Part IV Language, Speech and Hearing

Section 1 Speech Communication

Chapter 1. Speech Communication

Sponsors

Apple Computer, Inc.
C.J. Lebel Fellowship
National Institutes of Health (Grants T32-NS07040, R01-NS04332, R01-NS21183,[1] and P01-NS23734[2])
U.S. Navy / Naval Electronic Systems Command (Contract N00039-85-C-0254)
U.S. Navy - Office of Naval Research (Contract N00014-82-K-0727)

Academic and Research Staff

Professor Kenneth N. Stevens, Professor Jonathan Allen, Professor Morris Halle, Professor Samuel J. Keyser, Dr. Corine Bickley, Dr. Suzanne Boyce, Dr. Carol Chapin Ringo, Dr. Carol Y. Espy-Wilson, Dr. James R. Glass, Dr. Dennis H. Klatt, Dr. Sharon Manuel, Dr. Joseph S. Perkell, Michael Phillips, Dr. Stephanie Seneff, Dr. Stephanie Shattuck-Hufnagel, Dr. Mario A. Svirsky, Dr. Victor W. Zue

Collaborating Scientists

Dr. Xavier Furtado,[3] Dr. Richard S. Goldhor,[4] Dr. Robert E. Hillman,[5] Dr. Tatsuya Hirahara,[6] Eva B. Holmberg,[7] Dr. Haruko Kawasaki,[8] Dr. Harlan Lane,[9] Dr. Leah S. Larkey,[10] Dr. John Locke,[11] Dr. John I. Makhoul,[12] Dr. Shigeru Ono,[13] Dr. Noriko Suzuki,[14] Kasuya Takeda,[15] Jane Webster[16]

---

1 Under subcontract to Boston University.

2 Under subcontract to Massachusetts Eye and Ear Infirmary.

3 Visiting Scientist, Tata Institute of Fundamental Research, Bombay, India.

4 Sensimetrics Corporation.

5 Boston University.

6 Visiting Scientist, Advanced Telecommunications Research, Osaka, Japan.

7 MIT Staff Member and Research Scientist in Department of Speech Disorders, Boston University.

8 Voice Processing Corporation, Osaka, Japan.

9 Northeastern University.

10 Kurzweil Applied Intelligence.

11 Massachusetts General Hospital.

12 Bolt Beranek and Newman, Inc.

13 NEC, Kawasaki, Japan.

14 Visiting Scientist, Showa University, Tokyo, Japan.

15 Visiting Scientist, Advanced Telecommunications Research, Osaka, Japan.

16 Massachusetts Eye and Ear Infirmary.

Graduate Students

Abeer Alwan, Nancy Daly, Wil Howitt, Caroline Huang, Rob Kassel, Lori Lamel, Hong Leung, Jeff Marcus, Helen Meng, John Pitrelli, Mark Randolph, Lorin Wilde

Undergraduate Students

Peter Ihionu, Charles Jankowski, Tunay Kuru, Haruko Mitra, Christine Pao, Andy Shaw, Dave Whitney

Technical and Support Staff

Ann Forestell, Seth M. Hall, Katy Isaacs, Keith North, Vicky Palay

Part-Time Assistants/Special Projects

Katie Abramson, Nicholas P.T. Bateman, Laura Glicksman, Fred G. Kennedy, Joan Matelli, Hope Menard, Michael K. McCandless, Davin Wong

1.1 Processes and Representations Involved in Speech Planning

In earlier experiments we have explored two separate aspects of speech planning: 1) what can be learned from speech error patterns (particularly those that involve segmental errors) about the units and structures that speakers use in planning utterances, and 2) the nature of prosodic structures and the acoustic-phonetic correlates of these structures. While continuing to pursue both of these topics separately, we have begun to investigate their interaction. That is, on the assumption that speech production involves the integration of prosodic planning with segmental planning, we are focusing on how the prosodic structure of an utterance affects speech error patterns.

1.1.1 Segmental Errors and Phonemic Context Similarity

Error elicitation experiments comparing the error rate for consonants in words with identical vowels (*pat fan*) vs. different vowels (*pat fin*) shows that shared vowel context results in a 50 percent higher error rate for consonantal segments. If this increase results because vowels define the planning frame into which consonantal information must be integrated, we should not find the same effect for shared consonantal context on vowel error rate. That is, /æ/ and /I/ should participate in errors about equally often for *pat fin* and *pat pin*. Experiments are now under way to test this hypothesis.

