Speech Communication

Sponsors
C.J. LeBel Fellowship
Dennis Klatt Memorial Fund
Donald North Memorial Fund
National Institutes of Health (Grants R01-DC00075,
R01-DC01925,
R01-DC02125,
R01-DC02978,
R01-DC03007,
RO1-DC04331,
1 R29 DC02525,
T32-DC00038,
and National Science Foundation (SES-9820126).

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, Dr. Marilyn Chen, Dr. Lisa Lavoie, Dr. Chao-Yang Lee, Jennell Vick, Majid Zandipour, Dr. Margaret Denny, 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, 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, MIT Media Laboratory, Cambridge, Massachusetts.

Graduate Students
Lan Chen, Xuemin Chi, Laura Dilley, Tony Okobi, Neira Hajro, Annika Imbrie, Steven Lulich, Nicole Marrone, Xiaomin Mou, Chi-youn Park, Ariel Salomon, Atiwong Suchato, 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, Kushan Surana, Enrique Urena, Sumudu Watugala, Betty Yang, Yelena Yasinnik, Man Yin Yee, Wei-An Yu

Technical and Support Staff
Arlene E. Wint

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

1.1 Overview

The aim of the project on Lexical Access from Features is to develop a model of the process by which human listeners convert a continuous speech signal into a discrete sequence of words. The system contains a lexicon of words, each represented by a sequence of bundles of distinctive features, and these words are recognized through identification of acoustic landmarks in the signal, and extraction of acoustic cues for the features in the vicinity of these landmarks. The acoustic analysis is carried out with interaction with the lexicon, so that hypotheses about the words are constantly being updated and verified through an analysis-by-synthesis process. Several modules of this system have been completed or are under development as described in the following sections.

1.2 Toward modules for identifying place features for vowels: the features [low] and [high]

The goal of this project is to automatically detect the features [high] and [low] for vowel segments based on measurements from average spectra. A long-term and a short-term average spectrum are computed for all vowel regions in the utterance and these average spectra are used to estimate speaker-specific parameters such as average F0 and average F3 (an indicator of vocal tract length). These parameters are used to estimate F1 using a peak-picking process on the average spectrum at each vowel landmark. Preliminary results are derived from read connected speech for 738 vowels from 80 utterances (two male speakers, two female speakers). Speaker-independent logistic regression analysis using only average F0 and F1 determines the feature [high] with 73% accuracy and the feature [low] with 84% accuracy.

1.3 Module for place of articulation for stop consonants

This study is concerned with acoustic attributes used for identifying the place of articulation features for stop consonant segments, i.e., labial, alveolar, and velar. The acoustic attributes are selected so that they capture the information relevant to place identification, including amplitude and energy of release bursts, formant movements of adjacent vowels, spectra of noises after the releases, and some temporal cues.

An experimental procedure for examining the relative importance of these acoustic attributes for identifying stop place was developed. The ability of each attribute to separate the three places is evaluated by the classification error based on the distributions of its values for the three places,

and another quantifier based on F-ratio. These two quantifiers generally agree and show how well each individual attribute separates the three places.

Combinations of non-redundant attributes were used for the place classifications based on Mahalanobis distance. When stops contain release bursts, the classification accuracies are better than 90%. It was also shown that knowledge of voicing and vowel frontness contexts lead to a better classification accuracy of stops in some contexts. When stops are located between two vowels, information on the formant structures in the vowels on both sides can be combined. Such combination yields the best classification accuracy of 95.5%. By using appropriate methods for stops in different contexts, an overall classification accuracy of 92.1% is achieved.

Linear discriminant function analysis was used to address the relative importance of these attributes when combinations are used. Their discriminating abilities and the ranking of their relative importance to the classifications in different vowel and voicing contexts were obtained. The overall findings are that attributes relating to the burst spectrum in relation to the vowel contribute most effectively, while attributes relating to formant transition are somewhat less effective. The approach used in this study can be applied to different classes of sounds, as well as stops in different noise environments.

