in or elimination of one feature or groups of features.

Another source of interspeaker variability, which we shall not consider here, arises when different dialects are involved. Usually, however, it is possible to describe the phonetic differences between dialects in terms of a small set of rules operating on features.

Toward a Model for Feature-Based Recognition

How might a listener make use of features in decoding an utterance given the acoustic signal? Or, given the theme of this conference, how might we implement these ideas in a speech recognition system? The point of view we take here is that there are two stages to this process. The first stage is to identify the properties in the signal from which estimates of the features are made, and the second stage is to identify the lexical items from these properties. We imagine that testing for each property is carried out continuously through the speech signal. Most of the properties achieve maximum values or degrees of strength at particular points in time in the speech signal. These peak values of the properties define events in time within the signal. Some properties, however, maintain approximately constant strength over longer time intervals, and thus are identified with regions of time rather than with events in time. An example is the feature voiced, for which the acoustic correlate is the presence of low-frequency periodicity. (Other features are often active, and hence other properties are often present in the signal, when the feature voiced is implemented in English.) Also, there are some interrelationships between properties so that some properties cannot be extracted unless other properties are present. Thus the continuous speech signal is characterized by a series of signal streams, one corresponding to each property that is the acoustic correlate of a feature. For the most part, these signal streams consist of marks indicating brief time intervals or events, and these marks are labeled with the strength of the property. There is a tendency for these events corresponding to some groups of features to be approximately aligned, for example in the vicinity of a stop-consonant release.

We shall not discuss in detail the next stage of processing in which lexical items are accessed on the basis of these signal streams. Probably the most difficult and important problem to be solved is to determine a proper structure for the lexicon so that it can be accessed from these signal streams (or modified versions of these signals), given that these signals reflect the effects of redundancies and spreading phenomena of the type discussed above. There are several requirements for this structure: (1) in the feature representation, the notion that some features are redundant should be indicated in some manner; (2) while some features are aligned within the same segment, the representation should be structured to allow some flexibility in this alignment, possibly along lines of the tiered structure proposed by phonologists; (3) features or feature groups that are susceptible to spreading should be indicated so that assimilation phenomena may be accounted for in a natural manner.

[Supported in part by grants from the National Institute of Neurological and Communicative Disorders and Stroke and from the National Science Foundation.]

References

Chistovich, L.A., Sheikin, R.L., and Lublinskaja, V.V. (1970) "Centres of gravity and spectral peaks as the determinants of vowel quality," in B. Lindblom and S. Ohman (eds.), Frontiers of Speech Communication Research, Academic Press, London, pp. 143-157.

Chomsky, N. and Halle, M. (1968) The Sound Pattern of English, Harper and Row, New York.

Clements, G.N. (1985) "The geometry of phonological features," Phonology Yearbook, Vol. 2, Cambridge University Press, Cambridge.

Jakobson, R., Fant, G., and Halle, M. (1963) Preliminaries to Speech Analysis, MIT Press, Cambridge, MA.

Halle, M. (1985) "Speculations about the representation of words in memory," in V.A. Fromkin (ed.), Phonetic Linguistics, Academic Press, New York, pp. 101-114.

Stevens, K.N., Keyser, S.J., and Kawasaki, H. (1986) "Toward a phonetic and phonological theory of redundant features," in J. Perkell and D.H. Klatt (eds.), Variability and Invariance in Speech Processes, Erlbaum, Hillsdale, NJ.

MODELS OF PHONETIC RECOGNITION III: THE ROLE OF ANALYSIS BY SYNTHESIS IN PHONETIC RECOGNITION

Victor W. Zue

Department of Electrical Engineering and Computer Science and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract This paper proposes a recognition model that attempts to deal with variabilities found in the acoustic signal. The input speech signal is first transformed into a representation that takes into account known properties of the human auditory system. From various stages of this transformation, acoustic parameters are extracted and used to classify the utterance into *broad* phonetic categories. The outcome of this analysis is used for lexical access. The constraints imposed by the language on possible sound patterns should significantly reduce the number of word candidates. Finally, detailed acoustic cues will be utilized to select the correct word from the small set of candidate words.

Introduction

The task of phonetic recognition can be stated broadly as the determination of the transformation of the *continuous* acoustic signal into a *discrete* representation that can then be used for lexical access. In presenting my arguments, I will assume that words in the lexicon are represented by a set of phonological units. While the precise nature of these units, be they metrical feet, syllables, phonemes, or distinctive feature bundles, is not important for the present discussion, for the sake of consistency I will assume that words are expressed as strings of phonemes.

My proposed model of phonetic recognition makes use of broad phonetic analysis and language-specific constraints to reduce the number of lexical hypotheses, and to establish the context for further, detailed phonetic analysis. This is the third of a set of three papers from the MIT Speech Communication Group, expressing somewhat opposing views on the topic. Upon closer examination, however, there may not be as many differences as there are similarities. Like Klatt (these proceedings), I believe that the signal must be transformed into an acoustic, segmental description. However, I do not share his view regarding the feasibility of lexical access from short-time spectra, nor the use of a set of uniform distance metrics to measure phonetic similarities. Like Stevens (these proceedings), I believe in a representation based on distinctive features. However, I am increasingly frustrated by our inability to find invariance of these features in the acoustic domain, and thus I question the hypothesis that such invariance in fact exists.

Why Is Phonetic Recognition Difficult?

Phonetic recognition is difficult chiefly because the process of phonetic encoding in the acoustic signal is highly variable. Specifically, the acoustic realizations of a given phoneme can vary greatly as a function of context (Zue, 1985). On the one hand, different acoustic cues can signify the same underlying phonological representation. For example, the acoustic realization of the phoneme /t/ is drastically different in words such as "tea," "tree," "steep," "button," and "butter." On the other hand, the same acoustic cue can signify influences from different levels of the linguistic representation. For example, duration of a phoneme can be influenced by factors ranging from semantic novelty and syntactic structure to phonetic context and physiological constraints (Klatt, 1976). In order to perform phonetic decoding, a computer must extract and selectively attend to many acoustic cues, interpret their significance in light of other evidence, and combine the inferences to reach a decision. This is an immensely difficult task, given the incomplete state of our knowledge about the important acoustic cues and the ways they should be combined.

In addition to contextual variations, there are several other sources of variability that can affect the acoustic realization of utterances (Klatt, 1986). First, *acoustic variations* can arise from changes in the environment or in the position and characteristics of the transducer. Second, *within-speaker variations* can result from changes in the speaker's physiological or psychological state, speaking rate, or voice quality. Third, differences in sociolinguistic background, dialect, and vocal tract size and shape can contribute to *across-speaker variations*. Some of these variations may have little effect on phonetic distinctiveness, whereas others will have dire consequences. Successful phonetic recognition crucially depends on our ability to deal with all these sources of variability. Not only must we extract and utilize information from phonetic variations during recognition, we must also learn to disregard or deemphasize acoustic variations that are irrelevant.

Utilizing Constraints

The contextual variations observed in the speech signal can often be attributed to constraints imposed by the human articulatory mechanisms. For example, the motion of the formant frequencies during the production of the diphthong /a^y/ directly reflects the movement of the tongue from a low posterior position to a high anterior position. However, superimposed on such articulatory constraints is the knowledge possessed by a native speaker that certain gestures need not be as precise as others. In American English, for example, a speaker can choose to nasalize vowels at will, since the degree of nasality does not affect a phonetic decision. Similarly, a native speaker can produce a front, rounded vowel in place of a back, rounded vowel (as in the word sequence "two two") simply because the [+back] is a redundant feature for rounded vowels in American English.

Examples of such language-specific constraints are easy to find. The so-called *phonotactic* constraints govern the permissible phoneme combinations. There are also the *prosodic* constraints, limiting the possible stress patterns for a word. Knowledge about these constraints is presumably very useful in speech communication, since it enables native speakers to fill in phonetic details that are otherwise unavailable or distorted. Evidence of the usefulness of such language-specific knowledge can be gleaned from experiments in which phoneticians were asked to transcribe utterances (Shockey and Reddy, 1975). The transcription error was typically high when the utterance was from a language unknown to the transcriber, suggesting that "knowing what to expect" is important for phonetic decoding.

Large dictionaries have been used in several recent investigations into the magnitude of phonotactic and prosodic constraints for American English and other languages (Shipman and Zue, 1982; Huttenlocher and Zue, 1984; Carlson et al., 1985). All of these studies found that a broad phonetic representation roughly corresponding to manner of articulation of phonemes can often map words into equivalence classes with extremely sparse membership. In American English, for example, the expected value of the class size based on a six-category classification scheme was found to be 34, a reduction of more than two orders of magnitude from the size of the original lexicon. Results such as these suggest that a complete and detailed phonetic analysis of the speech signal not only is undesirable but may indeed be unnecessary. Broad phonetic analysis by its nature focuses on acoustic cues that are more invariant against contextual influences. That such a representation is also able to capture important phonological constraints imposed by the language suggests that large-scale lexical candidate reduction may be possible. Furthermore, because the exact phonetic context is specified by the candidate words, detailed phonetic knowledge can be used with greater confidence. If "tree" is a candidate word, then the verification process can use the predictive knowledge of the retroflexed context, as specified by the following /r/. The recognition algorithm will then be able to focus its attention on the detection of the retroflexed /t/ rather than a generic /t/.

A Phonetic Recognition Model

Figure 1 shows a possible recognition model incorporating some of the previously discussed ways of dealing with variability. The input speech signal is first transformed into a representation that takes into account known properties of the human auditory system, such as critical-band frequency analysis, dynamic range compression, temporal and frequency masking, adaptation and onset enhancement, and synchrony processing (see, for example, Seneff, 1985). From various stages of this transformation, acoustic parameters are extracted and used to classify the utterance into broad phonetic categories. The coarse classification also includes prosodic analysis that identifies regions where the speech signal is likely to be more robust. The outcomes of these analyses are used for lexical access. The constraints imposed by the language on possible sound patterns should significantly reduce the number of word candidates. Once the phonetic context has been established, detailed acoustic cues can then be used to select the correct answer from the small set of candidate words.

Note that the proposed recognition model is essentially a hypothesis-test, or analysis-by-synthesis, model. It has been proposed in the past for speech analysis (Bell et al., 1961) as well as for speech perception (Stevens and House, 1970). The

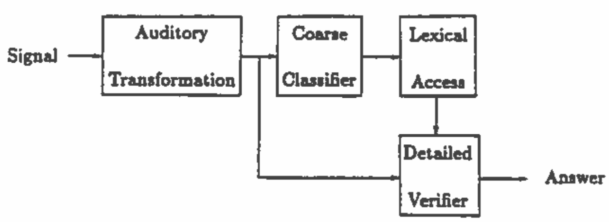


Figure 1: A Speech Recognition Model

A proposed speech recognition model that attempts to incorporate features for dealing with variabilities.

success of such a model relies heavily on the assumption that the number and the dimensionality of the hypotheses remain small. In our case, this is achieved through large-scale hypothesis pruning utilizing a proper set of constraints. Once the number of hypotheses becomes manageable, attention can be directed toward detailed acoustic cues that will enable us to make fine phonetic distinctions. The model is also computationally efficient since detailed acoustic cues are computed only when necessary. During verification, the acoustic cues can be determined in a prioritized manner as well. The computational savings, however, should be considered a side benefit; the primary appeal of the model stems from its ability to deal with variability. The coarse analysis is desirable because the resulting representation is relatively invariant across contexts and yet implicitly captures lexical and phonotactic constraints. Since detailed phonetic recognition is often error-prone, deferring this process will minimize error propagation.

To successfully implement such a model, mechanisms must be provided to insure that correct word candidates are not accidentally pruned and irretrievably lost. Errors of this sort occur for two reasons: either the coarse classifier makes a mistake or the lexicon does not anticipate a particular phonetic realization for the word by the speaker. This problem can be alleviated by permitting the lexical access procedure to accept reasonable insertions, deletions, and substitutions. If the errors are indeed reasonable, the correct word candidates should have better scores than the incorrect ones.

While the discussion leading to this model has focused on isolated words, the model can, in principle, deal with continuous speech as well. Instead of working with a set of word candidates, the verifier would deal with a *lattice* of word candidates. Provisions would then be made to determine and compare the relative goodness of words and word strings, subject to phonological, syntactic, and semantic constraints. Recent lexical studies using larger linguistic units such as syllables and metrical feet (Huttenlocher and Withgott, personal communication) show that these units exhibit constraints of similar magnitude. Using these large units may prove to be a more elegant way of accommodating continuous speech.

[Research Supported by DARPA under contract N00014-82-K-0727, monitored throught the Office of Naval Research.]

References

Bell, C. G., Fujisaki, H., Heinz, J. M., and Stevens, K. N. (1961), "Reduction of Speech Spectra by Analysis-by-Synthesis Techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725–1736.

Carlson, R., Elenius, K., Granstrom, B., and Hunnicutt, S. (1985), "Phonetic and Orthographic Properties of the Basic Vocabulary of Five European Languages," *Speech Transmission Laboratory Quarterly Progress Report*, STL-QPSR 1-2.

Huttenlocher, D. P., and Zue, V. W. (1984), "A Model of Lexical Access Based on Partial Phonetic Information," *Proc. ICASSP-84*, pp. 26.4.1–26.4.4.

Klatt, D. H. (1976), "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," *J. Acoust. Soc. Am.*, vol. 59, no. 5, pp. 1208–1221.

Klatt, D. H. (1986), "The Problem of Variability in Speech Recognition and in Models of Speech Perception," in *Variability and Invariance in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds., Hillsdale, NJ: Lawrence Erlbaum Assoc., pp. 300–319.

Seneff, S. (1985), "Pitch and Spectral Analysis of Speech Based on an Auditory Synchrony Model," Ph.D. Thesis, Massachusetts Institute of Technology.

Shipman, D. W., and Zue, V. W. (1982), "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," *Proc. ICASSP-82*, pp. 546–549.

Shockey, L., and Reddy, D. R. (1975), "Quantitative Analysis of Speech Perception," in *Proceedings of the Stockholm Speech Communication Seminar*, G. Fant, Ed., New York: John Wiley and Sons.

Stevens, K. N., and House, A. S. (1970), "Speech Perception," in *Foundations of Modern Auditory Theory*, J. Tobias and E. Schuber, Eds., New York: Academic Press.

Zue, V. W. (1985), "The Use of Speech Knowledge in Automatic Speech Recognition," *Proceedings IEEE*, vol. 73, no. 11, pp. 1602–1615.

ON THE AVAILABILITY OF DURATIONAL CUES

Thomas H. Crystal & Arthur S. House

Communications Research Division, Institute for Defense Analyses, Princeton, NJ 08540-3699 U. S. A.

ABSTRACT

Ongoing research to identify phones and measure their durations in recordings of read speech has resulted in the analysis of 10,300 phones produced by six talkers. The texts, the marking technique and some preliminary results were reported previously [2]. This report extends the earlier findings and tests for the presence of well-established durational cues cited in the literature. The analysis found, in general, that most of the cited effects are not clearly evident in continuous (read) speech signals. Some findings to be discussed are (a) completeness of stops; (b) stop variation in context; and (c) vowel lengthening.

INTRODUCTION

This is a progress report in an on-going program dealing with segmental durations in connected speech signals. An earlier report [2] described, in detail, the speech materials, talkers and methods of analysis. The study of these recorded materials has continued with an emphasis on segment durations and modeling of distributions. This report deals with measurements made on two scripts as spoken by six typical talkers—three from the original *slow* group (Nos. 1, 4 & 7; Table II [2]) and three from the *fast* group (Nos. 22, 34 & 43.) The scripts total approximately 600 words in 33 sentences [2, Appendix].

As before, the speech-sound segments of the readings were identified by studying a computer-graphics spectrogram and/or waveform display while simultaneously listening to the signal and by applying most of the standard criteria of acoustic and auditory phonetics. For stops and affricates the *hold* portions were measured (with occasional exceptions), as well as the plosive *release*. For stops with nonplosive releases—nasal, lateral, *etc.*—the released portion generally was included in the following segment. Word and pause boundaries are marked and have been used in the analyses.

SEGMENTAL DURATIONS

On the completeness of stops. A finding in Crystal & House [2] was the low percentage of "complete" stops (hold + plosive release) in the sample. Recently, stop closure duration was studied [6] with "sytematic conditions," using a corpus in which more than 95% of the stops were complete. Such a corpus may be very uncharacteristic of standard speech.

In this corpus the over-all frequency of occurrence of complete stops is 59%. There is a tendency for voiceless stops to be complete a higher percentage of the time than voiced stops (over-all, 65% *vs.* 51%), particularly in word-final position (42% *vs.* 18%.) As expected, word-initial stops are complete more often than word-final ones (85% *vs.* 33%.) There is a tendency, also, for velars to be complete more often than more fronted stops. Stop completeness is examined more closely in Table 1. The table entries display individual stops in various contexts, as indicated. *Caveat lector:* Validity is limited by small sample size and consequent atypical phoneme distributions!

The finding that plosions for /t/ and /k/ were always, essentially, measurable following /s/ is a little unexpected; they are considerably shorter, however, than the plosions of singletons. In the case of stops followed by /l/ or /r/, the plosion releases that occur generally are lallized or rhoticized. It is interesting to notice, also, that a higher percentage of plosions occurs in prepausal word-final stops (col. *F*) than in word-final stops in general.

Completeness of stops appears to be related to talking rate. The counts in Table 2 show that the *fast* talkers had about 10% lower completion than the *slow* talkers.

Differentiation of stop occlusion. Table 2 displays the duration of stop occlusions (holds) as a function of voicing characteristic and of place of articulation. (The results for *all* stops are highly similar.)

Table 1. Proportion (*Pr*) of occurrence of complete stops in various contexts. Symbols: *SI* = silence; # = word boundary; — = undefined context; *N* = total tokens in category. Six talkers; two scripts. [*A*] *#stop—*; [*B*] *#stop+l/r*; [*C*] *#s+stop+vowel*; [*D*] *#s+stop+r*; [*E*] *—stop#*; [*F*] *—stop#SI*.

		Position in Word						
		Any	*Initial*				*Final*	
Stop			*A*	*B*	*C*	*D*	*E*	*F*
p	*Pr*	.55	.88	.94	–	.34	.23	–
	N	108	42	18	0	6	43	0
t	*Pr*	.61	.84	.67	1.00	.96	.36	.49
	N	734	234	18	66	24	363	40
k	*Pr*	.77	.98	1.00	1.00	–	.69	.95
	N	310	137	17	6	0	73	21
b	*Pr*	.79	.79	.62	–	–	–	–
	N	208	206	29	–	–	0	0
d	*Pr*	.34	.81	.92	–	–	.18	.31
	N	471	116	12	–	–	320	51
g	*Pr*	.87	.93	.76	–	–	.34	1.00
	N	60	54	17	–	–	6	1
All	*Pr*	.59	.85	.79	*	*	.38	.48
	N	1891	798	111	*	*	838	163

*(voiced cognates do not occur)

Table 2. Analysis of hold portions of *complete* stops according to voicing characteristic (two left cols.) and place of articulation (three right cols.) Three *slow* and three *fast* talkers. Two scripts. *N* = number of tokens. *Dur* = duration in ms.

		Voicing		*Place*		
Talkers		*Voiced*	*V'less*	*Labial*	*Alveol.*	*Velar*
Slow	*Dur*	54	55	58	50	62
	N	202	388	120	322	148
Fast	*Dur*	54	50	56	48	53
	N	173	356	104	281	144
All	*Dur*	54	53	57	49	58
	N	375	744	224	603	292

The entries indicate that the hold portions of the *slow* talkers tend to be a few ms longer than those of the *fast* talkers. The durations of the holds of voiced and voiceless stops are not substantially different, contradicting experiments using citation forms [1] or words in a frame [10]. This confirms earlier observations [2] questioning the potential usefulness of a putative perceptual cue [1, 5] based on such a difference. (On the other hand, the average *plosions* of voiceless stops are about twice as long as those of voiced stops, as noted earlier by Zue [11].)

The effect of place of articulation is complicated. The average durations of the hold portions for the three (putative) places of stop articulation (right portion of Table 2), while not very different, show a definite tendency for alveolar stops to be shortest. The plosions give a different pattern, however, with duration increasing, on the average, as the point of contact moves from the lips to the velum. This results in total stop duration that, on the average, is about 80 ms for alveolars and labials and about 100 ms for velars.

O'Shaughnessy [9] measured durations of sounds in French words embedded in a sentence frame. In his materials labial stops were about 20 ms longer than lingual stops, with both types being considerably longer than the present results. He also has reported average stop (hold) durations for a read French passage [8] that are more comparable to the values in Table 3, but reports that voiceless stops are 10–15 ms longer, on the average, than voiced stops (63 ms *vs.* 76 ms.) Zue's [11] finding of longer releases for velars compared to labials and alveolars is supported in these materials, but his finding of longer hold portions for /p/ *vs.* /t/ and /k/ is not.

The corpus also contained 705 hold-only stops, *viz.*, without plosion *per se.* The average hold duration for these stops is the same, essentially, as that for complete stops, and the tempo-group differences are comparable.

The over-all conclusion, supported by [6], is that, in continuous speech, the hold portions of stop consonants are not strong indicators of voicing characteristic or place of articulation.

Vocalic variation. A contextual effect that is well-studied in English—lately in [6]—is the change of vocalic duration as a function of the voicing characteristic of the following consonant in the same syllable—the so-called *lengthening-before-voicing* effect. In [2] it was found for long (that is, *tense*) vowels preceding stops, but not for short (*lax*) vowels preceding stops, nor for either type of vowel preceding fricatives. In the present data the effect was investigated when the consonants are word-final and when they are word-final and prepausal, *viz.*, followed by a pause (but see *caveat* above.) Two general facts emerge: (1) the average duration of vowels preceding word-final prepausal consonants is considerably longer than that of vowels preceding word-final consonants in general, and (2) with the prepausal constraint, the data demonstrate the lengthening-before-voicing effect. The only exception noted was for the few cases of short vowels preceding fricatives. With this exception, there is an average 20-ms lengthening associated with vowels preceding prepausal voiced (*vs.* voiceless) obstruents. Without the prepausal constraint, however, the effect is not evident. It can be noted, also, that the progressive lengthening of short vowels before /t/, /s/, /n/, /d/ and /z/, pointed out in Lehiste [4], is not found in the present materials.

O'Shaughnessy [9] described two "strong" preconsonantal effects on vocalic duration in French: *lengthening* before voiced fricatives and *shortening* before voiceless obstruents. Neither effect is obvious in the present data, but there is a tendency for long vowels to lengthen before voiced fricatives. In [9] there also was a "weak" tendency for vocalic duration to vary inversely with vowel height. The present data confirm this for high (long) vowels (*viz.*, /i/ & /u/: $N = 379$) which are, on the average, shorter—108 ms—than other long vowels. However, the relation fails when mid (long) vowels (/e/ & /o/: $N = 318$, $Dur = 141$ ms) are compared to low (long) vowels (/a/ & /æ/: $N = 464$, $Dur = 132$ ms.)

Chen [1] reported that the lengthening usually attributed to the voicing characteristic of a postvocalic consonant functions across intervening sonorants separating a vowel and an obstruent (*sent vs. send.*) In his citation-form data, both sonorant and vowel were lengthened before a voiced, compared to a voiceless, obstruent. A rough test—long and short vowels, separately, before nasals and liquids followed by /p/, /t/, /k/, /č/, /f/, /s/ and their voiced cognates—shows the effect to be quite robust in the present data.