1.1.2 Prosodic Structures and Stress Shift

The phenomenon of stress shift (in which stress moves to the left in polysyllabic words like *thirteen* or *Japanese*, when they appear in phrases like *thirteen men* or *Japanese businessman*) has been attributed to stress clash (i.e., the presence of a strong syllable immediately on the right). Since many speakers of English do not agree that this shift occurs in their speech, we have begun to document its rate of occurrence, its acoustic-phonetic correlates, and the structural contexts that condition it in a number of speakers. Preliminary results suggest that stress shift is at least partially governed, not by local stress clash, but by more global aspects of prosodic phrasing.

1.1.3 Segmental Errors, Morpheme Position and Prosodic Structure

Final Position Protection Effect

Error elicitation experiments comparing CVC monosyllables in lists (*leap note nap lute*) vs. phrases (*from the leap of the note to the nap of the lute*) showed that consonants in word-final position are protected against errors in the phrasal condition, but not in the list condition. These results suggest that the

more complex structure of the phrasal utterances requires the speaker to formulate a more detailed representation, perhaps in terms of the onsets and rhymes of syllables, that protects final consonants while leaving the word-initial consonants vulnerable to errors. Further experiments are under way to determine whether it is the syntactic or the prosodic structure of the phrasal condition that is responsible.

Morpheme Position vs. Stress Position

Elicitation experiments as well as naturally-occurring speech errors have shown that word-onset prestressed consonants are particularly susceptible to errors. In a series of experiments that separated word-onset from pre-stress position (e.g., comparing the rate of p/f errors in *parade fad* vs. *repeat fad*), we have shown that this susceptibility is greater for pairs of word-onset consonants than for pairs of pre-stressed consonants. These initial results have now been replicated using a second set of words arranged in a different prosodic pattern, suggesting strongly that morpheme structure as well as prosodic structure plays a role in segmental planning representations at the point where segmental interaction errors arise.

1.2 Speech Physiology

1.2.1 Quantal Nature of Speech in the Production of the Vowels /i/, /a/ and /u/

Hypotheses about quantal articulatory targets for the vowels /i/, /a/ and /u/ were examined in light of some previously-published[17] and new data. The hypotheses predicted that for multiple repetitions of each vowel, a point on the tongue dorsum near the place of maximum constriction for the vowel should show more precise positioning (less scatter among repetitions) in a direction normal to the vocal-tract midline than tangent to the midline. Data were in the form of displacements of points on the tongue dorsum at the time of minimum tangential velocity during the vowel (the "articulatory target"), as pronounced in a variety of nonsense utterances. The previously-published data, from an x-ray microbeam facility, based on two speakers of American English, supported the hypotheses for /i/ and /a/; however, they contained context effects and did not include /u/. The new data, on a single additional speaker of American English, were obtained with an alternating magnetic field movement transducer system. Utterances were chosen to minimize the effects of context and included the vowel /u/. The new data also agree with the hypotheses, lending additional indirect support to certain predictions about the quantal nature of speech production.

1.2.2 Articulatory Movement Transduction

Work has been completed on further refinement of techniques for transduction of articulatory movements with the use of alternating magnetic fields. Additional testing devices have been constructed and the performance of a two-transmitter and of a three-transmitter system has been compared. We conclude that with careful mounting of transducers close to the midline and checking the alignment at the end of the experiment, it will be possible to obtain useful data with the three-transmitter system. Because of a number of advantages of the three-transmitter system (including ease of use, less intense magnetic fields, and less costly transducers), we have decided to use our latest implementation of the three-transmitter system. Final electronics changes have been completed, re-calibration is in progress, and new single-axis transducers have been ordered. A new application for the use of humans as experimental subjects is being prepared.

---

[17] *J. Acoust. Soc. Am.*, 77:1889-1895.

1.2.3 Glottal Airflow and Transglottal Air Pressure Measurements with Changes in Fundamental Frequency

Measurements on the inverse filtered airflow waveform and of average transglottal pressure and glottal airflow were made from syllable sequences in low, normal and high pitch for 25 male and 20 female speakers. Correlation analyses indicate that several of the airflow measurements are more directly related to voice intensity than to fundamental frequency. Results suggest that air pressure may have different influences in low and high pitch in the speech task which was used. An unexpected finding was that of increased pressure in low pitch. It is suggested that this result is related to maintaining voice quality, that is, avoiding vocal fry. In high pitch, increased pressure may serve to maintain vocal fold vibration. The findings suggest different underlying mechanisms and vocal adjustments for increasing versus decreasing F0 from normal pitch.