1.4 Module for the feature [nasal]

The focus of this project was the design, implementation, and evaluation of a set of automated algorithms to detect nasal consonants from the speech waveform in a distinctive feature-based speech recognition system. The study used a VCV database of over 450 utterances recorded from three speakers --- two male and one female. The first stage of processing for each speech waveform included automated 'pivot' estimation using a Consonant Landmark Detector. These 'pivots' were considered possible sonorant closures and releases in further analyses. Estimated pivots were analyzed acoustically for the nasal murmur and vowel-nasal boundary characteristics. For nasal murmur, the analyzed cues included observing the presence of a low frequency resonance in the short-time spectrum, stability in the signal energy, and characteristic spectral tilt. The acoustic cues for the nasal boundary measured the change in the energy of the first harmonic and the net energy change of the 0-350 Hz and 350-1000 Hz frequency bands around the pivot time. The results of the acoustic analyses were translated into a simple set of general acoustic criteria that detected 98% of true nasal pivots. The high detection rate was partially offset by a relatively large number of false positives --- 16 percent of all non-nasal pivots were also detected as showing characteristics of the nasal murmur and nasal boundary. The advantage of the presented algorithms is in their consistency and accuracy across users and contexts, and their potential with applicability to spontaneous speech.

1.5 Recognition of English vowels using a top-down method

A basic approach to speech recognition is first to obtain a cohort of possible words or word sequences, either from initial acoustic analysis of the signal or from higher-level knowledge. The validity of each member of the cohort is then evaluated through internal synthesis of the expected sound pattern (or some relevant aspects of that pattern), and assessing the degree to which this internally generated pattern matches the input sound pattern. The selected member of the cohort is the one providing the best match to the input sound pattern. To gain experience with the implementation of top-down processing, we have examined a task in which a database of several hundred C_1VC_2 English words was recorded by several speakers. Several algorithms were developed to recognize the vowels in these words when various kinds of information about the words were available. Vowel recognition scores were obtained for four conditions: (1) the vowel was identified from measurements of the formant frequencies in mid vowel, with no knowledge of the consonants; (2) the algorithm had knowledge of the identity of the consonants and there was some normalization of the first two formant frequencies across speakers, together with inclusion of the effects of the consonants on vowel durations; (3) the lexical constraint was imposed so that the identified word had to be in the lexicon; and (4) both conditions (2) and (3) were imposed.

The cues that were used were the automatically extracted vowel formant frequencies, and the (normalized) durations. Conditions (2) and (3) both gave significant gains in identification relative to condition (1), for which the identification score was 54 percent. Condition (4) gave the best vowel identification rate of 87 percent. Individual features were also evaluated; the feature that was identified with the lowest score was [tense/lax].

2. Speech Prosody

2.1 Some physiological correlates to the end of phonation at the end of an utterance

This research examines respiratory and acoustic events at the ends of utterances that coincide with the end of a breath. Glottal area is estimated from pressure and flow measures, and data are analyzed at the start of the final fall in signal amplitude and at the end of phonation. For four speakers of American English, the termination of glottal vibration results from an increase in glottal area by a factor near 2 to 4 and a 1-2 cm H_2O drop in alveolar pressure. Physiological correlates are examined for utterances that end both with regular and with irregular vocal fold vibration. Irregular vocal fold vibration for these speakers may be described as largely adducted (closed over the majority of the cycle) or as largely abducted (open over the majority of the cycle) with the latter being the dominant type at utterance termination. Acoustic cues can be used to distinguish between these two types without the need for obtaining other physiological measures. In some cases, there is a lack of contact between the folds with almost sinusoidal modulation of the airflow and others show the generation of an acoustic wave at both the collision and separation of the folds.

2.2 Alignment of F0 peaks and valleys

The occurrence of peaks and valleys of the f0 contour of an utterance on non-prominent syllables in American English (as on the *–ing* or *a-* in *reading again*) raise the question of how to label these inflection points. Analysis of samples from prosodically-labeled corpora of natural speech shows that sequences with an f0 peak on a weak syllable between them can occur quite commonly in continuous communicative speech. Informal listening suggests that the alignment of these f0 peaks with specific non-prominent syllables between the two accented syllables can change the perceived relative prominences of the accents. This observation is supported by results of perceptual experiments using synthesized f0 contours: the location of the peak in the weak syllable can shift the perceived strongest prominence from the initial syllable to the final syllable of a word like *lemonade* or *millionaire*. These findings illustrate the pervasiveness of f0 inflection points that are not aligned with syllables perceived as prominent, and suggest that alignment of the inflection point is a critical aspect of the specification of an intonational contour.

2.3 Timing of speech-accompanying gestures

The hypothesis that some of the hand and head movements produced during speaking are timed with respect to the prosodic structure of the utterance was tested, by comparing the timing of separately-labeled video and sound files from videotaped lectures. Word boundaries and prosody were labeled in the sound files using the ToBI system to specify location of pitch-accented words and syllables, and of intonational phrase boundaries. Gestures were labeled in the video files using a two-way distinction between discrete, 'stroke'-like movements and more continuous movements. The time location of movements was expressed in terms of frame location in a 30-frames-per-second video display; video frames corresponding to target attainment for the discrete gestures and to onset and offset for all gestures were then aligned with the word boundary time markings in the sound files. Preliminary analysis of samples from three professional male lectures suggests that (a) discrete gestures may be timed with respect to pitch accented syllables, and (b) gesture onsets and offsets may be timed with respect to intonational phrasing. If ongoing studies of additional utterances and speakers confirm these results, it will

provide evidence that speech planning models need to generate a speaking plan and a gesturing plan in tandem.

