Speech Communication

RLE Group
Speech Communication Group

Academic and Research Staff
Professor Kenneth N. Stevens, Professor Morris Halle, Professor Samuel J. Keyser, Dr. Joseph S. Perkell, Dr. Stefanie Shattuck-Hufnagel, Dr. Helen Hanson, Dr. Janet Slifka, Mark Tiede, Majid Zandipour, Dr. Margaret Denny, Dr. Satrajit Ghosh, Ellen Stockmann, Seth Hall.

Visiting Scientists and Research Affiliates
Dr. Takayuki Arai, Department of Electrical and Electronics Engineering, Sophia University, Tokyo, Japan.
Dr. Corine A. Bickley, Department of Hearing, Speech and Language Sciences, Gallaudet University, Washington, District of Columbia.
Dr. Suzanne E. Boyce, Department of Communication Disorders, University of Cincinnati, Cincinnati, Ohio.
Dr. Krishna Govindarajan, SpeechWorks International, Boston, Massachusettts.
Dr. David Gow, Department of Psychology, Salem State College, Salem, Massachusetts, and Department of Neuropsychology, Massachusetts General Hospital, Boston, Massachusetts.
Dr. Frank Guenther, Department of Cognitive and Neural Systems, Boston University, Boston, Massachusetts.
Dr. Andrew Howitt, Otolith, visit site at: http://www.otolith.com.
Dr. Robert E. Hillman, Department of Voice Surgery and Rehabilitation, Massachusetts General Hospital, Boston, Massachusetts.
Dr. Harlan Lane, Department of Psychology, Northeastern University, Boston, Massachusetts.
Dr. Sharon Y. Manuel, Department of Speech Language Pathology & Audiology, Northeastern University, Boston, Massachusetts.
Dr. Melanie Matthies, Department of Communication Disorders, Boston University, Boston, Massachusetts.
Dr. Richard McGowan, CReSS LLC, Lexington, Massachusetts.
Dr. Lucie Menard, Department of Linguistics and Language Education, University of Quebec, Montreal, Canada.
Dr. Rupal Patel, Department of Speech Language Pathology and Audiology, Northeastern University, Boston, Massachusetts.
Dr. Alice Turk, Department of Linguistics, University of Edinburgh, Edinburgh, United Kingdom.
Dr. Nanette Veilleux, Department of Computer Science, Simmons College, Boston, Massachusetts.
Dr. Lorin Wilde, Massachusetts.

Graduate Students

Tamas Bohm (Visiting Student, University of Hungary), Lan Chen, Xuemin Chi, Laura Dilley, Elisabeth Hon, Annika Imbrie, Steven Lulich, Nicole Marrone (Sargent College, Boston University), Xiaomin Mou, Tony Okobi, Chi-youn Park, Yoko Saikachi, Kushan Surana, Virgilio Villacorta, Julie Yoo, Sherry Zhao

Undergraduate Students

Akua Adu-Boahene, Jorge Alvarado, Flora Amwayi, Sheeva Azma, Nathan Boy, Meredith Brown, Bashira Chowdhury, Neil Desai, Priya Desai, Tiffany Dohzen, Mingyan Fan, Megan Foster, Mun Yuk Ko, JungEun Lee, Daniel Leeds, Shirley Li, Jonathan McEuen, Michael Mejia, Conor Murray, Elizabeth Park, Ketan Patel, Lars Plate, Zachary Rich, Margaret Renwick, Emyluz Rodriguez, Janet Ryu, Rodrigo Sanchez, Sanghamitra Sen, Dewang Shavdia, Danny Shen, Morgan Sonderegger, Robert Speer, Enrique Urena, Sumudu Watugala, Betty Yang, Yelena Yasinnik, Man Yin Yee, Wei-An Yu

Technical and Support Staff

Arlene E. Wint

1.   Constraints and Strategies in Speech Production

Introduction

The objective of this research is to refine and test a theoretical framework in which words in the lexicon are represented as sequences of segments and syllables and these units are represented as complexes of auditory/acoustic and somatosensory goals.   The motor programming to produce sequences of sensory goals utilizes an internal neural model of relations between articulatory motor commands and their acoustic and somatosensory consequences. The relations between those articulatory motor commands and the movements they generate are influenced by biomechanical constraints, which include characteristics of individual speakers' anatomies and more general dynamical properties of the production mechanism.   To produce an intelligible sound sequence while accounting for biomechanical constraints, speech movements are planned so that sufficient perceptual contrast is achieved with minimal effort.   There are individual differences in planning movements toward sensory goals that may be due to relations between production and perception mechanisms in individual speakers.

In a current project, the internal model is implemented as a neurocomputational model that is used to control a vocal-tract model (an articulatory synthesizer). The combined models provide the bases of hypotheses about the planning of speech movements. To test these hypotheses, we are conducting experiments with speakers and listeners in which we measure articulatory movements, speech acoustics, perception, and brain activation. We are manipulating speech condition, phonemic context and speech sound category and we introduce transient and sustained perturbations. We are also performing modeling and simulation experiments, in which we adapt the vocal-tract model to the morphologies of individual speakers.   We are testing properties of the neurocomputational model by using it to control the individualized vocal tract models in efforts to replicate those speakers' production data.

During this last year, two new studies were begun and ongoing studies have been continued or completed.

1.1 Development of facilities

Software has been acquired and used to generate natural-sounding synthetic speech stimuli for a battery of perceptual tests that will be used to measure speakers' auditory acuity. Scripts have been developed to facilitate the use of the synthesis software for the rapid generation of large numbers of stimuli in continua for vowel, stop and sibilant contrasts.   Software has been developed to record and evaluate responses in labeling and discrimination tasks.   Preliminary testing of these paradigms is underway.

1.2 Sensorimotor adaptation in the production of vowels

We have used a DSP board and its control software to modify the first formant frequency (F1) of vowels that were fed back to subjects (with an 18 ms delay) as they pronounced simple utterances.   The signal processing used LPC analysis and resynthesis to detect and shift F1 according to a custom-designed algorithm. Twenty subjects were run, 10 male and 10 female. They showed varying amounts of compensation (changing F1 in the opposite direction to that of the perturbation). We also tested the subjects' auditory acuity—measured as their ability to discriminate acoustic perturbations in F1—and found a significant relation across subjects between the amount of compensation and acuity. This result was compatible with the behavior of our neurocomputational model of speech motor planning. The model was used to control an articulatory synthesizer and simulate the relation between speaker acuity and amount of compensation.   This was done by varying the model's auditory acuity and observing the compensation produced in response to an F1 perturbation.