Table 3. Mean durations (*Dur*) and standard deviations (*SD*), in ms, for five "matched" pairs of back and front vowels preceding word-final stops and nasals grouped by place of articulation. N = number of tokens. Types in groups not equated.

	Back Vowels			*Front Vowels*		
Consonant Class	*N*	*Dur*	*SD*	*N*	*Dur*	*SD*
Labial	80	128	42	82	116	56
Dental	262	125	58	411	84	48
Velar	12	67	14	82	92	35

Another potential influence of consonantal context on vocalic duration is a place-of-articulation effect discussed by Fischer-Jørgensen [3] in which, before labials and dentals, back vowels > front vowels, but before velars, back vowels < front vowels. Data for examining this effect are presented in Table 3 (see *caveat* above.) For each consonant class the vowel category that, on the average, is longest, is the one predicted by Fischer-Jørgensen. The Fischer-Jørgensen study followed one by Maack [7], which claimed the relation "vowel+velar > vowel+dental > vowel+labial," but this relation does not hold in the present data. Further tests of vowels preceding word-final labial, dental and velar consonants using (1) 10 vowels (six long; four short) and (2) using all vowels occurring in the context resulted in an ordering by vocalic length that was the reverse of that described by Maack [7]. There are reports also on durational variation according to voicing characteristic for vowels *following* stops [3, 9]. Some of these phenomena are found in the present data.

REFERENCES

1. M. Chen, Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, 22, 1970, 129-159.

2. T. H. Crystal & A. S. House, Segmental durations in connected speech signals: Preliminary results. *J. acoust. Soc. Amer.* 72(3), 1982, 705-715.

3. E. Fischer-Jørgensen, Sound duration and place of articulation. *Z. Phonetik usw.* 17, 1964, 175-207.

4. I. Lehiste, Segmental features of speech. In *Contemporary issues in experimental phonetics*, N. J. Lass, ed. (Academic Press, 1976), pp. 225-239.

5. L. Lisker, Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, 33, 1957, 42-49.

6. P. A. Luce & J. Charles-Luce, Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *J. acoust. Soc. Amer.* 78(6), 1985, 1949-1957.

7. A. Maack, Die Beeinflussung der Sonantendauer durch die Nachbarkonsonanten. *Z. Phonetik usw.* 7, 1953, 104-128.

8. D. O'Shaughnessy, A multispeaker analysis of durations in read French paragraphs. *J. acoust. Soc. Amer.* 76, 1984, 1664-1672.

9. D. O'Shaughnessy, A study of French vowel and consonant durations. *J. Phonetics*, 9, 1981, 385-406.

10. L. J. Raphael, Durations and contexts as cues to word-final cognate oppositions in English. *Phonetica*, 38, 1981, 126-147.

11. V. W. Zue, *Acoustic characteristics of stop consonants: A controlled study.* Unpublished ScD dissertation, Massachusetts Institute of Technology, 1976.

USING STRESS INFORMATION IN LARGE VOCABULARY SPEECH RECOGNITION

Pierre Dumouchel, Matthew Lennig*
INRS-Télécommunications (Univ. du Québec)
3, Place of Commerce
Verdun, Québec CANADA H3E 1H6

Using stress information in a Markov source-based large vocabulary speech recognition system provides a way to examine a nonlocal cue which is generally poorly represented by the Markov source model. In this paper, we present an algorithm for estimating the stress pattern based on syllable durations and short-time energies. The output also gives the probability of the correctness of the estimated stress pattern. The parameters are first normalized in an attempt to reduce variability due to different linguistic contexts. The stress pattern is then estimated based on a statistical approach. After initial training, tests on a new word list yielded 95% correct detection of the syllable carrying the primary stress. Finally, inclusion of this algorithm in a large vocabulary isolated word recognition system contributes to its accuracy.

INTRODUCTION

The goal of this research is to devise an algorithm for the estimation of the stress pattern of a spoken word from its acoustic signal. Such an algorithm would serve as a component of a speech recognition system. Input to the stress pattern estimation algorithm consists of a word's hypothesized phonemic transcription with stress markers and the corresponding acoustic signal. The output is a probability estimate of the correctness of the hypothesized stress pattern assuming the segmental transcription is correct. Only duration and short-time energy are used as parameters.

The definition of stress differs depending on whether we regard it from the point of view of the speaker or from the point of view of the hearer. From the speaker's standpoint, stress may be defined in term of greater effort to produce a syllable. From the listener's standpoint, stress is manifested by duration, energy level and increased (or decreased) pitch. Moreover, stress information is not strictly localized in time but requires information from the surrounding syllables of the word. In other words, stress is a contrastive nonlocal cue which overlaps adjacent segments because it is expressed *relative to other segments*. In this work, since we are interested in speech recognition applications, we will adopt the listener's point of view.

The purpose of the stress pattern estimation algorithm is to sharpen the overall accuracy of a Markov source-based speech recognition system. The incorporation of this algorithm as a module in a recognition system will also provide a way to examine a contrastive nonlocal cue. Nonlocal cues are poorly represented in the framework of Markov models.

A published lexical stress detection technique due to Aull (1984) may be described as categorical since no confidence estimate of the decision correctness is made. Aull tries to find the primary stress syllable of the word and the remaining syllables are labelled by rules as either unstressed or reduced. The present paper explores a probabilistic lexical stress detection technique. First, a normalization is applied on the energy and duration parameters in an attempt to reduce the variability due to different linguistic contexts. Second, the algorithm finds the hypothesized primary stress syllable based on a statistical approach. Finally, the probabilities of the estimated stress pattern and the lexical stress pattern (as given by the Webster's Seventh New Collegiate Dictionary) are evaluated.

* also with Bell-Northern Research, Montréal, Canada

DESCRIPTION OF THE ALGORITHM

Duration, energy level and pitch of the syllable are phonetic correlates of stress. But stress is not the only phenomenon which exerts an influence on these parameters. Intrinsic phonetic characteristics, phonological context, prepausal lengthening and speaking rate may also affect them. Hence the lexical stress algorithm uses a series of fixed correction factors to compensate for each of these effects except stress. In this study, only the duration and energy level cues are used. Pitch is not employed due to the difficulty of extracting reliable fundamental frequency information. Since stress principally affects the vowel part of the syllable, we judge it to be sufficient to examine only this class of phonemes. By doing so, we avoid having to segment difficult classes of phonemes such as initial and final voiceless stops. Hence, the duration cue used by the stress algorithm is the duration of the vowel part. Similarly, the energy level cue is the average of energy level over the vowel. The phonemic segmentation is based on a Viterbi alignment technique.

Normalization of intrinsic phonetic characteristics is used to compensate for the intrinsic duration and intensity of the vowels. For example, for the same source power the high-front vowel i will generally be less intense than a low-back ɑ . Hence, compensation factors are proposed for the intrinsic phonetic characteristics to counter this variability. The compensation factors that we use come from two studies of Lehiste (1960, 1970). Similarly, phonological normalization is used to compensate for the influence of the adjacent phonemes on the duration of the vowel. For example, a vowel is longer if the syllable ends with a voiced stop rather than a voiceless stop. The phonological duration compensation factors come from the previously cited Lehiste study (1960). The phonetic description of the syllable is given by the dictionary. No phonological energy compensation factors are proposed. A fixed factor is proposed to compensate for prepausal lengthening. Finally, linear normalization of parameters, such that within a word the normalized durations sum to unity and the normalized energies sum to unity, acts as a compensation for speaking rate and overall speaking level effects.

Figure 1 shows the distribution of vowels based on stress type for a corpus of 135 two-syllable words read in isolation from a text by a male speaker. The symbol P stands for *primary stress*, S for *secondary stress*, U for *unspecified* stress as given by the dictionary. The unspecified stress syllable is one with no lexical stress marker and it corresponds either to a ternary stress syllable or an unstressed syllable. The vowels are represented by their normalized, compensated parameters. No evident demarcation between the unspecified and secondary stress syllables is seen. It appears from this figure that three regions can be identified: a region where the primary stress syllables predominate, another one where the unspecified stress and secondary stress syllables predominate, and finally an overlapping region where all the

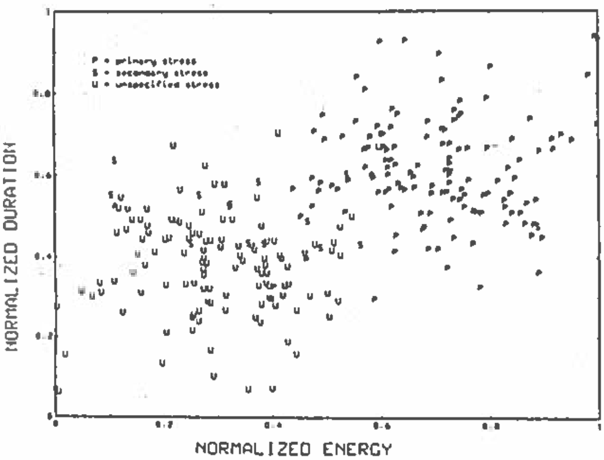


Fig. 1 Distribution of vowels.

types of stress are present. Similar figures are obtained for three and four syllable words but with different centers of gravity for each region. The difference between centers of gravity is due to the fact that we normalize the sum of each parameter to unity regardless of the number of syllables in the word. We conclude that a statistical approach is viable only to differentiate between the primary stress syllable and the other types of syllables (including the secondary and unspecified stress syllables). Furthermore, it appears that an additional normalization factor applied to each parameter for words containing more than two syllables can produce plots with centers of gravity similar to two-syllable ones. Based on these facts, the energy-duration space has been partitioned into 41 regions. The regions are enclosed by straight lines with slopes of minus one. Regions corresponding to the overlapping region are of smaller dimensions to achieve finer discrimination at the category boundary. We allocate to each region a probability denoting a specific type of stress. The probability is based on the frequency of appearance of a specific type of stress within a region compiled from a list of 220 polysyllabic words:

$$\Pr[stress = X \mid region = Y] = \frac{number\ of\ X\ in\ Y}{total\ number\ in\ Y}$$

A hand-smoothed version of results obtained with the above equation has been used. This is necessary to avoid unwanted effects of the relatively small size of the corpus such as a region not containing any data points. Finally, the estimated primary stress for the word is assigned to the syllable which has the highest probability of being primary. The probability of the maximum likelihood stress pattern is estimated as the average of syllable probabilities with respect to its estimated type of stress. We use the average of syllable probabilities instead of the multiplication of syllable probabilities since the latter incorrectly favors words with the smallest number of syllables. The lexical stress probability is determined in a similar way except the stress pattern is now the one proposed by the dictionary.

RESULTS

After initial training, tests on the same speaker reading a new word list yielded 95% correct detection of the primary stress syllable when compared to the lexical stress pattern. A list of 50 words pairs such as *PERfect—perFECT* (noun/verb) and a 220 polysyllabic words constitute the training word set. The test set contains 112 new polysyllabic words. The syllable distribution within the test corpus is the following: 66% two-, 23% three-, 10% four- and 1% five-syllable words. An examination of the errors reveals that of the 5% errors, three-fifths are due to incorrect phonemic segmentation produced by the Viterbi algorithm and one-fifth are due to a stress pronunciation of the word which differs from that of the dictionary. A final test which consists of examining the contribution of this algorithm in a large vocabulary speech recognition system has been performed. The recognizer uses hidden Markov models to hypothesize a list of words with their associated probabilities. During this test we modify the likelihood of each word derived from the acoustic data by the probability that the required lexical stress pattern is supported by the observed stress data. Results show that the rank of the correct word in the word hypothesis list improves by an average of 0.3 word positions when using stress information. This test is performed using 60 test words. However, for two-thirds of the list the correct word is already ranked first, so no improvement is possible. Excluding these top rank cases, the improvement amounts to an average of 0.9 word positions.

DISCUSSION

Lexical stress can be useful in recognition but its estimation is difficult because

- even in isolated word speech, word stress differs from the lexical stress pattern (1% of cases),
- the lexical secondary stress syllable is considered less stressed than the unspecified stress syllable of the same word in 30% of cases, based on a perceptual experiment with one subject on a list of 25 words.
- normalized duration and short-time energy parameters for secondary and unspecified stress form overlapping classes.

Hence an approach which attempts to find the primary, secondary and unspecified stress syllables of the word is excluded. However, an approach which consists of finding only the primary stress syllable is possible and can also offer a good constraint. By expressing the confidence of the detection probabilistically, the performance of the algorithm can be integrated with the results of the other recognition system modules. The technique described in this paper respects these constraints and the performance of the algorithm is extremely satisfying. However, the contribution of the algorithm to a large vocabulary speech recognition system is only a minor improvement in the rank of hypothesis. Further improvements are anticipated from a better match between relative likelihoods based on acoustic-model estimation and stress estimation.

REFERENCES

Aull, A.M., *Lexical stress and its application in large vocabulary speech recognition*, Master's thesis, Massachusetts Institute of Technology, 1984.

Peterson, G.E. and Lehiste, I., "Duration of syllable nuclei in English", *Journal of the Acoustical Society of America*, vol. 32 no. 6, June 1960.

Lehiste I., *Suprasegmentals*, The MIT Press, Cambridge, Massachusetts, Chapter 4, 1970.

CHARACTERIZING FORMANTS THROUGH STRAIGHT-LINE APPROXIMATIONS WITHOUT EXPLICIT FORMANT TRACKING

S. Seneff

Rm 36-549, Research Laboratory of Electronics, M.I.T., Cambridge, MA. USA 02139.

A new method for representing the formants of sonorant speech sounds is described. The method collapses the two-stage process of (1) formant tracking and (2) abstraction of rates and directions of formant movements into a one-step process of directly assigning straight-line segments to the resonance contours in the frequency-time space. The method resembles techniques used in vision research [1], and is also motivated by observations of specialized frequency-modulation detectors in the central auditory system [4]. The computational procedures are straightforward, leading to a description of the formant information for a given vowel by a list of oriented straight-line segments. The line segments are not assigned to particular formants, such as F_2. Instead, the recognition process is hypothesis-driven. For each vowel or diphthong to be recognized, a short description of expected ranges of frequency and orientation in the time-frequency dimensions for the first two formants is given. Feasibility of the method is demonstrated by applying it to the specific task of recognizing the vowels and diphthongs of American English in restricted context, spoken by multiple speakers.

OVERVIEW

It is generally accepted that the frequencies of the formants, particularly the first two formants, are the most important information leading to the identification of vowels. Formant movements are also necessary for identifying diphthongs and semivowels. As a result, a number of investigators have attempted to develop formant tracking algorithms, which assign spectral peaks to specific formants, such as F_1, F_2 and F_3. Once the formant tracks are available over time, it is possible to develop algorithms that detect high-level features, such as a rising formant over the second half of a vowel.

Our approach is to represent the formant information directly by a collection of straight line segments, thus bypassing the stage of formant tracking. The formant patterns are described by oriented lines which often overlap in time and/or frequency, and which collectively provide sufficient information for identification of the phonetic content. These line segments lead naturally to descriptions such as "rising formant", with the slope of the line conveying the degree of rise.

The spectral representation, the "pseudo spectrum," from which the line segments are abstracted is obtained using an auditory-based signal processing method, as described in [2]. The method typically yields enhanced peaks at formant frequencies with smooth transitions over time. For voices with a high fundamental frequency, the individual harmonics of the pitch are often resolved below the first formant, thus making it very difficult to track F_1 in the traditional way.

LINE FORMANT EXTRACTION PROCESS

The process to obtain a list of straight-line segments describing the formant patterns in a given sonorant segment of speech is illustrated in Figure 1. The pseudo spectrogram for the word "Burt", spoken by a male speaker, is shown in Part (a) of the Figure, with the frequency axis represented on a Bark scale. A nonlinear filter-and-quantize procedure defines "On" and "Off" contour regions in time and frequency, shown in Part (b). Each robust peak in a given pseudo spectral cross-section is allowed to vote for a best-fit line segment passing through its time-frequency location, restricted to stay within an "On" region, and oriented in one of 11 specified directions. The votes of the robust peaks are accumulated in a list giving information about the orientation, center-points in time and frequency, duration, and mean amplitude of each line.

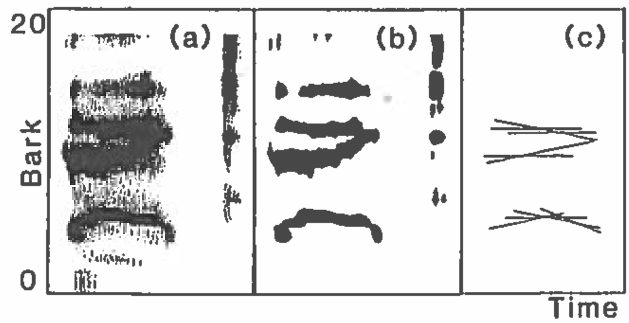


Figure 1: Illustration of Line-formant Abstraction Process (a) Pseudo spectrogram for word "Burt"; (b) One-bit enhanced spectrogram defining allowable regions for line segments; (c) Resulting line segments describing formants of vowel.

The next step is to consider collectively the list of candidate lines over a time interval defined by the unknown vowel's extent. Usually, several peaks vote for the same line or very similar lines. A heuristic algorithm was developed to collapse the list of lines into a new list, with "equivalent" lines merged into a single representative, which includes a count of the number of votes being merged. Finally, the list is further pruned, and line segments that appear to be insignificant are discarded. Elimination is based on threshold requirements for the number of votes, the minimum allowable duration, and the mean amplitude. The line segments that remain after pruning in the example are shown in Part (c) of the Figure.

The final step is to convert the list of line segments into a fuzzy descriptor format. The temporal extent of a given line is converted to a verbal description of its extent relative to the vowel end points, such as "first half". Similarly, the strength and orientation of the line are quantized to a small set of possibilities. Only the center frequency is retained as a number. Table 1 lists allowable categories for each item.

Orientation		Temporal		Strength
Rapid Rise	Rapid Fall	At Start	At End	Strong
Rising	Falling	First Half	Second Half	Medium
Slight Rise	Slight Fall	In Middle	Throughout	Weak
Steady				

Table 1: Categories for descriptors of line formants.

VOWEL RECOGNITION STRATEGY

The line formant representation was applied in a speaker-independent recognition task for the following 16 vowels and diphthongs of English, restricted to /bVt/ context: /i, e, yu, I, ε, æ, ɑ, ɔ, o, ʌ, ʊ, u, ɑ^y, ɑ^u, ɔ^y, ɝ/. The only step used for speaker normalization was to reference the center frequency *in*

3. REDEFINING THE SEGMENTATION PROBLEM

Within the context, defined above, speech-segmenting consists in researching an acoustic trajectory in the hope of tracking down targets, whether or not they are articulatorily met. As may be noticed, the larger problem of target identification can be made to pertain to acoustico-phonetic decoding, thanks to a grammar of distortions; as such a grammar of distortions can be inferred both from what is already known of co-articulation and from facts observed along the trajectory.

4. AN APPROACH THROUGH ANALYTICAL-MECHANICS

4.1. Usual Dimensions

Beside the already defined notions of velocity and acceleration, other dimensions can also be computed :

- curvature radius of the trajectory at point $M(t_n)$
- torsion of the trajectory at point $M(t_n)$

Whence it is possible to deal with the usual notions of rectilinear trajectory, stationary trajectory, etc. These notions can be extend over even longer temporal window-slits by associating, to each point $M(t_n)$, the variance-covariance matrix calculated over the m preceding points, using the m vectors $\{X_{n-m+1} \ldots X_n\}$. The first two proper directions (proper vectors) of this matrix can be assimilated to the directions of, respectively, the mean velocity vector $\underline{V}_n$ and the mean acceleration vector $\underline{G}_n$, on both of which the computations, aluded to above, can be run.

Now, if a mass is associated to point M, any directional alteration is the resultant of all forces applied to this point. It being assumed that clustering forces are frictionless, and that point M obeys strictly to the general laws of dynamics: point acceleration (whether positive or negative) is the resultant of attraction forces whose respective origins are the different targets --here considered as force fields.

4.2. Modelization

In order to extract interpretable path-portions from the trajectory, the following assumptions are made:

(a) the material point M moves towards one and only one target at a time,

(b) a target is considered met, whenever the trajectory becomes quasi-stationary,

(c) clustering forces are frictionless,

(d) the mean velocity $\underline{V}_n$ increases with speech output,

(e) a target is there but fails to be met, whenever the trajectory shows either a retrogression point or a sudden and marked directional change,

(f) around each target, there exists a force field the intensity of which decreases with speech output.

4.3. Experimentation

In an initial study, the p parameters of R^p to be retained were cues, otherwise used in speech analysis [Caelen et al, 81]. They are slow-variation cues, and thus the trajectories secured were sufficiently "smooth" to be meaningful. Over a preliminary corpus (isolated words pronounced by 10 speakers) the following observations were made: (Fig. 1)

(a) parameters are locally correlated according to phonemes; bringing out the existence of local clustering forces (or constraints). This should not come as a surprise, since we are dealing with co-articulation phenomena; but it allows (through intercorrelation-coefficient parameters) to quantify these phenomena.

(b) within a transitional phase between targets, the trajectory is quasi-linear (although this depends upon the coordinate system used).

(c) the trajectory is quasi-stationary whenever a target is met. A "Brownian movement" is then to be noticed around the target center.

(d) the trajectory does exhibit a directional alteration, if a target fails to be met.

(e) whenever speech-output rate becomes high, the number of such "failed" targets rises, while their mean reciprocal distances decrease.

(f) point M picks up speed as it leaves a target, and slows down as it nears the next one.

(g) there exists a grammar of distortions that makes it possible to superpose various speakers respective utterance trajectories.

4.4. Segmenting Automaton

On the basis of the preceding observations (a through g) it is possible, for the purpose of segmenting, to classify trajectories into three different types :

1 - "Brownian" trajectories (weak-amplitude motion about a target center) corresponding to a "target-met" detection procedure (TM).

2 - "Angular" trajectories (negative scalar product of mean velocities, retrogression point, slow down before odd point and speed up thereafter) corresponding to a "failed-target" detection procedure (FT). Note that the failed target always lies beyond the retrogression point.

3 - "Steady" trajectories (large curvature-radius, no odd point, maximum velocity reached about mid-course) corresponding to a transition-path detection procedure (T).

These three types of trajectory define the three different states assumed by an automaton whose transitional arcs are activated by TM, FT and T procedures.

5. CONCLUSION

The above makes it possible to look at segmenting, and subsequently at acoustico-phonetic decoding, from a new and maybe more advantageous angle : instead of researching discontinuity, we would resort to the formal instruments of mechanics (or data-analysis) to examine local variations in speech-trajectories that are represented in suitable spaces. Such a representation allows for an ascending description, from acoustics to phonology; while by-passing any a priori (even implicit) phonetic model. At the same time, it seems possible to find a grammar of distortions capable of superposing the several trajectories that correspond to one sequence uttered by several speakers. This kind of results, nevertheless, remains to be confirmed over large speech-corpuses and large numbers of speakers.

Acknowledgment: I wish to thank J.F. Malet (of California State University, Sacramento) for translating this text from French, and for suggesting various improvements as I wrote it.