2.4 Other prosody-related studies for English

Measurements of the duration and amplitude have been made for a number of examples of singleton and geminate consonants across word boundaries. For example, comparisons were made between *grey talks* (syllable-initial /t/), *great auks* (syllable-final /t/) and *great talks* (geminate /t/). Results for several speakers suggest that the duration of the release noise is shorter for the geminates, suggesting that either the intraoral pressure is smaller at the end of the geminate or the release is more rapid, or both.

The duration of a monosyllabic verb has been measured in various prosodic relations with a following weak syllable, e.g., *bake apples* (no weak syllable; *bake* forms a monosyllabic foot), *baking apples* (the affix –*ing* forms a two-syllable foot with the root *bake*), *bake us apples* (the weak pronoun *us* forms a clitic group with the preceding verb *bake*), and *bake an apple* (the weak article *an* does not form a clitic with the preceding verb *bake*, but may form a rhythmic foot). Preliminary results suggest that the affix –*ing* shortens the verb substantially in comparison with its duration in the control string *bake apples,* as shown by Lehiste et al. (1972), and the pronoun *us* often does as well, while the article *an* does not. Current analyses are testing the hypothesis that the shortening effect of *us* occurs only when this pronoun is integrated with the preceding verb into a Clitic Group, and fails to occur when *us* is produced as a separate prosodic element. Such a finding would provide further evidence for the claim that prosodic structure governs many aspects of duration variation.

2.5 Japanese repetition of normal and prosodically-modified English words

Native speakers of Japanese were asked to repeat 2 and 3-syllable English words, with varying lexical stress placements, spoken by a native speaker of American English. One group of words consisted of unaltered sonorant and semi-sonorant nouns, while the second group contained the same words, but with artificially flattened fundamental frequency (F0) and intensity contours. The accuracy of Japanese speakers in producing the prosodic characteristics of the English words was determined acoustically by analyzing the F0 contour, intensity contour, and syllable durations of their utterances. Production of the correct prosodic characteristics was affected by the number of syllables and place of the lexical accent, as well as by the individual subject's level of familiarity and understanding of the word. Furthermore, the results showed that subjects' accuracy was influenced by the prosodic modification of the words. (This research was carried out at the University of Tokyo in collaboration with Keikichi Hirose.)

3. Acoustic/Articulatory Studies of Speech-Sound Production

3.1 Interaction between aerodynamic forces and yielding vocal-tract walls for obstruent consonants

To produce obstruent consonants, an oral constriction must be formed and/or released in such a way that the proper sound source can be generated. The size of the constriction and the release trajectories of the constriction from a full closure for stop consonants are taken as the parameters important for the proper source generation, which can determine the intensity of the fricative sound as well as the duration and intensity of the frication noise of stop consonants. Because of the small size of the constriction and the fast release movement, these parameters have been estimated indirectly based on several observations.

Physical models have been developed to study the physics of the formation and release of the oral constriction. The effect of aerodynamic force and the mechanical state of the soft tissue articulatory structure used to form the constriction are major factors considered in the present

physical modeling. For fricative consonants, a two-step model of the constriction has illustrated that the air stream can provide a quantal effect on the final positioning of the soft tissue articulator to arrive at the size of the oral constriction capable of maintaining the proper intensity of the fricative sound source.

For stop consonants, a numerical finite element method is adopted to derive the release trajectory. When forming an air-tight closure, a proper amount of contact pressure between the contactor articulator is necessary to act against the buildup of intraoral pressure. The model has been used to obtain initial simulations of the contact pressure, the air flow, and the release trajectory, but further experimental data on the mechanism of the release of the closure are needed to obtain simulations that match these data more closely. The same finite element model will also be used to explore the mechanism for sustained vibration of the tongue tip in the current of air when producing trills.