1.3 fMRI study of sensorimotor adaptation in the production of vowels

We are using functional magnetic resonance imaging (fMRI) to investigate brain regions contributing to sensorimotor adaptation. A condenser microphone placed over the subjects' mouth in the bore of the scanner is being used to record the subjects' production. The recorded speech is then processed using custom software running on a DSP board and fed back to the subject with a delay of 18ms using electrostatic headphones. The signal processing uses LPC analysis and resynthesis to detect and shift the F1 frequency according to a custom-designed algorithm; the resynthesis uses the LPC residual as the source, so the voiced sounds fed back to the subject sound reasonably natural. Twelve subjects have been run and the data are being analyzed.

1.4 Control of tongue movements in acoustic and articulatory spaces

This project is investigating the planning of vowel-to-vowel tongue movements using two different control models: the above-mentioned model of speech motor planning in acoustic space, and motor trajectory planning through muscle activation control. The planning models are used to control a vocal tract model, which consists of an improved 2D biomechanical/physiological tongue model combined with simple second-order models of jaw rotation and translation, lip opening and protrusion and larynx height movements. When using motor trajectory planning for a vowel to vowel target, i.e. muscle activation goals, the vocal tract model simulates a subject's acoustic, articulation, and EMG data. To investigate control strategy in the acoustic space an adaptive controller that consists of two neural networks has been developed. One network, a forward model that maps muscle activation patterns on to acoustic space (64 dimensional spectrum in mel space), was trained with data extracted from 2500 simulations of the vocal tract model. The second network, an inverse model, maps acoustic trajectories on to muscle activation patterns. Currently we are running simulations of vowel-to-vowel movement based on acoustic space planning. The simulation results from the vocal tract model based on the two planning strategies will be compared with data from a speaker to whom the vocal-tract model's morphology has been adapted.

2.   Formulation and Implementation of a Model for Lexical Access in Running Speech

2.1 Progress in model development

In the model of speech perception that we are developing in the Lexical Access Project, we plan to assign probability estimates to landmarks and associated features. This will be a core part of the way in which the model assembles a cohort of possible words or word sequences. These probability estimates will also guide the top-down path which is used to select the best candidate from the cohort. The decision to move to a probabilistic-based structure requires the re-formulation of some existing modules as well as being a part of current work.

In particular, our algorithm for vowel landmark detection is being re-formulated to produce a probability estimate. The new version of the detector still generates all candidate vowel landmarks based on peaks in energy in the first formant region. For all candidates, a set of cues associated with a canonical vowel are measured. These include parameters that characterize the spectral shape (to detect the presence of a general formant structure) and parameters that characterize the type of sound source (periodic and higher intensity). Initial tests have been done with part of the TIMIT database, and results using logistic regression analysis indicate that about 95% of roughly 4000 candidate peaks are assigned a probability estimate that is >0.5 for vowel peaks and <0.5 for non-vowel peaks. The current focus is to continue to refine the cue set and to choose the most appropriate algorithm for estimating the probabilities.

Additionally, a module has been developed to hypothesis the presence of a vowel-vowel sequence when only a single vowel peak is present. Currently, this module only uses the cue of

vowel-region duration and is based on statistical analysis of all of the TIMIT train set. Future work on the module will be to condition the estimate based on a local characterization of speaking rate.

Another on-going project aims to automatically detect irregular phonation in continuous speech. Instances of irregular phonation will be used in the Lexical Access Project to construct a prosodic structure as well as to refine the detection of landmarks. A set of acoustic cues has been selected and classification has been implemented with several algorithms. When given short segments of either modal phonation or irregular phonation, classification accuracy exceeds 90%. The current focus is to estimate a confidence measure or probability estimate for classification as well as to explore means for detection in running speech.

2.2 Accessing the lexicon with partial features

The initial acoustic analysis of the signal yields partial estimates of the distinctive features for some of the segments of the input word sequence. A next step is to generate a cohort of words that is consistent with these estimated features, rank-ordered according to the confidence of these estimates. This approach cannot be directly applied to our system, which uses a different approach from most conventional speech recognition systems. These differences include: (1) the features of most recognizers are real numbers of statistical importance, but LAFF uses binary valued features based on the knowledge of speech; (2) each feature represents a different characteristic of speech, so different types of phonemes use completely different sets of features; (3) we should deal with varying numbers of features, and in many cases we end up with an incomplete set of features; (4) LAFF uses the idea of landmarks, instead of having a sequence of equally spaced time frames.

The first proposed method is to use a reinforcement strategy. This method was developed through logical steps from the following simple intuitively reasonable assumptions: (1) each feature increases or decreases the overall rating of the phoneme; (2) the rating should range between 0 and 1; (3) the rating should not depend on the ordering of features; (4) unidentified features should not affect the rating; and (5) no phoneme should be rated exactly 1. By converting the ratings into a mathematical problem and solving it we have a method to rate similarities between a given feature bundle and the phonemes in the lexicon. This method also derives estimates of the importance of the features by weighting each feature.

A revised probability method has been developed to deal with this type of problem. The LAFF system does not operate on a strict probabilistic basis. Although the features have some relations between each other, these relations cannot be completely considered in the feature extraction process. And in the perspective of the LAFF system, extracting and calculating every possible conditional probability is not useful. We use speech knowledge to deliberately ignore the less meaningful factors. This can greatly reduce the size of candidates, but it can be a nuisance in rating the probability. However, although there had to be several minor assumptions, such as independence of features, due to the system's characteristic mostly this method is based on a firm mathematical background and works as well as the reinforcement method.

