Part V Language, Speech and Hearing

Section 1 Speech Communication

Chapter 1. Speech Communication

Academic and Research Staff

Professor Kenneth N. Stevens, Professor Jonathan Allen, Professor Morris Halle, Professor Samuel J. Keyser, Dr. Krishna K. Govindarajan, Dr. David W. Gow, Dr. Helen M. Hanson, Dr. Joseph S. Perkell, Dr. Stefanie Shattuck-Hufnagel, Dr. Alice Turk, Dr. Reiner Wilhelms-Tricarico, Peter C. Guiod, Seth M. Hall, Glenn S. LePrell, Jennell C. Vick, Majid Zandipour

Visiting Scientists and Research Affiliates

Dr. Shyam S. Agrawal,[1] Dr. Corine A. Bickley, Dr. Suzanne E. Boyce,[2] Dr. Limin Du,[3] Dr. Anna Esposito,[4] Dr. Carol Espy-Wilson,[2] Astrid Hagen,[5] Dr. Robert E. Hillman,[6] Dr. Eva B. Holmberg,[7] Dr. Caroline Huang,[8] Dr. Harlan Lane,[9] Dr. John I. Makhoul,[10] Dr. Sharon Y. Manuel, Dr. Melanie L. Matthies,[2] Dr. Richard S. McGowan,[11] Dr. Pascal H. Perrier,[12] Dr. Yingyong Qi,[13] Dr. Lorin F. Wilde,[14] Dr. David R. Williams,[11] Karen Chenausky,[15] Jane W. Wozniak[2]

Graduate Students

Marilyn Y. Chen, Harold Cheyne, Jeung-Yoon Choi, Michael Harms, Mark A. Hasegawa-Johnson, David M. Horowitz, Hong-Kwang J. Kuo, Kelly L. Poort, Janet L. Slifka, Jason L. Smith, Walter Sun

Undergraduate Students

Barbara B. Barotti, Howard Cheng, Erika S. Chuang, Laura C. Dilley, Emily J. Hanna, Dameon Harrell, Mark D. Knobel, Genevieve Lada, Adrian D. Perez, Adrienne M. Prahler, Hemant Tanaja

Technical and Support Staff

Arlene E. Wint

---

1 CEERI Centre, CSIR Comlex, New Delhi, India.

2 Boston University, Boston, Massachusetts.

3 Institute of Acoustics, Chinese Academy of Sciences, Beijing, China.

4 International Institute for Advanced Scientific Studies SA (IIASS), Italy.

5 University of Erlangen-Nürnberg, Erlangen, Germany.

6 Massachusetts Eye and Ear Infirmary, Boston, Massachusetts.

7 MIT staff member and Massachusetts Eye and Ear Infirmary, Boston, Massachusetts.

8 Altech Inc., Cambridge, Massachusetts, and Boston University, Boston, Massachusetts.

9 Department of Psychology, Northeastern University, Boston, Massachusetts.

10 Bolt, Beranek and Newman, Inc., Cambridge, Massachusetts.

11 Sensimetrics Corporation, Cambridge, Massachusetts.

12 Institut de la Communication Parlée, Grenoble, France.

13 Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona.

14 Pure Speech, Inc., Cambridge, Massachusetts.

15 Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts.

Sponsors

C.J. Lebel Fellowship
Dennis Klatt Memorial Fund
National Institutes of Health
Grant R01-DC00075
Grant R01-DC01291
Grant R01-DC01925
Grant R01-DC02125
Grant R01-DC02978
Grant R01-DC03007
Grant R29-DC02525
Grant F32-DC00194
Grant F32-DC00205
Grant T32-DC00038
National Science Foundation
Grant IRI 89-05249[16]
Grant IRI 93-14967[17]
Grant INT 94-21146[18]

1.1 Studies of the Acoustics, Perception, and Modeling of Speech Sounds

1.1.1 Glottal Characteristics of Male Speakers

The configuration of the vocal folds during the production of both normal and disordered speech has an influence on the voicing source waveform and, thus, affects perceived voice quality. Voice quality contains both linguistic and nonlinguistic information which listeners utilize in their efforts to understand spoken language and to recognize speakers. In clinical settings, voice quality also relays information about the health of the voice-production mechanism. The ability to glean voice quality from speech waveforms has implications for computer-based speech recognition, speaker recognition, and speech synthesis and may be of value for diagnosis or treatment in clinical settings.

Theoretical models have been used to predict how changes in vocal-fold configuration are manifested in the output speech waveform. In previous work, we used acoustic-based methods to study variations in vocal-fold configuration among female speakers (nondisordered). We found a good correlation among the acoustic parameters and perceptions of breathy voice. Preliminary evidence gathered from fiberscopic images of the vocal folds during phonation suggest that these acoustic measures may be useful for categorizing the female speakers by vocal-fold configuration. Currently, that work is being extended to male speakers (also nondisordered). Because male speakers are less likely than female speakers to have posterior glottal openings during phonation, we expect to find less parameter variation among male speakers, in addition to significant differences in mean values, when compared with the female results. The data collected thus far are consistent with the theoretical models. Males tend to have lower average values of the acoustic parameters than female speakers. Although there are individual differences among the male speakers, the variation is smaller than that among female speakers. Voicing-source characteristics may play a role in the perception of gender.

