INVARIANT AUDITORY ATTRIBUTES AND A MODEL OF SPEECH PERCEPTION

James R. Sawusch

State University of New York at Buffalo, Buffalo, New York

ABSTRACT

A model of speech perception is outlined that incorporates auditory grouping processes and a set of invariant auditory attributes as the basis for the phonetic and lexical coding of speech. Data from priming studies indicates that the phonetic code for words includes the positional specification of each phoneme. Based on this and some considerations of how invariant attributes might support auditory event and speech perception, a set of invariant auditory attributes for perception is described. This set is supported by perceptual data and, in conjunction with the model, accounts for a number of phenomena (such as trading relations) in the speech literature.

1. INTRODUCTION

In any model of speech perception, there are three key elements that need to be addressed. The model must detail the nature of the auditory attributes or invariants that support perception. The nature of perceptual grouping or organization processes that bind cues from the acoustic signal together as a single entity and separate the acoustic components of one event from other events needs to be specified. The mapping process that converts auditory cues to words must also be detailed. In the sections that follow, a model designed to address these issues will be outlined. Once the model is outlined, some details of the auditory invariants that underlie speech and other auditory perception will be considered.

The model is set in an information processing framework [cf., 12]. Like the LAFS model of Klatt [6], it proposes that invariant auditory attributes are captured from a series of spectral sections. Unlike LAFS, these attributes are mapped onto an intermediate phonetic representation and then to words. Like the proposals of Fowler [3], the ultimate goal of perception is recognition of the object or event that produced the sound. The set of auditory attributes is designed to support the recognition of speech and nonspeech. Unlike Fowler's proposal, in this model perception is mediated by stages of processing. The model does contain a "speech mode" [cf., 7]. However, this mode is layered over a set of auditory coding processes [see 15].

2. MODEL OUTLINE

The model consists of a sequence of representations and transformations. After the transformation of sound into an internal spectral/temporal representation by the peripheral auditory system, a basic set of auditory features is extracted. These features include the amplitude envelope, periodic or aperiodic nature of the waveform, fundamental frequency, and the frequencies and amplitudes of peaks in the spectrum. As these features are extracted over time, they are grouped together in sets that represent a common source or event [cf., 1]. The grouping process at this stage is driven by local spectral and temporal information. All further processing is then performed within a set of features.

The next step in processing represents the extraction of a set of invariant auditory attributes from the local cues. This set of attributes constitutes the basic information that drives acoustic object or event recognition. That is, regardless of whether the sound source is music, bird calls, door slams or speech, the attributes capture the information that preserves object or event identity. One example of such invariant attributes is the set of frequency differences between adjacent spectral peaks at each point in time. The nature of this set of invariants will be elaborated below.

The third stage in the model maps the auditory attributes, over time, onto objects or events. In speech these events are phonetic features, phonemes, syllables and words. Since all of these units seem to have a role in perception [see 11, 17], the model should explain the role of each abstract speech unit. The model should also provide an account of adult perception that is compatible with data and theories of infant speech perception and the acquisition, by children, of the words of their language [5].

The model assumes that in an adult the process of transforming auditory attributes into words is a two step process. The first step is a mapping of attributes onto phonetic features while the second step maps features, over time, onto words. The mapping of attributes onto phonetic features might be accomplished within a connectionist architecture. In this case, a time delay net could be used to map sequences of invariant attributes onto phonetic features. Interactions within the net would reproduce many of the phonetic trading relations described in the literature [13]. In addition, information about the rhythm of speech as cued by acoustic/auditory features such as the rise and fall of the amplitude envelope would govern the scope of cue interactions in the net. That is, a syllable-like integration window or context would result from the use of the amplitude envelope as a perceptual grouping factor at this stage of processing.

Finally, a sequence of features, over time, is mapped onto the words of the listener's lexicon. Since the phonetic features contain positional information and were grouped according to the rhythmic structure of speech, potential word boundaries are marked in the phonetic feature information for use by the word recognition process [cf., 9].

3. PHONETIC CODING

At this point, some further elaboration of the phonetic feature representation is in order since it will influence the nature of the set of invariant attributes as well as how they are mapped onto phonetic units. The phonetic features here are not the abstract entities used in linguistic theories. Rather, these features are position specific and contain some allophonic detail. Based upon data from auditory priming experiments using lexical decision and naming tasks, Gagnon and Sawusch [4] proposed that the phonetic representation used in word recognition includes information about the syllable position of each phonetic element. Thus, a syllable initial /b/ and a syllable final /b/ would be coded as two separate entities. In the present model, each phonetic feature would include a positional specification within the syllable. For example, the stop manner feature would be coded as either an onset stop or an offset stop.