3.   Acoustic/Articulatory Studies of Speech-Sound Production

3.1 Acoustic characteristics of the consonant /ð/

Among the fricative consonants in syllable-initial position in running speech in English, the voiced nonstrident fricative /ð/ occurs with the greatest frequency. This consonant is unique in at least two ways: (1) it occurs almost always in function words like the, that, and those, in which the following vowel is frequently reduced; and (2) the acoustic attributes of the consonant show considerable variability depending on the preceding context. As part of a broader acoustic study of fricatives, we have been cataloguing the occurrence of various versions of /ð/ as a function of the preceding context, and we have been measuring several acoustic attributes of some of these versions. The data were obtained from the TIMIT recordings developed by the Spoken Language

Systems group in CSAIL at MIT.   Estimates of the various types of /ð/ were made through observations of spectrograms, waveforms, and informal listening, and did not use the labels provided with TIMIT.

When /ð/ is preceded by a voiceless obstruent consonant, it almost always surfaces as a stop consonant; when it follows a voiced obstruent, it sometimes is stop-like.   When it is preceded by a vowel or by a sonorant consonant it is usually implemented as a continuant.   We concentrate here on the acoustic attributes of /ð/ when it occurs as a stop consonant, and we compare these attributes with those of the voiced stop consonant /d/ when it occurs in a similar context.

Three acoustic attributes were measured for these consonants: (1) the frequency of the spectrum peak with the maximum amplitude in the noise burst at the consonant release; (2) a simple measure of the spectrum slope defined by the amplitude of burst spectrum in the frequency range above 2 kHz (Ahi) minus the maximum spectrum amplitude (Alo) below 2 kHz (in dB), and (3) the second formant frequency F2 at the time of the onset of voicing in the following vowel.   Each of these measures showed differences between /ð/ and /d/ in the expected direction (higher peak frequency, larger Ahi-Alo and higher F2 for /d/).   The frequency of the spectrum peak provided the best separation between the two consonants.   Future work will examine other possible cues as well as combinations of these cues.

3.2 Comparison of acoustics and perception of nasal consonants in two languages
In both English and Mandarin Chinese there is a phonemic inventory of three nasal consonants. In English, all of these consonants (/m/, /n/, and /ŋ/) can occur in postvocalic position, but only two (/m/ and /n/) occur prevocalically.   In Mandarin, the same two nasal consonants (/m/ and /n/) can appear prevocalically, and only the alveolar and velar nasals occur in coda position.   When these two postvocalic consonants follow the low vowel /a/, the vowel quality is modified so that the front vowel [æ] follows /n/ and the back vowel [ɑ] follows /ŋ/.   For English, on the other hand, which has two underlying low vowels, all four sequences [æn], [æŋ], [ɑn], and [ɑŋ] are possible. There are also more minor influences on vowel quality in Mandarin when the two nasal codas follow nonlow vowels.

For both English and Mandarin, therefore, the feature nasal is distinctive for consonants, as are the three places of articulation for these consonants.   However, there are different constraints on the contexts in which the consonants can occur, and there are also differences in the inventories of vowels in the two languages.   In the present study, we are examining how these language differences can influence the acoustic implementation and the perception of nasals in the two languages.   The goal is to develop a theoretical framework in terms of which these differences can be explained.

In an initial perception study, several examples of four VC sequences [Cɑn], [Cɑŋ], [Cæn], and [Cæŋ] were recorded by an English speaker, where C = several different initial stop consonants. These items were presented to several Mandarin-speaking listeners who were instructed to identify the Mandarin syllable /ɑn/ or /æŋ/ which most closely resembled the English syllable. Acoustic measurements of trajectories of the first two formants F1 and F2 were made for the English syllables and also for versions of the two Mandarin syllables produced by a native speaker.   The perception results showed that the English syllables /ɑŋ/ and /æn/ were identified as being similar to /ɑŋ/ and /ɑn/ in Mandarin.   However the responses to /ɑn/ and /æŋ/ were more variable, with no clear preference in either case.   Acoustic analysis of /an/ and /aŋ/ in Mandarin showed clear differences in the trajectories of F2, with F2 being higher for the vowel in /an/ and lower for the vowel in /aŋ/.   It is hypothesized that the vowel quality in the Mandarin syllables is modified in order to enhance the perceptual saliency of the consonantal place contrast.

3.3 Coarticulation of alveolar nasal consonants with following labial consonants
Several past studies have shown that when a syllable-final alveolar stop consonant is followed by a labial or velar stop, the acoustic and perceptual signature for the alveolar place of articulation is either deleted or is replaced by a labial or velar consonant.

An acoustic study has been initiated to examine similar consonant sequences when the syllable-final alveolar is a nasal consonant and the following consonant is a labial stop or a labial nasal (e.g., Dan bakes or Dan makes). Some of the acoustic attributes that distinguish place of articulation for nasals are different than those for stops, since nasal consonants lack the noise bursts that provide some of the cues for stop consonants. A database of sentences was designed to include a number of utterances containing the appropriate consonant sequences, as well as control sequences such as Tom bakes and Tom makes. A number of filler utterances were also used. Recordings of these utterances were made by several male and female speakers.

Various measurements of the formant frequencies were made in the vowel preceding the nasal consonant closure and in the time interval following the nasal. A general finding was that the formant movements in the Vn#b and Vn#m sequences were different from those in the Vm#b or Vm#m sequences, and that complete assimilation rarely occurred.

3.4 Enhancement and overlap in speech production
A broad goal of some of our research in speech communication has been to develop models of human production and perception of word sequences in utterances of running speech. These models assume that, in the memory of a speaker and a listener, words are represented as sequences of discrete segments, each of which consists of a series of bundles of discrete binary distinctive features. (Various other aspects of the words related to syntax and semantics are, of course, also represented in memory.)

It is well known, however, that the acoustic patterns for words and for the segments within words exhibit considerable variability depending on the context in which the segments and words occur, the rate and the style of speaking, prosodic aspects of the utterance, and characteristics of individual speakers. Among phoneticians, this lack of a clear correspondence between acoustic patterns on the one hand and the discrete features, segments, and words on the other hand have led to some skepticism in the use of distinctive features as representations from which acoustic patterns are derived in speech production, or as representations that are derived during access of lexical units in running speech.

