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).

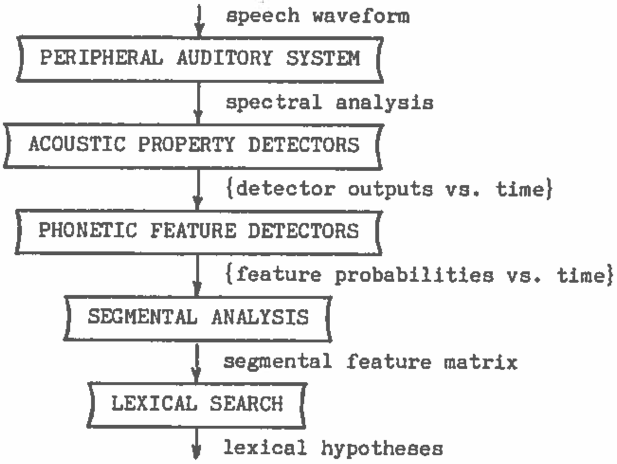
speech waveform
PERIPHERAL AUDITORY SYSTEM
spectral analysis
ACOUSTIC PROPERTY DETECTORS
{detector outputs vs. time}
PHONETIC FEATURE DETECTORS
{feature probabilities vs. time}
SEGMENTAL ANALYSIS
segmental feature matrix
LEXICAL SEARCH
lexical hypotheses

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.

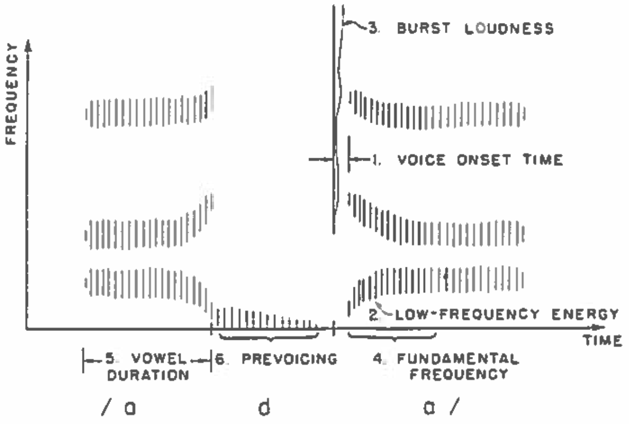

Figure 2. Six acoustic cues to voicing for plosives.

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.