3.2 Models of aspirated stops in English

The releases of unvoiced aspirated stops in English are typically modeled as having three phases: (1) transient, when the pressure behind the constriction is released and the resulting abrupt increase in volume velocity excites the entire vocal tract; (2) frication, when turbulence noise generated at the supraglottal constriction excites primarily the cavity in front of the constriction; and (3) aspiration, when turbulence noise generated near the approximating vocal folds excites the entire vocal tract. These phases are expected to overlap somewhat in time, but each is expected to be marked by the dominance of one type of excitation. Close examination of stop releases reveals that the aspiration phase (3) is more complicated than has been assumed. We are exploring the possibility that frication generated during the third phase may sometimes dominate the aspiration noise. This frication may be an extension of that generated at the original supraglottal constriction, or may be additional frication generated at a tongue-body or pharyngeal constriction formed in anticipation of the following vowel. Data for two female and two male subjects have been analyzed. Results suggest that two subjects follow the classical model, but other two subjects produce a mix of frication and aspiration during the third phase. This additional frication appears to be generated at the original supraglottal constriction; that is, the spectral characteristics are similar to those of the burst. Nevertheless, listeners do not have trouble with identification. We suggest that speakers can choose between using an extended burst or formant transitions to provide enhancing cues to place of articulation.

3.3 Obstruent effects on F0

It has been observed that in American English when a vowel follows an obstruent consonant, the fundamental frequency in the first few tens of milliseconds of the vowel is influenced by the voicing characteristics of the consonant. Perception experiments have determined that these changes in F0 provide cues to a listener concerning the consonantal voicing characteristics. We refer to these F0 effects as obstruent intrinsic F0, or obstruent IF0. In an effort to more thoroughly understand obstruent IF0, target words consisting of (s) C_1VC_2 were inserted into a carrier phrase, where C_1 is the target consonant, (either an obstruent or a nasal consonant) and C_2 is the nasal consonant /m/. Three intonation contours were used so that each target syllable was recorded with a high pitch accent (H*), a low pitch accent (L*), and no pitch accent (no PA). The initial nasal consonant provided baseline data. Three female subjects and three male subjects have been analyzed.

In the H* environment, F0 at vowel onset is significantly higher than the baseline when the obstruent is voiceless; when the obstruent is voiced, F0 tends to be nearly the same as the baseline. In the L* environment, for both voiced and voiceless obstruents, F0 tends to be slightly different from the baseline, but not significantly. When there is no pitch accent on the target syllable, the results vary by subject, although the differences between the obstruents and the

baseline are still not as great as for the unvoiced obstruents in the H* environment. We note, however, that there can be considerable variability among subjects.

We propose that the lack of obstruent IF0 effects in certain pitch-accent environments is evidence of sometimes conflicting needs to control vocal-fold stiffness. The vocal folds must be stiffened to increase pitch in H* environments, and likewise they cannot be stiffened in L* environments. Since the defining gesture to implement the feature for obstruents is the stiffness of the folds, the need to stiffen the folds for voiceless obstruents corresponds with a H* environment, while the need to not stiffen the folds corresponds with a L* environment. However, there will be a conflict when voiceless obstruents occur in a L* environment, and when voiced obstruents occur in a H* environment. It appears that in these instances, the demands of the prosodic level mask those of the segmental level, and the defining gesture for the obstruents is overridden. Fortunately, enhancing gestures are used to inhibit or facilitate vocal-fold vibration, as necessary, and thus acoustic cues for the feature are preserved despite the loss of the defining gesture.

3.4 Possible physical basis for some vowel categories

A model of acoustic coupling between the oral and subglottal cavities predicts discontinuities in vowel formant prominences near resonances of the subglottal system. One discontinuity occurs near 1300-1500 Hz, suggesting the hypothesis that this is a quantal effect dividing speakers' front and back vowels. Recordings of English vowels (in /hVd/ environments) for several male and female speakers were made, while an accelerometer attached to the neck area was used to capture the subglottal waveform. Statistics on our subglottal resonance measurements are given and compared with prior work. Qualitative agreement is shown between the resonator model and diphthong data with time-varying F2 for several speakers. Comparison of the second vowel formant and second subglottal formant tracks across all speakers, analysis of the formant spaces spanned by each speaker's vowel data, and a survey of vowel formant data for a sample of the world's languages support the possibility that a speaker's second subglottal resonance divides front and back vowels. These results have possible implications for theories of vowel inventory structure.