1.2.4 Phonatory Function Associated with Vocal Fold Lesions

A combination of non-invasive acoustic and aerodynamic measures were used to objectively compare and contrast vocal function among patients having one of four major types of vocal-fold lesions: nodules, polyps, polypoid degeneration and contact ulcers. Measures included: 1) vocal intensity and fundamental frequency, 2) parameters derived from inverse filtered approximations of the glottal airflow waveform, and 3) estimates of transglottal pressure and average air flow. Preliminary results show that a combination of measures of fundamental frequency, transglottal pressure and maximum airflow declination rate appear to differentiate among the types of lesions studied. These results are interpreted as evidence that different underlying phonatory mechanisms are associated with different types of hyperfunctionally-related vocal fold lesions.

1.3 Speech Recognition

The overall objectives of our research in machine recognition of speech are:

1. To carry out research aimed at collecting, quantifying, and organizing acoustic-phonetic knowledge, and
2. To develop techniques for incorporating such knowledge, as well as other relevant linguistic knowledge, into speech recognition systems.

During the past year, progress has been made on several projects related to these broad objectives.

1.3.1 Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition

Phonetic recognition can be viewed as a process through which the acoustic signal is mapped to a set of phonological units used to represent a lexicon. Traditionally, researchers often prescribe an intermediate, phonetic description to account for coarticulation. This research investigates an alternative approach whereby this phonetic-level description is bypassed in favor of directly relating the acoustic realizations to the underlying phonemic forms. In this approach, the speech signal is transformed into a set of segments which are described completely in acoustic terms. Next, these acoustic segments are related to the phonemes by a grammar which is determined using automated procedures operating on a set of training data. Thus important acoustic regularities that describe contextual variations are discovered without the need to specify a set of preconceived units such as allophones.

The viability of this approach depends critically on the ability to detect important acoustic landmarks in the speech signal and describe these events in terms of an inventory of labels that captures the regularity of phonetic variations. In last year's report, we described a procedure that enables us to embed important acoustic landmarks in a multi-level structure called a dendrogram, in which information ranging from coarse to

fine is represented in a unified framework. We also described how a set of acoustic labels for these segments can be determined through a hierarchical clustering algorithm.

We have continued this line of investigation, this time focusing on relating these acoustic forms to the underlying phonemic descriptions. An analysis of the clustering algorithm on data from all phonemes indicates that it is possible to assign an acoustic segment to one of a small set of acoustic categories, each having a meaningful phonetic interpretation. An examination of the realizations of weak voiced fricatives and velar stop consonants indicates that it is possible to determine a finite number of regular acoustic forms which capture consistent contextual dependencies, some of which have not been described previously by phoneticians. Additionally, there is evidence that these regularities can generalize across sets of phonemes with similar phonetic features.

1.3.2 Phoneme Duration Models for Continuous Speech Recognition

Durations of speech sounds often serve as a primary perceptual cue for a number of distinctions necessary to convey some of the linguistic content of an utterance. Examples of these distinctions include voiced versus voiceless consonants, phrase-final versus non-phrase-final syllables, and stressed versus unstressed or reduced vowels. However, the abundance of factors that affect segment duration and our lack of understanding of their interactions have made it difficult to utilize durational cues for speech recognition. The goal of this study is to develop an understanding of duration factors, such as stress, speaking rate and cluster reduction and to model their effects and interactions in such a way as to facilitate the use of duration cues for speech recognition.

Our preliminary experiments have shown that many of the factor interactions are sufficiently irregular that they are not modeled accurately by the traditional additive and multiplicative models described in the literature. We are therefore developing a hierarchically-structured model for the discrete-valued factors such as phonetic context and sentence-final lengthening and the interactions among these factors. The modelling process is helped greatly by the availability of thousands of phonetically labelled sentences.

The models that we have developed so far have reduced standard deviations to approximately 30 ms for vowels and 25 ms for consonants in multiple-speaker continuous speech databases, compared to 50 ms for vowels and 30 to 35 ms for consonants for the one-prediction-per-phoneme model. Because our standard errors are approaching the just-noticeable differences for durational perception for phonemes, we are approaching the point where modelling is sufficiently accurate to contribute to duration-based phonetic distinctions.