BIBLIOGRAPHIC REFERENCES

[Abry et al, 85] C. Abry, C. Benoit, L.J. Boë and R. Sock, Un choix d'événements pour l'organisation temporelle du signal de parole, Proceedings GALF-CNRS XIVèmes JEP, Paris, 1985, pp.133-135

[Caelen et al, 81] J. Caelen and G. Caelen-Haumont, Indices et propriétés dans le projet ARIAL II, Proceedings GALF-CNRS, Processus d'encodage et de décodage phonétique, C. Abry, J. Caelen, J.S. Liénard, G. Perennou and M. Rossi eds., 1981, pp. 129-143

[Cohen, 81] J.R. Cohen, Segmenting speech using dynamic programming, JASA, Vol 69, n°5, 1981

[Fant, 73] G. Fant, Speech sounds and features, MIT Press, Cambridge, 1973

[Fujimura, 81] O. Fujimura, Temporal organization of articulatory movements as a multidimensional phrasal structure, Phonetica, Vol. 38, 1981, pp. 66-83

[Jakobson et al, 51] R. Jakobson, G. Fant and M. Halle, Preliminaries in speech analysis, MIT Press, Cambridge, 1951

[Ladefoged, 71] P. Ladefoged, Preliminaries to linguistic phonetics, The University of Chicago Press, 1971

[Lindblom, 83] B. Lindblom, Economy of speech gestures, in: P.F. MacNeilage ed., The production of speech, Springer-Verlag- Heidelberg, 1983

[Rossi, 83 M. Rossi, Niveaux de l'analyse phonétique: nature et structuration des indices et des traits, Speech. Com., Vol. 2, n° 2-3, 1983, pp. 91-106

[Zue, 83] V. Zue, The use of phonetic rules in automatic speech recogition, Speech. Com., Vol. 2, n° 2-3, 1983, pp. 181-186

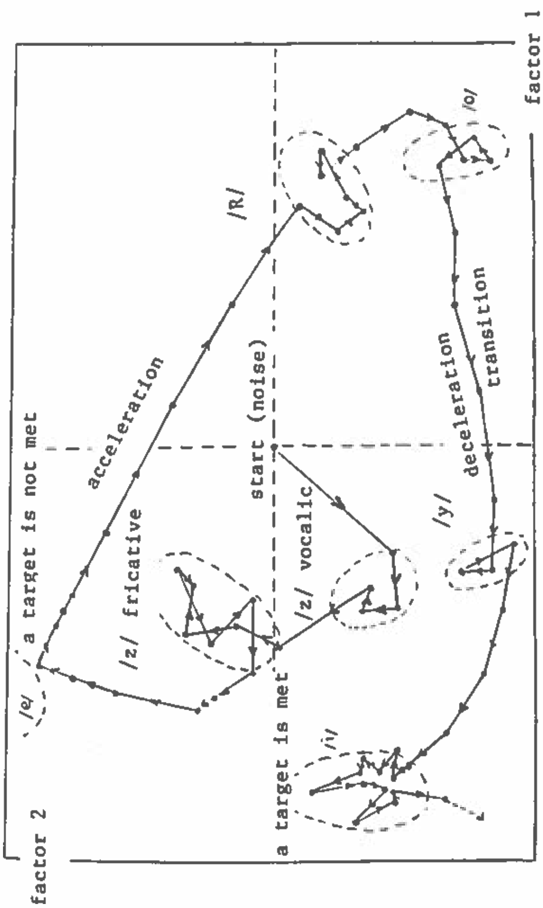


Fig. 1: Trajectory of the words "zéro-huit" /zeRo yi/

INVARIANCE DES SPECTRES DE PAROLE PAR ANALYSE DES CORRELATIONS CANONIQUES.

Adaptation d'un Système de Reconnaissance de Mots Isolés à de nouveaux locuteurs.

K.Choukri*,** ,G.Chollet** ,Y.Grenier**

* Laboratoires de Marcoussis, CRCGE. Route de NOZAY, 91460 Marcoussis.
** ENST-SYC, CNRS UA 820, 46 rue Barrault, 75013 Paris, France.

RESUME:

Cet article décrit une technique d'adaptation d'un dictionnaire de formes de référence à de nouveaux locuteurs, dans le cadre d'un Système de Reconnaissance Automatique de la Parole (SRAP). Elle se base sur l'hypothèse d'une corrélation maximale entre les espaces spectraux du locuteur standard et du nouveau locuteur pour déterminer un espace commun où les spectres respectifs sont invariants. Une application à la reconnaissance des dix chiffres montre les améliorations qu'elle apporte.

ABSTRACT:

Various speaker normalization and adaptation techniques of a knowledge data base or reference templates to new speakers in automatic speech recognition (ASR) have been studied during last years. This paper focusses on a technique for learning spectral transformations, based on a statistical analysis tool (Canonical correlation analysis), to adapt a standard dictionary to arbitrary speakers which does not require prior knowledge about them. The proposed method permits to improve speaker independance in Large vocabulary ASR. Application to an isolated digit recognizer improved a 70% correct score to 87%.

1. Introduction:

La représentation mathématique du signal de parole est déduite de l'onde acoustique acquise dans différents environnements (microphone, bruit ambiant, ...). La production de la parole (vibrations des cordes vocales et transmission par le conduit vocal) dépend fondamentalement des caractéristiques physiologiques et articulatoires des locuteurs, de l'influence des contraintes sémantiques, syntaxiques et lexicales (compétence et aptitude linguistiques), de l'état physique du locuteur (fatigue, émotion, ...) ainsi que d'autres facteurs paralinguistiques.

Ces différences expliquent la variabilité inter-locuteur observée dans le signal de parole. On observe aussi une variabilité intra-locuteur, mais beaucoup moins importante, ce qui explique les meilleures performances des systèmes dépendents du locuteur par rapport aux systèmes indépendents des locuteurs. Cela explique aussi le biais introduit dans les mesures de distance spectrale.

Pour réaliser des systèmes de reconnaissance indépendents du locuteur, plusieurs axes de recherche sont actuellement explorés. On distingue trois grandes directions. La première tente d'atténuer l'influence de la variabilité inter-locuteur en augmentant le nombre d'archétypes associés à chaque son dans le dictionnaire de référence, de telle sorte que tous les locuteurs représentatifs de la population d'utilisateurs fassent partie des locuteurs d'apprentissage. Pour y parvenir on utilise différents artifices tels que chaines de Markov, analyse discriminante, Clustering,...

La seconde méthode cherche des traits invariants aussi bien au niveau articulatoire, acoustique que perceptuel et ne garde que ces paramètres pour la représentation de la parole.

La première technique est telle que l'acquisition, la sélection et le codage des références deviennent vite une longue et coûteuse procédure. En outre le dictionnaire de références résultant occupe une place mémoire substantielle et on ne fait pas appel à des données spécifiques à la parole. La seconde technique, quoique plus attrayante, a encore besoin de quelques années de recherche avant d'être operationnelle, en élaborant un modèle de l'influence des caractéristiques du locuteur et de ses habitudes articulatoires sur le signal observé.

Ce papier concerne la troisième technique qui est l'adaptation d'un SRAP de base à chaque nouvel utilisateur.

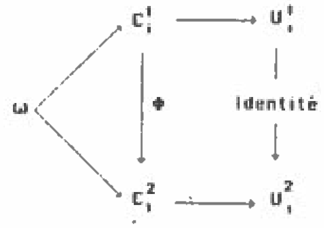
Chaque individu écoutant pour la première fois la parole d'un locuteur inconnu a souvent besoin de s'adapter à la nouvelle voix (ou d'adapter son appareil de perception), et les premiers mots d'un dialogue n'apportent guère d'informations que celles nécessaires à cette adaptation. D'une façon similaire on peut envisager une adaptation de SRAP basé sur le dictionnaire spécifique à un locuteur, à d'autres locuteurs sans acquérir leur dictionnaire spécifique respectif. Ceci permettra d'utiliser des algorithmes qui ont fait leurs preuves et dont on connait les performances. Par ailleurs cette procédure peut être accomplie d'une façon dynamique (Choukri et al.,1986), c'est à dire incorporée dans un système en configuration réelle d'exploitation ou d'utilisation.

2. Principe de l'adaptation de SRAP au locuteur:

Beaucoup d'auteurs qui s'intéressent aux problèmes dûs à la variabilité du signal de parole cherchent à normaliser des paramètres utilisés dans sa représentation. Il tiennent compte de paramètres articulatoires tels que la longueur du conduit vocal ou d'autres paramètres tels que les positions relatives des formants.

Le principe de la méthode est basé sur le constat qu'un même "son", produit par différents locuteurs, est interprété de manière identique par les personnes qui l'entendent malgré la variabilité inter-locuteur. On peut donc envisager un espace où des sons phonétiquement identiques seront représentés par des modèles identiques (Choukri et al.,1986).

Si on considère des cepstres sur une échelle Mel (MFCC) comme paramètres représentant la manifestation acoustique de chaque mot, l'espace associé à chaque locuteur est donc, dans un premier temps, un espace cepstral où la variabilité inter-locuteur s'exprime pleinement. Si on considère un son ω (mot, syllable, ...), on peut schématiser ces constats par la figure suivante où $\{C_i^j\}_i$ représente une succession de vecteurs cepstraux associée au locuteur j (Grenier,1980), (Grenier et al.,1981):



Production/Perception de la parole

Le problème de l'adaptation sera résolu si on arrive à déterminer les références $\{C_i^2\}$, associés au nouveau locuteur (2) à partir de celles associées à un locuteur standard (1). Il va de soi que nous ne connaitrons jamais - à moins de refaire un apprentissage sur le locuteur 2 - les références exactes mais uniquement une estimation de celles-ci.

Au lieu de chercher des transformations directes $C_2^i = \Phi(C_1^i)$, on se propose de chercher des transformations qui permettent de définir l'espace commun U. Pour cela on va partir d'un échantillon représentatif des espaces paramètriques C_1 et C_2, par exemple une phrase code ou un nombre limité de mots. Ensuite on va déterminer les projecteurs P_N et P_S tels que les projections soient identiques.

On se contentera dans un premier temps de transformations linéaires qui donnent des spectres projetés aussi proches que possible au sens d'un critère d'erreur. Si on choisit le critère des moindres carrés l'erreur de projection se traduit par l'équation (1):

$$J = \sum_i (u_i^1-u_i^2)^T (u_i^1-u_i^2) \qquad (1)$$

Il est facile de montrer à partir de cette équation qu'on peut minimiser l'écart entre les spectres projetés si et seulement si la corrélation entre les spectres associés est maximale, ce que réalise l'Analyse des Corrélations Canoniques (Golub,1970), en fournissant les projecteurs P_N et P_S en question (Choukri et al.,1986).

L'analyse canonique a pour but d'étudier la position relative d'un nuage de points par rapport à un autre (dans notre cas chaque nuage représentera l'échantillon d'un espace spectral d'un locuteur). Elle recherche des couples de variables, formés d'une combinaison des variables du premier nuage et d'une combinaison du second, les plus corrélés possible. Elle permet ainsi de définir un espace paramètrique où les projections de ces nuages coincident au mieux (au sens d'un critère d'erreur), qui sera alors une sorte d'espace "typologique" des deux locuteurs. On parle alors d'invariance des spectres par analyse des corrélations canoniques.

3. Procédure d'adaptation:

Pour valider notre propos on se propose d'appliquer cette méthode dans le cadre d'un système de reconnaissance de mots isolés avec un vocabulaire des dix chiffres.

Le spectre est paramétrisé avec 6 coefficients MFCC par trame. Durant la phase d'apprentissage chaque chiffre est prononcé une fois par un locuteur standard pour obtenir le dictionnaire de référence. La reconnaissance se fera grâce à des algorithmes de comparaison dynamique classiques, la détection de début et fin de mot est réalisée manuellement pour éviter toute erreur de détection pendant l'évaluation de cette méthode.

La première phase de la procédure d'adaptation consiste à acquérir et à aligner temporellement un échantillon représentatif de l'espace spectral associé à chacun des deux locuteurs. Il se pose alors le problème du choix de cet échantillon: que doit-on faire prononcer au nouveau locuteur comme "phrase code"?.

Dans une évaluation préliminaire cet échantillon sera réduit à un mot (le dixième du vocabulaire). les meilleurs mots semblent ceux qui reflètent le mieux la structure de l'espace phonétique (meilleure distribution dans le plan des premiers axes canoniques). Un logiciel d'analyse des corrélations canoniques permet alors de définir le nouvel espace de projection.

Grâce à ce logiciel on détermine la base génératrice du nouvel espace sur laquelle on projette le dictionnaire associé au locuteur standard pour obtenir le nouveau dictionnaire. On se retrouve dans le cas d'un "système monolocuteur" et on reprendra les algorithmes du système de base.

4. Evaluation:

Pour l'évaluation de cette méthode on dispose d'un corpus de 130 mots (comprenant les dix chiffres) prononcés par 100 locuteurs une seule fois. On cherche à évaluer la méthode dans le cadre d'un système monoréférence en insistant sur la variabilité inter-locuteur.

Des tests préliminaires ont pour but d'évaluer le système non-adapté en mono-locuteur croisé: le dictionnaire est obtenu grâce à un locuteur standard et on le teste sur des locuteurs choisis parmi les autres. Ensuite avec les mêmes données on évalue le système après adaptation.

Les taux de reconnaissance sont présentés en donnant les "bons" candidats qui sont reconnus en première position ou dans les deux premières positions avec l'intervalle de confiance correspondant à une probabilité d'erreur de 5%). Le taux de reconnaissance d'un système multi-référence utilisant les techniques de clustering (Syril) est de 93% en première position (Flocon et al.,1984).

Taux de reconnaissance et intervalle de confiance		
	première position	deux premières positions
non adapté	70% (68,73)	84% (81,86)
adapté	87% (84,89)	92% (91,94)

5. Conclusion:

Ce papier montre une adaptation de dictionnaires de formes à de nouveaux locuteurs. Une application à des systèmes mono-référence montre que les taux de reconnaissance sont améliorés de quelque 17%. Ce résultat reste à confirmer dans le cadre des systèmes Multi-références et de vocabulaire plus grands (130 mots).

6. Références:

Choukri, K. , Chollet, G. & Grenier, Y. (1986), Spectral transformations through canonical correlation analysis for speaker adaptation. in Proc. ICASSP, Tokyo (to be published).

Flocon, B. and Briant, N. (1984), SYRIL: système temps réel de reconnaissance de mots isolés indépendant du locuteur, 4ème congrés AFCET RFIA, Paris.

Golub, G.H. (1970), Matrix decomposition and statistical calculations. in Statistical computation, Edited by Milton, R.C. & Nelder, Y.A. (Academic press), PP. 365-397.

Grenier, Y. (1980), Speaker adaptation through canonical correlation analysis. in Proc. ICASSP, Denver, pp.888-891.

Grenier, Y., Miclet, L., Maurin, J.C. & Michel, H. (1981), Speaker adaptation for phoneme recognition. in Proc. ICASSP, Atlanta, pp.1273-1275.

GRAMMATICAL COMPONENTS AND MACRO-PROSODY : QUANTITATIVE ANALYSIS TOWARD STATISTICAL CORRELATIONS

G. Caelen-Haumont

Laboratoire C.E.R.F.I.A., UA-Cnrs N°824
Université P. Sabatier
118 Route de Narbonne 31062 Toulouse Cedex France

ABSTRACT

This study fits within the scope of the natural understanding of texts. Already known, simplified grammatical (syntactic, semantic) models of linguistic analysis have been either adapted or elaborated upon, in order to verify the hypothesis according to which there exist actual traces of abstract grammatical levels within the prosodic continuum of speech.

A per-speaker statistical file was compiled, containing both (1-syntactic, 2-semantic, 3-pragmatic) parameters issuing from the above models, and phonetico-prosodic parameters that are specific to melodic, energetic and temporal (including pauses) registers. Such a file makes it possible, if we resort to correlation analysis, to secure a quantitative appreciation of variability in the strategies adopted by speakers.

While anticipating an analysis of statistical correlations, the present article states the contents of the various analytical levels involved in the segmentation and labelling of a prosodic data-base.

1. INTRODUCTION

The problem we turn to is very aptly described by Hirst (1983) : "A deeper reason [for the elusiveness of intonation] comes from the fact that an adequate description of intonation needs to take into account not simply the phonology of the language, but also the syntax and the semantics, as well as the interfaces between the grammar and 'the real world' constituted by phonetics and pragmatics." Initially touched upon by Kellenberger (1932), this domain has since often been explored; particularly, within the last few years, in generative phonology --viz., Chomsky and Halle (1968), Liberman (1975), Liberman and Prince (1977) in the United States, and by Hirst (1983 a,b), Dell (1984), Dell and Vergnaud (1984) in France.

The present paper does not deal at all with any theoretical excercise in generative phonology; instead, as a follow up on previously published preliminary work [Caelen-Haumont 1985], it reports on a linguistic analysis (for syntactic, semantic, pragmatic and prosodic components) that was run in an experimental attempt to relate text structures to prosodic ones, by means of a prosodic data-base. The categories yielded by this linguistic analysis are used as labels in the prosodic data-base; eventually, either they are symbols (alphabetic ones) involved in the computation of various averages, or they are the addresses of event-parameters (e.g., pause duration). Therefore, the parameters involved in correlation analysis issue either from computations run at those addresses, or from numerical categories involved in labelling.

2. LINGUISTIC ANALYSIS

2.1. Text Analysis

This involves three different components.

2.1.1. Pragmatic Component

The 3 successive reading instructions determine different relationships between the linguistic signs imbedded in the text and their human users (reader to human/computer listener); a three-grade scale being thus defined on the pragmatic axis : instructions 1 through 3.

2.1.2. Syntactic Component

The model text is limited to a set of sentences without subordinate clauses. The syntactic component is limited to a morphological analysis, as well as to an analysis of the syntactic complexity.

Through morphological analysis (1st level), a phonetic item (acoustical realization phase, phoneme, syllable or word) can be identified by locating it with respect to sentence boundaries, or to boundaries of groups (this term being, here, conceived of as designating a unit that pertains to the next deeper level, beyond the surface structure). Or again, a phonetic item can be identified by locating it within these groups. A further distinction (2nd level) is made by specifying whether a word is mono- or pluri-syllabled and, in this latter case, whether a syllable is initial, final or intermediate within a word. Words with a final /ə/ are in effect no problem, since the syllable that can actually be stressed can be counted as the real final syllable; provided the subsequent consonant, or consonantic group is also counted as part of it, and the /ə/ as a post-final phoneme. A third phase of analysis involves two facets : 1/ a description of how a word appertains grammatically --i.e., whether it is a "lexical" or a "grammatical" word, sometimes referred to as a "tool" word-- and 2/ an identification of 2 constituents having specific prosodic properties (coordinating conjunction and clitic).

All these different items of information can be recombined in such a way as to suggest 18 different 2-character codes. This code is illustrated on fig. 1.

At this stage of analysis, the depth-degree of a group within the constituent structure of a sentence, is not taken into account; major and minor groups being lumped together. This reinforced type of structure analysis tackles syntactic complexity.

Unlike morphological analysis, which proceeds by means of symbolic designation of elements, the coding procedure we describe here is quantitative. As is done with the semantic-complexity analytic model, quantification of syntactic complexity is performed by means of a procedural graph.

In its present stage, the syntactic model emphasizes deep structure at the expense of surface structure : despite their actual diversity, relations among the infra-syntagmatic units that make up the group have all been given the same weight (i.e., +1).

The kind of analysis, herein described, has no claim to being exhaustive. It purports, instead, to recognize and quantify more or less complex constituents or processes of syntax; whether, in the process of either coding or decoding linguistic units, such complexity is a matter for grammatical theory or for psycho-linguistics. In any event, this complexity is to be perceived at different levels of analysis. At the level of structure , the deeper a constituent is thought to be --and subsequently the more extent the sentence-- the more weight is ascribed to it : the heaviest weight, in the sentence, being ascribed to the P-level constituent --i.e., the final one-- while the skipping rate, from one hierarchical level to the next, is taken to be equal to 1.

The syntagmatic-relation module describes relations among constituents, in three different locations : definite end of syntagma, relative end of syntagma followed by a coordinated or a subordinated syntagma (respective weights for this three situations : +3, +2, +1). Finally, the model is sensitive to constituent order, and displacement within the structure is ascribed a +2 weight. Figure 1 shows an example of syntactic-complexity quantification that is obtained through adding the module weights, described above, to eachother.

2.1.3. Semantic Component

This study also attempted to quantify the semantic complexity of the lexical items in text, by means of a new analytic model. This complexity is analysed from the point of view of any person insofar as he is considered outside his own speciality domain. This model is otherwise explained [Caelen-Haumont, 1986] and applied to textual analysis.

The model sought to describe the semantic effect, not the means of achieving it. In this matter, although they participate to the elaboration of meaning, the syntactic structuration processes have not been made explicit. The actual application range of this model is not the sentence but the text. The method, used, assumes both the intra- and inter-lexical components to be textwide dimensions; two dialectical poles in between which meaning is generated, in the course either of writing, reading, listening, or of analyzing the text for meaning. The analytic model consists of three modules :

1- intra-lexical analytic module :

a/ lexical-item register : fundamental, standard or specialized but vulgarized, specialized (respective weights : +1, +4 and +7)

b/ referent : concrete, concrete/abstract for items with two different acceptations (e.g., "combination"), abstract or imaginary (weights 0, +2, +4).

c/ specifying an essence : 1/ "state" or spatial notion of structure, 2/ relational link between concrete or abstract objects, 3/ "process" or temporal notion of evolution, 4/ combination of both (example: the lexeme "addition") with respective weights : 0, +1, +1, +1. These notions are independent from syntactic categories.

d/ designating of something in nature: "substance" or nature of the designated object and "attribute," quality of the latter. In turn, substance is subdivided into either spatial or temporal type categories (example : perfective vs. imperfective for "process"); these two notions possibly neutralizing eachother or combining together.

The "attribute" category covers the distinction between intrinsic and extrinsic attribute, and it applies to both types of substance, contemplated in their own pecularity.

At the outcome of this analytic level, quantification is obtained through repeatingly adding 0 or +1 weights.

2- transition module

This causes a lexeme to change category according to context; it either simplifies or complexifies (respective weights : +1 and +2, example : abstract to concrete (+1)).

3- inter-lexical analytic module

It encompasses various lexical networks both of form and of content. Form : repeating the term commands either -1 or -3 weight, depending on its register, as defined above. Content :

a/ use in the figurate possible (no figurate or cliché, lexicalized figurate, living figurate --respective weights : 0, +3, +5).

b/ occurrence of a lexical field (belonging to the field or initiating it, changing field, weights : 0, +2).

At the outcome of the procedural graph, each lexical item is given a weight (in the range : 1 to 25) which is held to be a quantitative (though subjective) assessment of its complexity of meaning and, followingly, part of the complexity of meaning of the whole text. Example on figure 1.

2.2. Prosodic Analysis

2.2.1. Phonetic Aspect

Concerning this aspect, two dimensions are considered : the phonemic and the infra-phonemic.

On the phonemic level, 43 labels are made available; beside the pause, various allophones. On the infra-phonemic level, the notions of realization phase and of "intonemes" are combined to yield 9 one-character codes. These structure up the phonemic space that has already been pre-segmented into "phones" (see 1., above); on the one hand, in terms of realization phases --set-in, sustained, caudal-- based on acoustic-cue behavior and, on the other hand, in terms of beginning and end of specific intonemes, spotted on the melody curve. In the present work, only continuity-intonemes have been retained and, for the sake of generalization, both maxima and minima of all final vowels of lexical words (as well as adjacent phonemes within the syllable, whenever necessary) have been coded, even in the case of weak or zero tonal variations.