As part of our current research, we are proposing the outlines of a model of speech production that derives the principal attributes of the speech signal from a discrete phonological representation in terms of bundles of distinctive features. There are three major aspects of this model:

(1) The distinctive features are defined by "quantal" relations between particular articulatory parameters and the acoustic properties that are generated when these parameters are manipulated. For a given distinctive feature, for a particular range of values of a defining articulation there is a corresponding defining acoustic property that is relatively insensitive to these changes in articulation. There is a universal inventory of such features, and a given language selects from this universal set.

(2) In a given language, the defining gesture involved in the implementation of a particular distinctive feature is often enhanced by the speaker through the introduction of additional articulatory gestures that contribute to the perceptual saliency of the contrast for the feature. The introduction of these enhancing gestures may depend on the segmental or prosodic context in which the feature occurs.

(3) In running speech, there is often overlap of the articulatory gestures that implement the distinctive features, so that the acoustic consequences of these gestures are weakened or even masked. This gestural overlap may reduce the perceptual saliency of some features.

3.5 Possible physical basis for some vowel categories

As a part of our general interest in quantifying articulatory/acoustic relations in speech production, we have been examining the role of tracheal resonances in influencing the possible acoustic characteristics of vowels that have formant frequencies close to these resonances. In particular, it has been proposed that the second tracheal resonance (at about 1400 Hz for adults) provides a natural dividing line between back and front vowels in language. An attempt to generate a vowel with a second formant frequency close to this line will lead to a F2 spectrum prominence that is on one side or the other of the line. We have made a number of measurements of formant frequencies and amplitudes as well as subglottal resonances for utterances in which the vowel is a diphthong such as [ai]. For such a vowel, F2 usually passes through the region where the second subglottal resonance is expected. In past work we have observed a perturbation in F2 and in the amplitude of the F2 prominence in the vicinity of the subglottal resonance for most speakers.

We have recently been examining in more detail the influence of the second tracheal resonance on the acoustic properties of these vowels. The new approach is based on the fact that acoustic coupling between the supraglottal vocal tract and the subglottal system is time-varying, since the acoustic path connecting these two systems through the glottis varies as the glottal area fluctuates periodically during phonation. During the initial part of the cycle of glottal vibration, which is initiated by the glottal closing movement, the glottis is either quite narrow or is completely closed. Later in the cycle, the glottis is more open. It is expected, therefore, that the influence of the subglottal system on the second vocal-tract resonance would be greater during the second part of the glottal cycle when the glottis is more abducted. We have reanalyzed the acoustic recordings of the diphthong [ai] for several speakers, using a time window of 4 milliseconds for the spectrum analysis. Spectra were obtained in both the initial and final portions of the glottal cycle, for a series of cycles. For most utterances the amplitude and frequency of the F2 prominence showed a greater abrupt change or discontinuity when the spectrum was sampled in the second part of the cycle, i.e., during the open phase of the cycle. This method of examining the fine structure of the glottal cycle may have application to the studies of individual differences in phonation and for investigation of disorders in laryngeal function.

3.6 A role for tracheal resonances in speech perception

Acoustic coupling between the vocal tract and the trachea results in the introduction of pole-zero pairs corresponding to resonances of the uncoupled trachea. If the second formant (F2) passes through the second tracheal resonance (T2) a discontinuity in amplitude occurs. This work explores the hypothesis that the F2-T2 discontinuity affects how listeners perceive the distinctive feature [back] in transitions from a front vowel (high F2) to a labial stop (low F2). In utterances of this kind, F2 at the beginning of the vowel is expected to be above the point of the F2/T2 discontinuity and near the end is expected to decrease to a value below this discontinuity. We synthesized two versions of an utterance ("apter") with a T2/T2 discontinuity at different locations in the initial VC transition. Subjects heard portions of the utterance with and without the discontinuity, and were asked to identify the utterance. Results show that the presence of the F2/T2 discontinuity facilitated the perception of frontness in the vowel. Discontinuities of the F2/T2 sort are proposed to play a role in shaping vowel inventories in the world's languages [K.N. Stevens, J. Phonetics 17, 3-46, (1989)]. Our results support a model of lexical access in which acoustic discontinuities subserve phonological feature identification. (This research was carried out by graduate students in three groups: Asaf Bachrach, Department of Linguistics and Philosophy, Steven Lulich, Research Laboratory of Electronics, and Nicolas Malyska, Lincoln Laboratory.)

4. Studies of Speech Development and Speech Disorders

4.1 Experiments on improving electrolarynx speech

For individuals whose larynx has been excised, phonation can be simulated using an electrolarynx (EL). The electrolarynx generates a simulation of the voice source with a buzzer-like device that is pressed against the neck. The sound radiated from the lips is reasonably intelligible, but is unnatural for several reasons: (1) the fundamental frequency (F0) is fixed; (2) it is difficult to produce voiceless obstruent consonants; (3) the waveform of the excitation is quite different from the normal glottal waveform; and (4) the point of excitation of the vocal tract is not at the end of the tract.

Lack of F0 movement is one of the more severe reasons for lack of naturalness. The goal of the present study is to devise a procedure for controlling the F0 contour with the EL device. This research is carried out in collaboration with the Voice Surgery and Rehabilitation laboratory of the Massachusetts General Hospital.

The first step in this research has been to develop the ability to synthesize EL utterances with the Klatt synthesizer (KLsyn) in our laboratory. With proper adjustment of the synthesis parameters, we were able to generate reasonable replicas of recorded EL speech available from the Voice Surgery and Rehabilitation laboratory at MGH. With this synthesizer, it was then possible to modify the F0 control to simulate the prosody of natural speech. Informal judgments of the quality of synthesized versions of EL speech with modified F0 contours indicated a significant increase in naturalness. In the course of doing the synthesis, it was observed that there was considerable variation in the amplitude of the EL speech, and this variation was incorporated in the synthesis, primarily by manipulating the parameter AV in the synthesizer. Acoustic analysis of a number of EL utterances showed that there was always a significant variation in amplitude during the vowels, particularly near the termination of an utterance.

In addition to having a database of EL utterances from a number of subjects who had their larynx removed, the group at MGH also had recordings of the same utterances from the same speakers, obtained before the laryngeal surgery. Acoustic analysis of these utterances showed that there was some correlation between the fluctuation of F0 and the amplitude. An algorithm was developed to calculate F0 fluctuations based on the amplitude fluctuations in the EL speech. Informal listening to these utterances with automatically generated F0 contours showed substantial improvement in naturalness of the speech.

