# 16.0 Speech Communication

### Academic and Research Staff

Prof. K.N. Stevens, Prof. J. Allen, Prof. M. Halle, Prof. S.J. Keyser, Prof. V.W. Zue, Dr. T. Carrell, Dr. C. Chapin Ringo, M. Cohen, Dr. B. Delgutte, Dr. R. Goldhor,[1] Dr. B. Greene,[2] Dr. R.E. Hillman,[3] E.B. Holmberg,[4] D.H. Kaufman, Dr. H. Kawasaki,[5] Dr. D.H. Klatt, Dr. L.S. Larkey ,[6] N. Lauritzen, Dr. J. Locke,[7] Dr. J.I. Makhoul[11], Dr. S. Manuel, Dr. P. Menyuk ,[8] Dr. J.L. Miller,[9] Dr. J.S. Perkell, Dr. D. Pisoni,[10] Dr. P.J. Price,[11] Dr. S. Seneff, Dr. S. Shattuck-Hufnagel, Dr. L. Wheeler, K. Yoshida,[12]

### Graduate Students

A. Alwan, M. Anderson, C. Bickley, N. Daly, S. Dubois, C. Espy-Wilson, J. Glass, C. Huang, R. Kassel, L. Lamel, H. Leung, L. Pastel, J. Pitrelli, M. Randolph, T. Wilson

*C.J. Lebel Fellowship*
*National Institutes of Health (Grants 5 T32 NS07040 and 5 R01 NS04332)*
*National Science Foundation (Grant 1ST 80-17599)*
*U.S. Navy - Naval Electronic Systems Command Contracts (N00039-85-C-0254, N00039-85-C-0341, N00039-85-C-0290)*

## 16.1 Studies of the Acoustics and Perception of Speech Sounds

### 16.1.1 Stop and Fricative Consonants

Several studies of the production, acoustics, and perception of stop and fricative consonants are in progress. The aim of these studies is to specify the articulatory and

---

1 Staff Member, Kurtzweil Applied Intelligence

2 Visiting Scientist

3 Assistant Professor, Department of Speech Disorders, Boston University

4 Research Scientist, Department of Speech Disorders, Boston University

5 Staff Member, Voice Processing Corporation

6 Staff Member Kurtzweil Applied Intelligence

7 Director, Neurolinguistics Laboratory, Massachusetts General Hospital, Institute of Health Professions

8 Professor of Special Education, Boston University

9 Associate Professor, Department of Psychology, Northeastern University

10 Visiting Scientist

11 Staff Member, Bolt, Beranek and Newman, Inc.

12 Visiting Scientist

acoustic correlates of the features that distinguish between aspirated and unaspirated consonants, between consonants produced with different places of articulation along the vocal tract, and between voiced and voiceless cognates of these consonants. These consonants have the common property that their production involves the generation of turbulence noise at some region along the vocal tract. The sound that is radiated from the lips is dependent on the properties of the turbulence noise source in the vicinity of the constriction and on the filtering imposed on this source by the vocal-tract shape.

In one set of experiments we are examining the mechanism of production of sound when turbulence is generated in the airflow through a constriction in a tube. The dimensions of this mechanical model are selected to simulate some aspects of the vocal tract during the production of fricative consonants. We are extending the results of previous experiments by investigating a variety of configurations and orientations of the constriction and of an obstacle downstream from the constriction in order to be able to predict more precisely the location, amplitude, and spectrum of the turbulence noise source for different vocal-tract configurations. One outcome of these experiments is that the source is often distributed through a region that extends several centimeters along the vocal tract, and that there are variations in the spectrum for the components of the source located at different places within the region.

Another set of experiments consists of acoustic and perceptual studies of aspiration, i.e., the sound produced in the vicinity of the glottis when the glottis is spread. One finding is that the onset of voicing following an interval of aspiration shows an enhancement of the low-frequency spectrum amplitude, similar to that for breathy vowels. Preliminary perceptual experiments indicate that this characteristic of the onset of voicing influences listener judgements of the distinction between aspirated and unaspirated consonants in English.

In other experiments we have been looking at the acoustic and perceptual attributes that distinguish velar consonants from alveolars and labials. In terms of distinctive feature theory, we are examining the acoustic and perceptual correlates of the feature of compactness, or, equivalently, the feature of anteriority as commonly used in phonology. The findings indicate that compact consonants require a spectral prominence in the midfrequency range that matches approximately in frequency and amplitude to the prominence formed by the second or third formant in an adjacent vowel.

