

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 divided 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 hydro-mechanic 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, amplified in 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-

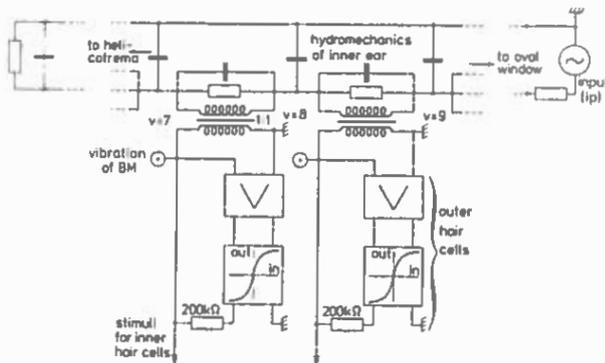


Fig. 1: Schematic diagram of a peripheral preprocessing model containing nonlinear active feedback.

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.

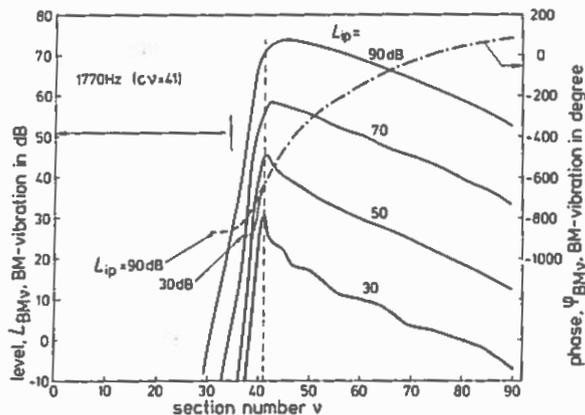


Fig. 2: Level L_{BMV} of the voltage equivalent to basilar membrane vibration and its phase ϕ_{BMV} as a function of the section number v corresponding to place along the basilar membrane. Parameter is the input level L_{ip} of the 1770-Hz tone.

The so far more generally described characteristics of the model are shown quantitatively in Fig. 2. The level L_{BMV} and the phase ϕ_{BMV} corresponding to the level of the vibration of the basilar membrane and to its phase are plotted as a function of v , 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 $cv=41$ 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 v and level independent. The two phase-place patterns show an expected behaviour of strong phase lag with decreasing v which depends near the characteristic place cv on input level L_{ip} .

The effect of compressing the dynamic range is most clearly seen in the relation between level L_{BMV} 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 existence 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 extremely steep rise, the level-dependent 3dB bandwidth which corresponds for normal speech level of 60dB to a Δv 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 for 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

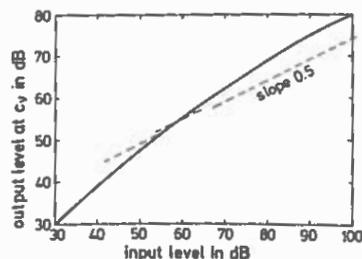


Fig. 3: Input-output relation expressed in 1770-Hz tone levels measured at the input and at $cv=41$.

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'_v , 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 et al., 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 effectiveness of the use of specific loudness-critical band rate-time patterns.

Another 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 strength 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

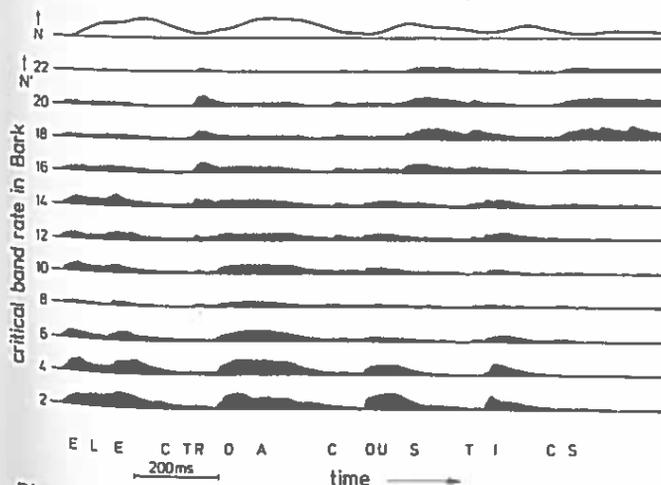


Fig. 4: Example of a specific loudness-critical band rate-time pattern. Total loudness $N(t)$ on top.

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 (Suchowersky, 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 divide 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 non-linear devices or at least combinations of linear and nonlinear devices, 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.

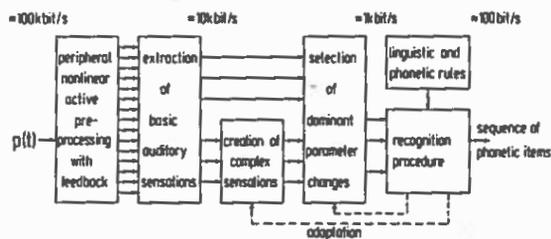


Fig. 5: Blockdiagram of a speech recognition system based on cochlear preprocessing and psychoacoustics.

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 non-auditory 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 (f_0) varies in speech. Perceptual data indicate that the auditory system is not bothered by changes to f_0 , but processing strategies used in speech recognition, such as linear prediction, filterbank analysis, and the synchrony spectrum are seriously perturbed as f_0 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 (f_0), which has been assigned a different constant value for each stimulus. The stimuli were devised to illustrate the problem of estimating the frequency (F_1) and level (A_1) of the first formant as fundamental frequency changes.

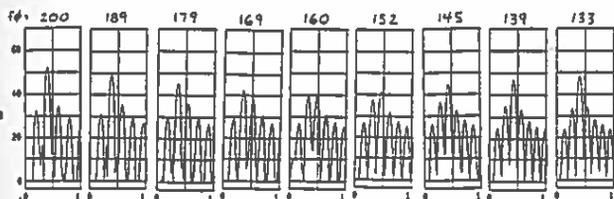


Figure 1. DFT magnitude spectra of 9 synthetic vowel stimuli varying only in f_0 .

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 ($F_2=1800$ Hz, $B_2=140$, $F_3=2900$, $B_3=240$, $F_4=3800$, $B_4=350$), 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 F_1 , 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 F_1 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 F_1 , 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	f_0	F_1	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 f_0 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; Mosen, 19xx). Stimuli C and F differ by 63 Hz according to this model, which is 16 percent of F_1 . This difference would be easily audible because the JND for F_1 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 F_1 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 F_1 from stimulus C to F, which is a 13 percent change. Also, there is a slight bias toward overestimating F_1 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 f_0 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 f_0 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 f_0 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 f_0 automatically invokes a kind of vowel-normalization process such that vowels spoken at higher f_0 are assumed to come from shorter vocal tracts (Miller, 1953; Fujisaki and Kawashima, 1968; Carlson, Granstrom and Fant, 1970; Schwartz, 1971; Slawson, 1968; Traummuller, 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 f_0 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 f_0 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 f_0 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 existence 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-creeaky 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 FO 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 FO 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.
- Trautmüller, H. (1982), "Perceptual Dimension of Openness in Vowels", J. Acoust. Soc. Am. 69, 1465-1475.