1.1.2 Vowel Amplitude Variation During Sentence Production

During the production of a sentence, there are significant changes in the amplitudes of the vowels and in the spectrum of the glottal source, depending on the degree of prominence that is placed on each syllable. The sources of such variations are of interest for the development of models of speech production, for articulatory synthesis, and for the design of methods for accounting for variability in speech recognition systems. These sources include changes in subglottal pressure, fundamental frequency, vocal-tract losses, and voicing-source spectral tilt. Through a study of both the acoustic sound pressure signal and estimates of subglottal pressure from oral pressure signals, we have examined some of the sources of variation in vowel amplitude and spectrum during sentence production.

We have found that there are individual differences, and perhaps gender differences, in the relationship between subglottal pressure and sound pressure level, and in the degree of vowel amplitude contrast between different types of syllables. Such differences may contribute to the perception of a speaker's individuality of gender. However, an overall trend can also be noted and may be useful for

---

[16] Under subcontract to Boston University.

[17] Under subcontract to Stanford Research Institute (SRI) Project ECU 5652.

[18] U.S.-India Cooperative Science Program.

speech synthesis: Subglottal pressure can be used to control vowel amplitude at sentence-level and main prominences (i.e., phrasal stress), but vowel amplitude of reduced and non-nuclear full vowels may be manipulated by adjustment of vocal-fold configuration.

1.1.3 The Perception of Vowel Quality

Vowel quality varies continuously in the acoustic realization of words in connected speech. In recent years, it has been claimed that vowel quality drives lexical segmentation. According to this view listeners posit a word boundary before all unreduced syllables. Moreover, it has been suggested that these unreduced syllables are perceived categorically. To address this claim, synthetic speech continua were generated between schwa and four unreduced vowels [i], [ei], [o], and [ɪ] by manipulating the values of *F1*, *F2*, and *F3*. Identification studies employing these continua have provided mixed evidence for a clear categorical function for vowel quality classification. However, analysis of reaction time (RT) data collected during this task suggests that the perception of tokens that lie between categories requires additional processing effort. It might be concluded, then, that this classification is not automatic. Analysis of discrimination data using these continua shows a peak in discrimination sensitivity at identification boundaries that is consistent with categorical perception. RT data from this task also show a peak in the amount of effort needed to make discriminations of tokens near the classification boundary. Additional studies are planned to determine if vowel quality perception still appears categorical in the perception of connected speech when processing resources may be limited by the competing demands of continuous word recognition and sentence processing.

1.1.4 Comparison of Acoustic Characteristics of Speech Sounds Produced by Speakers Over a 30-year Period

In collaboration with Dr. Arthur House of the Institute for Defense Analysis in Princeton, New Jersey, we have been analyzing speech recorded by three adult males on two occasions 30 years apart (1960 and 1990). The speech materials were the same for the two sessions. They consisted of a large number of bisyllabic utterances of the form the /hə' C_1VC_2/. The consonants C_1 and C_2 were usually the same, and included all the consonants of American English; the utterances contained 12 different vowels V. Almost all combinations of consonants and vowels were included.

A number of acoustic measures were made on these utterances, such as formant frequencies, fundamental frequency, properties relating to glottal source characteristics and to nasalization, spectra of sounds produced with turbulence noise, and duration characteristics. Many of these measures showed little or no changes for any of the speakers over the 30-year period. These included vowel formant frequencies, fundamental frequency, vowel and consonant durations, and characteristics of nasal vowels and consonants. In some cases, these changes were significant, but they were small. For example, there was an upward shift in fundamental frequency for all speakers, ranging from 3 to 9 Hz on average.

There were significant differences in the downward tilt of the spectrum of the glottal souce for two of the speakers, as measured by the spectrum amplitude in the third-formant region relative to the amplitude of the first harmonic. The change in this measure was 15-20 dB for these speakers. In one speaker, this change was accompanied by a 6 dB reduction (on average) in the amplitude of the first-formant (*F*1) prominence, indicating an increased *F*1 bandwidth due to incomplete glottal closure. These changes in the glottal spectrum resulted in modifications in the long-term average spectrum of the speech. All speakers showed a reduction in the spectrum of aspiration noise at high frequencies, and for one speaker there was a reduction in high frequency energy for strident fricatives.

1.1.5 Formant Regions and Relational Perception of Stop Consonants

Listeners are quite adept at restoring speech segments that have been obliterated by noise. For example, if in running speech the /š/ in "ship" is replaced by noise, listeners report hearing "ship" even though the acoustic cues for /š/ are not present. The current set of experiments investigates whether listeners can restore the first formant (*F*1) in stimuli where *F*1 is missing, based on the information present in the rest of the formant structure. Does the listener substitute a default value for *F*1 or treat the second formant (*F*2) as the effective "*F*1"?