Current work is directed towards obtaining more detailed documentation of listener judgments of this EL speech with F0 modification. A related question that is being addressed is: What is the mechanism that leads to variation in the amplitude of EL speech?

4.2 Acoustic characteristics of sounds produced by children

We have carried out further acoustic analysis and interpretation of stop consonants of children as they develop in the age range 2-4 years. Five principal findings from these data, which compare children's productions and adult productions, are summarized as follows:

1. Based on acoustic measurements of spectra of stop bursts, it appears that the positioning of the primary articulator (labial, alveolar, velar) by the children is correct.

2. The tongue body adjustment (as opposed to positioning of the primary articulator) shows some differences from normal adult values. For example, transitions of the second formant frequency (an indicator of tongue body movement) for alveolar and velar consonants are not as well separated as they are for adults.

3. The release mechanism for the primary articulator differs from that of adults, as evidenced by the temporal shape of the burst in the initial part of the consonant release.

4. The children are still developing the appropriate glottal adjustments and intraoral pressure for stop consonant production, including vocal fold stiffness and glottal spreading.

5. The children have less control over subglottal pressure, as evidenced by high variability in amplitude measurements.

Some of these differences are presumably due to anatomical and physiological differences between the children and adults, some arise from continuing development of general "postures" for speech production such as subglottal pressure and adjustment of stiffness of tissue surfaces, and others arise from continuing development of the control and coordination of particular structures in producing sound sequences.

5. Effects of Hearing Status on Adult Speech Production

Introduction
This work is continuing and extending our program of research on postlingual deafness and the role of hearing in speech production. We are characterizing the speech production of adults who were deafened postlingually as children or adults and have had varying degrees of experience with auditory prostheses; and we are describing the changes in speech communication that take place in these deaf adults when they receive cochlear implants. We aim to contribute to the research literature on the role of hearing and hearing loss in speech production; specifically to the body of knowledge concerning the effects of long and short-term changes in auditory feedback on speech, including (i) the deterioration of speech in long-term deafness, (ii) the effects of conditions for speech communication, such as environmental noise and visible articulation, (iii) the effects of age at hearing loss and its relation to later speech production and cortical activation in relation to age at hearing loss, and (iv) audio-visual integration in speech production. During this year, nine studies were completed. These results are described below. (Some of these studies were described in less final form in previous RLE reports.)

5.1 Effects of short- and long-term changes in auditory feedback on vowel and sibilant contrasts
Vowel and sibilant productions were elicited from a core group of eight speakers with adult-onset hearing impairment before and at one month and one year after they received cochlear implants. Measures were taken with auditory feedback and also without it (implant microphone turned off). Normal-hearing controls also read the word lists, with auditory feedback masked and unmasked. Compared to controls, the postlingually deaf speakers produced smaller vowel and sibilant contrasts pre-implant. Once implanted, they recognized more phonemes and produced more distinct phoneme contrasts after a year's prosthesis use compared to just one month. Their contrasts decreased when their auditory feedback was interrupted in the two time samples.

5.2 Effects of bite blocks and hearing status on vowel production
The core group of implant users and hearing controls read elicitation lists of /hVd/ syllables with and without bite blocks and auditory feedback. Recording sessions were held pre-implant and one month and one year after, and once for controls. Pre-implant, long-standing absence of auditory feedback was associated with heightened dispersion of vowel tokens, which was inflated further by inserting bite blocks. The restoration of some hearing by the implant reduced dispersion. Deaf speakers' vowel contrasts were reduced compared to controls. Insertion of bite blocks reduced them further because of the speakers' incomplete compensation. A year of implant use expanded the vowel space measured with auditory feedback during elicitation, both with and without the bite blocks. The improvement from pre-implant to one year supports the inference that models of speech production must assign a role to auditory feedback in error-based correction of feedforward commands.

5.3 Effects of deafness and length of implant use on mapping allophones into a single auditory goal

In hearing speakers, /r/ allophones vary little in the primary acoustic cue, lowered F3, despite marked differences in their tongue shapes. We hypothesized that deaf speakers would show larger allophonic variation than controls and that auditory feedback from a cochlear implant would reduce that variation. Seven of eight of the postlingually deaf speakers exhibited an acoustic variation significantly greater than that of the hearing speakers' pooled. After a year of auditory prosthesis, however, their allophonic variation was comparable to that of the controls. In terms of our neurocomputational model, a year's experience of auditory feedback with the implant enabled speakers to reduce the size of their auditory goal regions for /r/ into a normal range. Implant users' perception of stop and stop+/r/ blends also improved from one month to one year post-implant. However, their productions of the corresponding contrast distances were still reduced compared to controls.

5.4 On the structure of phoneme categories in listeners with cochlear implants

Presented with synthetic speech continua that implemented a vowel and a sibilant contrast, the core group of implant users yielded phoneme labeling, discrimination, and within-category perceptual structure (based on goodness ratings) that were anomalous at one month post-implant. Labeling and goodness ratings were also measured at one year post-implant; labeling had improved with a year's prosthesis use, within-category structure had not. Compared to controls, the implant users' vowel and sibilant labeling slopes were substantially shallower. Their sensitivity to subtle phonetic differences within phoneme categories was about half that of the controls. Prototypes for the sibilant contrast were not as differentiated as they were among the controls.

5.5 Effects of speaking condition and hearing status on vowel production in postlingually deaf adults with cochlear implant

A different group of cochlear implant users produced nine American English vowels in three time samples (prior to implantation, one month, and one year after implantation); in three speaking conditions (clear, normal, and fast); and in two feedback conditions after implantation (implant processor turned on and off). Despite longstanding deafness, speakers produced differences in spectral contrast and duration between clear and fast conditions in the range of those found for normal-hearing controls, leading to the inference that maintenance of these distinctions may not be influenced by hearing status. Instead, the difference between fast and clear speech may simply reflect a difference in movement durations, as slower movements are larger and larger movements result in greater contrast distances. Vowel contrast did not change from pre-implant to one month, presumably because speakers had not had enough experience with hearing from the implant to adequately retune their auditory feedback system, needed to improve feedforward commands. After a year of implant use, contrast distances had increased relative to pre-implant with processor on but also with it off, indicating improvement in feedforward commands.