Finally, we are carrying out theoretical and experimental studies to determine the acoustic correlates of voicing for stop and fricative consonants. The theoretical analysis suggests that the initiation of voicing in these consonants requires careful coordination and control of articulatory and laryngeal movements, and that simultaneous generation of voicing and strong turbulence noise for stops and fricatives is somewhat incompatible. We are beginning the collection of acoustic data that we will attempt to interpret in terms of the theoretical model.

### 16.1.2 Vowel Perception

We are carrying out several experiments that are attempting to elucidate the mechanisms whereby human listeners process vowel sounds. Some current experiments are investigating: 1) the perceptual interpretation of spectral details in the vicinity of the

first formant; 2) the perceptual interpretation of two closely-spaced spectral prominences.

The perceptual interpretation of spectral details in the vicinity of the first formant *(F1)* remains an important unresolved research topic. Aspects of the spectrum below 1000 Hz are thought to signal: 1) tongue height, as implied by the location of the first formant frequency; 2) the distinction between oral and nasalized sounds; and 3) the contrast between normal and breathy phonation. Confounding the disentanglement of these factors is: 1) variation in the fundamental frequency of the voice (a high fundamental frequency implies that relatively few harmonic samples are available as cues to deduce the shape of the vocal-tract transfer function); 2) variation in the source spectrum due to changes in glottal state (the relative amplitudes of harmonics are affected by the rapidity of glottal closure and other details of the vibration pattern); 3) presence of tracheal resonances when the glottis is sufficiently open; and 4) source-tract interactions such as an increase in first formant bandwidth during the open phase of a period for low vowels. Research on these issues has focused on the perceptual location of the first formant frequency as fundamental frequency varies. Using carefully designed synthetic speech stimuli, it has been shown that the listener is able to recover the "true" location of *F1* in spite of harmonic/formant interactions that are such as to fool linear prediction and other energy-based methods of formant estimation. It is not easy to explain this behavior in terms of peripheral mechanisms of audition because individual harmonics are resolved under the stimulus conditions studied. Accounting for the perceptual data in terms of simple processing strategies is not yet possible, but if the central auditory system is provided with the frequencies and amplitudes of harmonics near *(F1)*, a moderately complex calculation could yield *(F1)* with minimal error.

The experiments concerned with the perceptual interpretation of closely spaced spectral prominences involve the matching of single-formant synthetic vowels against multiformant vowels with various formant locations. The results are in general agreement with previous work, which shows the match of a single formant to a vowel with two closely spaced formants tends to be located at an effective "center of gravity" of the two formants, but the conditions under which this behavior occurs are somewhat different from earlier findings. For example, for high back vowels, matching tends to be to the first formant frequency, whereas for nonhigh back vowels, the matching is intermediate between the first and second formant frequencies. The data also suggest that there is a rather abrupt shift in matching behavior as the formant pattern for the target vowel crosses a boundary between back and front vowels.

### 16.1.3 Analysis of a Female Voice

Waveforms and spectra of a set of speech materials collected from a single female speaker have been studied. This pilot effort, an attempt to determine some of the acoustic characteristics that differentiate male and female voices, may lead to improved synthesis of female voices.

Data have been examined on breathy onsets, breathy offsets, glottalized onsets, glottalized offsets, and vowel spectra over extended intervals involving large changes to fundamental frequency. Harmonic spectra reveal a significant breathy (noise) component in the mid and high-frequency region of the spectrum. Locations of spectral zeros are consistent with inferences that the open quotient remains at about 50% over

much of the data, but decreases for glottalized sounds and increases in the vicinity of a voiceless consonant or breathy offset. Turbulence spectra for [h], measured with a long-duration window to minimize statistical variation, indicate the presence of extra poles and zeros at relatively fixed locations across vowels. These probably reflect acoustic coupling to tracheal resonances. Glottal attacks are often accompanied by a brief interval of turbulence noise - presumably a manifestation of pharyngeal frication.

### 16.1.4 Alternative Analysis Procedures

Theoretical and experimental studies of the acoustic properties at the release of stop and nasal consonants show that rapid spectral changes occurring over time intervals of a few milliseconds can carry information about some of the features of these consonants. Conventional spectral analysis tools are often unable to resolve some of the details of these rapid changes. We have been exploring whether application of the Wigner distribution to these signals can provide insight into acoustic attributes at the consonant release that are not well represented by conventional methods. Procedures for calculating and displaying the Wigner distribution have been implemented, and comparisons between this representation and other spectrographic representations for a variety of speech and speechlike sounds are being made.