1.3.3 Phonetic Classification From Automatically Discovered Acoustic Attributes

Past attempts at phonetic classification often make use of parameters such as formant frequencies and other spectral and temporal measurements. While these parameters are clearly useful, their extraction is prone to error, and thus can lead to poor phonetic classification results. Over the past year, we experimented with a slightly different approach, in which we first define a set of generic property detectors based on our knowledge of acoustic phonetics. To determine the optimal settings for the free parameters, we first compute the classification performance on a large set of training data for all combinations of the parameter settings. We then search for the maximum on the surface defined by the classification performance.

The parameter settings that correspond to the maximum are chosen to be the optimal settings. An attribute can also be used in conjunction with other attributes, or to derive other attributes.

We believe that the procedure described above is an example of successful knowledge engineering in which the human provides the knowledge and intuition, and the machine provides the computational power. Frequently, the settings result in a parameter

that agrees with our phonetic intuitions. Our experience with this procedure suggests that it is able to *discover* important acoustic parameters that signify phonetic contrasts, without resorting to the use of heuristic rules.

Once the attributes have been determined, they are selected through an optimization process. Classification is achieved using conventional pattern classification algorithms. To evaluate the phonetic classification results, we compared the labels provided by the classifier to those in a time-aligned transcription. We have performed the evaluation on two separate databases, as summarized in table 1. Performance was measured on a set of 38 context-independent phone labels. For a single speaker, the top-choice classification accuracy was 77 percent. The correct label is within the top three nearly 95 percent of the time. For multiple and unknown speakers, the top-choice accuracy is about 70 percent and within the top three over 90 percent of the time.

Table 1. Summary of speaker-dependent and -independent phonetic classification results

Database	No. of Training Sentences	No. of Training Speakers	No. of Test Sentences	No. of Test Speakers	Top-Choice Accuracy (%)	Top-Three Accuracy (%)
1	510	1	210	1	77	95
2	1500	300	225	45	70	90

1.4 Studies of the Acoustics, Production, and Perception of Speech Sounds

We have continued our studies of the acoustic characteristics of several different classes of speech sounds, the mechanisms of generation of these sounds, and the perception of the sounds. Our experimental work is examining both the canonical acoustic properties of these sounds in simple contexts and the modifications that are observed in fluent speech.

1.4.1 Fricative Consonants

Using a database containing a number of fricative consonants in different vowel environments, we have measured the spectra of the frication noise, the formant transitions, and the characteristics of the glottal output at the onsets and offsets of the fricatives. These data have led to the formulation of improved rules for the synthesis of fricative consonants. The rules include a specification of the time-varying characteristics of the glottal source at the vowel-fricative and fricative-vowel boundary, as well as the formant transitions and appropriately shaped envelopes of the amplitudes of the glottal and frication sources.

Studies of fricative consonants have also included further experiments and acoustic measurements aimed at delineating the mechanisms of voicing and devoicing for these consonants. (This work is being carried out in collaboration with Dr. Sheila Blumstein of Brown University.) The acoustic measurements show that formant transitions at fricative-vowel and vowel-fricative boundaries are weak or sometimes nonexistent for voiceless fricatives, but are clearly evident for voiced fricatives. Perceptual experiments have verified the importance of the transitions and of glottal vibration in 20-30 ms intervals near these boundaries for determining the voicing characteristics of fricative consonants.

Another component of our research into the nature of fricative consonants has included an experimental study of the acoustic characteristics of postalveolar fricatives of Polish. These particular fricatives are of some interest in phonetic theory, since there has been some question concerning their classification within a theoretical phonetic framework. We measured the spectra and the formant transitions of the relevant fricative consonants in a number of utterances of two Polish speakers. We used these data, together with existing

x-ray and palatographic data from the production of these sounds, to estimate the distinctive articulatory and acoustic attributes of the fricatives. Our tentative conclusion was that the two postalveolar fricatives of Polish can be classified in terms of existing features *anterior, distributed*, and *back*.

1.4.2 Nasal Consonants

Theoretical analysis and acoustic data have shown that abrupt changes of 15-20 dB in spectrum amplitude (at mid- and high-frequencies) occur at the implosion and release of nasal consonants. These discontinuities can be attributed to a rapid movement of a zero in the transfer function of the vocal-nasal tract. This movement occurs as the principal sound output shifts from the nose to the mouth.

1.4.3 Stop Consonants

Our acoustic studies of stop consonants have led to a detailed specification of the characteristics of the aspiration noise and of the glottal source at voicing onset, following the release of voiceless aspirated stop consonants. Spectral data sampled at the onset of voicing for aspirated consonants indicate a breathy component to the glottal vibration. Perceptual experiments in which these aspects of the onset are manipulated through synthesis demonstrate that this kind of modification provides cues to voicelessness of the consonant.