The implication of this for the extraction of auditory attributes is that no single invariant or set of invariants is used by humans in the perception of all variations of a phoneme or feature. Rather, the invariant attributes themselves are part of the cue to the position of the feature in the syllable. Thus, the focus in our search for invariant attributes has been to examine the acoustics of stops before the vowel in a syllable and stops following the vowel separately. No common invariant across position has yet been found which is consistent with perceptual data [see 4].

4. INVARIANT ATTRIBUTES

The set of invariant auditory attributes must meet a number of constraints. First, they must represent a set of acoustic attributes that are not specific to speech. Rather, this set should be capable of supporting all auditory object or event recognition, including speech and music. This does not imply that speech perception, music recognition and the classification of a sound as a door slamming shut are all variants of the same perceptual processes. The process of mapping invariant attributes onto phonetic features described previously is a speech specific coding process. If implemented in a connectionist network, the weights on the connections would be the result of learning and represent a "speech mode" of processing. A similar learning process would be involved in the perception and recognition of other sounds.

A second requirement for this set is that they meet the requirements described by Sawusch and Dutton [16] for a formal, computational model or metric for perception. The attributes must represent a robust set in which the information supporting phonetic coding is preserved in spite of variation in talker, talking rate and the speech context. The attributes should also support perception even when degraded and should not lead to a sudden failure of perception in a noisy environment. The attributes should support graceful degradation. Finally, the attributes must be formally specified or computable in a manner that does not require intelligent guidance.

To illustrate the nature of invariants for sound recognition, consider the phonetic dimension of place of articulation for consonants and vowels. Miller [8] and Syrdal and Gopal [18] have proposed that the frequency differences between adjacent peaks in the spectrum capture the essential properties necessary for vowel recognition. Forrest, Weismer, Milenkovic, and Dougall [2] proposed that the statistical moments of the spectrum (mean, variance, skewness and kurtosis) capture a sufficient set of qualities for perception of the voiceless fricatives and stops. Sawusch and Dutton [in press] examined both of these alternatives for voiced stops and vowels. They found that the peak difference metric did not degrade gracefully while the statistical moments metric was not as robust as desired. However, these failures were largely complimentary so that a hybrid of both proposals might be sufficient to capture human perceptual capabilities.

In the present model, both the statistical moments of the short term spectrum and the frequency differences between adjacent peaks would be computed. In addition, the amplitude differences between adjacent peaks in the spectrum are also computed. The rate of zero crossings, the rate of change of overall amplitude and a set of source attributes such as the degree of periodicity in the spectrum and the fundamental frequency are also a part of the information represented here. These properties would be computed for each temporal section of the waveform on a continuous or running basis. At this point the representation of sound is still continuous and has not yet been segmented.

The statistical moments is a generalized description of the spectrum that subsumes the spectral tilt cue that has been proposed as an invariant to stop place. In a study of the efficacy of changes in spectral tilt as a cue to stop place, Richardson and Sawusch [14] found that changes in spectral tilt did not predict human listener classification of synthetic syllables. In a subsequent analysis of these syllables, it was found that both changes in the frequency differences between spectral peaks and changes in the statistical moments do predict listener's classification responses. Further tests of the efficacy of the statistical moments and peak differences as cues to stop consonant place are now in progress.

Certain properties are not represented at this level. The duration of an acoustic segment or the duration of a change in one of the attributes listed above, such as the rise-time, are not directly represented in the set of attributes. Instead, these are recovered by the process of mapping the static attributes, over time, onto phonetic features. There is also no single invariant for voice onset time (VOT). Rather, VOT is a composite that represents the mapping of source attributes, the amplitude differences between adjacent peaks and other attributes onto a position specific voicing feature. The processing of a nonspeech analog to VOT such as tone onset time [10] produces classification and discrimination results similar to speech because the nonspeech stimuli contain attributes in common with speech which are mapped (in the context of an experiment) onto categories correlated with speech categories.

Conversely, nonspeech experiments and speech experiments with the same stimuli sometimes produce differences in category boundary placements. This reflects a speech mode with connections from attributes to phonetic features that include learned contextual dependencies. In the context of a single experiment, no such detailed learning would take place for a nonspeech distinction. Consequently, comparisons between speech and nonspeech perception reveal both commonalities in processing based on a single set of auditory attributes and subtle differences due to the different learning histories between speech and nonspeech groups.