### 16.1.5 Children's Speech

Studies of the properties of the speech of children in the age range of one to three years old show substantial differences from the speech of adults. Some of these differences can be attributed to differences in dimensions and configurations of airways and mechanical structures of the larynx and vocal tract, and to differences in motor control capabilities. We are examining how these differences impose constraints on the sounds produced by children. Among the topics being studied through theoretical analysis and through acoustic measurements are: 1) the influence of the reduced dimensions of the airways on the spectral characteristics of children's utterances; 2) the effect of the substantial differences in dimensions of the laryngeal structures on vocal-fold vibration; and 3) the constraints imposed on temporal properties of children's speech by their reduced respiratory dimensions and ranges and their ability to control and coordinate rapid movements of articulatory and other structures.

## 16.2 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.

## 16.2.1 Signal Representation for Acoustic Segmentation

The task of phonetic recognition can be stated broadly as the determination of the mapping of the acoustic signal to a set of phonological units (e.g., distinctive feature bundles, phonemes, or syllables) used to represent the lexicon. In order to perform such a mapping, it is often desirable to first transform the *continuous* speech signal into a *discrete* set of acoustic segments. During the past year, we have explored a number of alternative acoustic segmentation algorithms, and have compared the results based on several signal representations.

The objective of the segmentation algorithm is to establish stable acoustic regions for further phonetic analysis. One of the more promising segmentation algorithms adopts the strategy of measuring the similarity of a given spectral frame to its immediate neighbors. The algorithm moves on a frame-by-frame basis, from left to right, and attempts to associate a given frame with its immediate past or future, subject to a similarity measure. Acoustic boundaries are marked whenever the spectral vector changes affiliation from past to future. The algorithm makes use of the fact that certain acoustic changes are more significant than others and that the criteria for boundary detection often change as a function of context. It can self-adapt to capture short regions that are acoustically distinct. In addition, the algorithm's *sensitivity* to acoustic changes can also be controlled so that the resulting acoustic description can be as broad or as detailed as is desired.

Using this acoustic segmentation algorithm, we performed a set of experiments exploring the relative merits of five different spectral representations for acoustic segmentation. The five spectral representations are as follows:

1. **Wideband**: The spectral vector is obtained by applying a 6.7-ms Hamming window to the speech waveform.

2. **Smoothed narrowband**: The spectral vector is obtained by applying a 25.6-ms Hamming window to the speech waveform, followed by smoothing with a 3-ms window in the cepstral domain.

3. **LPC**: The spectral vector is obtained from a 19th-order LPC analysis on a 25.6-ms segment of speech.

4. **Critical band**: The spectral vector is obtained from the first stage of an auditory model and represents the outputs of a bank of critical-band filters.

5. **Hair cell**: The spectral vector is obtained from intermediate outputs of the same auditory model and represents the outputs of the hair-cell/synapse transduction stage. The envelope response of the filter outputs corresponds to the mean-rate response of neural firing.

For each spectral representation, a 39-dimensional spectral vector is computed once every 5 ms. The array of spectral vectors is the only information used for acoustic segmentation. To evaluate the effectiveness of the spectral representations for acoustic segmentation, the sentences are transcribed phonetically and the transcriptions are time-aligned with acoustic landmarks. For each experiment, the output of the acoustic segmentation is compared with the hand transcription, and the numbers of extra and missed boundaries (*insertions* and *deletions*) are tabulated. By adjusting the sensitivity

parameters, one can bias the algorithm to favor segment insertion or deletion. Using a database of 1,000 sentences spoken by 100 talkers, we tested each spectral representation in more than 24 experiments covering a wide range of insertion and deletion errors. The "best" result for each spectral representation was defined to be the one that minimizes the *sum* of the number of inserted and deleted segments. A comparison of the results shows that the hair-cell spectral representation is superior (with the lowest total insertion and deletion rate - 25%), followed closely by the critical-band and LPC representations. These representations were consistently better than the discrete Fourier transform representations, by 3% to 4% on average. We view the results as lending support to the speculation that signal representation based on auditory modeling can potentially benefit phonetic recognition.

### 16.2.2 Acoustic Evidence for the Syllable as a Phonological Unit

Phonetic recognition is difficult partly because the contextual variations of the acoustic properties of speech sounds are still poorly understood. Traditionally, such systematic variations have been described in the form of context-sensitive rules. More recently, it has been suggested that rules that make reference only to the local phonemic environment may be inadequate for describing allophonic variations. Instead, one may have to utilize larger phonological units such as the syllable in order to describe such regularities. While evidence in support of the syllable as a relevant unit in the formulation of acoustic-phonetic rules has come from diverse sources, direct acoustic evidence of its existence has been scarce. During the past year, we have conducted a durational study of stop consonants, with the goal of determining the possible role of syllable structure on their realizations.