Subjects were asked to identify synthetic consonant-vowel (CV) stimuli where *F*1 had a transition that rose into the vowel, had a flat transition, or was absent. For rising *F*1 transitions, listeners heard /b*a*/ or /d*a*/ depending on the *F*2 transition, and stimuli with a flat *F*1 transition were heard as /*a*/. For stimuli where *F*1 was absent and *F*2 had a rising transition, listeners perceived either /b*a*/ or

/dɑ/ depending on the *F3* transition; and for stimuli with *F*1 absent and a flat *F*2 transition, listeners perceived /ɑ/. This experiment suggests that *F*2 is perceived as "*F*1" when *F*1 is missing. A second experiment tried to determine if there is a frequency limit to listeners' treatment of *F*2 as "*F*1" by varying *F*2 in the missing *F*1 stimulus. The results from the second experiment imply that listeners treat *F*2 as "*F*1" when *F*2 is below approximately 1350 Hz, but treat *F*2 as "*F*2" for *F*2 greater than 1350 Hz.

The results from these experiments suggest that there is a region where listeners expect *F*1 to exist, and they will perceive "*F*1" within this region even for unusually high values of "*F*1". Thus, for stimuli with missing *F*1, listeners treat *F*2 as "*F*1" up to 1350 Hz. However, once *F*2 crosses this boundary, listeners treat *F*2 as "*F*2", suggesting an implicit value for "*F*1".

1.1.6 MEG Studies of Vowel Processing in Auditory Cortex

We are collaborating with researchers at MIT and the University of California at San Francisco in a brain imaging study to determine which attributes of vowels give rise to distinctive responses in auditory cortex. The technique that is employed is magneto-encephalography (MEG), a method that uses passive, non-invasive measurements of magnetic fields outside the head to infer the localization and timing of neural activity in the brain. One measure derived from MEG is the M100 latency—the first peak in the magnetic activity after the onset of the stimulus, which usually occurs near 100 ms. Previous studies using three-formant vowels showed that the M100 latency varied as a function of vowel identity and is not correlated with the fundamental frequency (*F*0).

To simplify the issue of which components influence the M100 latency, this experiment studied the M100 latency for single-formant synthetic vowels, /ɑ/ and /u/, for both a male *F*0 (100 Hz) and a female *F*0 (170 Hz). The single formant corresponds to *F*1. This M100 latency was compared to the latency determined for the three-formant vowels and to the latency for two-tone complexes that matched *F*0 and *F*1 in the single formant vowels. The latencies across the three conditions showed similar trends. The latency varied with *F*1 (or the tone corresponding to *F*1) and not *F*0 (or the tone corresponding to *F*0). The similarity to the three-formant stimuli suggests that the most prominent formant, *F*1, is the primary cause of the M100 latency.

1.1.7 Synthesis of Hindi

In collaboration with Dr. Shyam Agrawal, of the Central Electronics Engineering Research Institute of New Delhi, India, we have initiated a project on speech synthesis of Hindi. The aim of the New Delhi group is to develop a rule-based system for the synthesis of Hindi. As one component of this work, we have synthesized examples of voiced and voiceless aspirated consonants in Hindi. In order to achieve natural-sounding aspiration noise for Hindi, it was necessary to modify the spectrum of the aspiration noise in the Klatt synthesizer to enhance the amount of low-frequency energy in the aspiration source.

1.2 Speech Production Planning and Prosody

Our work in speech production planning focuses on the role of prosodic structure in constraining phonetic and phonological modification processes and speech error patterns in American English. The development of labelling systems and of extensive labelled databases of speech and of speech errors provides resources for testing hypotheses about the role of both prosodic and morpho-syntactic structures in the production planning process. This process results in substantial differences in a given word or segment from one utterance to another, with changes in adjacent phonetic context, speaking rate, intonation, phrasing, etc. An understanding of these systematic variations, and the structures that govern them, will not only improve our models of human speech processing, but will also make possible more natural synthesis and improved recognition algorithms.

1.2.1 Prosodic Structure and Glottalization

Drawing on a large corpus of FM radio news speech that we have labeled for prosodic structure using the ToBI system, we replicated Pierrehumbert and Talkin's (1992) finding that glottalization of vowel-initial words was more likely at prosodically significant locations such as the onsets of intonational phrases and pitch accented syllables, in a larger speech corpus with the full range of normal phonetic and intonational contexts. Further, we extended their finding to show that reduced-vowel onsets are more likely to glottalize at the onset of a Full Intonational Phrase than at the onset of an Intermediate Intonational Phrase, providing support for the distinction between these two constituent levels. We also reported profound differences among speakers both in the rate of glottalization for

word-initial vowels and in the acoustic shape preferred by each speaker.

This same labeled speech corpus also allowed us to examine how intonational phrase structure influences the placement of optional prenuclear pitch accents before the last (or nuclear) pitch accent of the phrase. Earlier work showed that speakers tend to place the first accent of a new intonational phrase on an early syllable in late-main-stress words, as in "The MASSachusetts instiTUtions." More recently we found that phrase-medial accents avoid clash with accented syllables both to the left and to the right, demonstrating the avoidance of left-ward clash and providing further support for the role of intonational phrase structure in production planning.