5.6 Effects of masking noise on vowel and sibilant contrasts in normal-hearing speakers and postlingually deafened cochlear implant users

The role of auditory feedback in speech production was investigated by examining how speakers modify phonemic contrasts in response to increases in the noise-to-signal (N/S) level in that feedback. The core group of cochlear implant users and normal-hearing controls pronounced utterances containing vowels and sibilants while hearing their speech mixed with noise at seven equally-spaced levels over their tolerated ranges. Average vowel duration and SPL rose with increasing N/S, as expected. Overall vowel contrast rose initially then fell with increasing N/S. The initial rise in vowel contrast is interpreted as part of speakers' attempts to maintain clarity under degraded acoustic transmission conditions. As N/S increases further, speakers no longer detect phoneme contrasts in their auditory feedback, and economy of effort presumably becomes an increasingly predominant influence, leading to contrast decrements. The implant users' vowel results approached the controls' pattern with a year of implant use.

5.7 Audiovisual integration in normal-hearing and hearing-impaired subjects

In listeners with normal hearing, the sight of a speaker articulating a syllable can influence the syllable that is heard, most observably when the normally coupled sight and sound of the syllable have been separated and the sight paired with a different speech sound. This study investigated such audiovisual integration of the spoken syllable ("the McGurk effect") in hearing-aid users and in normal-hearing controls. We hypothesized that hearing-aid users would rely more on visual input and thus their percepts would be biased more toward the visual stimulus when the visual and auditory components are mismatched. Both groups observed computer-presented syllables that paired three plosives with three vowels under three conditions: auditory-only, visual-only, and audiovisual. Participants labeled each stimulus according to the consonant perceived. There was no difference between the groups in the degree to which they integrated the audio and visual components of mismatched stimuli. However, when integration did not occur, the hearing-aid users gave significantly more visual responses while the normal-hearing group gave significantly more auditory responses.

5.8 Cortical networks underlying audio-visual speech perception in normal-hearing and congenitally deaf individuals

Functional magnetic resonance imaging (fMRI) was used to investigate brain activity underlying audiovisual speech perception in congenitally deaf persons and normal-hearing controls. Subjects were presented with three different types of speech stimuli: auditory-only, visual-only, and audiovisual. The stimuli were isolated vowels or CVCV syllables, presented in different blocks. In the audio-only condition, deaf subjects showed less cortical activity than controls, as expected. In the visual-only condition, deaf subjects showed distinctly more activity in the right hemisphere than controls and far more activity in the frontal lobes, including Broca's area. In the audiovisual condition, controls showed more activation in auditory cortical areas than in the visual-only condition but for deaf subjects, auditory cortex was heavily activated in both conditions. Analyses of neural connectivity between brain regions in the visual-only condition for both groups of subjects suggest that a pathway through the fusiform gyrus is crucial for transforming visual inputs into auditory cortical activation.

5.9 Timing of SPL and contrast changes in response to a modification of auditory feedback

The timing of changes in parameters of speech production was investigated in six cochlear implant users by switching their implant microphones off and on a number of times in a single experimental session. The subjects repeated four short $/dV_1n'SV_2d/$ utterances (S=/s/ or /ʃ/) in semi-random order. The changes between hearing and non-hearing states were introduced by a voice-activated switch within 20 msec of V_1 onset; the number of utterances between switches was varied to minimize subject anticipation of the switches. Measures included, for the vowels, SPL, duration, F0, and contrast distance; for the sibilants, spectral mean and contrast distance. Changes in parameter values were computed by averaging multiple tokens, lined up with respect to the switch. Five of the six subjects changed at least one of these measures in response to changes in hearing state. For the first utterance produced post-switch, the initial vowel (V_1) was unchanged, while the following vowel (V_2) in the same utterance did change in relation to vowels in the utterances immediately before the switch. This was true of switches in either direction, on-to-off or off-to-on. In terms of our model, abrupt changes in the availability of auditory feedback led to adjustments in feedforward commands observed in the second syllable following the switch.

5.10 Control of voice-onset time in the absence of hearing: A review

The relation between partial or absent hearing and control of the voicing contrast has long been of interest to investigators, in part because speakers who are born deaf characteristically have great difficulty mastering the contrast and in part for the light it can cast on the role of hearing in the acquisition and maintenance of phonological contrasts in general. This review of the literature observes that both prelingually and postlingually deaf speakers have a tendency to reduce the

difference between voiced and voiceless VOT. The separation of the cognate VOTs can be improved when some hearing is restored with a cochlear implant. Both populations also present anomalies in speech breathing that can hinder the development of intraoral pressures and transglottal pressure drops that are required for the production of the VOT contrast. The successful management of VOT further requires critical timing among phonatory and articulatory gestures, most of which are not visible, rendering the VOT contrast a particular challenge in the absence of hearing.

6. Speech Prosody

6.1 The distribution of phrase-final lengthening
We have looked more closely at the well-known phenomenon of duration lengthening at the ends of spoken phrases in American English, and found that (contrary to earlier assumptions) it is not limited to the final syllable of the phrase. Although most of the boundary-related duration lengthening occurs in the final syllable, additional lengthening (approximately 10%) is also seen on the main-stress syllable of phrase-final words. Critical cases that demonstrate this additional main-stress-syllable lengthening are words in which the main-stress syllable is not the phrase-final syllable, as in *Madison* and *Trinidad*. This finding suggests that models of boundary-related duration adjustment must be more complex than presently envisioned, to produce naturally-patterned synthesized utterances. (Work carried out with Dr. Alice Turk, Department of Linguistics, University of Edinburgh.)

6.2 Spoken prosody and speech accompanying gestures
Our extensive experience with prosodic labeling using the ToBI system for marking pitch accents and boundary tones allows us to investigate the timing of speech events and gestures of the head and hands that accompany speaking. In particular, we ask whether these head and hand movements align with the prosodic structure of spoken utterances, and provide markers for disfluencies. If so, it would suggest both the psychological reality of prosodic structure, and a production planning process that coordinates gestures of the vocal tract with those produced by other motor systems.