Our database consisted of some 5,200 stops collected from 1,000 sentences. Phonemic transcriptions, including lexical stress and syllable markers, were provided and aligned with the speech waveforms. The stops were categorized according to their position within syllables (e.g., *syllable-initial-singleton, syllable-final-affix,* etc.) and marked according to their local phonemic context. Segment durations were measured and the stops were classified as released, unreleased, or deleted on the basis of their duration and voice onset time (VOT). In the analysis of these data, including the examination of VOT and other durational measurements, we found substantial effects due to syllable structure. For example, the probability of a stop being released, unreleased, or deleted is largely determined by its position in a syllable template. Even if released, the distributions of the VOT differ substantially depending again on the position of the stop within a syllable.

Our plans are to continue these experiments by completing our investigation of the stop consonants and expanding to other classes of sounds. Eventually we would like to develop a computational framework that incorporates contextual knowledge in phonemic decoding.

## 16.3 Speech Synthesis

An extensive manuscript reviewing text-to-speech conversion for English has been submitted and accepted for publication in the Journal of the Acoustical Society of America. The paper traces the history of research, the nature of the scientific problems,

some solutions developed for the Klattalk system (a text-to-speech system developed at MIT several years ago), and areas where future research is needed. A companion tape recording provides examples of 30 milestones in the development of synthesis capabilities.

## 16.4 Speech Planning

We have continued our study of the phonological planning process for speech production using speech error analysis as a tool, and have expanded the scope of this investigation to include the analysis of acoustic phonetic correlates of lexical and phrasal stress or prominence.

### 16.4.1 Error Studies

Earlier work has shown that when single-consonant speech errors occur, they are more strongly influenced by shared word position than by shared stress or syllable position. Word-onset consonants are particularly susceptible to interaction errors, as in "sissle theeds" for "thistle seeds." Patterns of errors elicited by tongue twisters like "parade fad foot parole" and "repair fad foot repeat" confirm this finding, for both word-list and phrasal stimuli. We are currently extending the elicitation experiments to determine: 1) whether an earlier finding that phrasal stimuli protect word-final consonants against errors is due to the syntactic or prosodic structure of the phrases, or to some interaction between these two effects; 2) whether the similarity of adjacent vowels influences the probability that two consonants will interact (i.e., will there be more /t/-/k/ errors for "tan cats" than for "ten kits"); and 3) whether different types of single-segment errors are governed by different constraints and therefore can be presumed to occur at different points during the planning process (e.g., exchanges like "boting vooths" for "voting booths" appear to be more sharply limited to word-onset positions than do anticipatory substitutions like "boting booths."

### 16.4.2 Prosody Studies

The phenomenon of Stress Shift has been described as the leftward movement of lexical stress near the end of a polysyllabic word, to prevent stress clash with the following word. For example, missisSIPpi become MISsissippi MUD, for some speakers, and siaMESE becomes SIamese CATS. In the early stages of a study designed to determine what proportion of speakers shift their stress, when they do so, and what the acoustic correlates are, it has become clear that the data suggest quite different hypotheses about the nature of this apparent shift. While the study is still in progress, we can state these hypotheses as follows: 1) the large pitch change that has sometimes been associated with lexical stress in the literature is in fact associated with phrasal prominence instead; 2) Stress Shift is what occurs when phrasal prominence is not placed on a content word, and reflects the lack of a large pitch change on the lexically-stressed syllable (rather than an actual shift in the location of that pitch change to an earlier syllable, as sometimes claimed); and 3) failure to place the phrasal pitch change on the lexically stressed syllable of a word occurs in environments where there is NO stressed word following, for many speakers. (For example, the syllable "-ra-" may have no substantial pitch rise in "TRANSport operations"; similarly, there may be no substantial pitch rise on any of the syllables in "mississippi mud" in the utterance "But

I HATE mississippi mud.") These observations cast some doubt on the hypothesis that Stress Shift occurs to prevent stress clash with a following syllable.

Taken together, these initial formulations suggest that the survey of acoustic-phonetic correlates of lexical and phrasal prominence now under way may reveal distinctions between these two kinds of prosodic phenomena. These distinctions will be useful in evaluating models of the production planning process.