2.2.2. Prosodic Relief-Map

The tonic-stress structure is analyzed according to the traditional key-points, based on position and quantity criteria : onset, pre-tonic, tonic and post-tonic vowels. With an aim to testing the influence of stressed-vowel position upon prosodic quantity (cf. notion of metrical structure in generative phonology), both types of vowels located between attack and stress have been numerically coded in decreasing order, down to the pre-tonic --coded 1. An illustration of phonetic labelling (phonemic, infra-phonemic and prosodical levels) is given figure 1.

3. CONCLUSION

The syntactic, semantic, pragmatic and prosodic components supplied a set of alphabetic and numerical labels. These were used to code the linguistic units (infra-phonemic items to sentences) or events of a prosodic data-base. A base containing prosodic data was set up on LSI 11-73 from a corpus handled as follows : 10 speakers reading a 45-word text, under 3 successive, increasingly demanding sets of instructions --i.e., 1/ natural and intelligible reading, 2/ very intelligible reading, and 3/ very very intelligible reading for the computer. This made for 30 uttered texts. Once segmented and labeled the 30 data-files were fed into other stastitical files that were set up through automated extraction of parameters deemed relevant --e.g., items of syntactic complexity, pragmatic situations, prosodic values (Fo, energy, duration) at certain points of the statement that are localized through the linguistic item addresses. By facilitating various types of data-analysis --e.g., of correlations [Caelen et alii,1985 a,b]-- this prosodic data-base opens up a possibility of working on the verification of various hypotheses concerning the presence, within speech and more specifically within loud reading, of grammatical-structure cues of a syntactic, semantic and pragmatic type.

Acknowledgment: I wish to thank J.F. Malet (of California State University, Sacramento) for meticulously translating this text from French.

BIBLIOGRAPHIC REFERENCES

[Caelen-Haumont,85] G. Caelen-Haumont. Dialogue homme /machine et aspects temporels des stratégies de locuteurs de type phonétique, syntaxique et sémantique,, Paris, 1985, pp. 183-187.

[Caelen-Haumont,85] G. Caelen-Haumont, Propositions

pour un modèle d'analyse sémantique simplifiée de la complexité des signifiés, Galf-Cnrs XVèmes JEP Proceedings, Paris, 1986.
[Caelen et alii, 85] J. Caelen et N. Vigouroux, A multi-level acoustic and phonetic base : from facts to knowledge, Symposium Franco-suédois Proceedings, Grenoble, 1985.
[Caelen et alii, 85] J. Caelen et N. Vigouroux, Segmentation automatique de la parole, Galf-Cnrs XIVèmes JEP Proceedings, Paris, 1985, pp. 152-155.
[Chomsky et alii, 68] N. Chomsky et M. Halle, The sound pattern of English, Harper and Row, New-York, 1968.
[Dell et alii, 84] F. Dell, D.J. Hirst, J-R. Vergnaud, La forme sonore du langage : la nature des représentations en phonologie, Hermann, Paris, 1984, p.95-116.
[Hirst, 83] D. Hirst, Structures and Categories in Prosodic Representations, in A. Cutler & D.R. Ladd (eds), Prosody : Models and Measurements, Springer, Berlin, 1983, pp. 93-109.
[Hirst, 83] D. Hirst, Interpreting intonation : a modular approach, Journal of Semantics 2: 2, 1983, pp. 171-181.
[Kellenberger, 32] H. Kellenberger, The influence of Accentuation on French Word Order, Princeton, 1932.
[Liberman,75] M. Liberman, The Intonational system of English, PH. D., Massachussetts Institute of Technology, distribué par Indiana University Linguistics Club, Bloomington, Indiana, 1975.
[Liberman et alii,77] M. Liberman and A. Prince, On stress and linguistic rhythm, Linguistic Inquiry 8.2, 1977, pp. 249-336.

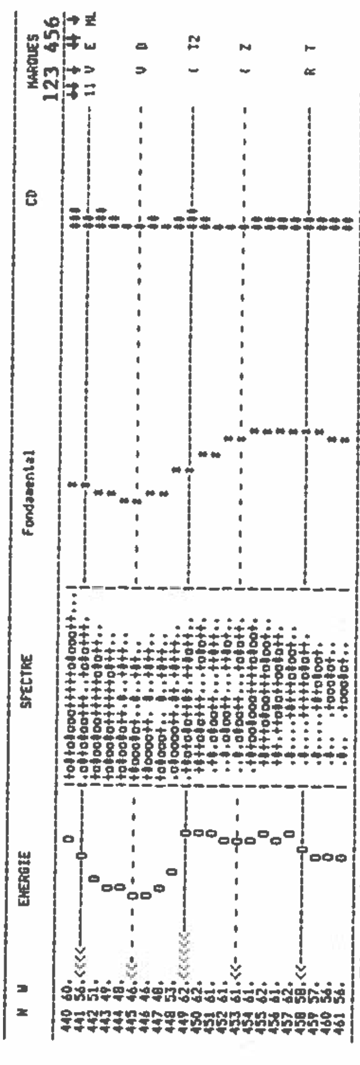


Figure 1 : An illustration of the prosodic data-base labels

1 : Semantic complexity quantification of lexical items

2 : Syntactic complexity quantification of lexical items

3 : Phonemic labels
V —> /v/
(—> / /
R —> /R/

4 : Acoustical realization phase
+ Intoneme
E set-in phase
T sustained phase
Q caudal phase
D beginning of intoneme on Q
Z end of intoneme on Q

5 : tonic-stressed structure
2 second vowel before the tonic vowel

6 : sentence morphological structure
ML monosyllabled lexical word within a syntagma

ORGANIZATION OF PHONEMIC SPACE REPRESENTED BY THE UNITS OF SPECTRA AND SPECTRAL CHANGES

Katsuhiko Shirai and Kazunori Mano

Department of Electrical Engineering, Waseda University
3-4-1, Ohkubo, Shinjuku-ku, Tokyo 160, JAPAN

ABSTRACT

This paper describes a method of organization of phonemic space for phoneme recognition. Phonemic space is obtained by clustering speech spectra and spectral changes. Power change, LPC cepstral coefficients and the differences of LPC cepstral coefficients are used to represent the characteristics of the spectral contour and spectral change. The efficiency is shown by an experiment of phoneme recognition.

INTRODUCTION

There are many factors which make it difficult to extract phonemic features precisely. Some of the factors are as follows.

(1) In continuous speech, boundaries between adjacent phonemes are uncertain and it is difficult to segment correctly.

(2) There are many variations in phoneme patterns.

(3) As the characteristics of phonemes exist not only in spectral contours but also in spectral changes, both static and dynamic properties in speech signals must be considered as acoustic features.

Vector quantization (VQ) method is an efficient method to encode speech signals[1]. We have used the VQ technique as a clustering method to extract phonemic features frame by frame[2][3]. In this paper, an organization of phonemic spaces with a VQ technique is discussed and we consider the relation between acoustic features represented by VQ codes and their phonemic features which belong to the clusters of the VQ codes.

REPRESENTATION OF ACOUSTIC FEATURES AND PHONEMIC FEATURES FOR CLUSTERING

Acoustic Features

Acoustic features defined in each frame are LPC cepstral coefficients called Level 1 feature, changes of LPC cepstral coefficients called Level 2 feature and power change. The Level 1 feature is calculated in a frame and denoted by the following.

Level 1 feature : $(C1(1),\ldots,C1(n))$,

where n is the order of LPC analysis. The Level 2 feature and the power change are defined as the differences between the parameters in the first half and the second half of the frame. If the LPC cepstral coefficients in the first half and the second half are denoted by $(C21(1),\ldots,C22(n))$ and $(C22(1),\ldots,C22(n))$ and the powers P1 and P2, the Level 2 feature and the power change in the frame are defined as follows.

Level 2 feature : $(\Delta C2(1),\ldots,\Delta C2(n))$,

where

$$\Delta C2(i) = C21(i)-C22(i).\ (i=1,\ldots,n)$$

Power change :

$$\Delta P = (P2 - P1) / P1$$

The Level 1 feature shows a spectral contour which represent a static property in a frame. The Level 2 feature corresponds to the change of the spectrum. This feature is efficient to describe the precise movements of spectrum in a frame, especially in transient parts of speech such as consonant-to-vowel(CV) sounds. The power change shows a global changes such as the change from silence or unvoiced sound to voiced one.

Phonemic Features

A label called a frame label which is composed of three phonemic symbols is assigned to each frame by visual inspection before clustering. For example, if a frame belongs to a transient part of speech /.pa/, where /./ means silence, the frame labels such as /..p/, /.pp/, /ppa/, /paa/ or /aaa/ are sequentially yielded according to the position of the frame. The frame label of /..p/, means that the frame contains silence /./, in more than half part of the frame and a sound of /p/ is following the silence in the frame. The /aaa/ means the frame exists only in vowel part, that is, the frame is almost stationary.

CLUSTERING METHOD BASED ON VQ ALGORITHM

Phonemic features are related to acoustic features by clustering. The main reason of using clustering method is that it makes the speech frames into groups which have both acoustically and phonemically similar properties. Each frame is characterized by code numbers of the produced cluster and the frame labels in the cluster.

As for the clustering, vector quantizer design method which is a slightly modified one proposed by Linde, Buzo and Gray[1] is adopted. The modified points are that the centroids to be split are determined by considering kinds of the frame labels for effective distributions of centroids. That is, more centroids are assigned to the clusters which have a lot of kinds of frame labels and less centroids to the clusters which have only one or two frame labels. By this modification, the quasi-optimality of the VQ method is not kept any more, but it is more useful to extract phonemic features.

For example, if a cluster has the frames which have the same frame labels, the centroid of the cluster is not split in the preceding procedure because the phonemic features of the cluster is sufficiently represented by the frame label. Such clusters appear in stationary parts. On the other hand, if a cluster has various kinds of frame labels, the phonemic features in the domain of the cluster are not described by the centroid and it means that more centroids are necessary to obtain phonemically unified clusters. Such clusters mainly exists in transient parts.

ORGANIZATION OF PHONEMIC SPACE

The above clustering method is applied to each set of frames to organize phonemic space.

Before clustering, all the speech frames are classified into three parts called the ascending, flat and descending parts by the degree of power change _ P in each frame. The ascending part contains such sounds like consonant-to-vowel, silent-to-consonant or vowel-to-stronger vowel. The flat part contains almost stationary parts of vowels, nasals and fricative consonants. In the descending part, the sounds such as vowel-to-consonant, vowel-to-silence or vowel-to-weaker vowel are contained. By this pre-classification, it is possible to avoid grouping of the frames which have entirely different frame labels, even in the case that the acoustic distortion between the frames is small.

Clustering is performed in each part of the three parts with Level 1 features and Level 2 features, respectively and six codebooks composed of centroid vectors and sets of frame labels are produced. The phonemic space is organized by the distributions of the centroids and the frame labels which belong to the corresponding cluster in each part and each level.

EXPERIMENT OF PHONEME RECOGNITION

For an evaluation of the phonemic space which is represented by codebooks and frame label sets, an experiment of phoneme recognition is carried out. Figure 1 shows the diagram of extracting phonemic features. When a frame is analyzed, the power change is calculated and one of the part number of the power change is assigned to the frame and according to the LPC cepstrum and the cepstral differences, codes of Level 1 and 2 are given to the frame. Output of the frame labels is obtained from the intersection of the sets of frame labels in Level 1 and 2. By symbolic processing the sequences of the frame labels, phoneme sequences are produced.

The result of the cumulative recognition rates of phonemes for one male speaker is shown in Figure 2. In the experiment, the codebooks are generated from 800 syllables and 100 city names are used for the recognition. The sampling frequency is 12.5[kHz]. The frame length in Level 1 is 32[ms] and the interval of analysis is 16[ms]. The number of VQ codes in each part is about 256. In Fig.2 the phoneme recognition rates are about 91% in vowel sounds and 73% in consonants in the first candidate. Within 3 candidates, the rates increase to 99% in vowels and 89% in consonants.

CONCLUSION

A method to make phonemic space base on the spectrum and the spectral difference was proposed. The efficiency of this method was evaluated.

REFERENCES

[1] Linde, Y., Buzo, A. and Gray, R. M., "An algorithm for Vector Quantizer Design," IEEE Trans. Commn., Vol.COM-28, No.1, pp.84-95, January 1980.

[2] Shirai, K. and Mano, K., "A Clustering Experiment of the Spectra and the Spectral Changes of Speech to Extract Phonemic Features", Signal Processing, Vol.10, No.3, April 1986.

[3] Shirai, K. and Mano, K., "Feature Extraction of Phonemes by clustering the spectra and the spectral changes in continuous speech. Proc. of IASTED Inter. Symp. of Applied Signal Processing and Digital Filtering, pp.201-204, June 1985.

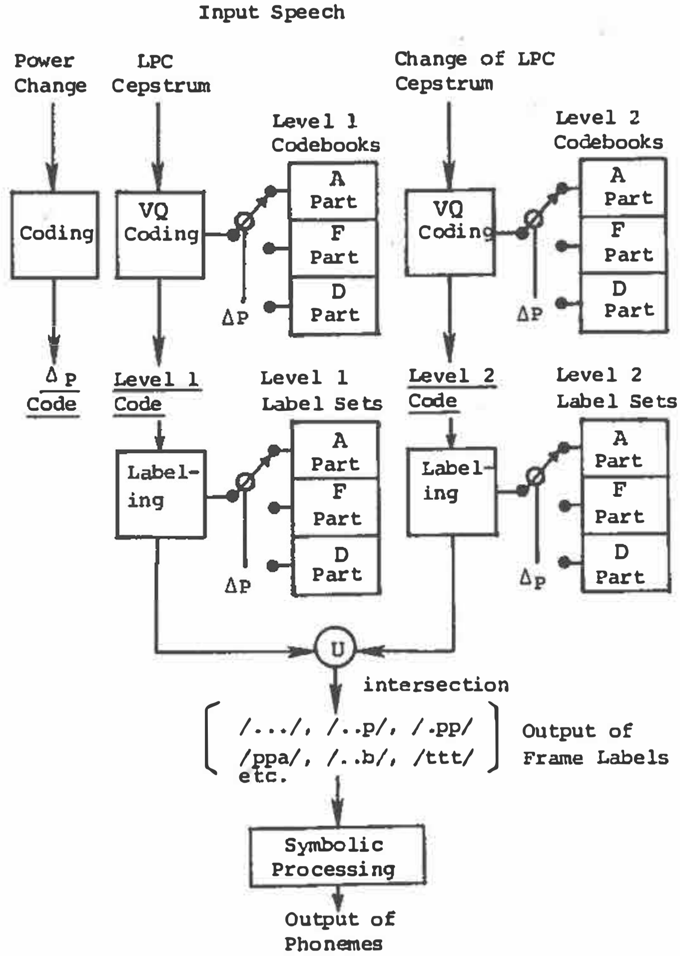


Fig.1 Diagram of Extracting Phonemic Features

A Part : Ascending Part
F Part : Flat Part
D Part : Descending Part

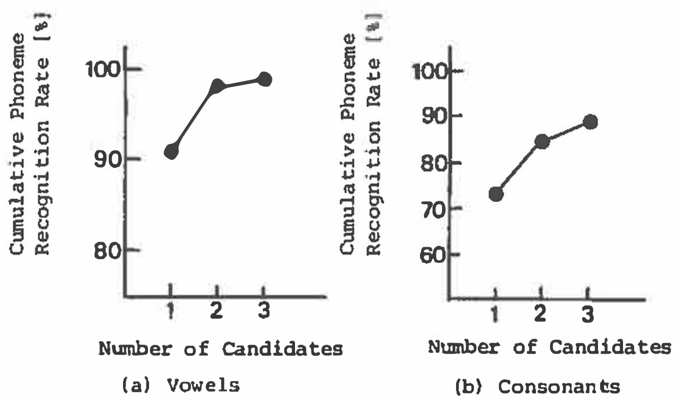


Fig.2 Cumulative Recognition Rate of Phonemes

SPEECH RECOGNITION BY USE OF WORD DICTIONARY WRITTEN IN LINGUISTIC UNIT

Ken'iti KIDO, Shozo MAKINO, Michio OKADA, Satoshi MORIAI and Tetsuo KOSAKA

Research Center for Applied Information Sciences
Tohoku University, Katahira, Sendai 980, JAPAN

INTRODUCTION

We have carried out the researches on speaker independent recognition of words[1] by use of word dictionary which is composed of the sequences of phonemic symbols. The phonemic symbols are derived from linguistic representation of Japanese language. In the system, the spoken word is transformed into the sequence of phonemic symbols and the item of the word dictionary most similar to the input sequence is chosen as the recognition output. That is, the system uses the phoneme as the linguistic unit for the recognition of word.

SPEECH RECOGNITION SCHEME

The unit in speech recognition can be classified into two groups: one is based on articulatory model and the second one is not so. The purely acoustical units and the units which refer to the characteristics of auditory organ belong the second group. And the size of unit is also divided into several groups: least one is the segment of speech which is the minimum unit to express word or speech and the maximum one is the word. Figure 1 shows the hierarchical relation between those units. The thick lines between two boxes in Fig. 1 denotes the relations which are considered to be important but difficult to formulate.

Figure 2 shows the schematic diagram of speaker independent spoken word recognition system we have developed. In the system, the recognition is to find out the item of word dictionary which corresponds to the input speech. And the system is equipped with word dictionary which contains all the words to be recognized.

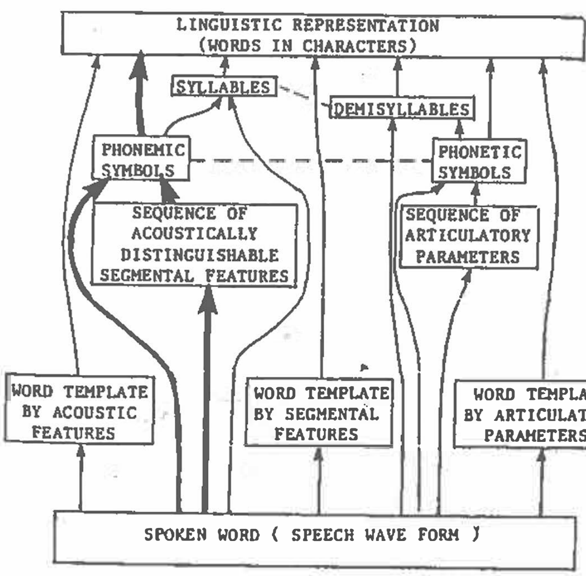


Fig. 1 Hierarchy in the unit of representation of spoken words for speech recognition

In the system, the input speech is transformed into a sequence of phonemic symbols. And the similarity of the content of word dictionary to the input speech is computed for every item. The recognition output is the dictionary item of maximum similarity.

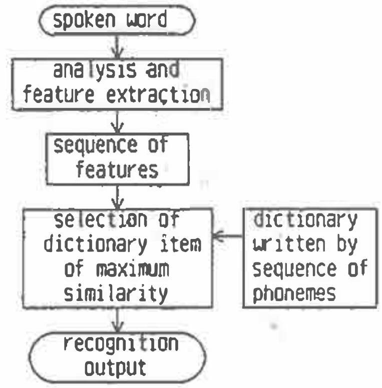


Fig. 2 Schematic diagram of the spoken word recognition system

PHONEME AS LINGUISTIC UNIT

The most important problem in such the system is how to describe the contents of word dictionary. If the contents are described by phonemic symbols, it may be very simple to make the word dictionary especially in Japanese as all the Japanese words are in the form "CVCVCV..." where C denotes the consonant and V the vowel. But the transformation of speech into the sequence of phonemic symbols is not easy because the acoustic characteristics of speech segment does not always correspond to the phonemic symbol which are derived from the linguistic representation.

If the contents are the standard patterns composed of acoustic features directly obtained by analyzing the spoken words, it would be easy to transform the input speech to the patterns for the comparison with the standard patterns. But, a lot of computation is necessary for making the standard patterns common to all the possible speakers especially in the case of large vocabulary.[2]

And there is intermediate system[3] in which the word dictionary is composed of the sequences of acoustic features which are defined by classifying the words uttered by a number of speakers. The classification is based on the differences in acoustic characteristics of speech segments. Such the features may be able to express the acoustic characteristics of words more exactly than the phonemic symbols. The phonetic transcription may be exactly carried out using such the features, and we call the features as the phonetic features in this paper. The transformation of the input speech into the sequence of the phonetic features may be easier than the transformation into the sequence of phonemic symbols. But, a lot of computation and a number of speech samples will be necessary for making the word dictionary composed of such the phonetic features and it may be difficult problem to compose a set of phonetic features which can be used for many vocabulary regardless of speakers.

Therefore, we have used the phonemic symbols for the description of dictionary items and now we are trying to use the acoustic features of segments to derive the sequence of phonemic symbols.

CONVERSION OF SPEECH INTO PHONEMIC SYMBOLS

The input speech is passed through a 29 channel band pass filter bank which is composed of single tuned circuit of Q=6 and the center frequencies are at every 1/6 octave between 250 Hz and 6 300 Hz. The power of every channel is computed for every frame of 10 ms duration and logarithmically transformed.

Eight features are extracted by using the discriminant filters which are designed by use of speech samples of 212 words uttered by 10 male and 10 female speakers. Figure 3 shows the examples of the solution weight vectors for eight discriminant functions. Another feature is the logarithmic spectrum summation which is the sum of logarithmic power of all the channels.

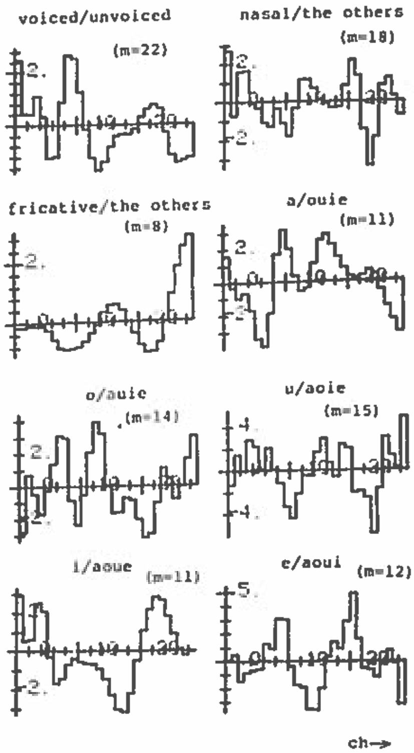


Fig. 3 Examples of solution weight vectors for extracting the eight features

Table 1 Discriminant functions

function X_1/X_2	X_1	X_2
voiced/ unvoiced	/a/,/o/,/u/,/i/, /e/,/j/,/w/,/m/, /n/,/n/,/N/,/b/, /d/,/g/,/r/,/z/	/h/,/s/,/c/ /p/,/t/,/k/
nasal/ the others	/m/,/n/,/n/,/N/	/a/,/o/,/u/,/i/,/e/ /j/,/w/,/b/,/d/,/g/ /r/,/z/,/h/,/s/,/c/ /p/,/t/,/k/
fricative/ the others	/z/,/h/,/s/,/c/	/a/,/o/,/u/,/i/,/e/ /j/,/w/,/m/,/n/,/n/ /N/,/b/,/d/,/g/,/r/ /p/,/t/,/k/
a/ouie	/a/	/o/,/u/,/i/,/e/
o/auie	/o/	/a/,/u/,/i/,/e/
u/aoie	/u/	/a/,/o/,/i/,/e/
i/aoue	/i/	/a/,/o/,/u/,/e/
e/aoui	/e/	/a/,/o/,/u/,/i/

5 vowels, 2 semi vowels, 15 consonants

The functions of the discriminant filters for the eight features are listed in Tab. 1. The phoneme boundaries are assumed to be the frame where the weighted sum of absolute values of the first order time-derivatives of the features takes maximum value exceeding a threshold. The frames of unvoiced and voiced plosives are detected using the discriminant filters. The primary phoneme recognition is carried out for every assumed segment using the outputs of discriminant filters and the standard patterns for phonemes which are made using the 212 spoken words.