(a) Alignment of gestural 'hits' with pitch accents. Some speech-accompanying gestures end very abruptly, with a short sharp stop (or hit) that can be precisely located in the speech stream. When speech samples from video-taped lectures are independently labeled for their acoustic landmarks and prosody (from the acoustic wave form) and for the video frame that corresponds to each gestural hit (from the silent video display), and the two sets of labels are integrated, we find that head and hand hits appear to be timed to occur with phrasally-prominent syllables, i.e. they occur on a pitch-accented syllable or on an immediately-following reduced syllable. This suggests that speakers plan their vocal and non-vocal gestures together, and raises interesting questions about which pitch accents receive gestural marking and which ones do not.

(b) Marking of speech errors with facial gestures. In addition to studying the alignment of head and hand gestures with pitch accents, we have investigated the distribution of eyebrow hits in speech. While these speech-accompanying gestures did not align systematically with pitch accents (in a lecture sample from one speaker), they did occur preferentially at locations that involve speech errors, such as 'ungovernmentally sh---sanctioned' or 'har---artists had…'. Although this combined pattern does not hold for all speakers, it raises the possibility that speakers may provide specific gestural cues to help listeners recognize the occurrence of disfluencies that might otherwise disrupt the communicative process.

6.3 Prosody labeling workshop and expanded tutorial
We organized an NSF-sponsored workshop on ToBI labeling of the prosody of spontaneous speech (August 2004), with experienced ToBI labelers from a number of countries as invited

attendees.   The results of those discussions have been incorporated into a new introductory tutorial for training apprentice labelers, a discussion website, and a project comparing labeler accuracy for several different prosody labeling systems. We expect these resources to increase the pool of trained prosody labelers, as well as encouraging on-line discussion of challenging research issues in prosody.

6.4 Spectral correlates of lexical-prosody
This study is part of a broader project that seeks to develop a quantitative model of the influence of certain prosodic contexts on the acoustic and articulatory attributes of full vowels (i.e., vowels that are not reduced).   Native speakers of American English were asked to name bisyllabic visualizable nouns containing a stressed vowel and an unstressed full vowel. Subjects uttered short phrases containing these words. In one group of phrases the main accent in the phrase was on the stressed vowel in the target word. In a second group the main accent was on an adjective preceding the target word. Various acoustic measures were made on the vowels. These measures were selected to provide indications of laryngeal adjustments, duration, and amplitude as determined by subglottal pressure and by glottal configuration. Results show that within the category of full vowels, unstressed and stressed vowels can be distinguished by syllable/vowel durations and spectral tilt. Spectral tilt (SpT) is an acoustic measure related to the degree of glottal spreading. Stressed full vowels have longer duration and less SpT. Distinction between unaccented and accented stressed vowels can be made by a measure of amplitude of voicing (AV), F0 (pitch), and intensity contour differences. Accented stressed vowels have higher pitch, and greater AV and intensity. These results suggest that there are acoustic correlates to lexical stress that can be used to determine the stressed syllable of a word, regardless of whether or not it is pitch accented.

7. Data-Sharing Initiative

We have begun participating in the MIT Libraries DSpace initiative, with the goals of (1) sharing databases developed in our lab, and (2) making available electronic versions of theses written by lab members.

Researchers who are attempting to develop a set of principles governing the various acoustic and articulatory manifestations of distinctive features must rely on empirical data obtained from utterances produced by different talkers. In the course of our research, we have prepared and used several such databases. These databases include citation forms of words or syllables with different phonetic contexts, a variety of speakers, different prosodic contexts, different speaking styles, both read and casual speech, etc.   In addition to citation forms of words and read sentences and paragraphs, the database includes naturally spoken utterances in a situation where pairs of talkers were solving a puzzle by means of dialogue. The database also includes a few CV and CVC utterances produced as part of a cineradiographic study of speech production. The latter utterances consist of words containing different consonants and vowels produced by one talker. As a part of this database, digitized midsagittal images of individual frames of the radiographs are included.

These databases will be of interest to researchers who are studying the issues of variability that we address in our work. A specific benefit is that researchers can jumpstart a research project by either basing it entirely on our data, or by using the data to do pilot studies. Based on the results of the pilot studies, the researchers can determine if it is worthwhile to record new data. Data sharing also enables researchers to compare the results of new analysis techniques to older methods by using the original data. This technique is often used in the speech-recognition community, in which a known database provides a sort of reference that one can easily use to test recognition algorithms. It is then easy to compare new results to older results appearing in the literature. In our lab, we frequently get requests from researchers who would like to use some of our databases for the reasons just given: to use as pilot data and for testing new analysis techniques.

The databases will also be useful for education purposes, particularly in laboratory-based courses or classes that require projects involving speech analysis. Educators often look for small databases that illustrate a particular point (e.g., databases of consonant-vowel syllables to compare formant transitions between the consonant and the vowel). Likewise, for students doing term projects (by necessity quite short-term), a ready-made database is extremely helpful.

Our second goal is to make available materials that are normally not easy to obtain, such as copies of theses or unpublished manuscripts. In the first phase of this project, we are creating electronic versions of theses written by current or former members of the Speech Communication Group. In later phases we will work on other materials, such as working papers, presentation slides, and unpublished manuscripts.

Publications

Journal Articles, Published

J.S. Perkell, M.L. Matthies, M. Tiede, H. Lane, M. Zandipour, N. Marrone, and E. Stockmann, "The Distinctness of Speakers' /s-ʃ/ Contrast is Related to their Auditory Discrimination and Use of an Articulatory Saturation Effect," *J. Speech, Language and Hearing Res.* 47: 1259-69 (2004).

J.S. Perkell, F. Guenther, H. Lane, M.L. Matthies, E. Stockmann, M. Tiede, and M. Zandipour, "The Distinctness of Speakers' Productions of Vowel Contrasts is Related to their Discrimination of the Contrasts," *J. Acoust. Soc. Am.* 116: 2338-44 (2004).

A. Nieto-Castanon, F. Guenther, J.S. Perkell, and H. Curtin, "A Modeling Investigation of Articulatory Variability and Acoustic Stability during American English /r/ Production," *J. Acoust. Soc. Am.* 117: 3196-3212 (2005).