1.2.2 Detection and Analysis of Laryngealized Voice Source

An algorithm for the automatic detection of laryngealized regions, developed for German at the University of Erlangen, has been found to perform well for read American English speech but not for a corpus of spontaneous telephone quality speech. An extensive labelling effort was undertaken for the spontaneous speech to determine the causes of this result, which was unexpected since the algorithm performed well for both read and spontaneous band-limited German speech. The labelling has revealed that perceived glottalization is not always directly correlated with evidence for irregular pitch periods in the signal in English, and we are currently exploring this paradox. The four corpora labelled for both laryngealized voice source and prosodic structure (i.e., read speech and spontaneous speech in both American English and German) will permit us to compare the distribution of glottalized regions with respect to prosodic structure in the two languages. Preliminary results support earlier claims that syllable-final /p t k/ do glottalize in spontaneous American English speech, but not in German.

1.2.3 Development of a Transcription System for Other Aspects of Prosody

Work by Couper-Kuhlen in the 1980s and by Fant, Kruckenberg, and Nord in the 1990s suggests that speech may be characterized by structure-based manipulations of a regular rhythmic structure. Moreover, work in the British framework of intonational analysis in the 1950s suggests that repeated use of similar *F*0 contours is an important aspect of the speech signalling system. We have developed a transcription system for these two aspects of spoken language, along with an on-line training tutorial, and are currently evaluating the reliability of the transcription system across listeners, as well as the relationship between these two aspects of speech and the intonational phrase structure and prominence patterns captured by the ToBI labeling system. Preliminary analysis shows that perceived rhythmic beats are often coincident with pitch accented syllables, but not always, and that rhythmic patterns can set up expectations that lead the listener to perceive a beat even in a silent region of the speech wave form.

1.2.4 Other Work

We continue to develop the MIT Digitized Speech Error Corpus, the computer categorization system for speech errors captured in written form, and extensive corpora of labeled speech, including phonetic, syntactic and prosodic labels for the same utterances. We published a tutorial to summarize our examination of the evidence for prosodic structure in the acoustic-phonetic and psycholinguistic literature. This will be followed by a closer examination of the evidence for the smaller constituents in the prosodic hierarchy, such as the mora, the syllable and the prosodic word, and the evidence for rhythmic constituents such as the Abercrombian cross-word-boundary foot.

1.3 Studies of Normal Speech Production

1.3.1 Experimental Studies

During this year, we completed most of the data collection, signal processing, and data extraction on four interrelated experiments. The experiments are designed to test hypotheses based on a model of speech production. According to the model, the speaker operates under a listener-oriented requirement to produce an intelligible signal, using a production mechanism that has dynamical properties that limit kinematic performance. Various strategies are employed to produce speech under these constraints, including the modification of clarity by using values of kinematic parameters that will produce sufficiently distinctive acoustic cues, while minimizing the expenditure of articulatory effort.

In the four studies, we have made kinematic and acoustic measures on the same ten subjects using a multiple single-subject design, and we are obtaining intelligibility measures of the subjects' speech from an independent group of listeners. The four studies are: (1) The clarity constraint

versus economy of effort: speaker strategies under varying clarity demands; (2) clarity versus economy of effort: the relative timing of articulatory movements; (3) kinematic performance limits on speech articulations; and (4) interarticulator coordination in the production of acoustic/phonetic goals: a motor equivalence strategy. In study one, speech is produced in the following conditions: normal, fast, slow, clear, clear+fast and casual. The other studies employ subsets of these conditions. On the basis of previous recordings and analyses of data from two pilot subjects, we made improvements in utterance materials, recording protocols and algorithms for data extraction.

For the eight subsequent subjects, data were collected and derived from the acoustic signal and the movements of points on the lips, tongue and mandible, as transduced by an electromagnetic midsagittal articulometer (EMMA) system. Spectral and temporal measures are being extracted from the acoustic signal, and articulatory kinematic measures (displacements, velocities, accelerations, durations) are being extracted from the movement signals. Measures of "effort" are being calculated from the kinematic parameters. These measures are being compared with each other and they will be compared with results of tests of intelligibility and prototypicality of phonetic tokens. Initial results show effects of speaking condition, articulator, vowel, consonant manner and vowel stress. The pattern of results differs somewhat across speakers.

1.3.2 Physiological Modeling of Speech Production

In order to achieve practical computation times for a finite-element simulation of the behavior of the tongue and for a future, more complete vocal tract model, it will be necessary to use parallel computation. For this purpose work has been done on the theoretical aspects of a parallel implementation. In the resulting proposed algorithm, the tongue is split into interconnected domains, usually containing several finite elements, and the state variables in each sub-domain are updated by one processor. The continuity requirement at the boundaries between the domains is simplified to require that the state variables of two interfacing domains are equal only at nodes in the interfaces. Therefore, in theory, it is possible that different computational methods can be used in the different domains. The continuity requirements are maintained by Lagrange multipliers (one per equality constraint), whose computation is achieved by several processors, each acting on the nodes of one inter-domain surface. The resulting algorithm still leaves much to be desired, since at each time-step, interchange of information between neighboring domains would be required. On the other hand, this scheme should be useful when we interface two different models: a new finite element tongue model and a mandible and hyoid model that is based on rigid body movements. The latter model was implemented by other researchers with whom we plan to collaborate.