After correcting errors by the errorcorrection rules, the secondary phoneme recognition is carried out. Here, the nasals and the voiced and unvoiced plosives are recognized.

WORD RECOGNITION USING LINGUISTIC UNIT

In the word recognition part, a number of sub-items are generated referring to the confusion matrices of phoneme recognition for initial-, mid- and final positions of words. The confusion matrices includes the probabilities of insertion, omission and substitution of phoneme. The computation of similarity between the phonemic sequence with top three recognition results and each sub-item follows. The dynamic programming algorithm is used to reduce the time for similarity computation.

The dictionary item having maximum similarity to the input sequence is chosen as the recognition output.

RECOGNITION EXPERIMENTS

Word recognition experiments were carried out using the speech samples used to design the discriminant functions, standard patterns and confusion matrices and the same 212 words uttered by the other 30 males and 20 females. Table 2 shows the summary of the results.

Table 2 Word recognition score

Training set	10 males	93.7%	Average
	10 females	91.3%	92.4%
Test set	30 males	87.0%	Average
	20 females	89.6%	88.1%

CONCLUSION

This paper describes the use of phoneme as the linguistic unit of speech in the spoken word recognition system for a large vocabulary. In the system, the phoneme recognition is first carried out and the word dictionary item with the maximum similarity to the sequence of recognized phonemes is chosen as the recognition output. The score of word recognition is 92.4% in the experiment which is much higher than that of phoneme recognition (75.9%) due to the utilization of word dictionary as the linguistic information source. Studies on phoneme recognition is now continued to improve the word recognition. The vocabulary to be recognized can easily be altered and expanded by changing the dictionary item from key board.

LITERATURES

1) H.Suzuki,S.Itahashi and K.Kido: TheEffectiveness of Utilization of Word Lexicon in Recognition of Japanese Spoken Language, Proc.1967 Conference on Speech Communication and Processing,B12,p.128(1967)
2) S.Chiba: Recognition of Spoken Words,J.Info.Proc. Jpn.19-7(1978)
3)N.Sugamura,S.Furui: Large Size Vocabulary Spoken Word Recognition by Use of Pseudo-phoneme Standard Patterns,J.IECE.Jpn,65-D,8(1982)

DURATIONAL CONSTRAINTS FOR NETWORK-BASED CONNECTED DIGIT RECOGNITION

Marcia A. Bush

Schlumberger Palo Alto Research, 3340 Hillview Ave., Palo Alto, CA 94304, USA. [1]

This paper examines the influence of durational constraints on recognition accuracy in an acoustic-phonetically based, speaker-independent connected digit recognizer. The constraints are expressed using a set of finite-state pronunciation networks, together with specifications of minimum and maximum allowable durations for network primitives. The recognizer was tested on a corpus of 1232 5-digit and 7-digit strings, with and without a priori knowledge of string length. Recognition accuracies ranged from 33.9% to 94.6% and from 91.6% to 96.8%, for unknown and known string lengths, respectively, depending on the particular durational constraints incorporated in the network models.

INTRODUCTION

The word models used in the connected digit recognizer described here consist of a set of finite-state pronunciation networks, in which primitive branches correspond to meaningful acoustic-phonetic units (Table 1). Unlike networks based on the hidden Markov model formalism, these word models allow for the convenient expression of acoustic-phonetic constraints which are manifest over portions of an utterance longer than a single time frame. One example of such a constraint is segment duration. [2]

This paper examines recognizer performance as a function of the minimum and maximum allowable durations for primitives in two types of network: 1) a baseline network formed by simply connecting in parallel the isolated digit models shown in Table 1; and 2) a set of networks which incorporate additional paths representing prepausal lengthening for the digits *oh* and *eight*. Constraints on minimum duration were found to have the greatest influence on recognition accuracy, particularly when recognition was performed without a priori knowledge of digit string length. Prepausal durational constraints proved useful in reducing a common class of digit insertion errors.

The digit recognizer incorporates a set of generalized acoustic pattern matchers and a dynamic programming search in addition to the pronunciation network models. Details of the recognition framework, and of signal preprocessing, are provided in [1] and [2].

CORPUS

The corpus used in the recognition experiments consisted of the adult-talker, 5-digit and 7-digit subset of the training portion of Texas Instruments' multi-dialect connected digits database [3]. The utterances of half of the talkers (27M, 29F, 1232 tokens) in this subset were used for training the recognizer and the utterances of the remaining half (28M, 28F, 1232 tokens) were used for testing. These two corpora will be referred to as TRNA-57 and TRNB-57 respectively.

An initial version of the recognition system was trained on 616 handlabelled 5-digit strings from TRNA-57, and run over the entire TRNA-57 corpus [2]. The segmentations generated for correctly identified tokens in this experiment defined a set of bootstrapped training data, which were used in all of the experiments reported here. Statistics on minimum and maximum segment duration were collected for both the handmarked and bootstrapped data, and used in specifying the durational constraints in the network models.

RESULTS

Table 2 shows recognition data for corpus TRNB-57 using the baseline network (unknown string length) and various constraints on segment duration. As indicated in the first three columns, recognition accuracy ranges from 33.9% when the minimum allowable duration is a single frame (10 msec), as in first-order hidden Markov models, to 93.2% using the minimum durations for the bootstrapped training data. During the bootstrapping experiment, very short durations (i.e., those falling in the bottom 5% of the distributions for each segment type) were penalized, with the result that minimum durations for the bootstrapped training data were typically 1 to 2 frames longer than for the handmarked utterances. The main effect of prohibiting very brief segments is to reduce the number of digit insertion errors from 1407 to 33.

Not surprisingly, constraints on mimimum segment duration have a less dramatic effect on recognizer performance when string length is known a priori. As shown in the first two columns of Table 3, recognition accuracy increases from 91.6% with minimum allowable durations of a single frame to 96.8% using the bootstrapped minima.

In the experiments just described, the maximum allowable segment duration was 1.5 times that observed for the bootstrapped data. Comparison of columns 3 and 4 in Table 2, and of columns 2 and 3 in Table 3 indicate that imposing tighter constraints on maximum segment duration (i.e., the bootstrapped maxima) has virtually no effect on recognition accuracy with the baseline network.

Table 4 shows recognition data for corpus TRNB-57 using networks which require prepausal lengthening for the digit *oh* (column 1) or for both *oh* and *eight* (columns 2-4). These networks were motivated by the observation that the most consistent errors using the baseline network were *oh* and *eight* insertions following the third digit of a 7-digit string. (Presumably, talkers used a "telephone number" grouping in producing these tokens.) *Oh*'s were most often inserted after the digits *oh*, *two* and *zero*, and *eight*'s after *two*, *three* and *eight*. Prepausal lengthening was required for each of the vocalic segments in the two digits, with the degree of lengthening estimated from the two sets of training data.

Incorporating prepausal lengthening for the digit *oh* serves to reduce the number of *oh* insertions from 19 to 10 relative to the baseline situation (Table 5, columns 1 and 2), and to increase overall recognition accuracy from 93.0% to 93.8% (Table 2, column 4, and Table 4, column 1.) Adding

[1] After 1 Aug 86: Division of Engineering, Box D, Brown University, Providence, RI 02912, USA.

[2] As used in this paper, the term *segment* refers to the acoustic-phonetic primitives listed in Table 1.

prepausal lengthening for *eight* reduces the number of *eight* insertions from 17 to 11 (columns 2 and 3, Table 5) and increases overall accuracy to 94.2% (Table 4, column 2).

Virtually all of the prepausal *oh* and *eight* insertions which remain after these two network modifications occur following the digits *two* and *three*. Several of these errors can be eliminated by increasing the maximum allowable durations for the vocalic portions of *two* and *three* from 1.0 to 1.5 times their bootstrapped values (Table 5, column 4), increasing overall recognition accuracy to 94.6% (Table 4, column 3). (Additional *eight* insertions can be eliminated by allowing a noisy or breathy "release" segment after these same two digits.) Allowing looser maximum durational constraints for all segments results in a small decrease in recognizer performance (Table 4, column 4), in contrast to experiments with the baseline network.

SUMMARY

The experiments described above illustrate the importance of appropriate durational constraints for high-accuracy network-based connected digit recognition. Modeling duration in the current system is facilitated by the use of network primitives corresponding to meaningful acoustic-phonetic units.

REFERENCES

[1] M. Bush and G.Kopec, "Network-based connected digit recognition using explicit acoustic-phonetic modeling", in *Proceedings, 1986 IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, Japan (Apr 1986).

[2] M. Bush and G.Kopec, "Network-based connected digit recognition", submitted for publication in *IEEE Transactions on Acoustics, Speech and Signal Processing* (Mar 1986).

[3] G. Leonard, "A database for speaker-independent digit recognition", in *Proceedings, 1984 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA (Mar 1984).

Digit	Network Primitives
oh	OW1 OW2 OW3
1	WAH1 WAH2 N
2	(TS) TR UW1 UW2
3	TH RIY1 RIY2
4	F AOR1 AOR2
5	F AY1 AY2 V
6	S IH KS KRS
7	S EH V AX N
8	EY1 EY2 (TS) (TR)
9	NI AY1 AY2 NF
zero	Z IYR1 IYR2 ROW1 ROW2

Table 1: Network primitives for the baseline pronunciation network. Parentheses indicate optional segments.

Segment Duration: Minimum Maximum	1 frame 1.5xBS	HM 1.5xBS	BS 1.5xBS	BS BS
% correct	33.9	86.2	93.2	93.0
string length errors	800	118	47	49
matches	7270	7306	7328	7338
substitutions	121	79	49	47
insertions	1407	124	33	43
deletions	1	7	15	7

Table 2: Recognition data for corpus TRNB-57 using the baseline network and various constraints on segment duration. Unknown string length. HM = handmarked TRNA-5, BS = bootstrapped TRNA-57.

Segment Duration: Minimum Maximum	1 frame 1.5xBS	BS 1.5xBS	BS BS
% correct	91.6	96.8	96.8
string length errors	–	–	–
matches	7257	7350	7349
substitutions	122	42	43
insertions	13	0	0
deletions	13	0	0

Table 3: Recognition data for corpus TRNB-57 using the baseline network and various constraints on segment duration. Known string length. BS = bootstrapped TRNA-57.

Segment Duration	BS, except as noted			
Prepausal Minimum Lengthening	OW 1.75xBS	OW - 1.75xBS EY - 1.5xBS		
Maximum Lengthening	None	None	UW,RIY 1.5xBS	All 1.5xBS
% correct	93.8	94.2	94.6	93.8
string length errors	40	34	30	39
matches	7338	7338	7338	7328
substitutions	47	47	47	49
insertions	34	28	24	25
deletions	7	7	7	15

Table 4: Recognition data for corpus TRNB-57 using networks with prepausal lengthening and various constraints on segment duration. Unknown string length. BS = bootstrapped TRNA-57.

Segment Duration	BS, except as noted			
Prepausal Minimum Lengthening	None	OW 1.75xBS	OW - 1.75xBS EY - 1.5xBS	
Maximum Lengthening	None	None	None	UW,RIY 1.5xBS
oh	19	10	10	7
8	17	17	11	10

Table 5: *Oh* and *eight* insertion errors for corpus TRNB-57 for various networks and constraints on segment duration. Unknown string length. BS = bootstrapped TRNA-57.

SPEECH RECOGNITION BASED UPON A SEGMENT CLASSIFICATION AND LABELLING TECHNIQUE AND HIDDEN MARKOV MODEL

W. A. Mahmoud and L. A. M. Bennett

Department of Electrical and Electronic Engineering
University College of Swansea,
Swansea SA2 8PP U.K.

1 Abstract

A new structure for isolated-word speech recognition via vector quantisation (VQ) is described, namely the segment classification and labelling technique (SCLT). The proposed recognizer requires the generation of separate codebooks for the acoustically dissimilar events and then the merging of them to produce a single reference codebook. Three major acoustic events were considered, namely voiced, unvoiced and silence (V/U/S). The results show that the proposed structure has the capability of reducing the degradation of VQ in speech recognition and provides a better set of observations for the hidden Markov model (HMM).

2 Introduction

Two very important speech modelling techniques have been applied to speech recognition. They are vector quantisation (VQ) of the linear predictive coding (LPC), which is used for representing the short-term spectral characteristics of speech, and the Hidden Markov Model (HMM), which can be used for representing the long-term statistical characteristics of speech. The VQ generates an ordered set of reference codewords, referred to as the codebook, which represents a partitioning of the acoustic space in the domain of the speech being quantized. The HMM treats any speech utterance as a sequence of random observations generated according to a particular underlying law of the HMM. The underlying law is estimated in the form of the generation of a given utterance from a given set of observations by making a maximum likelihood estimation. The random observations can be in various forms, one of which is quantised LPC vectors.

While enjoying certain advantages, however, VQ has the drawback of reducing recognition accuracy. Recently the authors successfully proposed a method for effectively reducing this degradation called the Segment Classification and Labelling Technique (SCLT) [1]. The SCLT classifies the training data into three classes; voiced, unvoiced or silence. Then it generates a separate codebook for each of these classes before producing a single reference codebook. It is interesting to use these codebooks as the random observations for HMM. Various codebook sizes (16,32,64,128 and 256) have been used for quantizing the LPC vectors and testing the performance of our systems. The performance is also compared with both VQ/DTW and SCLT/DTW alpha-numeric recognition systems which share the same LPC quantizer and testing data.

3 The Segment Classification and Labelling Technique (SCLT)

In the first step of the SCLT, the training speech sequence is required to be classified into three major classes namely, V/U/S. In the second step, the separate data of each class is used to generate the corresponding codebook using the VQ algorithm [1]. In the final stage of this technique a reference codebook of desired size will be formed from the three separate codebooks following a combination (merging) criterion.

A novel approach for detecting the VUS classes was used, in which a spectral characterization of each of these signals was obtained during clustering of the training data, using a K-mean algorithm similar to the VQ algorithm. This method of classification was used for the following reasons: (1) Since it uses the VQ algorithm it does not need to implement a new algorithm for the application under consideration. (2) It does not require the calculation of any other feature other than that used in the analysis of the application. (3) It gives an acceptable discrimination accuracy.

Since the aim was to apply the SCLT to speech recognition, then the table look-up method used here necessitated the need for a criterion for merging these codebooks. Thus such a combination criterion should result in a single reference codebook, so allowing the calculation of the distance of the matrix for its codewords in the usual way of VQ. In such a criterion, codebooks of different codeword counts were combined to form the desired reference codebook. The question that arises now is how to make the most efficient use of this combination. From the actual counts, it was observed experimentally that the number of voiced vectors was approximately twice that of each of unvoiced and silence. This population of voiced vectors satisfied the principle of giving a higher representation for them in the reference codebooks. Therefore, in the following tests a voiced codebook of twice the size of the unvoiced and silence was attempted. Thus to form a reference codebook of size 64, a voiced codebook of size 32 was combined with a codebook of 16 unvoiced codewords and a codebook of 16 codewords of silence.

4 The Hidden Markov Model (HMM)

The idea of representing speech events by HMM's has been used in several speech processing systems. In the HMM we assume that each word model has N-states (where N=5 is used here) and is characterised by a state-transition matrix A and a symbol-probability matrix B. The model parameters (i.e. A and B elements) are estimated from a training sequence of two versions of the vocabulary for each speaker and used to calculate the probability of the observation set given a particular model M. Re-estimation formula due to Baum-Welch was used to iteratively adjust the A's and B's elements until the probability of the observation sequences conditioned on the parameter values stopped increasing significantly, or when some other stopping criterion is met (e.g. the number of iteration exceeded some limit). The recognition procedure used was the Viterbi algorithm.

5 The Database used in the Evaluation

Ten speakers, five male and five female, generated the database. Each speaker was asked to read out as isolated words a list of five versions of the alphabet in random order and ten versions of the randomly ordered English digits (0-9). The VQ and the SCLT algorithms training data were collected from one version of the vocabularies for each speaker. A Hamming window of 256 points at 75% overlap was used. The isolated words are first processed by a 12 poles LPC analysis using the autocorrelation method and Durbin's recursion to form sequences, of LPC vectors. These sequences are then quantized by a VQ and an SCLT. The distance measure used is the minimum prediction residual of Itakura. The outputs of the VQ and SCLT algorithms are then divided into two exclusive sets, one for training the HMM and the other for testing.

6 Comparison of the Performance of VQ and SCLT Recognizers

To evaluate the effectiveness of the SCLT-produced codebooks a series of isolated-word recognition tests were carried out in independent mode for the digit vocabulary and in adaptive mode for the alphabet vocabulary. 50 versions of each word, with an equal number of male and female speakers, were used in creating two reference templates for the independent mode, where a new method of creating reference templates was used [2]. To make the most use of the alphabet data, the letters of each version were assumed to be templates and compared to the letters of the other versions of the vocabulary that were assigned as test words.

Fig. 1 compares the recognition results for the different codebooks, generated by the VQ and SCLT using the method of combination described before, for the digits and the alphabet vocabularies. An examination of these results show that; first, the SCLT reference codebooks gave a lower recognition error rate in comparison with all VQ conventional codebooks for both vocabularies. Second, from Fig. 2 for the Hidden Markov Model recogniser, it is clear that the SCLT codebooks have lower error rates in comparison with the VQ codebooks of the same size. Thus, the SCLT reference codebooks provides a better observation sequence for the HMM than that of VQ codebooks. Generally speaking, the above results suggest that it may be better to quantise other acoustically dissimilar events in addition to V/U/S with a codebook that is formed from separate codebooks.

7 References

1. Mahmoud, W.A. and Bennett, L.A.M., "The Distortion Measure of the Segment Classification and Labelling for different window lengths" Proc. of IEEE International Electrical and Electronics Conference (ELECTRONIC'85), Canada, 7-9 October 1985.

2. Mahmoud W.A. and Bennett, L.A.M., "Creating Reference Patterns Via a Vector Quantization Algorithm", Proceeding of the 2nd International Conf. on Advances in Pattern Recognition and Digital Techniques, India, pp 31-45,6-9 Jan. 1986.

8 Acknowledgement

The authors wish to acknowledge the kind assistance and helpful discussion of Dr. S. Levinson, AT and T Bell Laboratories in USA.

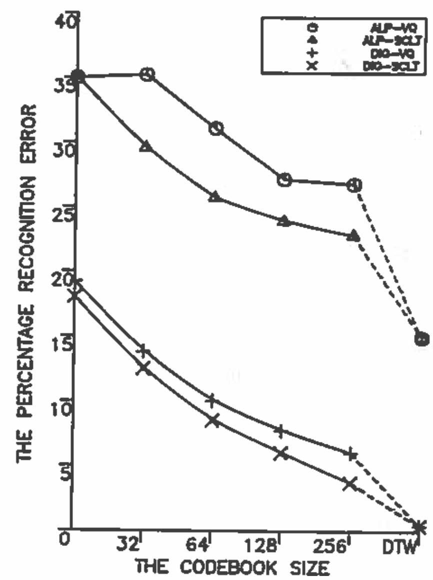


Fig.(1) Average Recognition Error for both vocabularies using different VQ technique as a function of codebook size.

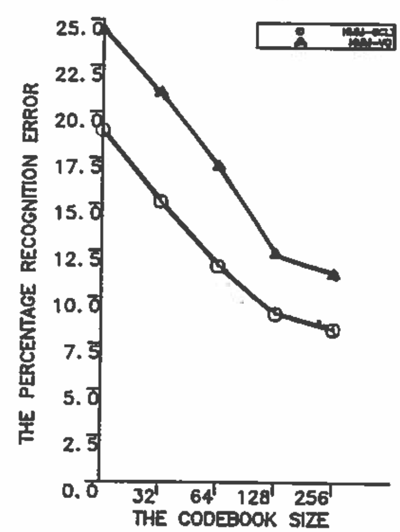


Fig. (2) The average recognition Error for the digit vocabulary using HMM on different VQ techniques.

THE EFFECT OF LPC ORDER ON THE PERFORMANCE OF VECTOR QUANTIZATION IN ISOLATED-WORD RECOGNITION

W. A. Mahmoud and L. A. M. Bennett

Department of Electrical & Electronic Engineering
University College of Swansea
Singleton Park
Swansea SA2 8PP
United Kingdom

1 ABSTRACT

The Vector quantization (VQ) of LPC spectra has been applied to cocoding and also very recently to speech recognition as a means of reducing memory requirements for the storage of reference templates and of reducing computation time. This paper examines the effect of the LPC order (P) on the distortion measure and on the performance of the VQ algorithm in isolated-word speech recognition (IWSR).

2 INTRODUCTION

Linear predictive coding (LPC) coefficients have become the most powerful and predominant features for representing the speech signal. The number of LPC spectra required to describe the words of a vocabulary is very high. The basic concept of VQ is to classify these LPC spectra by comparing them with vectors in a codebook. The goal of a VQ algorithm is to minimise the distortion measure associated with the classification procedure.

Several factors that effect the distortion have been studied including the initial codebook, the multiplying factors type of distance measure etc. In this paper the effect of an important factor, the LPC order P, on the distortion as well as on the performance of VQ in IWSR is considered. A speech recognition system has been developed in software on a 68K mini-computer using Dynamic Time Warping (DTW) and Vector Quantization (VQ). The recognition error rates against codebook sizes of 4,...,128 codewords have been obtained and compared for five different values of LPC order P.

3 THE VQ ALGORITHM

Assume that a training set of V vectors in the form of gain normalised autocorrelation terms is given. It is desired to find a codebook of size C codewords such that the average distortion measure (distance) (DS (C)) of a vector in the training set from the closest codeword is minimised, thus:

$$DS(C) = \underset{(C)}{\text{Min}} \left| \frac{1}{V} \cdot \sum_{i=1}^{V} \underset{1<m<C}{\text{Min}}\, d(v_i, c_m) \right| \qquad [1]$$

where $d(v_i, c_m)$ is the LPC distance between the training vector v_i and the codeword c_m. The log likelihood distance measure of Itakura is used.

4 THE EXPERIMENTAL BACKGROUND

Seven speakers, five male and two female, generated the database of spoken Arabic digits (0-9). In one session each speaker was asked to contribute ten digits as isolated utterances. In a second session each speaker was asked to read out as isolated-words a list of one hundred digits in random order. The first training set was used to generate the codebook of sizes of 2, 4,....,128 codewords using the above vector quantization algorithm for five different values of P. A fourth-order antialiasing elliptic filter with cut-off of 4.8 kHz was used together with 12-bit ADC and a sampling rate of 10 kHz.

5 THE EFFECT OF LPC ORDER P ON THE DISTORTION MEASURE OF VQ

To obtain useful results with vector quantization it is important to understand the relationships and the effects of the choice of order of the LPC model on the distortion measure.