3.5 Formant shift in nasalization of vowels

The formant shifts of vowels in the context of a nasal consonant have been measured, and experiments have been carried out to determine whether human perception is able to compensate for such shifts. According to acoustic theory, nasal coupling causes a modification on the spectrum, including formant frequency shift. The first goal of this study is to confirm that the formant frequencies actually shift due to nasalization. Based on several measurements of formant frequencies of various vowels in nasal contexts, we confirmed that the first formant (F1) tends to shift in a more central direction when nasalized. In English, vowels should be perceived as the same phoneme regardless of nasalization. In other words, listeners might have the capability to compensate for such formant shifts. The second goal of this study is to examine this compensation effect by a perceptual experiment. For stimuli, we synthesized a nonnasal vowel AEO that has the same formant frequencies as a nasalized vowel AEN. Another nonnasalized vowel AE1 was synthesized with F1 shifted upwards in frequency to correspond to the vowel that would be generated in the context of a stop consonant. A continuum was then synthesized from AEO to AE1. We conducted a vowel identification test in which the possible responses were /æ/ and /E/ (2AFC). Results show AE1 is more correctly identified as /æ/ than AEO, which suggests the existence of the compensation effect.

4. Acoustic Characteristics of Sounds Produced by Children

We have continued to collect data on the production of stop consonants of children as they develop in the age range 2-4 years. From acoustic analysis of utterances in this database, our aim is to make inferences about how the children develop the ability to manipulate and coordinate

the various articulators that are involved in producing the stop consonants. Acoustic events at stop consonant releases were sampled from the utterances of several children over a time interval of six months. These acoustic patterns were compared with those obtained from adult female speakers.

At the beginning of the 6-month period, when the children are younger, they show several differences relative to the adults: at consonant release they often exhibit multiple noise bursts, they show longer burst duration for voiced stop releases, the voice onset time (VOT) for voiceless stops is greater, and the variability in VOT is greater. At the end of 6 months, these and other differences from adult productions decreased, but in most cases were still outside of the adult range. During this 6-month period, it appears that the children have improved their ability to coordinate articulatory and laryngeal gestures, and to achieve a more normal posture for the tongue blade and lips when producing consonants that involve an increased pressure in the mouth.

5. 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.

A five-year renewal has been awarded by NIDCD for this research, in which the internal model is implemented as a computational neural model that is used to control a biomechanically based vocal-tract model. The combined models provide the bases of hypotheses about the planning of speech movements. To test these hypotheses, we are preparing to conduct experiments with speakers and listeners in which we measure articulatory movements, speech acoustics, perception, and brain activation. We will manipulate speech condition, phonemic context and speech sound category and we will introduce transient and sustained perturbations. We will also perform modeling and simulation experiments, in which we adapt the vocal-tract model to the morphologies of individual speakers. We will test properties of the computational neural model by using it to control the individualized vocal tract models in efforts to replicate those speakers' production data.

During this last year, a number of steps have been taken to prepare for the planned research, and studies from the previous funding period have been continued or completed.

5.1 Development of facilities and techniques to conduct the planned research.

We have made the following upgrades to our facilities: purchasing and integrating a new digital video camera and DVD recorder; purchasing and integrating a new "speech feedback switch" that enables us to deliver masking noise to subjects with voice-activated switching under computer control; ordering a new transducer array for our EMMA system; and beginning work on porting our real-time data acquisition software to the newest Microsoft .NET development environment.

5.2 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 equilibrium point 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. In order to simulate more realistic tongue movements, the genioglossus muscle is now divided into three segments: posterior, middle and anterior, instead of the previous two. We have also replaced the original λ model of muscle activation with a more straightforward muscle model, combined with an adaptive controller. For further model development we have collected data from two English speakers, including electromyography (EMG) records from the extrinsic tongue muscles and measures of acoustics and articulation. The configuration of the vocal tract model has now been adapted to one of the English speakers with the use of magnetic resonance images of the subject's vocal tract. To further test different control strategies, we plan to use the new vocal-tract model to rerun simulations of vowel-to-vowel movements and compare model output with data from the speaker.

5.3 Effects of perturbations with a bite block and masking noise, saturation effects and perceptual abilities on vowel contrast and variability.

We recorded the acoustic signal and articulatory movements for productions of the vowels /i, I, E, æ, ɑ/ (in hVd context) from 11 subjects under four possible combinations of perturbation: with a bite block, with masking noise, with both perturbations, and in a normal speaking condition (no perturbation). There were reductions in vowel contrasts due to perturbation with a bite block and perturbation with masking noise. Among the high vowels, dispersion around vowel targets in F1, F2 space increased with a bite block. Presumably, with a more open mandible, subjects were less able to brace the lateral edges of a stiffened tongue blade against the walls of the hard palate. Across subjects, masking noise alone caused a reduction in a measure of overall vowel contrast, AVS (average spacing among all vowel pairs). Subjects with high perceptual acuity (measured in another study) showed a further reduction in AVS when a bite block was added to the noise condition. AVS increased toward normal when the noise was removed. There was an unexpected negative correlation between subject perceptual acuity and AVS reduction from the bite block plus noise. In other words, the vowel contrasts of speakers with more acute vowel perception seemed more vulnerable to perturbation with a bite block. Also unexpectedly, there was no significant relation between the effects of noise and subject acuity. Some high-acuity subjects were good compensators; others were not. Further acoustical analyses are planned to evaluate the perturbation effects.