Journal Articles, Accepted for Publication

S.J. Keyser, and K.N. Stevens, "Enhancement and Overlap in the Speech Chain," *Language,* forthcoming*.*

H. Lane, M. Denny, F.H. Guenther, M. Matthies, L. Ménard, J. Perkell, E. Stockmann, M. Tiede, J. Vick, and M. Zandipour, "Effects of Bite Blocks and Hearing Status on Vowel Production," *J. Acoust. Soc. Am*., forthcoming.

H. Lane, & J.S. Perkell, "Control of Voice-Onset Time in the Absence of Hearing: A Review," *J. Speech Lang. Hear. Res*., forthcoming.

A.E. Turk, and S. Shattuck-Hufnagel, "Word-Boundary-Related Duration Patterns in English," *J.Phonetics*, forthcoming.

K.N. Stevens, "The Acoustic/Articulatory Interface," Invited review for *J. Acoust. Soc. Japan*, forthcoming.

Journal Articles, Submitted for Publication

H. Lane, M. Matthies, M. Denny, F.H. Guenther, J. Perkell, E. Stockmann, M. Tiede, J. Vick, and M. Zandipour, "Effects of Short- and Long-term Changes in Auditory Feedback on Vowel and Sibilant Contrasts," submitted to *J. Ear & Hearing*.

H. Lane, M. Denny, F.H. Guenther, M. Matthies, J. Perkell, E. Stockmann, M. Tiede, J. Vick, and M. Zandipour, "On the Structure of Phoneme Categories in Listeners with Cochlear Implants," submitted to *J. Speech Lang. Hear. Res*.

C.Y. Lee, and K.N. Stevens, "Strident Fricatives in Mandarin Chinese: From Acoustics to Articulation and Features," submitted to *J. Phonetics*.

N.L. Marrone, H. Lane, E.S. Stockmann, J. Perkell, F.H. Guenther, and S. Ghosh, "Audio-Visual Integration in Normal-Hearing and Hearing-Impaired Subjects," submitted to *J. Speech Lang. Hear. Res*.

L. Ménard, T. Balkany, M. Denny, H. Lane, M. Matthies, J. Perkell, M. Polak, E. Stockmann, M. Tiede, J. Vick, and M. Zandipour, "Effects of Speaking Condition and Hearing Status on Vowel Production in Postlingually Deaf Adults with Cochlear Implant," submitted to *J. Acoust. Soc. Am*.

J. Slifka, "Some Physiological Correlates to the Irregular and Regular Phonation at the End of an Utterance," submitted to *J. Voice*.

Book/Chapters in Books

F. Guenther, and J. Perkell, "A Neural Model of Speech Production and its Application to Studies of the Role of Auditory Feedback in Speech," in *Speech Motor Control in Normal and Disordered Speech*, eds: B. Maassen, R. Kent, H.F.M. Peters, P. Van Lieshout & W. Hulstijn (Oxford University Press, 29-50, 2004).

S. Shattuck-Hufnagel, "Prosodic Structure and Surface Phonetic Variation: Implications for Models of Speech Production Planning," in *Proc. Laboratory Phonology 8*, ed. L. Goldstein (2004).

Slifka, J. "Respiratory System Pressures at the Start of an Utterance," In *Dynamics of Speech Production and Perception*, ed., P. Divenyi, (IOS Press, 2005).

K. Stevens, Z. Li, C-Y. Lee and S.J. Keyser, "A Note on Mandarin Fricatives and Enhancement". In *From Traditional Phonology to Modern Speech Processing*, eds., G. Fant, H. Fujisaki, J. Cao, and Y. Xu, (Beijing: Foreign Language Teaching and Research Press, 393-403, 2004).

K.N. Stevens, "Features in Speech Perception and Lexical Access", In *The Handbook of Speech Perception*, eds., D. Pisoni and R. Remez, (Cambridge, MA: Blackwell Publishers, 125-155, 2005).

Meeting Papers, Presented

A. Bachrach, S. Lulich, and N. Malyska, "A Role for Tracheal Resonances in Speech Perception," paper presented at the 149th Meeting of the Acoustical Society of America, Vancouver, Canada, May 16-20, 2005.

A.K. Imbrie, "Acoustical Study of the Development of Stop Consonants in Children," paper presented at the 149th Meeting of the Acoustical Society of America, Vancouver, Canada, May 16-20, 2005.

M.K. Tiede, F.H. Guenther, J.S. Perkell, M. Zandipour, G. Houle, and D.J. Ostry, "Perturbation and Compensation in Speech Acoustics using a Jaw-Coupled Robot," paper presented at the 148th Meeting of the Acoustical Society of America, San Diego, California, November 15-19, 2004.

J.J. Yoo, F.H. Guenther, and J.S. Perkell, "Cortical Networks Underlying Audio-Visual Speech Perception in Normally Hearing and Hearing Impaired Individuals," paper presented at the 148th Meeting of the Acoustical Society of America, San Diego, California, November 15-19, 2004.

Meeting Papers, Published

V. Villacorta, J.S. Perkell, and F.H. Guenther, "Relations between Speech Sensorimotor Adaptation and Acuity," J*ournal of Acoustical Society of America, 117* (*4*), 2618-9, 2005.

Y. Yasinnik, S. Shattuck-Hufnagel, and N. Veilleux, "Gesture marking of disfluencies in spontaneous speech." To appear in *Proc. of the Disfluency in Spontaneous Speech Workshop*, Aix-en-Provence, France, 2005.

J.J. Yoo, F.H. Guenther, and J.S. Perkell, "Cortical Networks Underlying Audio-Visual Speech Perception in Normally Hearing and Hearing Impaired Individuals", *Proceedings of the Workshop Plasticity in Speech Perception*, London: UCL Centre for Human Communication, June 15-17, 2005.

Theses

L. Dilley, *The Phonetics and Phonology of Tonal Systems*, PhD thesis, Harvard-MIT Division of Health Sciences and Technology, 2005.

A.Imbrie, *Acoustical Study of the Development of Stop Consonants in Children*, PhD thesis, Harvard-MIT Division of Health Sciences and Technology, 2005.

M. Sonderegger, *Subglottal Coupling and Vowel Space: An Investigation in Quantal Theory*, S.B. thesis, Department of Physics, MIT, 2004.