An investigation of oral and glottal gestures and acoustics of underlying /t/ in English has been carried out in collaboration with Eric Vatikiotis-Bateson of Haskins Laboratories. Records of the sound, the glottal adjustments, and contact patterns between tongue blade and palate show that, in syllable-final position, /t/ is often produced with glottalization. The timing between the glottal closure and the tongue-blade closure can be variable, depending on the speaker, the speaking rate, and the following phonetic context. Glottal closure frequently occurs before oral closure, and oral closure is often omitted, with the result that formant transitions at consonant offset are either weak or absent.

1.4.4 Vowels

In a continuing series of experiments concerned with the acoustics and perception of vowels, we are examining several influences of context on vowels in fluent speech, particularly the effects of selected consonantal contexts on vowel production. We are focusing on consonant contexts that appear to have an especially strong influence on vowels, such as the consonants /ℓ/, /r/ and /w/, which impose constraints on the tongue body position. The aim is to quantify these influences and to account for them in models of lexical access. A data base of utterances designed especially for this study has been developed, and measurements of vowels in this database are being collected.

1.4.5 Distinctive Features and Lexical Access

As an outgrowth of our investigations of the distinctive features and their acoustic correlates, we have been developing a model of lexical access based on features. Lexical items are specified in terms of patterns of distinctive features, expressed in a form that is somewhat modified from the conventional matrix representation. The lexicon is accessed by extraction from the stream of sound properties that bear a one-to-one relation with the distinctive features. An advantage of using features for lexical representation and access is that many (if not most) of the modifications that occur in the sound pattern of a word when it is produced in fluent speech can be characterized as shifts in just one or two features, with other features remaining unchanged. The geometrical arrangement of the patterns of features in the lexical representation, and the strategy for extracting properties from the signal, are designed to take advantage of the fact that some acoustic properties define events or landmarks in the signal, and these landmarks identify regions in the sound where other properties are to be extracted.

1.4.6 Phonetic Theory

Three papers with a theoretical orientation

have been completed in the past year.[18] One of these is a reexamination of evidence for a quantal theory of speech. This theory takes the view that the inventory of sounds in language is based in part on quantal or nonmonotonic relations between acoustic and articulatory parameters, or between auditory and acoustic parameters. The second paper explores the concept that the acoustic manifestations of particular distinctive features can be strengthened or enhanced if those features are implemented simultaneously with selected other features. The third paper is a review of several recently proposed models of speech perception, including both phonetic and feature-based models and statistically-oriented models.

1.5 Computer Facilities

The Klatt speech synthesizer (KLSYN) that is used in a number of laboratories for generating synthetic speech has been modified and enhanced in several ways. The new version (KLSYN88) includes a more flexible glottal source, with various parameters available to generate different male and female glottal waveforms, together with an additional pole-zero pair (to simulate acoustic coupling to the trachea). The capabilities of the synthesizer are now sufficiently broad that it is being applied in acoustic studies of speech using an analysis-by-synthesis paradigm. This technique allows a spoken utterance to be described in terms of the time variation of the synthesis parameters that are needed to match the utterance instant-by-instant.

For more efficient analysis of multi-channel signals from speech production experiments, our two VaxStation II/GPX workstations have been upgraded with faster processors. They have been combined into a functional system which includes A/D, D/A, mass storage, hard copy, a signal interface panel, a new version of the MITSYN signal processing languages, and software for data management and statistical analysis. A number of procedures for signal processing and data analysis have been implemented on the new facility. Text processing and E-mail functions, formerly performed on a DEC System 20, have been implemented on two new VaxStation 8-user workstations, with attached disks, laser printers, terminal servers and terminals. The facility is being augmented with several IBM PS/2 personal computers obtained through a grant to RLE from IBM for use in statistical and graphical data analysis, bibliographic data management and text processing. All of these components are being integrated into a Local Area Vax Cluster to provide a uniform environment for experimental control, signal processing, data analysis, communications, and manuscript preparation.

---

[18] D.H. Klatt, "Review of Selected Models of Speech Perception," in *Lexical Representation and Process,* ed. W.E. Marslen-Wilson, MIT Press, in press; K.N. Stevens, "On the Quantal Nature of Speech," *J. Phonetics,* in press; K.N. Stevens and S.J. Keyser, "Primary Features and their Enhancement in Consonants," *Language* 65 (1):81-106 (1989).