5. CONCLUSION

The outline of a model and elaboration of some aspects of the auditory coding of speech here is the beginning of a process of building a theory of speech perception. Among the issues that remain to be addressed are how phonotactic constraints and phonological knowledge should be incorporated into the model [see 19]. A further elaboration of the acoustic cues that guide perceptual grouping at each stage of the model is also needed. Simulations using a connectionist network to model the mapping of auditory attributes to phonetic features proposed here are also under way to test the sufficiency of the set of attributes. Finally, we are running a series of experiments in which the acoustic properties of synthetic stimuli are varied to examine how the attributes described here are extracted in perception and assess their relative roles in speech recognition. These data, together with theoretical elaborations, should lead to a more complete specification of the auditory to phonetic coding of speech.

6. REFERENCES

[1] BREGMAN, A. S. (1990), "*Auditory scene analysis*", Cambridge, MA: MIT Press.

[2] FORREST, K., WEISMER, G., MILENKOVIC, P., & DOUGALL, R. N. (1988), "Statistical analysis of word-initial voiceless obstruents. Preliminary data", *Journal of the Acoustical Society of America, 84*, 115-123.

[3] FOWLER, C. A. (1990), "Sound-producing sources as objects of perception: Rate normalization and nonspeech perception", *Journal of the Acoustical Society of America, 88*(3), 1236-1249.

[4] GAGNON, D. A., & SAWUSCH, J. R. (1990), "Priming and perceptual units in auditory word recognition", Paper presented at the 31st meeting of the Psychonomic Society, New Orleans, LA.

[5] JUSCZYK, P. W. (in press), "Infant speech perception and the development of the mental lexicon", In H. C. Nusbaum, & J. C. Goodman (Eds.), *The transition from speech sounds to spoken words: The development of speech perception.* Cambridge, MA: MIT Press.

[6] KLATT, D. H. (1989), "Review of selected models of speech perception", In W. Marslen-Wilson (Ed.), *Lexical representation and process.* Cambridge, MA: MIT Press.

[7] LIBERMAN, A. M., & MATTINGLY, I. G. (1985), "The motor theory of speech perception revised", *Cognition, 21*(1), 1-36.

[8] MILLER, J. D. (1989), "Auditory-perceptual interpretation of the vowel", *Journal of the Acoustical Society of America, 85*(5), 2114-2134.

[9] NAKATANI, L. H. & DUKES, K. D. (1977), "Locus of segmental cues for word juncture", *Journal of the Acoustical Society of America, 62*(3), 714-719.

[10] PISONI, D. B. (1977), "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops", *Journal of the Acoustical Society of America, 61*, 1352-1361.

[11] PISONI, D. B. (1982), "In defense of segmental representations in speech processing", In A. S. House (Ed.), *Project SCAMP 1981: Acoustic phonetics and speech modeling*. Princeton, NJ: Institute for Defense Analysis, Communications Research Division.

[12] PISONI, D. B., & SAWUSCH, J. R. (1975), "Some stages of processing in speech perception", In A. Cohen, & S. G. Nooteboom (Eds.), *Structure and process in speech perception*. New York: Springer-Verlag.

[13] REPP, B. H. (1982), "Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception", *Psychological Bulletin, 92*(1), 81-110.

[14] RICHARDSON, K. H. & SAWUSCH, J. R. (1989), "The perception of stop consonants: Spectral tilt revisited", Paper presented at the 117th meeting of the Acoustical Society of America, Syracuse, NY.

[15] SAWUSCH, J. R. (1986), "Auditory and phonetic coding of speech", In E. C. Schwab, & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines, Vol. 1*. New York: Academic Press.

[16] SAWUSCH, J. R., & DUTTON, D. L. (in press), "Computational metrics for place of articulation information in consonants and vowels", In J. Charles-Luce, P. Luce, & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production and development*. Norwood, NJ: Ablex.

[17] STUDDERT-KENNEDY, M. (1976), "Speech perception", In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics*. New York: Academic Press.

[18] SYRDAL, A. K., & GOPAL, H. S. (1986), "A perceptual model of vowel recognition based on the auditory representation of American English vowels", *Journal of the Acoustical Society of America, 79*(4), 1086-1100.

[19] WANNAMACHER, C. A. & SAWUSCH, J. R. (1989), "The role of phonological permissibility in the phonetic coding of speech", Paper presented at the 118th meeting of the Acoustical Society of America, St. Louis, MO.

7. ACKNOWLEDGMENTS

I would like to thank Dawn Dutton, Deborah Gagnon, Laurie Garrison, Kevin Richardson and Carol Wannamacher for their helpful discussion and critique of the ideas presented here. This research was supported by NIH grant R01DC00219 to the State University of New York at Buffalo.