In the planned collaboration, we will combine a new tongue model with the existing mandible/hyoid-bone model. For this purpose, a modified "lambda" muscle control model has been designed for implementation as part of the tongue modeling. It is an extension of the lambda model that was originally designed for the control of limbs moving as rigid bodies. While in the original lambda model there are only two parameters describing the muscle state, length and velocity, the corresponding variables have first to be computed for each muscle in the tongue model. In this computation, the length and velocity parameters are replaced by the average stretch and stretch velocity of the muscle that is modeled as a continuum. Both variables can be obtained by integration over the finite elements.

In addition, progress has been made on implementing an algorithm for parallel computation and an overall plan for a new model building method.

1.3.3 Improvements in Facilities for Speech Production Research

We have completed the refinement of facilities for gathering articulatory movement and acoustic data. We made further progress on a facility for performing perceptual tests. In collaboration with a radiologist from the Massachusetts Eye and Ear Infirmary, we have developed techniques for obtaining high-quality MRI images of the vocal tract. We developed software to facilitate the extraction of area function data from the MR images. We also implemented additional algorithms for articulatory and acoustic data extraction, and we implemented the main component of our program for EMMA data extraction in MATLAB, in anticipation of changing computer hardware and software platforms.

1.4 Speech Research Relating to Special Populations

1.4.1 Speech Production of Cochlear Implant and Bilateral Acoustic Neuroma (NF2) Patients

We began work on a new NIH-funded project, "The Role of Hearing in Speech: Cochlear Implant Users." This research is an extension of work done previously under other funding sources. In this initial year, we refined previously-used paradigms, devised new sets of utterance materials, and established new procedures for recruiting patients. We have made four pre-implant recordings on our first new implant patient. She has received her implant and we will soon begin a series of post-implant recordings. We have also made recordings of two normally hearing speakers. We have analyzed a substantial part of the data from these recordings. We also have given extensive training in techniques for recording and data analysis to two staff members who are new to this research.

Auditory Feedback and Variation in Sound Pressure and Fundamental Frequency Contours in Deafened Adults

Sound pressure level (SPL) and fundamental frequency (*F0*) contours were obtained from four postlingually deafened adults who received cochlear implants and from a subject with Neurofibromatosis-2 (NF2) whose hearing was severely reduced following surgery to remove an auditory-nerve tumor and insert an auditory brainstem implant. SPL and *F0* contours for each phrase in passages read before and after changes in hearing were averaged over repeated readings and then normalized with respect to the highest SPL or *F0* value in the contour. The regularity of each average contour was measured by calculating differences between successive syllable means and averaging the absolute values of these differences. With auditory feedback made available, the cochlear implant user with the least contour variation pre-implant showed no change, but all of the remaining speakers produced less variable *F0* contours and three also produced less variable SPL contours. In complementary fashion, when the NF2 speaker had her auditory feedback severely reduced, she produced more variable *F0* and SPL contours. The results are interpreted as supporting a dual-process theory of the role of auditory feedback in speech production, according to which one role of self-hearing is to monitor transmission conditions, leading the speaker to make changes in speech "postures" (such as average sound level and speaking rate) aimed at maintaining intelligibility.

1.4.2 Speech Respiration and Changes in Hearing Status

We have analyzed speech respiration parameters from complete series of longitudinal recordings of seven of our former subjects who gained some hearing from cochlear implants, from three NF2 patients (who experienced hearing loss), and one control. We have also conducted an extensive search of the literature for normative data. The results from the implant subjects confirm our earlier finding that parameters change toward normal with the acquisition of some hearing. The hearing control maintained respiratory parameter values at about the same levels, and the NF2 patients behaved variously. We are currently examining changes in SPL as a potential confounding variable. As part of this study, we have also processed rate measures on all these subjects. In general, implant users increased speaking rate post-implant, the hearing control showed no change, and the NF2 patients behaved variously.

1.4.3 Voice Source Analysis of Speakers with Vocal Fold Nodules

In collaboration with Dr. Robert Hillman at the Massachusetts Eye and Ear Infirmary, we have been carrying out an analysis of vowels produced by several patients with vocal-fold nodules. The goal is to develop models that will help to interpret the acoustic and aerodynamic observations for these vowels, and hopefully to explain differences in the glottal source for these patients compared with a normal group of speakers. Previous work has shown that nodule subjects required an increased subglottal pressure of 4-6 dB in order to achieve vowel sound pressure levels comparable to those produced by normals. Other acoustic and aerodynamic measures have shown that the amplitude of the first harmonic in relation to the amplitude of the first-formant peak is increased for the nodule subjects, suggesting an increased first-formant bandwidth and possibly an enhanced amplitude of the first harmonic and an increased spectrum tilt extending into the *F1* region.