5.4 Sensorimotor adaptation in the production of vowels.

We are using a DSP board and its control software to modify the first formant frequency (F1) of vowels that are fed back to the subject (with an 18 ms delay) as the subject pronounces simple utterances. 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 during *training* utterances sound reasonably natural. Insert headphones are used to effectively prevent the subject from hearing air-conducted formant structure. During *test* utterances, the subject hears masking noise instead of his or her own speech through the apparatus. The noise level is high enough to mask the formant structure of the vowels without being annoying or damaging. Twenty subjects have been run, 10 male and 10 female. The subjects show varying amounts of compensatory adaptation (changing F1 in the opposite direction to that of the perturbation), both in training and test utterances, with larger amounts during training utterances. Synthetic stimuli have been constructed forming two vowel continua, which include some of the production utterances, that will be used to measure the subjects' perceptual acuity. We plan to examine the relation across subjects between the amount of compensation and degree of perceptual acuity, to test the hypothesis that subjects with better acuity will show greater amounts of sensorimotor adaptation.

5.5 Responses to unexpected perturbations of mandible movements: Formant trajectory measurements

In order to further investigate the role of auditory feedback in speech motor control, we have run pilot perturbation experiments on three subjects, in which we used a computer-controlled robot to interrupt mandible closing movements during the vowel /E/ in "see red". Differences were observed between formant trajectories in perturbed vs. control utterances, and compensatory formant trajectory adjustments were observed within about 100 msec of the onset of the perturbation. In support of this experiment, we are developing improved dental appliances for coupling the robot to the subject's mandible and for stabilizing the subject's head. These new appliances are designed to interfere less with the subject's speech and facilitate data acquisition procedures.

6. 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, data collection and analysis has continued for 11 separate studies that cover these topics, and the following interim results have been obtained.

6.1 The interacting effects of clarity demands and auditory feedback on acoustic-phonetic contrasts in postlingually deafened cochlear implant users

Decreasing the signal-to-noise ratio (SNR) in the auditory feedback loop has been shown to increase parameters in speech that are important for intelligibility, such as syllable-to-syllable modulation of F0 and sound pressure level (SPL). This study employed a task in which masking noise was gradually added to the feedback loop, from the level of detection of the noise to a level intended to completely mask speech feedback. Using this technique of increasing the demand for clarity allowed us to compare the responses of adults who receive cochlear implants (CI) (at one-month and one-year post-processor activation) to normal-hearing (NH) adults. Noise, mixed with normal levels of speech feedback, was delivered over headphones for NH participants and through the headpiece of a research sound processor for the CI users. The SNR was gradually decreased over seven steps as the speakers produced ten repetitions of two vowel contrasts (E/℘ and i/u). Both speech output level (dB SPL) and vowel contrast distance (in Hz in the F1, F2 plane) were measured at all noise levels. Both groups gradually increased SPL with decreasing SNR. Vowel contrast distance decreased with decreasing SNR in the NH group, whereas this change was not noted in CI users after one-month of experience with a CI. After a year of experience with a CI, however, these speakers showed decreased contrast distances associated with decreased SNR, much like those observed in the NH speakers. The results lead to the inference that greater experience hearing with a CI results in modifications in speech sound contrasts more like those of NH speakers in response to decreasing SNR.

6.2 Changes in the categorical perception of speech sounds following experience with a cochlear implant

Categorical perception experiments with synthetic speech stimuli were conducted with four postlingually-deafened adults one-month and one-year following fitting with a cochlear implant (CI), as well as with four normal-hearing (NH) adults. The synthesized stimuli formed continua corresponding to two phonetic contrasts, from an extreme /s/ to an extreme /Σ/ (in terms of

spectral median and symmetry), and from an extreme /i/ to an extreme /u/ (in terms of the first three formant frequencies). Data for each continuum were gathered in two sessions. In the first session, subjects identified each of the stimuli as one of the two possible phonemes. In the second session, these stimuli were rated for goodness, to assess the internal structure of the phoneme categories. All of the NH listeners evidenced well-defined phoneme categories with systematic variation in goodness ratings within each category. After one month of experience with prosthetic hearing, all of the CI users had poorly defined categories for both continua and associated disorganized quality ratings. Following one year of experience, three of the CI users evidenced defined categories on at least one of the two continua, as well as category boundaries similar to those observed in the NH listeners. These results indicate that greater experience hearing with a CI facilitates performance on speech perception tasks more like that of NH listeners.