To evaluate the effects of P on the performance of the VQ algorithm, a series of two sets of tests were run. These series of experiments consisted of the design of the VQ for the Arabic-digit vocabulary for two Hamming window lengths, 12.8 msec and 19.2 msec. Five different values of P were used, which were, 8, 10, 12, 14 and 16. Fig. 1 shows the distortion measure of VQ for these values of P as a function of codebook size for the 12.8 msec Hamming window. A similar result was achieved for the 19.2 msec window. It is clear from these plots that the distortion measured increases as P increases. This is understandable, because when the value of P increases, more of the details of the spectrum are included in the LPC spectrum.

This was observed from the plots of their LPC spectra where the 8 and 10-pole models give much smoother spectra than the 12 to 16-pole models for a given frame of speech. Therefore the distance measure between two LPC spectra for smaller values of P will be smaller than that between two spectra of higher P.

The natural question now is how many poles should one use in fitting the model for the data acquisition system under consideration? There is no direct answer to the above question as far as the distortion measure is concerned. Therefore to understand better the effect of P on the VQ algorithm it is necessary to study its performance outside the training algorithm.

6 THE EFFECT OF P ON THE PERFORMANCE OF VQ IN IWSR

To evaluate further the effect of P on the performance of the VQ outside the training data a series of recognition tests were carried out. Two sets of experiments were performed on the DTW/VQ and LPC/DTW isolated word recognizers.

(a) The first set of experiments was performed for the Hamming window of 12.8 msec length. The speaker-dependent mode of recognition was used, hence each speaker was treated separately. The 100 digits of each speaker were used as templates once and as tests next, hence a total of 10,000 crossing were performed for each speaker.

(b) In the second test the above experiments were repeated for a Hamming window of 19.2 msec.

The templates and the tests were quantized to a VQ codebook of size of 4,8, ..., 128 codewords and the recognition performance was compared. In this way some direct results are obtained for recognition error rate againsts VQ codebook size. The results of the first tests are given in Fig. 2, which shows plots of error rate versus codebook size. The results of the second test are given in Fig. 3.

From inspection of these plots it is clear that the 8 and 10-pole models provided insufficient recognition accuracy. For the 12 to 16-pole models the recognition error rate was acceptable and there was a slight difference in the recognition accuracy for them, particularly for codebook sizes of 32 and more. As a practical matter, it is generally desirable to use the minimum number of poles necessary to model accurately the significant features of the signal. Therefore, it was decided that the 12-pole model was sufficient for the recognizer under consideration.

7 CONCLUSIONS

This paper has studied the effect of P on the distortion measure and performance of VQ in IWSR. Two sets of experiments were run on a database of 700 isolated digits from 7 speakers. Increasing P increases the VQ distortion, however it improves its performance outside the training data. The results strongly suggest that, for the VQ recognizer under consideration, the minimum, value of P should be 12. This reinforces the result reported elsewhere for the DTW recognition systems, that the minimum P should be 2 poles for each kHz band of the filter plus two extra poles.

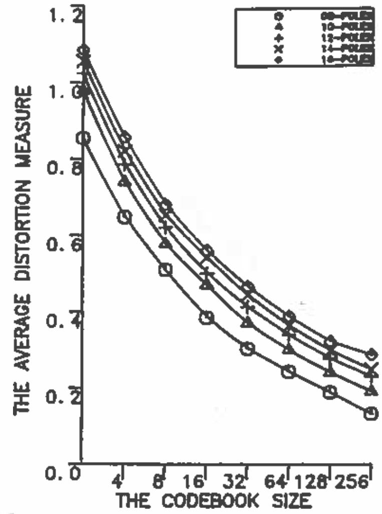


Fig.(1) The distortion measure for five LPC order, for the 12.8 msec window, as a function of codebook size.

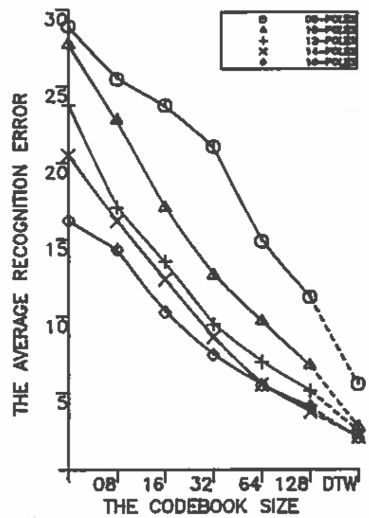


Fig.(2) The Average Recognition Error using a window of 12.8 msec, for five values of P as a function of codebook size.

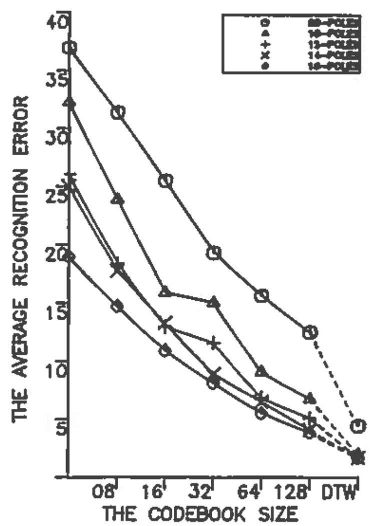


Fig.(3) The Average Recognition Error using a window of 19.2 msec, for five values of P as a function of codebook size.

ACOUSTIC/PHONETIC TRANSCRIPTION USING A POLYNOMIAL CLASSIFIER AND HIDDEN MARKOV MODELS

Alfred Kaltenmeier

AEG Research Institute, Ulm, Germany

ABSTRACT

This paper describes a module for acoustic/phonetic transcription in a continuous speech understanding system. This module segments input utterances into sequences of phone classes which belong to one of six broad phonetic categories. In a higher system level such segment sequences are used to hypothesize possible word candidates from a lexicon.
This module is hierarchically implemented in two stages: a polynomial classifier for a frame-by-frame classification of phone classes followed by a segmentation stage using Hidden Markov Models (HMM) of phone class segments.

INTRODUCTION

This paper describes an acoustic/phonetic module for a continuous speech understanding system which is being developed within the framework of the European Community ESPRIT Project No. 26.
Since continuous speech recognition presupposes an unlimited vocabulary, units smaller than words must be used for recognition. In our system two kinds of small phonetic units are used: phonemes and diphones on the one hand /1/ and phone classes (plosives, fricatives, etc.) on the other. The number of phone classes to be distinguished is low (5 to 10) whereas the number of phonemes and diphones is much higher (100 to 200). Using this double representation of phonetic units, the recognition part of our system can be effectively implemented in three levels (Fig. 1).

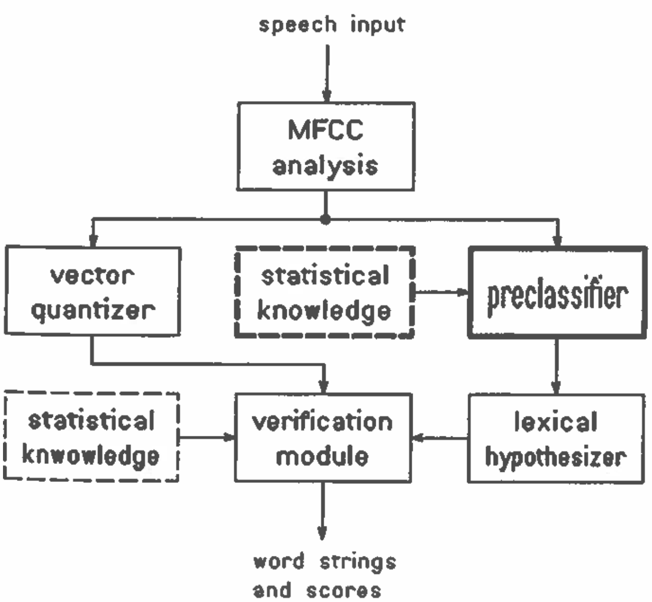


Fig. 1 Block diagram of the recognition stage in the continuous speech understanding system

The data reduction block in the first level computes mel-frequency cepstral coefficients (MFCC) as parametric representations of speech frames /2/.
For the second level, cepstral vectors form the input to both a vector quantizer and a preclassifier which hypothesizes phone classes. Vector quantization reduces the amount of data while preserving all information needed to correctly classify the various sounds. The preclassifier transforms speech signals into broad phonetic categories and in the process computes segment boundaries and likelihoods, too. The statistical knowledge consists of a coefficient matrix for polynomial classification and HMMs of phone class segments, phone class durations, rules, and error models for smoothing/segmentation.
In the third level the preclassifier output is used to extract a reduced number of word candidates from a word lexicon. This reduced set of word candidates is then verified and scored by the verification module which uses HMMs of phonemes and diphones as statistical knowledge.

PRECLASSIFIER MODULE

A hierarchical organization of the acoustic/phonetic transcription can greatly reduce the number of computations required in the word verification module. To this end, the selected set of phone classes must guarantee a high selectivity between the words in the lexicon while at the same time preserving a high reliability in the preclassification. Detailed investigations have shown that these two opposing requirements can be best met using six phonetic categories which are labeled as follows:

pl: plosives and silence
fr: fricatives and affricates
ln: sonorants (liquids and nasals)
fv: front vowels
cv: central vowels
bv: back vowels

The preclassifier is implemented in two stages (Fig. 2). The first stage consists of a polynomial classifier followed by a decision quantizer, both performed frame by frame. The classifier estimates the likelihoods that a cepstral vector belongs to each of the predefined phone classes by evaluating the following matrix product:

$$\underline{d} = A \cdot \underline{x} \qquad (1)$$

where $\underline{d}$ is a decision vector containing estimated likelihoods, A is a coefficient matrix, and $\underline{x}$ is a vector which contains linear, quadratic, and cubic terms of cepstral vector components.

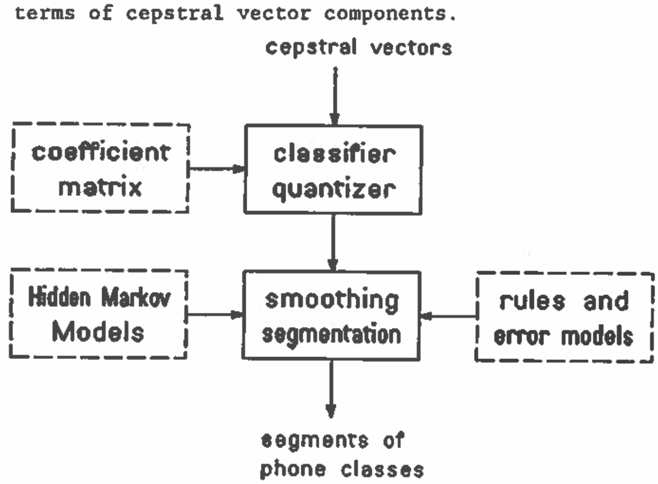


Fig. 2 Block diagram of the preclassifier with its associated statistical knowledge sources

Since determing the coefficient matrix A requires a large amount of computation and storage and a very large speech data base, it is computed off line with automatically labeled speech data . Due to the large computational and storage requirements, this classifier must be speaker-independent if it is to be at all practical.

The classification according to eq. (1) is implemented in a two-level structure. First we estimate the likelihoods of three combined classes (pl+fr, ln+fv, and cv+bv), which are then separated in their respective subclasses in a second level. This hierarchical structure increases performance and requires less computation than a parallel structure which estimates all six classes simultaneously.

Along with the estimated likelihoods the classifier produces a reliability score. This score represents a unique decision for one class, a decision for two of the six classes, or a reject if no reliable decision can be made. According to this score, a decision vector is quantized and transformed into a symbol. We have one symbol for the reject, 6 symbols for unique decisions, and 15 symbols for all possible two-case decisions or 22 symbols in all. Since the first stage of the preclassifier module transforms a cepstral vector into a symbol, it can be viewed as a vector quantizer which incorporates phonetic information. Hence, at the output of this stage an utterance is represented by a sequence of symbols, which then have to be smoothed and segmented by the second stage of the preclassifier.

Such a sequence may contain local irregularities (corresponding to spurious decisions, particularly during transitions) which have to be smoothed out in order to correctly segment an utterance. Using a simple fixed-length majority voting filter for smoothing is not very effective because this does not take statistical information on segment durations into account. Better segmentation results are obtained by statistical decoding using HMMs of phone classes and transitions as well as information on phone class durations.

Fig. 3 illustrates the complete preclassification process. The example used here is the German time phrase 'neun Uhr drei' (9:03) with following phoneme /diphone and phone class descriptions:

phon./diph.: - n o Y n U R d dr r a I -
phone clas.: pl ln bv fv ln bv cv pl ln cv fv pl

The first row of Fig. 3 shows the phoneme/diphone segments which were manually labeled for this example (transition segments are not shown). The second

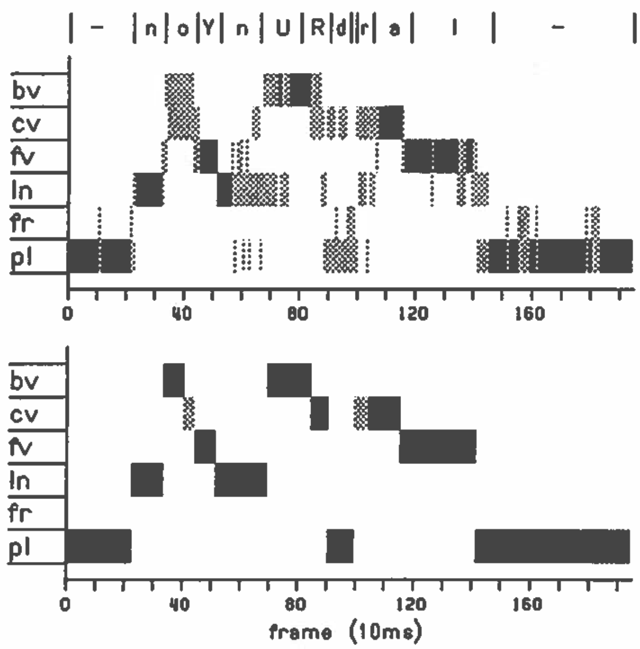


Fig. 3 phone class segmentation of the German time phrase 'neun Uhr drei' (9:03)

row shows the output of the first preclassifier stage. Dark areas are unique decisions, and shaded areas are two-class decisions. The third row shows the result of the segmentation. Obviously there are two errors in the segmentation (shaded areas). The first back vowel 'o' is split into two short bv and cv segments, and the liquid 'r' is merged with the following central vowel 'a'.

In order to reduce the number of such segmentation errors we will implement both a set of rules which directly include the speech signal energy in the segmentation process and also statistical models for the most frequent preclassification errors. A frequent error, for example, is the smoothing out of a short sonorant segment between two vowels. However, such a missing segment can be easily recovered using the energy contour which shows a clear dip in the sonorant segment.

Error models which define the likelihoods of context-dependent preclassification errors will be used to generate alternative segmentations. In order to evaluate error models some experiments with a larger preclassified data base are in progress.

The output of the preclassifier are error modeled phone class sequences forming the input to the next recognition stage. In this stage the lexical module first generates syllabic segments from the phone class sequences. Then syllabic segments are used to select a set of word candidates which are possible in the given part of an utterance.

PRECLASSIFIER PERFORMANCE

This section summarizes the preliminary performance of two preclassifiers which were computed from Italian and German speech data. The classifiers were trained with 720 words from four Italian speakers and 500 words from two German speakers.

The frame-by-frame classifications in the first stage of the preclassifiers have quite low error rates between 3% and 6%. Only segments labeled as stationary phonemes are considered here because its difficult to define an error rate during transitions. Error rates were obtained using the 'k best of six classes' rule, where k = 1 for unique decisions and k = 2 for two-case decisions.

The segmentation is based on the Viterbi algorithm; rules and error models have not yet been implemented. For isolated words we had an segment error rate of about 10%. Using the German preclassifier we made some additional experiments with 100 connected digit strings and 100 five-word sentences. With this material, which did not belong to the training data, we had segment error rates of about 12% for connected digit strings and about 16% for the sentences, respectively. By applying energy information and errors models, error rates can be decreased and the reliability of the preclassifier further improved.

Using the preclassifier approach described above, speech signals can be reliably segmented into six broad phonetic categories which minimize the ambiguity in the lexicon access.

LITERATURE

/1/ Cravero M., Pieraccini R., Raineri F. Definition and Evaluation of Phonetic Units for Speech Recognition By Phonetic Units Int. Conf. ICASSP86, April 1986, Tokyo

/2/ Davis S. B. and Mermelstein P., Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences IEEE Trans. ASSP, Vol. 28, Nr. 4, August 1980

MODELISATION AUTOREGRESSIVE EVOLUTIVE ET RECONNAISSANCE DE LA PAROLE

G.Boulianne, G. Chollet *, et Y. Grenier *
INRS-Télécommunications (Univ. du Québec)
3. Place du Commerce
Verdun, Québec CANADA H3E 1H6

Le signal de parole est caractérisé par une alternance de zones spectralement assez stables, entrecoupées de régions transitoires. Les systèmes de reconnaissance proposés par le passé reposent sur des propriétés de stabilité spectrale et de stationnarité; ils obtiennent des performances mitigées pour les régions de transition. Une représentation de ces régions par modèle AR évolutif, valide sur toute la durée d'une région transitoire, est proposée. Les coefficients du modèle dépendent du temps et s'expriment sur une base limitée de fonctions temporelles. Cette méthode de représentation est appliquée à la reconnaissance de segments transitoires C-V extraits de parole naturelle, et comparée à des méthodes plus classiques.

INTRODUCTION

La modélisation autorégressive est bien connue en traitement de la parole sous le nom de prédiction linéaire [1]. Elle oblige à un compromis entre précision et stationnarité qui consiste à découper le signal en fenêtres d'une dizaine de millisecondes.

La modélisation AR évolutive, telle que mise au point par Y.Grenier [2], n'exige pas la stationnarité du signal et de ce fait est mieux adaptée aux régions transitoires de la parole. Le développement d'un espace de représentation adéquat et d'une métrique adaptés à la représentation évolutive fait l'objet de ce travail.

MODELISATION AR EVOLUTIVE

Le modèle AR d'ordre p s'écrit habituellement:

$$y_t + a_1 y_{t-1} + \dots + a_p y_{t-p} = b_0 \epsilon_t \qquad (1)$$

Si le processus n'est pas stationnaire, les coefficients a_i deviennent dépendants du temps et sont appelés coefficients évolutifs:

$$y_t + a_1(t-1) y_{t-1} + \dots + a_p(t-p) y_{t-p} = b_0(t) \epsilon_t \qquad (2)$$

Leur expansion sur une base de m fonctions du temps s'écrit

$$a_i(t) = \sum_{j=0}^{m-1} a_{ij} f_j(t) \qquad (3)$$

et rend possible leur calcul [2]. Les a_{ij} sont appelés *composants invariants* du modèle évolutif. En représentant les fonctions de la base sous la forme d'un vecteur $F(t) = [f_0(t)\ f_1(t) \dots f_{m-1}(t)]$, le modèle évolutif $M(t)$ est obtenu par

$$M^T(t) = AF^T(t) \qquad (4)$$

où la matrice A est formée des composants a_{ij}. La stationnarité n'étant plus nécessaire, un modèle évolutif peut être calculé pour un segment de parole arbitrairement long.

* E.N.S.T., Paris, France

COEFFICIENTS EVOLUTIFS DU CEPSTRE

Pour un modèle stationnaire, les coefficients cepstraux se déduisent des coefficients de prédiction grâce à une relation récursive ([1]). Les coefficients cepstraux évolutifs sont définis par une extension de cette relation:

$$c_i(t) = a_i(t) + \sum_{k=1}^{i-1} \frac{k-i}{i} c_{i-k}(t) a_k(t) \qquad (5)$$

L'expansion des $c_i(t)$ sur une base de m fonctions orthogonales sur l'intervalle τ permet de dériver une approximation des *composants cepstraux invariants*:

$$c_{iq} = a_{iq} + \sum_{k=1}^{i-1} \frac{k-i}{i} \sum_{r=0}^{m-1} c_{(i-k)r} \sum_{s=0}^{m-1} a_{ks} f_{rsq} \qquad (6)$$

pour $i = 1, \dots, p$ et $q = 0, \dots, m-1$, et où les constantes f_{rsq} peuvent être précalculées:

$$f_{rsq} = \frac{\int_\tau f_r(t) f_s(t) f_q(t) dt}{\int_\tau f_q^2(t) dt} \qquad (7)$$

Le filtre de prédiction doit être stable afin de garantir un comportement raisonnable des composants cepstraux. La stabilisation d'un modèle est effectuée selon la technique de [3], qui consiste à évaluer le polynôme $A(z)$ sur un cercle de rayon supérieur à 1.

DISTANCE ENTRE MODELES EVOLUTIFS

Des essais préliminaires sur la distance euclidienne entre spectres logarithmiques, coefficients de prédiction,de réflexion, et du cepstre ont montré les mêmes tendances pour le modèle évolutif que celles observées par [1] et [4]. La métrique euclidienne sur les coefficients cepstraux a été retenue pour les tests de reconnaissance.

La distance euclidienne entre deux segments de parole décrits par deux trajectoires de paramètres, i.e. deux suites A et B de N points à p dimensions s'écrit habituellement:

$$d(A,B) = \sum_{n=0}^{N-1} \sum_{i=1}^{p} (b_i(n) - a_i(n))^2 \qquad (8)$$

L'équivalent pour deux trajectoires évolutives décrites par des composants invariants est

$$d_c(A_c, B_c) = \sum_{q=0}^{m-1} h_q \sum_{i=1}^{p} (b_{iq} - a_{iq})^2 \qquad (9)$$

où les coefficients h_q dépendent de la base de m fonctions.

$$h_q = \int_\tau f_q^2(t) dt \qquad (10)$$

ANAMORPHOSE TEMPORELLE

Les segments sont modélisés sur l'intervalle τ, subissant une normalisation linéaire du temps. Des déformations non-linéaires peuvent être obtenues directement dans le domaine des composants invariants. Soit la transformation temporelle $t' = u(t)$. Si Γ est une matrice de transformation dont les éléments sont

$$\gamma_{ij} = \frac{\int_\tau f_i(u(t)) f_j(t) dt}{\int_\tau f_j^2(t) dt} \qquad (11)$$

un modèle transformé s'exprimera en fonction de la base originale et de composants transformés $A' = A\Gamma$:

$$M^T(u(t)) = AF^T(t') = A\Gamma F^T(t) = A'F^T(t) \qquad (12)$$

En paramétrisant $u(t)$ par un polynôme $a_0+a_1t+a_2t^2+ ... + a_dt^d$, la matrice devient:

$$\Gamma = \{\gamma_{ij}\} = \{\gamma_{ij}(a_0, a_1, ...a_d)\} \qquad (13)$$

La transformation optimale est celle qui minimise la distance entre un modèle B et un modèle A anamorphosé:

$$d_c^* = \min d_c(A\Gamma, B) = \min \ \Delta(A, B, a_0, ..., a_d) \qquad (14)$$

avec des contraintes qui restreignent aux transformations plausibles: positivité de la pente (pas d'inversion du temps), degré peu élevé du polynôme $u(t)$, intervalle transformé situé dans l'intervalle τ. Le problème se résoud par un algorithme d'optimisation non-linéaire classique.