### 16.4.3 Cross-Language Studies

For both the speech error and prosody studies, we are beginning to extend our investigations into other languages, in a series of collaborative studies.

## 16.5 Physiology of Speech Production

### 16.5.1 Articulatory Movement Transduction

Work has been completed on an alternating magnetic field system for transducing midsagittal-plane movements of the tongue, lips, mandible and velum. The system consists of two transmitter coils mounted in an assembly that fits on the head and a number of bi-axial transducer-receiver coils that are mounted on articulatory structures, with fine lead wires that connect to receiver electronics. Output voltages from each receiver are digitized and converted to Cartesian coordinates using algorithms implemented with the MITSYN signal processing languages. The system was tested extensively and was found to meet design specifications. It tracks as many as 9 points (2 fixed---for a maxillary frame of reference-and 7 movable) with a resolution of better than .5 mm at a bandwidth from DC to 100 Hz. This performance is maintained: in the presence of dental fillings, up to 30 degrees of transducer tilt, several degrees of transducer "twist" and with off-midsagittal plane transducer placements of up to .5 cm. Three experiments have been run with human subjects, demonstrating that the system can be used to gather extensive data on articulatory movements. This work is reported in R.L.E. Technical Report No. 512.

The system does have several disadvantages which make it cumbersome to use; therefore, before embarking on an extensive series of experiments, we are exploring an alternative, theoretically advantageous three-transmitter design. Prior testing of a three-transmitter system had shown it to be less accurate than the-two transmitter system, most likely because of asymmetries in the magnetic fields generated by the transmitters. The field geometry of the original transmitters has now been explored in detail, documenting its asymmetrical nature. A computer simulation of the three-transmitter system has been implemented, allowing us to optimize the system design. The behavior of the simulation suggests that with symmetrical fields, the three-transmitter system would equal the performance of the two-transmitter system, with much greater ease of use. A transmitter which should generate symmetrical fields has been designed and is under construction. A modest amount of further testing will enable us to choose between the two systems, allowing us to begin extensive data gathering in the near future.

### 16.5.2 Data Processing

Two Digital Equipment Corporation VAX Station engineering workstations have been acquired for physiological data processing and analysis. Peripherals are being acquired for multichannel A/D and D/A conversion, data storage and graphical hard copy. An updated version of the MITSYN signal processing languages is being implemented (under subcontract) for the VMS operating system, and software for graphical and statistical analysis of data is on order. This new facility will make it possible to efficiently digitize and analyze the large number of channels of data generated by the alternating magnetic field movement transducer system.

### 16.5.3 Token-to-Token Variation of Tongue-Body Vowel Targets, Coarticulation and\Articulatory-to-Acoustic Relationships

Using the new alternating magnetic field movement transducer system, we have examined token-to-token variation of vowel targets for a midsagittal point on the tongue body of a single speaker of American English. The subject pronounced multiple repetitions of nonsense utterances of the form /bV1CV2b /, in which V = /i/,/u/,/a/ and C = /b/,/h/, with stress on the second syllable. For each vowel (V1 and V2) in each environment, a scatter plot of articulatory "target" locations (at the time of minimum tangential velocity) in the midsagittal plane was generated from all tokens having the same context. In general, the vowel-target distribution for multiple repetitions of the same utterance is elongated. When V1 and V2 are different from one another, the long axis of the distribution for one vowel in the utterance points toward the location of the target for the other vowel, providing a "statistically-based" demonstration of context-dependence of articulatory targets. When V1 = V2 the long axis is approximately parallel to the vocal tract midline at the place of maximum constriction for the vowel, suggesting that movement to the vowel target location is sensitive to the differential effects on vowel acoustics of change in degree of constriction vs. change in constriction location.

### 16.5.4 Anticipatory Coarticulation: Studies of Lip Protrusion

Work has continued on testing a "hybrid model of anticipatory coarticulation" in which gesture onset times and spatial characteristics are context-dependent, and there is "co-production", i.e., overlapping and summation of multiple influences on articulatory trajectories (see R.L.E. Progress Report No. 128). Lip protrusion movements and the acoustic signal for the vowel /u/ (embedded in carrier phrases) have been recorded from four speakers of American English, and plots have been generated in which movement events for multiple individual tokens can be examined in relation to interactively-determined times of acoustic events (sound segment boundaries). Initial qualitative examination of trajectories indicates a considerable amount of token-to-token variation in lip protrusion movements for most utterances. Movement event times and durations of various articulatory and acoustic intervals are currently being analyzed statistically to test the hybrid model.