We have been introducing some modifications into a conventional two-mass model of the vocal folds in order to simulate some aspects of vocal-fold vibration for nodule subjects. Two kinds of modifications have been attempted: changes in the coupling stiffness between the upper and lower masses (in an attempt to simulate the effect of the increased

stiffness of the nodule), and an increase in the static separation between the folds (simulating the lack of complete closure when nodules are present). These modifications lead to changes in the threshold subglottal pressure needed to initiate phonation and in the amplitude of the airflow pulses for a given subglottal pressure. The model behavior shows some of the attributes of the glottal source for the subjects with nodules.

1.4.4 Preliminary Studies of Production s and sh by Some Dysarthric Speakers

We have begun a detailed study of the fricative consonants /s/ and /š/ produced in word-initial position by four dysarthric speakers with different degrees of severity. The speakers ranged in age from 38 to 61, three have cerebral palsy and one has cerebellar ataxia. The intelligibility of words beginning with these sounds, measured in an earlier study, was found to be relatively low on the average. It was assumed that this low intelligibility was due to improper placement and shaping of the tongue blade against the hard palate and possibly due to inappropriate control of the intraoral pressure.

On the basis of acoustic and modeling studies for fricatives, including our own recent research in this area, it has been established that, for normal speakers, the spectrum for /s/ in most phonetic environments has a main prominence in the range of the fourth formant or higher. The spectrum for /š/ has a prominence in the *F3* range, as well as one or more additional prominences in the *F4* range or higher. The peak amplitudes of the spectral prominences at *F4* or higher (called *Ah*) and for the *F3* range (called *Am*) were measured for the two fricatives that occurred in initial position for several words. For each word, the amplitude *A1* of the spectral prominence corresponding to the first formant in the following vowel was also measured. This vowel amplitude was used as a reference against which the fricative amplitude was specified. All these amplitudes for each utterance were obtained by averaging the spectrum over a time interval of about 20 ms. The differences *A1-Ah* and *A1-Am* were then determined for each word, and averages were calculated for the s-words and the sh-words for each speaker. Similar data were obtained for a normal speaker.

There appeared to be a reasonable correlation between the word intelligibility and the acoustic measure of "stridency" for the s-words. When the high-frequency amplitude in the fricative is weak, the intelligibility is low. For the sh-words there is a similar trend for the mid-frequency amplitude, but it is not as regular. Other factors besides measurements of the average noise spectrum shape are clearly contributing to loss of intelligibility. These data are preliminary in two ways: (1) the values and ranges of the differences for a number of normal speakers should be given rather than values for one speaker; and (2) the data for the dysarthric speakers are based on only a small number of utterances.

The decreased amplitude of the frication noise for the fricatives produced by the dysarthric speakers probably is a consequence of incorrect placement and shaping of the tongue blade when it is raised toward the hard palate. Either the constriction is not sufficiently narrow, or the tongue blade is placed too far forward. For some utterances, there was strong vocal-fold vibration, which caused a strong low-frequency peak in the fricative spectrum. In these cases, it is possible that there was insufficient intraoral pressure to cause enough turbulence noise near the tongue-blade constriction.

1.5 Models of Lexical Representation and Lexical Access

1.5.1 Quantitative Error Models for Classification of Stop Consonants

The general aim of this project is to develop a model that will identify consonantal features in running speech. The goal is for the model to have a performance similar to that of a human listener. Implementation of the proposed model requires that certain spectral and temporal measurements are made in the vicinity of landmarks where rapid spectral changes occur in the speech signal. In the case of landmarks at stop-consonant releases, these measurements are intended to locate spectral peaks that represent particular vocal-tract resonances, and how these resonances change with time. In an initial experiment using a corpus of sentences, a series of such measurements were made by hand by individuals who know how to interpret the spectra and to avoid labeling spurious spectral prominences. The measurements were made on a large number of stop consonants, for which the time of release was specified in advance. A classification system was implemented, and the number of errors in identifying the consonants from these hand-made measurements was relatively small. Algorithms that were developed for making the measurements automatically were somewhat error-prone. However, it was possible to evaluate the error or uncertainty in these algorithms and to predict the difference between the classification errors based on the two types of measurements.

This exercise has emphasized the importance of quantifying the uncertainty in measurements of acoustic parameters and incorporating this knowledge in models for classifying features.