6.3 Brain imaging, hearing loss and audiovisual speech perception

We have collected fMRI data from 10 normal hearing subjects and 8 prelingually deafened subjects while performing a speech perception task in which subjects view a speaking face without audio input, view a speaking face with audio input, listen to audio input without visual input, or view a blank screen. Patterns of activation found for normal hearing subjects viewing speech without audio (compared to viewing a blank screen) are consistent with previous studies, showing bilateral activation of several occipital, temporal and parietal lobe areas (including visual cortex, angular gyrus, fusiform gyrus, and auditory cortex) as well as premotor areas in the frontal cortex. The pattern of activity for deaf subjects is similar, though with less premotor activation. Overall, there is distinctly more activity in the right hemisphere for visual-only conditions (for both vowels and CVCVs) in deaf subjects than in normal hearing subjects. Interestingly, deaf subject activation in the visual-only case mimics normal hearing subject activation in the audio-visual case. We are currently performing effective connectivity analyses, including dynamic causal modeling and structural equation modeling, to investigate the connectivity between brain regions in the different conditions. Preliminary results indicate stronger connectivity between visual cortex and the angular and fusiform gyri during visual-only conditions as compared to audio-visual conditions, as well as stronger connectivity between the angular and fusiform gyri and the higher-order auditory cortex. Weaker connectivity is found between primary and higher-order auditory cortex in the visual-only condition. We are currently performing further analyses to investigate any changes in connectivity with premotor cortex during the visual-only condition vs. the audio-visual condition (to investigate the possibility that motoric information is more important for perceiving visual-only speech) as well as differences in connectivity between deaf and normal hearing subjects in the different conditions.

Journal Articles, Published

H. Cheyne, H. Hanson, R. Genereux, K. Stevens and R. Hillman, “Development and Testing of a Portable Vocal Accumulator”. J. Speech, Language and Hearing Res. 46(6): 1457-67 (2003).

A.O. Okobi and K. Hirose, “Acoustic Analysis of English Lexical-Prosody Reproduction by Japanese Speakers”. IEICE Technical Report 103: 37-42 (2003).

J. Slifka, “Respiratory Constraints on Speech Production: Starting an Utterance,” J. Acoustical Society of America 114(6): 3343-53 (2003).

Journal Articles, Accepted for Publication

J.S. Perkell, M.L. Matthies, M. Tiede, H. Lane, M. Zandipour, N. Marrone, and E. Stockmann, “The Distinctness of Speakers’ /s-Σ/ Contrast is Related to their Auditory Discrimination and Use of an Articulatory Saturation Effect,” *Journal of Speech, Language and Hearing Research*, forthcoming.

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

Journal Articles, Submitted for Publication

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

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

J.S. Perkell, F.H. Guenther, H. Lane, M.L. Matthies, E. Stockmann, M. Tiede, & M. Zandipour, "The Distinctness of Speakers' Productions of Vowel Contrasts is Related to their Discrimination of the Contrasts," submitted to *J. Acoustical Society of America*.

J. Slifka, "Some Physiological Correlates to the End of 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," to appear in ed. L. Goldstein (ed.), *Proc. Laboratory Phonology 8*, 2004.

K.N. Stevens, "Features in Speech Perception and Lexical Access", to appear in eds., D. Pisoni and R. Remez, *The Handbook of Speech Perception*, (Blackwell Publishers, 2004).

K.N. Stevens, and H.M. Hanson, "Voice Acoustics," in ed. R. Kent, *MIT Encyclopedia of Communication Disorders (MITECD*), (Cambridge MA: MIT Press, 63-67, 2003).

K. Stevens, Z. Li, C-Y. Lee and S.J. Keyser, "A Note on Mandarin Fricatives and Enhancement". Festschrift for Prof. Z. Wu, forthcoming.

Meeting Papers, Presented

T. Arai, "Formant Shift in Nasalization of Vowels," paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

H.A. Cheyne, R.E. Beaudoin, T.E. von Wiegand, & K.N. Stevens, "One-hand Control of a Speech Synthesizer," paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

X. Chi, & M. Sonderegger, “Subglottal Coupling and Vowel Space,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

L.C. Dilley, & M. Brown, “Distinct Relative F0 Levels Elicit Categorical Effects in F0 Maximum and Minimum Alignment,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

L.C. Dilley, & M. Brown, “Are Categorical Production Effects in F0 Extremum Alignment Related to the Perceived Relative Pitch of Syllables?” Paper presented at the Conference From Sound to Sense: 50+ Years of Discoveries in Speech Communication, MIT, Cambridge MA, June 11-13, 2004.