SEGMENTATION

L'évaluation des techniques précédentes est faite sur des segments transitoires. Leurs frontières ont été définies comme étant les points de pente maximum d'une fonction de stabilité:

$$\nu(n) = \frac{-1}{4}\sum_{i=1}^{3}\sum_{j=-2}^{j=+2} |r_i(n) - r_i(n+j)| \qquad (15)$$

Les $r_i(n)$ sont les coefficients de réflexion d'une fenêtre n. Les régions considérées non-stationnaires sont ainsi celles où la dérivée seconde de la stabilité est positive. Cette définition ne comporte pas de seuils arbitraires ou dépendants du signal.

EXPERIENCES ET RESULTATS

Les expériences ont porté sur des transitions consonnes-voyelle de langue française. Dix séries des 18 syllabes /le/ /ʀe/ /je/ /we/ /ye/ /ve/ /fe/ /se/ /ʃe/ /ʒe/ /me/ /ne/ /pe/ /te/ /ke/ /be/ /de/ /ge/ prononcées par un seul locuteur adulte mâle, en ordre aléatoire, ont été filtrées, numérisées à 12 bits, puis segmentées en régions instables. Parmi ces régions, 175 situées immédiatement avant le noyau vocalique stable ont été extraites et modélisées avec 16 pôles et 4 fonctions de base (polynômes de Legendre). Ces dimensions sont basées sur l'optimisation du critère d'Akaike. Les composants invariants de prédiction ont ensuite été transformés en composants invariants cepstraux. Environ 19% des modèles ont dû être stabilisés. Les modèles cepstraux évolutifs obtenus ont été soumis à quatre expériences:

1. Les dix séries ont été divisées en deux moitiés de 5 séries; la distance euclidienne entre chaque segment d'une moitié et tous les autres segments de l'autre moitié a été évaluée, sans anamorphose. Cette procédure a produit 175 tests de reconnaissance.

2. La procédure 1 a été répétée en introduisant une anamorphose de degré $d = 2$ dans l'évaluation de la distance.

3. Chaque série a été comparée à des références obtenues en combinant les segments des 9 autres séries ("leave one out").

4. La procédure 3 a été répétée avec, pour chaque série, des références auxquelles elle avait participé. Ainsi les segments testés avaient servi à l'apprentissage.

Taux de reconnaissance (175 tests)			
Expérience	Rang du premier correct		
no.	1	2	3
1	57%	66%	70%
2	57%	68%	69%
3	58%	69%	78%
4	72%	82%	88%

La comparaison des expériences 1 et 2 montre que le taux de reconnaissance n'est pas modifié sensiblement par l'anamorphose. Seulement 16% des erreurs commises dans l'une ne le sont pas dans l'autre. L'anamorphose ne dégrade pas la capacité de discrimination de la mesure de distance, mais il ne semble pas y avoir d'avantage à l'utiliser pour des segments aussi courts. L'algorithme de création de références, mis en évidence dans l'expérience 3, se révèle efficace.

Les candidats aux erreurs les plus fréquentes se retrouvent parmi les segments qui ont dû subir une stabilisation. 48% des erreurs sur ceux-ci proviennent d'une confusion avec un autre segment stabilisé. On peut s'attendre à une amélioration marquée du taux de reconnaissance si une autre méthode de stabilisation peut être mise au point, qui n'aplatisse pas l'enveloppe spectrale.

Ces résultats peuvent être comparés aux expériences de [5] sur des séries consonnes-voyelle françaises similaires; 12 systèmes de reconnaissance disponibles sur le marché européen avaient alors obtenu un taux de reconnaissance situé entre 40% et 85%.

CONCLUSIONS

Cette étude montre qu'il est possible de développer, pour la modélisation évolutive, des techniques semblables à celles dont on se sert en prédiction linéaire. L'espace peut être muni d'une métrique utilisable pour la reconnaissance et de transformations permettant une anamorphose temporelle. Les résultats obtenus permettent déjà d'identifier les points faibles des techniques développées, sur lesquels devraient s'attarder de futurs travaux.

BIBLIOGRAPHIE

Markel J.D and A.H Gray, *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

Grenier,Y., " Time-dependent ARMA Modeling of Nonstationary Signals", *IEEE Trans. Acoust. Speech and Sign. Proc.*, vol. ASSP-31 no.4, aug. 1983, pp. 899-911.

Haskew J.R., Kelly J.M., Kelly R.M. and T.H. McKinney, "Results of a Study of the Linear Prediction Vocoder", *IEEE Trans. Comm.*, vol. COM-21 no. 9,sept. 1973, pp.1008-1014.

Davis S.B. and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoust. Speech and Sign. Proc.*, vol. ASSP-28 no.4, aug. 1980, pp. 357-366.

Chollet G.F., Astier A.B.P, and M. Rossi, "Evaluating the Performance of Speech Recognizers at the Acoustic - Phonetic Level", *Proc. ICASSP*, Atlanta 1981. pp.758-761.

NEW NON-SUPERVISED LEARNING METHODS FOR SPEAKER ADAPTATION

Osamu Kakusho and Riichiro Mizoguchi

The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, 567 Japan.

ABSTRACT: An inter-related phoneme template system is proposed together with its two nonsupervised learning algorithms. Their efficiency is verified through some computer experiments of word recognition.

1. INTRODUCTION

This paper is concerned with automatic speaker adaptation for speaker independent recognition. A new phoneme template system composed of inter-related phoneme templates is proposed[1] along with two efficient non-supervised learning algorithms. One is based on the selection of the inter-related phoneme templates from a set of templates prepared beforehand. The other is based on the creation of new templates appropriate for each speaker. The former algorithm is performed in "on-line" mode, that is, the selection is made every time a word is uttered. It is useful for rapid adaptation. The latter is performed in "batch" mode, that is, the creation is made after a reasonable amount of words are obtained. Although the adaptation is done one or two days after the first usage, almost complete adaptation can be made in this learning algorithm. The performance of these two non-supervised learning algorithms is verified by computer simulation of a word recognition system.

2. INTER-RELATED PHONEME TEMPLATES

2.1 Construction method

step 1: For each speaker, make augmented feature vectors of the dimensionality 5d by combining every feature vector of the dimensionality d of the frame corresponding to Japanese five vowels.

step 2: Apply k-means method[3] to the augmented vectors and obtain representative vectors of the clusters (one from each cluster).

step 3: Decompose the representative vectors into the original form, each of which is considered as a template of a vowel.

2.2 An example of the inter-related phoneme template

Fig. 1 shows four inter-related templates (pentagons) represented in a two dimensional space composed of the first and second formants frequencies. Speech samples are drawn from the isolated vowels uttered by ten male adults. The vertices of each pentagon are template patterns.

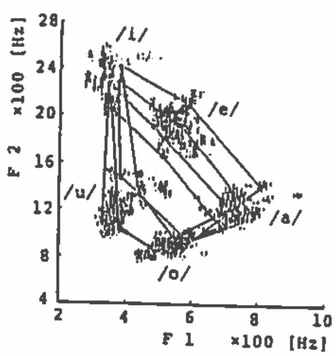


Fig. 1 Some examples of inter-related phoneme templates.

3. NON-SUPERVISED LEARNING METHOD OF ON-LINE TYPE

3.1 Algorithm

Let use-count of a template be defined as a number of input patterns which match best with the template. And let use-count of an inter-related template be defined as a sum of the use-count of the templates contained in it. Then, we have the following learning(selection) algorithm.

step 1: Calculate the use-count of all the templates.

step 2: Select the inter-related template of the maximal use-count.

This algorithm is based on the selection of the templates according to the use-count which are obtained without using the identities of the input patterns. Therefore, it is a non-supervised learning algorithm. The selection can be performed in any time period and in any scheme.

3.2 Evaluation

3.2.1 Vowel recognition

a) Speech samples

Japanese five vowels uttered consecutively like /ieaou/ by 15 male adults were analyzed with LPC method (10kHz sampling, auto-correlation, order 12, and hamming window of length 20ms with shift interval 10ms). Each speaker uttered a sequence of vowels five times. Two of them were used for template construction (600 frames in all, 600 = 15men x 5vowels x 8frames), and the rest of them were used for learning and recognition (675 frames in all, 675 = 15men x 5vowels x 9frames).

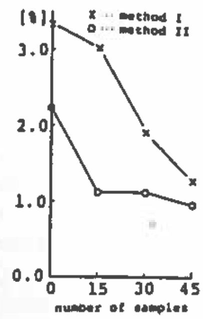


Fig. 2 Learning process(1).

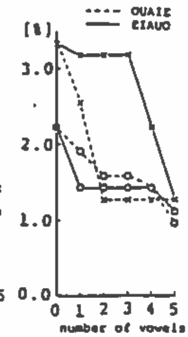


Fig. 3 Learning process(2).

Table 1. Recognition rates.

spkr	bef. L.	aft. L.	
	Rc %	RIR %	Rc %
KI	100	50	98
SN	98	24	100
MA	92	80	96
KO	98	10	98
YA	90	22	98
YU	98	-25	96
YM	98	16	100
AVE.	96.3	25	98.0

b) Recognition method

Recognition is made using the template matching in the 4-dimensional Fischer space constructed based on the samples for template construction.

c) Experiment

As described in a), the speakers used for template construction and recognition are identical, so this is a closed recognition as to the speaker.

Ten inter-related templates were obtained according to the method described in section 2. Fig. 2 depicts the learning process of vowel recognition. The vertical axis shows the error rates(%) and the horizontal one the number of vowels given to the system. The letter "x" denotes the error rate of the conventional templates and the letter "o" that of the proposed template. In order to see the effect of the order of vowels on the learning performance, simulation was done for two kinds of sequences. The error rates corresponding to the sequence /ouaie/ are depicted by broken line and those corresponding to /eiauo/ are depicted by solid one. Selection of the templates is done as follows: Every time when learning of a vowel is done, for the conventional templates, one template corresponding to the vowel is chosen from the templates. For our template, on the other hand, six inter-related templates are chosen after the learning of the first vowel is done. Four, three, two and one inter-related template are chosen after the learning of the second, third, fourth and last vowel is done, respectively. It is seen from Fig. 2 that learning of our templates does not depend on the sequence of vowels, while that of the conventional ones depends largely on the sequence.

Fig. 3 shows the relation between the error rates and the number of learning samples given to the system until it makes the final selection of the template. This figure demonstrates that learning of the inter-related templates is much faster than that of the conventional ones. Some advantages of the non-supervised learning method of the inter-related template are summarized below.

1) Adaptation is fast.
2) Learning process is stable.
3) Learning process is reliable.

3.2.2 Word recognition

Vocabulary of the system is composed of Japanese ten digits 0(/rei/) through 9(/kyu/). Open recognition as to seven male adults was done, where eight inter-related templates were prepared before hand. Table 1 shows the recognition rates for seven speakers. RIR parameter represents the ratio of the improved recognition rate of vowels contained in the digits. This results shows the effectiveness of our non-supervised algorithm.

4. NON-SUPERVISED LEARNING OF BATCH TYPE

The learning method proposed above is based on the selection of a template from a set of them prepared in advance. Therefore, performance of the learning depends on the speakers, that is, adaptation(selection) is done successfully only when at least one template appropriate to the speaker is stored in the system. When no such template is stored, however, much improvement can not be attained by the learning. In order to make the learning more effective, another learning method is proposed in this section.

The learning method creates new templates appropriate to the speakers rather than selection of them. To do this, the algorithm needs a reasonable amount of sample words. Consequently, adaptation to a speaker is made one or two days after his first use of the system. This is why the algorithm is said to be of batch type.

4.1 Algorithm

The block diagram of the total system is shown in Fig. 4, in which a block surrounded by broken line corresponds to the proposed learning algorithm.

4.1.1 Clustering of input words

A clustering method [2] is applied to respective sets of words uttered by a speaker. Since the clustering algorithm requires only a distance matrix as input data, it is easily executed.

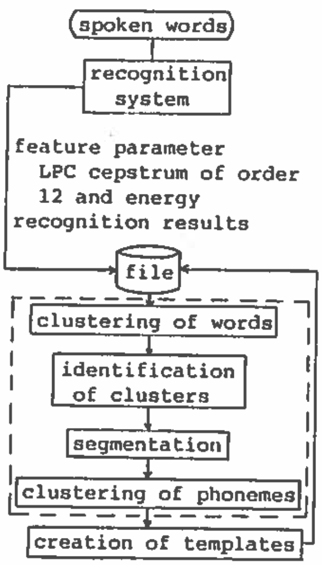


Fig. 4 Block diagram of the total system.

4.1.2 Identification of the categories of the clusters

Clusters obtained above are labeled according to the majority rule using the labels given by the system itself. There are two alternatives of the treatment of the minorities in the rest of the operations:

A) Reject them and
B) Relabel them to the category of the majority.

In the case of A), the new templates are created by using only words supposed to be recognized successfully by the system. In the case of B), on the other hand, such words that are supposed to be mis-recognized are also used for creating new templates.

4.1.3 Segmentation

Segmentation of every word is done by using the energy and label associated with it.

4.1.4 Creation of the inter-related templates

According to the operation thus far, a set of frames labeled one of the Japanese vowels are obtained. New templates are created from these frames by clustering procedures shown below.

step 1: K-means method of number of clusters 8 is applied to the whole set of frames.

step 2: For each vowel, count the number of frames belonging to respective clusters.

step 3: For each vowel, select major groups until they cover 80% of population of the vowel. And consider them as the templates of the corresponding vowel.

step 4: Register the template as an inter-related template.

Table 2 Recognition rates(%).

spkr	T-1	T-2	T-3
A	98.0	94.0	100
B	91.8	97.9	91.8
C	98.0	100	100
D	93.9	98.0	100
E	90.0	92.0	98.0
F	98.0	98.0	98.0
G	68.8	91.3	93.5
H	76.0	84.0	94.0
AVE	89.4	94.4	96.9

4.2 Evaluation

Performance evaluation of the proposed learning method was done in word recognition of 32 words. Phoneme templates were obtained from the words uttered by 31 male adults. Fifteen times utterances of ten words in the vocabulary made by eight speakers other than the 31 speakers were used for the evaluation. Ten utterances were used for initial recognition and collection of data for the non-supervised learning. The final recognition was done by using the rest of 5 utterances.

Table 3 Improvement of vowel recognition(%).

speaker	1->2	1->3
F	3.7	29.6
G	54.4	57.9

We have three sets of templates:

T-1: 31 templates before learning
T-2: Template created using category identification A) in step 2.
T-3: Template created using category identification B) in step 2.

The results are shown in Tables 2 and 3. Table 2 shows the recognition rates of the respective speakers and Table 3 shows the values of RIR. It is seen from both tables that the batch type non-supervised learning algorithm attains much improvement especially for the speakers having low initial recognition rates. Furthermore, performance of T-3 is slightly better than that of T-2. This is because T-3 is constructed from the mis-recognized words as well as recognized ones.

5. CONCLUDING REMARKS

Inter-related phoneme templates have been proposed together with two types of non-supervised learning algorithms. The results of the computer experiment has demonstrated the efficiency of them and shown the possibilities of this application to the real world situations.

(References)

[1] Mizoguchi, R., et al.:"Word recognition system for unspecified people based on inter-related phoneme templates", Trans. of the IECE Japan, Vol. J67-A, 6, pp. 572-579, 1984.

[2] Mizoguchi, R., et al.: "A nonparametric algorithm for detecting clusters using hierarchical structure", IEEE Trans., PAMI-2, 4, pp. 292-300, 1980.

[3] Anderberg, M.R.:"Cluster analysis for applications", Academic Press, 1973.

ON THE ROBUSTNESS OF PHONETIC INFORMATION IN SHORT-TIME SPEECH SPECTRA

Meg Withgott and Marcia A. Bush[1]

Stanford University, Center for the Study of Language and Information, Stanford, California 94305, USA

Schlumberger Palo Alto Research, 3340 Hillview Avenue, Palo Alto, California 94304, USA

Abstract: Speech recognition techniques which take fixed-time slices as input to a matcher face the task of mapping from arbitrary pieces of the physical signal to abstract linguistic units. This paper examines the reliability with which individual vector-quantized LPC spectra can be mapped to various sets of acoustic-phonetic classes. The database for the experiments consisted of approximately 130,000 spectra from a pre-labeled corpus of 616 5-digit strings, and classification was performed on the basis of a maximum likelihood decision rule. Classification accuracy, when the same database was used for training and testing, ranged from 94.0% for a simple voiced-voiceless distinction to 42.7% for a set of 45 acoustic-phonetic classes used in earlier connected digit recognition experiments [1,2].

Introduction

It is commonly accepted that the variability inherent in speech makes it difficult to recognize linguistic units such as allophones directly from sequences of short-time spectra. This observation has, in part, motivated work on broad phonetic classification schemes, in which an initial labeling of the recognition vocabulary is made on the basis of presumably robust acoustic-phonetic categories which then is used to identify subsets of the vocabulary for more detailed acoustic processing. Studies have shown that, for instance, a coarse-grained classification based on manner of articulation reduces a 20,000-item wordlist into approximately 100 phonetic cohorts (i.e., wordlist sublists) [3]. Relatively little quantitative data are available, however, to determine whether classification strategies designed and tested on the basis of abstract phonetic or phonemic considerations are actually useful in labeling large corpora of speech signals. Similarly, little is known about trade-offs between classification accuracy and the granularity of the labeling scheme.

This paper examines the reliability with which individual vector-quantized LPC spectra can be mapped to three types of acoustic-phonetic classes: one based on manner of articulation; a second based on multidimensional distinctive features (see e.g. [4]); and a third "system-specific" type influenced both by knowledge of the classifier's front end and of acoustic characteristics of individual classes in the recognition vocabulary.

Procedure

The database for the experiments consisted of 129,812 spectra from a pre-labeled corpus of 616 5-digit strings. The connected-speech utterances were spoken by 56 adult talkers (27M, 29F) from 22 geographically defined dialect groups, and form a subset of the training portion of Texas Instruments' connected digits database [5]. The initial label set comprised 45 acoustic-phonetic classes used in earlier connected digit recognition experiments [1,2]. Labeling was done primarily by hand, with simple durational rules for automatically dividing diphthongs and certain sonorant and word-boundary regions.

Signal preprocessing consisted of digital downsampling of the TI data from 20 KHz to 8 KHz (i.e., a 4 KHz bandwidth) and preemphasis by first-differencing. Short-time spectra were computed using an 11-pole LPC analysis, with a 25.6 msec Hamming widow and a 10 msec frame rate, and were vector quantized to a size 1024 codebook.

Classification of spectra was performed using a maximum-likelihood decision rule and, in these preliminary experiments, the same database was used for training and testing.

Classification Schemes

As noted above, three classification schemes were examined. Each involved grouping the initial 45-label set into smaller numbers of acoustic-phonetic categories. The grouping was complicated slightly by the fact that the initial labeling of the data was partially automated and thus not completely phonemic (e.g., glides typically included a short portion of the adjacent vowel). Such phenomena were uniform, however, across the three classification schemes.

With respect to the first classification, based on manner of articulation, label sets of size 4 (silence, fricative, nasal, vowel) and 6 (silence, weak fricative, strong fricative, nasal, glide and vowel) were used.

The second, multidimensional classification employed diverse distinctive features so that a given label represents a vector of cross-classified values. In contrast, manner forms a unidimensional classification. Figure 1 shows a distinctive feature tree corresponding to the complete [-sonorant] subset of the distinctive feature categories. Such trees yield relatively coarse-grained classes at the top nodes and finer-grained classes as the tree is descended. A binary partitioning of the initial label set led to the [+/-sonorant] distinction.

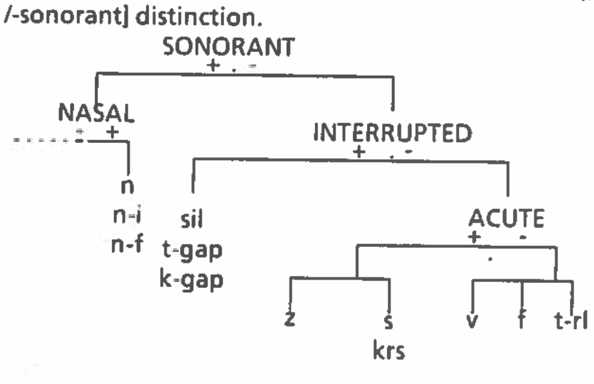


Figure 1: Distinctive Feature Tree for "consonants"

A partial tree for the third scheme, which is system-specific and multidimensional, is shown in Figure

2. As noted above, this classification strategy takes into account both characteristics of front-end processing and acoustic characteristics of individual acoustic-phonetic classes in the recognition vocabulary. For example,

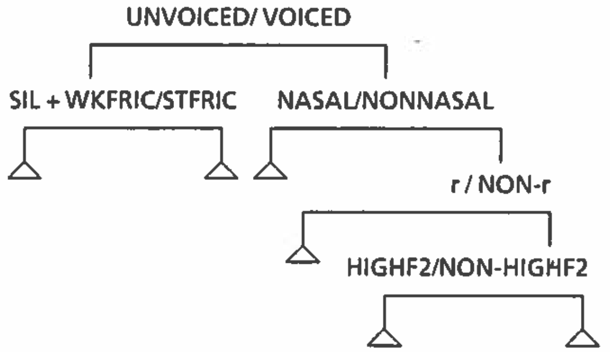


Figure 2: Partial Tree for the system-specific classification

weak fricatives and silent intervals are collapsed into a single class because they are difficult to discriminate on the basis of LPC spectra alone. On the other hand, the release portions of the [t]'s in the digits 2 and 8 are classified as strong and weak fricatives, respectively, on the basis of context-dependent acoustic manifestations.

Results and Discussion

Figure 3 shows overall classification accuracy (i.e., the percentage of short-time spectra correctly classified) as a function of number of acoustic-phonetic categories for the three classification schemes. Percentages are similar across the classification schemes when small numbers of categories are used. (For the purpose of comparison, a fourth classification with arbitrary six-way partitions was created and found to exhibit classification accuracy of 48.4%).

Number of Categories	manner	multiple features	task-specific
2		93.5	94.0
4	84.6	84.6	87.0
6	79.0	73.7	79.2
10		67.4	73.5
21			64.3
45			42.7

Figure 3:. Overall classification acuracy (percent correct) versus number of acoustic-phonetic categories for the three classification schemes.

An advantage of multidimensional classifications, such as the feature-based and system-specific classifications, as opposed to a unidimensional classification such as manner, is that they support a selective traversal down one or more branches of a classification tree. The choice of whether to collapse or differentiate categories can therefore be determined on the basis of the lexicon, or the discriminability of individual classes.

Figure 4 shows overall classification accuracy as a function of the branch traversed for the system-specific scheme, and shows, for example, that a 9-way classification determined by a broad unvoiced class being more finely-differentiated was equal to the performance of a 6-way classification when the voiced branch was descended. The same advantage does not

Number of Categories	branch traversal	
	unvoiced	voiced
3	89.0	92.0
4	84.6	88.8
6		82.3
9	83.0	
10		74.2

Figure 4. Overall classification accuracy (percent correct) for system-specific scheme as a function of the branch traversed.

show up in a 3-way or 4-way comparison, and thus classification accuracy depends both on how categories are sub-divided and on how many sub-divisions are formed. We are also able to note that combining categories representing relatively broad classes with categories containing a single segment type which proves to be highly discriminable in the vocabulary of interest (e.g., the early vocalic region in 4 (AOR1) in this database) can be advantageous.