1.5.2 Identifying Features in Isolated Consonant-Vowel Syllables

A test for the validity of a model for lexical access is a comparison of the errors made by the model with the errors made by human listeners. One relatively simple way to make such a comparison is to use consonant-vowel syllables in noise as input to a listener or to a model that is intended to recognize the consonants, and to compare the errors that are made in identification of the consonant features. A preliminary experiment of this kind has been carried out using a set of 10 CV syllables with the consonants /p t b d f s v z m n/ followed by /*a*/. Various levels of white noise were added to the syllables. The syllables were processed in four ways: listeners identified the consonants, spectrograms of the syllables (with noise) were made and the consonants were identified by a spectrogram reader, the syllables were processed by an algorithm that attempted to identify certain manner features of the consonants (sonorant and continuant), and an HMM recognition system was trained to recognize the noise-free syllables. Principal outcomes of this work were: (1) the types of errors made by the HMM system were quite different from those made by listeners; (2) automatic identification of manner features in noise showed many more errors than listeners, indicating that improvement in manner detection in noise is necessary; and (3) errors in place of articulation in noise are similar for listeners and for spectrogram reading.

1.6 Locating Glides in Running Speech

One of the initial steps in the model for lexical access that we are proposing is a procedure for identifying landmarks in the speech signal. These landmarks are of three kinds: (1) points in the speech signal where there are abrupt discontinuities, representing points where consonantal constrictions in the vocal tract are created or released; (2) regions in the signal that identify vowel prominences; and (3) regions in the signal where the energy is a minimum but there are no abrupt discontinuities. This last type of landmark is created by glides. An algorithm has been developed for locating glide landmarks in speech. The algorithm is based on locating regions in time where there are minima in amplitude, minima in first-formant frequency, and rates of change of these parameters that are constrained to be in particular ranges. The overall recognition results were 88 percent for glide detection and 91 percent for nonglide detection.

1.6.1 Enhancement Theory and Variability

In the model of lexical access that we and others have proposed, lexical items are organized into segments, and the segments, in turn, are specified in terms of hierarchically arranged binary features. Many of these features are defined in articulatory terms. That is, each feature or group of features specifies instructions to particular articulators. It is understood, however, that, in addition to specifying articulatory instructions, a feature also has acoustic correlates, although these acoustic correlates for a particular feature may depend on other features in the same segment. It is further recognized that the strength of a particular acoustic correlate that signals the presence of a feature in a segment can be enhanced by recruiting articulatory gestures in addition to the one specified by the feature. That is, more than one articulatory gesture can contribute to enhancement of the acoustic contrast corresponding to the + or - value of a feature. (An example is the partial lip rounding that is used to produce the palatoalveolar /s̃/, presumably to enhance the contrast with /s/). Our view is that this enhancing gesture (rounding in this example) is graded, and does not have the status of a lexically represented feature. Because enhancement can be graded, variability can exist in the acoustic manifestation of a feature, in addition to the variability from other sources.

This enhancement process is most likely to be brought into play for a given feature when there exists a gesture that can indeed help to strengthen the acoustic attribute generated by the primary articulator receiving instructions from the feature. It is also invoked when the acoustic differences between competing similar segments are minimal. We have been examining a number of enhancement processes from this point of view, and we have noted that enhancement is especially utilized to strengthen the voiced-voiceless contrast for obstruents, the contrast between different consonants produced with the tongue blade, and the height and front-back contrast for vowels.

1.6.2 The Special Status of Word Onsets in Word Recognition and Segmentation

Word onsets have been hypothesized to play a critical role in word recognition and lexical segmentation in the perception of connected speech. In order to understand why onsets appear to have this special status, onsets were examined from the per-

spectives of work in acoustic phonetics, phonological theory, and behavioral studies of word recognition. A review of research in acoustic phonetics suggested that onsets provide a particularly rich source of segmental featural information due to a constellation of articulatory and allophonic factors. A cross-linguistic survey of phonological rules conditioning featural change or deletion revealed a striking asymmetry in which onsets are generally insulated against change which occurs widely in other positions. Finally a review of behavioral research demonstrated that listeners have a lower tolerance for the loss or alteration of word-initial segments than for other segments in word recognition. This pattern coincides with a tendency for sensitivity to lexical effects over the course of a word. Together, these results suggest that onsets: (1) provide particularly robust featural information; (2) are more transparently interpretable in word recognition than non-onsets due to their lack of phonologically conditioned surface variation; and (3) appear to be particularly well-suited to drive lexically-mediated processes that facilitate the perception of words when non-onset information is difficult to recover or interpret due to acoustic or representational underspecification. In the context of a model of segmentation in which segmentation is accomplished as a by-product of word recognition, these observations may account for acoustic-phonetic nature and distribution of putative cues to word boundary.

1.7 Publications

1.7.1 Journal Articles

Dilley, L., S. Shattuck-Hufnagel, and M. Ostendorf. "Glottalization of Word-Initial Vowels as a Function of Prosodic Structure." *J. Phonetics* 24: 423-444 (1996).

Hanson, H.M. "Glottal Characteristics of Female Speakers: Acoustic Correlates." *J. Acoust. Soc. Am.* 101: 466-481 (1996).

Matthies, M.L., M. Svirsky, J. Perkell, and H. Lane. "Acoustic and Articulatory Measures of Sibilant Production with and without Auditory Feedback from a Cochlear Implant." *J. Speech Hear. Res.* 39: 936-946 (1996).

Perkell, J.S. "Properties of the Tongue Help to Define Vowel Categories: Hypotheses Based on Physiologically-Oriented Modeling." *J. Phonetics* 24: 3-22 (1996).

Shattuck-Hufnagel, S., and A. Turk. "A Prosody Tutorial for Investigators of Auditory Sentence Processing." *J. Psycholing. Res.* 25(2): 193-247 (1996).

Stevens, K.N. "Critique: Articulatory-Acoustic Relations and their Role in Speech Perception." *J. Acoust. Soc. Am.* 99: 1693-1694 (1996).

Wilhelms-Tricarico, R. "A Biomechanical and Physiologically-Based Vocal Tract Model and its Control." *J. Phonetics* 24: 23-28 (1996).

1.7.2 Published Meeting Papers

Perkell, J.S., M.L. Matthies, R. Wilhelms-Tricarico, H. Lane, and J. Wozniak. "Speech Motor Control: Phonemic Goals and the Use of Feedback." *Proceedings of the Fourth Speech Production Seminar: Models and Data* (Also called the First ESCA Tutorial and Research Workshop on Speech Production Modeling: From Control Strategies to Acoustics) 1996, pp. 133-136.

Stevens, K.N. "Understanding Variability in Speech: A Requisite for Advances in Speech Synthesis and Recognition." Special Session 2aSC of ASA-ASJ Third Joint Meeting, Speech Communication for the Next Decade: New Directions of Research, Technological Development, and Evolving Applications, Honolulu, Hawaii, 1996, pp. 3.1-3.9.

1.7.3 Chapters in Books

Halle, M., and K.N. Stevens. "The Postalveolar Fricatives of Polish." Festschrift for Osamu Fujimura. Berlin: Mouton de Gruyter, 1997, pp. 177-194.

Perkell, J.S. "Articulatory Processes." In *Handbook of Phonetic Science.* Eds. W. Hardcastle and J. Laver. Blackwell: Oxford, 1997, pp. 333-370.

Perkell, J.S., and M.H. Cohen. "Token-to-Token Variation of Tongue-body Vowel Targets: The Effect of Context." Festschrift for Osamu Fujimura. Berlin: Mouton de Gruyter, 1997, pp. 229-240.

Stevens, K.N. "Articulatory-Acoustic-Auditory Relationships." In *Handbook of Phonetic Sciences.* Eds. W. Hardcastle and J. Laver. Blackwell: Oxford, 1997, pp. 462-506.

Stevens, K.N. "Models of Speech Production." In *Encyclopedia of Acoustics*. Ed. M. Crocker. New York: John Wiley, 1997, pp. 1565-1578.

Wilhelms-Tricarico, R., and J.S. Perkell. "Biomechanical and Physiologically-Based Speech Modeling." In *Speech Synthesis*. Eds. J.P.H. van Santen, R.W. Sproat, J.P. Olive, and J. Hirshberg. New York: Springer, 1996, pp. 221-234.

1.7.4 Journal Articles and Book Chapters to be Published

Chen, M., and R. Metson. "Effects of Sinus Surgery on Speech." *Arch. Otolaryngol.* Forthcoming.

Govindarajan, K.K. "Relational Perception of Formants and Stop Consonants: Evidence from Missing First Formant Stimuli." *J. Acoust. Soc. Am.* Forthcoming.

Hillman, R.E., E.B. Holmberg, J.S. Perkell, J. Kobler, P. Guiod, C. Gress, and E.E. Sperry. "Speech Respiration in Adult Females with Vocal Nodules." *J. Speech Hear. Res.* Forthcoming.

Lane, H., J. Wozniak, M.L. Matthies, M.A. Svirksy, J.S. Perkell, M. O'Connell, and J. Manzella. "Changes in Sound Pressure and Fundamental Frequency Contours Following Changes in Hearing Status." *J. Acoust. Soc. Am.* Forthcoming.

Perkell, J.S., M.L. Matthies, H. Lane, R. Wilhelms-Tricarico, J. Wozniak, and P. Guiod. "Speech Motor Control: Segmental Goals and the Use of Feedback." Submitted to *Speech Commun.*

Shattuck-Hufnagel, S. "Phrase-level Phonology in Speech Production Planning: Evidence for the Role of Prosodic Structure." In *Prosody: Theory and Experiment: Studies Presented to Gosta Bruce.* Ed. M. Horne. Stockholm: Kluwer. Forthcoming.

Svirsky, M.A., K.N. Stevens, M.L. Matthies, J. Manzella, J.S. Perkell, and R. Wilhelms-Tricarico. "Tongue Surface Displacement During Obstruent Stop Consonants." *J. Acoust. Soc. Am.* Forthcoming.

1.7.5 Theses

Hasegawa-Johnson, M.A. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification.* Ph.D. diss., Dept. of Electr. Eng. and Comput. Sci., MIT, 1996.

Sun, W. *Analysis and Interpretation of Glide Characteristics in Pursuit of an Algorithm for Recognition.* S.M. thesis, Dept. of Electr. Eng. and Comput. Sci., MIT, 1996.