Summary

Multidimensionality appears to be a desirable trait of classification systems for applications in automatic speech recognition. This is because the identity and grain-size of the classes can be determined freely both by what features are the most useful for discriminating lexical items, and by what classes prove to be the least confusable for a particular classifier.

1. After Aug 86: Division of Engineering, Box D, Brown University, Providence, RI 02912, USA

References

[1] Bush,M. 'Durational constraints for network-based connected digit recognition,' (This volume).

[2] Bush,M., and G. Kopec. 'Network-based connected digit recognition,' submitted to *IEEE Transactions on Acoustics, Speech and Signal Processing*, March 1986.

[3] Shipman,D.W. and V.W. Zue. 'Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems,' *1982 IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, France.

[4] Fant, G. *Speech Sounds and Features*. MIT Press, 1973.

[5] Leonard, G. 'A database for speaker- independent digit recognition,' *1984 IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, CA.

DISCRIMINATION OF VOICED PLOSIVES USING TRANSITION PROPERTIES OF THE LPC CEPSTRUM PARAMETERS

Y.Yamashita, M.Yanagida, R.Mizoguchi and O.Kakusho

The Institute of Scientific and Industrial Research, Osaka University, 8-1, Mihogaoka, Ibaraki, 567 JAPAN.

ABSTRACT: In this paper, a discrimination method of voiced plosives is proposed using two sets of parameters; one, that describes the transition of acoustic parameters by fitting their transition loci with regression lines and the other, the acoustic parameters themselves at the beginning point of the transition. The gradient of the regression lines is found to be effective for discriminating among voiced plosives.

1. INTRODUCTION

It is generally assumed that the acoustic properties of voiced plosives lie both near the burst and in the transition part to following vowels.[1] Proposed here is a method for discriminating voiced plosives by employing both instantaneous properties near the burst point and dynamic properties in transient part. The transition properties of a consonant followed by a vowel are especially dependent on the following vowel. In this paper a following-vowel dependent discrimination method is adopted and analysis periods are adjusted for each following vowel. Its performance is evaluated on isolated syllables uttered by 38 male adults.

2. FITTING THE PARAMETER TRANSITION WITH REGRESSION LINES

2.1 Analysis and Discrimination of Voiced Plosives

The acoustic parameters of voiced plosives change drastically at the burst and during the succeeding short period. However, the variations of the parameters in slow transition parts are expected to be sufficiently described with regression lines, i.e. the transition of each acoustic parameter can be approximated with a line. the number of analysis frames to be fitted by regression lines is fixed here to be ten regardless of the frame shift interval.

The LPC analysis of order 12 is performed on 10 successive frames in the transition part starting at the burst point to the following vowel. The analysis start point is defined by time delay Td from the burst point and is determined according to the following vowel together with the frame shift interval Ts. The short-time energy and the LPC cepstrum coefficients are obtained for 10 frames, where the short-time energy is expressed in dB normalized by the short-time power of the following vowel part.

The time series of each parameter is approximated with a regression line. The gradients of the regression lines are employed as the parameters to describe transition properties, and the short-time energy and LPC cepstrum parameters of the first frame are used as those for instantaneous properties. Therefore, 26 parameters in all (13 first frame parameters and another 13 gradient parameters) are employed for mutual discrimination voiced plosives.

Speech samples employed here are 15 isolated CV syllables; /b/, /d/ and /g/ as the leading consonants followed by Japanese vowels /a/, /e/, /i/, /o/ and /u/, uttered by 38 males in a large anechoic chamber. (The total number of utterance is 15x38=570, i.e. 114(=570/5) for each following vowel) The speech samples were quantized at 10 ksamples/sec with 12 bit accuracy after low-pass filtering of 4.5kHz cut-off frequency and -260 dB/oct suppression characteristics.

The discrimination score is evaluated by the leaving-one-out method. In this paper, the Fisher space[2] is employed to reduce the dimension of parameters from 26 down to 2. The discrimination is performed as decision by majorities in 3-nearest neighbors on the 2-dimensional Fisher space.

2.2 Investigation on Analysis Periods

The optimal analysis periods, which are determined by the window length for one frame analysis (Tw), the location of the initial analysis frame (Td) and the frame shift interval (Ts), are investigated for each following vowel assuming that the burst points were detected by visual inspection, and the following vowels, a priori known. The investigation range for each parameter is as follows;

Tw = 10, 15 and 20msec.
Ts = 1, 2, 3, 5 and 7msec.
Td = -12 through 12 by 2msec step.

The discrimination test is performed under 195 (=3x5x13) combinations of analysis conditions.

The discrimination performance was compared among the window length of 10, 15 and 20msec. The discrimination results remain almost the same regardless of the window length. Therefore, the window length is fixed to 20msec in the rest of this paper considering the stability of analysis.

The total length for each CV syllable to be analyzed is determined by the frame shift interval Ts. The analysis start point is identified by Td which is the time delay relative to the burst point. Positive Td means that the analysis is started at Td msec after the burst. The optimal analysis period which yields the best discrimination score for each following vowel is shown in Fig. 1. Table 1(a) shows the best discrimination scores under condition that the burst points were detected by visual inspection, and the following vowel, a priori known.

Table 1 compares the discrimination score with and without employing the gradient parameters of the regression lines. It is recognized from Table 1 that introduction of the gradient parameters improves the discrimination score by 5% on average for the five following vowels. Fig. 2 shows the comparison of the distribution of the phoneme templates on the 2-dimensional Fisher space for both the discrimination schemes with and without parameters for following vowel /u/. The clusters of the with-gradient case have less overlaps and are clearly separated one another compared to the without-gradient case. The Fisher ratio is improved from 3.3 to 12.5.

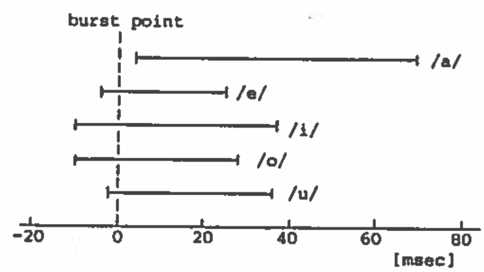
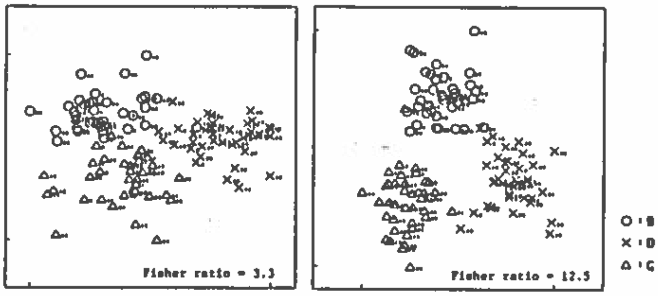


Fig.1 The optimal analysis period for each following vowel.

Table 1 Discrimination score of voiced plosives(I).
Burst point detection : by visual inspection
Following Vowels : a priori known

Gradient Parameters	Discrimination Score(%) Following Vowel /a/	/e/	/i/	/o/	/u/	average
(a)used	96	89	92	95	93	93
(b)not used	94	83	84	91	89	88



(a)without the gradient parameters (b)with the gradient parameters

Fig.2 A comparison of sample distribution on the Fisher space (following vowel /u/).

3. DETECTION OF BURST POINTS AND RECOGNITION OF FOLLOWING VOWELS

In the previous section, it is assumed that the burst points are detected by visual inspection and that vowels, a priori known. This section describes a recognition system of voiced plosives with automatic detection of burst points and automatic recognition of following vowels.

3.1 Automatic Detection of Burst Points

Automatic detection of burst points is realized by using the distance measure of the LPC cepstrum parameters between a lame pair of frames, where two frames start at the same point, and end at different points. The lengths of the long and the short frames are 20 and 17msec, respectively. The correct detection rate of the proposed method for burst point detection is evaluated under the following criterion. If the difference of the detected point and the real burst point is less than 3msec, the detection is presumed to be correct. Under the criterion above, the score is 92% in average.

3.2 Recognition of Following Vowels

Recognition of following vowels is realized also employing decision by majorities on the Fisher space projected from a 16-dimensional space scaned by the LPC cepstrum parameters. For test samples, 5 frames near the center of vowel part are analyzed and assigned to one of the five Japanese vowels according to decision by majorities in 5-nearest neighbors in the Fisher space. Then, the final decision is made again by majorities in the result of the successive five frames. The recognition score of the following vowels by this algorithm is 99% using leaving-one-out method.

4. AUTOMATIC DISCRIMINATION OF VOICED PLOSIVES

In this section, voiced plosives are automatically discriminated. The following vowel is first recognized, and the Fisher space is automatically chosen among those prepared for the five Japanese vowels separately. The analysis periods are determined referring the burst point detected automatically.

Table 2(a) shows the recognition scores based on the automatic identification of the following vowels. The notation of the recognition unit C in Table 2 indicates that the recognition score is that concerning the consonants only, regardless of recognition error concerning the vowels. In case of recognition unit CV, the score is recognition including the vowel identification, i.e. the score means recognition rate of CV syllables. Although the score is a little bit worse than Table 1, however, the score is over 90% in average. Table 2 shows the comparison of the score obtained by using the gradient parameters and that without using them. From this Table, it can be said that introduction of the gradient parameters improves the score by 5% for automatic CV syllable recognition.

Table 2 Discrimination score of voiced plosives(II).
Burst point detection : automatic
Following vowel recognition : automatic

Gradient Parameters	Rec. Unit	Discrimination Score(%) Following Vowel /a/	/e/	/i/	/o/	/u/	average
(a)used	C	93	90	91	95	89	92
	CV	93	88	91	95	89	91
(b)not used	C	91	82	82	89	87	86
	CV	90	82	82	89	85	86

Table 3 Discrimination score of voiced plosives by the following vowel independent discrimination system.
Burst point detection : automatic

	Following Vowel /a/	/e/	/i/	/o/	/u/	average
Score(%)	85	80	84	84	88	84

5. FOLLOWING-VOWEL INDEPENDENT DISCRIMINATION

As described above, a Fisher space is prepared for each following vowel for vowel-dependent discrimination aiming at improvement in discrimination ability. In order to justify the vowel-dependent discrimination, a following-vowel independent discrimination is tried on the speech data. Table 3 shows the scores of following vowel independent discrimination. The test samples are classified into three categories with decision by majorities in 5-nearest neighbors in a 4-dimensional Fisher space. The vowel-dependent scheme is proved to be very effective for discrimination of voiced plosives.

6. CONCLUSION

A following-vowel dependent discrimination system for voiced plosives is proposed employing additional parameters for describing the dynamic properties. The dynamic properties are extracted as the gradients of regression lines which approximate the transition of the acoustic parameters The short-time energy and the LPC cepstrum parameters are employed as the acoustic parameters here. The speech samples are CV syllables uttered by 38 male adults with voiced plosives for C and Japanese vowels for V.

The analysis periods are adjusted according to following vowels. In case that the burst points are automatically detected and following vowels are recognized by the system, discrimination score is 91%. It is proved to be effective for discrimination of voiced plosives

(1) to introduce gradient parameters of regression lines to describe the dynamic property, and

(2) to adjust the analysis condition for each following vowel and adopt a following-vowel dependent algorithm.

REFERENCE

[1] Tanaka,K : "A parametric representation and a clustering method for phoneme recognition --Application to stops in a CV environment", ASSP-29, 6, pp.1117-1127 (1981).
[2] Duda,R.O. et al. : "Pattern Classification and Scene Analysis", John Wiley, NY, pp.114-121 (1973).

TEXT INPUT USING SPEAKER-ADAPTIVE CONNECTED SYLLABLE RECOGNITION

Yoichi Takebayashi, Hiroyuki Tsuboi, Shouichi Hirai, Hiroshi Matsura and Tsuneo Nitta

TOSHIBA Corporation, 1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki 210, JAPAN

This paper describes a speech recognition system for large vocabulary text input. The system recognizes connected Japanese syllables by both continuous pattern matching and speaker-adaptation based on the Multiple Similarity(MS) method. The recognition algorithm consists of syllable boundary detection, vowel and consonant recognition and lexical verification. The reference pattern vectors adapt to each speaker by K-L expansion through covariance matrix modification. Recognition experiments on a 17,877 word Japanese vocabulary showed 92.6% accuracy for 10 male 4,400 phrase utterances.

INTRODUCTION

While many speech recognition systems have been developed in the last decade, few word recognition systems have been accepted for text input application owing to poor accuracy and limited vocabulary. DP matching is a prevailing technique for word pattern matching, but it's not practical enough except for speaker-dependent small-vocabulary word recognition. The Multiple Similarity(MS) word pattern matching method is extremely powerful, but limited to a speaker-independent small vocabulary[1]. Likewise, the multi-template method is not applicable to a large vocabulary. Several word recognition systems based on probabilitistic model have been developed[2], but they require a lot of computation for large vocabulary recognition. On the other hand, the phoneme or syllable based recognition methods, the syntactic methods, are absolutely required for both continuous speech recognition and practical large vocabulary recognition[3]. However, the accuracy of the phonological units has been insufficient due to no effectual training algorithms. While rule-based speech recognition method is being studied to achieve full use of speech knowledge[4], automatic learning is still a open problem in AI research.

In this paper, an approach to achieving a large vocabulary word recognition system is first described. Then the proposed system is presented concerning acoustical and phonetic and lexical representations, continuous pattern matching and speaker adaptation. Finally experimental results are shown.

APPROACH

In order to attain a practical large vocabulary word recognition system or voice-activated word processor, we focus on the following points as design concepts:

a) High recognition accuracy
b) Strong and automatic speaker adaptation mechanism
c) Ease of utterance for novice users
d) Hardware realization and LSI implementation.

Taking these points into account, we have developed new connected syllable recognition and speaker adaptation methods[5]. The recognition system--consisting of syllable segmentation, vowel recognition and consonant recognition--employs MS calculation and acoustic labeling on a time-continuous frame by frame basis. We introduce a promising MS based approach because of the reliability and accuracy of the MS method in speaker-independent word recognizers[1] and character readers[6]. Conventional pattern matching methods, like DP matching, are so sensitive to pattern variation that they cannot be applied to syllable recognition. Rule-based phoneme recognition systems are being developed to utilize speech-specific knowledge. However, the learning mechanism (automatic knowledge acquisition) is still poor to date. Hence the pattern recognition oriented approach is much more promising than the rule-based one for implementing the speaker adaptation mechanism. The continuous MS matching is suitable for hardware realization. Also the vowel and consonant pattern vectors are reasonably represented by considering their inherent properties. In addition, our MS based adaptation method has a huge capacity to represent the phoneme pattern variability in detail--a large degree of freedom, therefore it is robust and reliable in regard to pattern variation and distortion. While the connected syllable approach is restrictive, the continuous MS matching and adaptation methods are applicable to further continuous speech recognition research. The recognition system demands user's cooperation, that is, clear utterance for connected syllable recognition. Our main purpose for developing this system is that novice users can input a lot of data more comfortably and efficiently by using this recognizer than keyboard.

Figure 1 shows a newly developed recognition system.

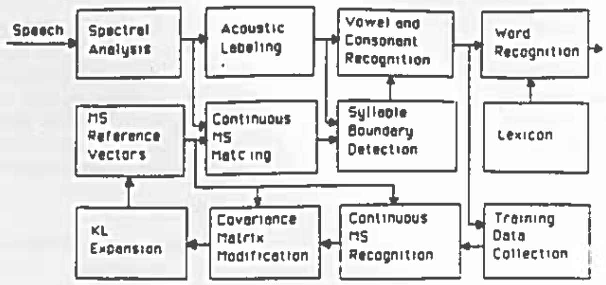


Fig.1 Blockdiagram of Recognition and Adaptation System

RECOGNITION AND ADAPTATION ALGORITHMS

Acoustic Representation

Input signal is converted into a 12-bit digital signal at 12kHz-sampling frequency. Spectral analysis is done by 3 sets of 4-pole digital band-pass filters. These filter outputs are squared and smoothed over 16ms. frames, and converted into logarithmic ones and then sampled at every 8ms. The overall energy is simultaneously obtained every 8ms. The 16-channel filter and 8-channel filter outputs are fed into vowel and consonant MS calculation, respectively. The 4- channel filter outputs are used for acoustic labeling.

The MS method utilizes the structure of pattern variation on pattern space for each category. Therefore parametric analysis, like LPC, is not used. Instead, non-parametric filter bank analysis is used. Modeling in pattern space based on the MS method is more reasonable and effective than that in speech signal for speech recognition.

Continuous Multiple Similarity Method

The MS method has been theoretically derived and experimentally proved to be powerful and effective by several optical character readers [5] and speaker-independent word recognizers[1]. The telephone speech recognizer accomplished a high performance in spite of significant pattern distortion. However, it cannot directly apply to phoneme or syllable recognition, as phoneme or syllable patterns have much less information than word utterance patterns. In order to obtain accuracy, real-time processing and adaptation mechanism, we propose the continuous MS pattern matching method. This method, based on the time-continuous MS calculation every 8 ms., is suitable for hardware realization. The problem in applying the MS method to connected syllable is how to represent the vowel and consonant feature vectors as N-dimensional feature vectors.

Vowel and Consonant Pattern Vector Representation

Each Japanese syllable has either one of five vowels or a syllabic nasal. The vowel is more durable and stable than the consonant. Therefore vowel recognition is a crucial component of all the recognition system. Considering these points, we represent the vowel pattern as a 16-dimensional vector(one frame 16 channel frequency spectrum) for the continuous vowel MS calculation.

As contrast with the vowel, the consonant part is not stable and inherently characterized by time-variant spectral patterns. Therefore we represent the consonant pattern vector as a multiple-frame time-frequency spectrum, not as a one-frame spectrum. The consonant 64-dimensional vectors, generated by 8-channel frequency spectra over 8 frames, have 128ms. duration and are continuously matched by consonant reference vectors every 8ms.

Acoustic Labeling

Although the continuous MS pattern matching might work considerably well for both vowel and consonant recognition, we also introduce the acoustic labels in order to complement the MS values. A similar acoustic labeling was effectively employed in the telephone speech recognition system[1]. The 4-channel spectrum and overall energy are fed into the labeling processing.

Syllable Boundary Detection

Loose syllable boundaries(start and endpoints) are needed as clues for vowel and consonant recognition, as the highly efficient and stable continuous matching is employed. These points are determined by not only a time series of overall energy, 4-channel spectrum and acoustic label but also syllable duration constraints. Syllable recognition accuracy depends significantly upon the syllable detection performance.

Syllable(Vowel and Consonant) Recognition

Vowel region is estimated by both the loose syllable boundary information and acoustic label sequences. Then vowel recognition is carried out by using a time series of vowel similarities and acoustic labels in the estimated vowel region. A segmented input syllable is classified to one of 6 vowels(/a/,/i/,/u/,/e/,/o/,/N/). The 15 entries of MS reference vectors are prepared for the accurate vowel recognition.

The vowel recognition result focuses the syllable candidates on ones that include the recognized vowel. It can significantly lighten the computation load and also consonant pattern variation based on co-articulation effect. The consonant region is determined by the syllable boundary, vowel recognition result and acoustic labels. Then consonant recognition is realized by using a time series of consonant MS values. The simplest recognition way is where the consonant category with the maximum MS value within the region is regarded to be a recognized consonant(syllable) as a result. The second and the third rank candidates with likelihood are also obtained by using their MS values for lexical verification at next stage.

Speaker Adaptation

While the proposed speaker adaptive recognition system works without training, recognition accuracy can dramatically increase after the sophisticated adaptation[5]. Most traditional adaptation methods, based on the multi-template technique or some statistical learning method like perceptron or linear discriminants, are not clear and not structural from lack of speech knowledge utilization. In contrast, the MS based adaptation and recognition methods positively utilize the structure of pattern variability. Namely, the speaker-adapted reference vectors of each category represent a specific speaker's essential pattern distribution, to accomplish robustness and reliability. An important problem is how to extract the training pattern vectors from the whole speech pattern. We consistently introduce the continuous MS matching not only for recognition but also adaptation. Training patterns including a vowel or consonant part are approximately extracted in terms of the acoustic labels and syllable boundaries. Then the continuous MS calculation is done on these patterns. Subsequently, the fixed training pattern vectors are extracted from the frames with the greatest MS values. Next, the covariance matrices are modified by these vectors. Finally the K-L expansion of the covariance matrices generates the reference pattern vectors. As the learning progresses, the extracting position can change to successively precise positions. Thus stable and robust reference pattern vectors can be obtained at the adaptation stage as the user utilize the recognizer more and more. Both enormous capacity and knowledge acquisition mechanism are remarkable advantages of the proposed method.

Word/Phrase Recognition

Dealing with only clearly spoken connected syllables, the system ignores possibility of the syllable insertion and deletion. Thus the lexicon for phrase(word) recognition is simply represented using syllables. Word or phrase recognition is carried out by lexical verification between the syllable candidates with their likelihood and the lexicon. For real-time processing, lexical search space reduction is made by using three preceding syllable candidates. As the lexical processing is quite simple, further research is necessary to improve the word recognition performance.

EXPERIMENTAL RESULTS

A 10 male training data set(50 samples per consonant, 100 samples per vowel) was collected for 101 Japanese syllables for adaptation of each speaker. Another test data set including 4,400 phrases(9,230 syllables) was collected for evaluation of large-vocabulary recognition at the speed from 3 to 4 syllables per second. Table 1 shows the accumulated syllable recognition scores for both data sets. The accumulated scores, 97.8% and 100% suggest the stability and robustness due to a large capacity of the MS based method. More than 99.0% vowel recognition accuracies were obtained for the same data. Table 2 gives the phrase recognition score for a 17,877 word vocabulary. While simple lexical matching is used, the phrase(word) recognition score is considerably high because of the high syllable recognition accuracy. The results also demonstrate the reliability of the MS based continuous matching and adaptation.

	Training Data Set (50,500 syllables)	Test Data Set (9,230 syllables)
Best Candidate	98.9%	91.4%
3 Best Candidate	100.0%	97.7%

Table 1 Syllable Recognition Score

(after lexical verification using 3-best-candidate)	Test Data Set (9,230 syllables 4,400 phrases)
Phrase Recognition Rate	92.6%

Table 2 Phrase Recognition Score

CONCLUSION

A text input recognition system using connected syllable recognition has been developed for novice keyboard users. The system employs both continuous matching and speaker-adaptation based on the MS method. The experimental results have shown that the proposed system is accurate enough to act as a practical voice-activated word processor or large vocabulary data entry system. Since the dominant computation of the MS recognition and learning methods is multiplication-accumulation, a real-time machine can be easily realized by LSI implementation.

REFERENCES

[1] Y.Takebayashi,et all.,"Telephone speech recognition using a hybrid method" Proc.fifth ICPR,pp.1232-1235,1984

[2] F.Jelinek,"The development of an experimental discrete dictation recognizer",Proc.IEEE,vol.73,no.11,pp.1616-1624,1985

[3] M.J.Hunt, et al.,"Experiments in syllable-based recognition of continuous speech",Proc.ICASSP'80,pp.880-883,1980

[4] V.W.Zue,"The use of speech knowledge in automatic speech recognition", Proc.IEEE,vol.73,no.11,pp.1602-1615,1985

[5] H.Tsuboi,et al.,"The connected syllable recognition based on Multiple Similarity method",Proc.ICASSP'86,1986

[6] K.Sakai,et al.,"An optical Chinese character reader",Proc. Third IJCPR,pp.122-126,1976