A.K. Imbrie, A.K., “Acoustical Study of the Development of Stop Consonants in Children,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

H. Lane, & J. Perkell, “VOT and Hearing Impairment,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

X. Mou, “Obstruent-Sonorant Consonant Sequences --- Analysis by Synthesis,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

K.G. Munhall, M. Tiede, J.S. Perkell, A. Doucette, and E. Vatikiotis-Bateson, & J.S. Perkell, “Sensorimotor Control of Speech Production: Models and Data,” keynote lecture at the Conference on Motor Speech, Albuquerque, NM, March 18-21, 2004.

J. Perkell, F. Guenther, H. Lane, M. Matthies, E. Stockmann, M. Tiede, & M. Zandipour, “Cross-subject relations between measures of vowel production and perception,” paper presented at the XVth International Congress of Phonetics Sciences, Barcelona, August 3-9, 2003.

J. Perkell, M. Matthies, F. Guenther, H. Lane, E. Stockmann, M. Tiede, & M. Zandipour, “Relationship between Perceptual Ability and the Effects of Perturbations on Produced Vowel Contrasts, “ paper presented at the Conference on Motor Speech, Albuquerque, NM, March 18-21, 2004.

M. Renwick, S. Shattuck-Hufnagel, & Y. Yasinnik, “The Timing of Speech-Accompanying Gestures with respect to Prosody,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

S. Shattuck-Hufnagel, N. Veilleux, A. Brugos, & R. Speer, “F0 Peaks Aligned with Nonprominent Syllables in American English,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

J. Slifka, “Automatic Detection of the Features [high] and [low] in a Landmark-Based Model of Speech Perception,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

A. Suchato, “Classification of Stop Consonant Place of Articulation,” paper presented at the 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

M. Tiede, J. Perkell, M. Zandipour, M. Matthies, & E. Stockmann, E., "A /t/ or Not a /t/: Apical Tongue Gestures in the 'Perfect Memory' Sequence," paper presented at the 6th International Seminar on Speech Production, Sydney, Australia, December 7 -10, 2003.

J.C. Vick, J.S. Perkell, H. Hanson, H. Lane, M. Matthies, N. Marrone, & F. Guenther, "Changes in the Categorical Perception of Speech Sounds Following Experience with a Cochlear Implant," Conference on Implantable Auditory Prostheses, Pacific Grove, California, 2003.

J.C. Vick, J.S. Perkell, E. Stockmann, M. Zandipour, H. Lane, & M. Tiede, "The Effect of Masking Noise on Acoustic-Phonetic Contrasts in Post-Lingually Deafened Cochlear Implant Users," 146th Meeting of the Acoustical Society of America, Austin, Texas, November 10-14, 2003.

V. Villacorta, & J. Perkell, "Sensorimotor Adaptation to Acoustic Perturbations in Vowel Formants," 147th Meeting of the Acoustical Society of America, New York, NY, May 24-28, 2004.

Meeting Papers, Published

T. Arai, "History of Chiba and Kajiyama and Their Influence in Modern Speech Science," *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communication,* MIT, Cambridge Massachusetts, pp. C115-C120, June 11-13, 2004.

F. Guenther, and J. Perkell, A Neural Model of Speech Production," *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge Massachusetts, pp. B98-B106, June 11-13, 2004.

H.M. Hanson, and K.N. Stevens, "Models of Aspirated Stops in English," *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 783-786, Barcelona, August 3-9, 2003.

A.O. Okobi, and K. Hirose, "Japanese Repetition of Normal and Prosodically-Modified English Words," *Proceedings of International Conference on Speech Prosody*, Nara, Japan, pp. 37-40, March 23-26, 2004.

J. Perkell, F. Guenther, H. Lane, M. Matthies, E. Stockmann, M. Tiede, & M. Zandipour, M., "Cross-Subject Relations Between Measures of Vowel Production and Perception," *Proceedings of the 15th International Congress of Phonetics Sciences*, pp. 439-442, Barcelona, August 3-9, 2003.

J.S. Perkell, M.L. Matthies, F.H. Guenther, M. Tiede, M. Zandipour, E. Stockmann, and N. Marrone, "Sensory Goals for Speech Movements: Cross-Subject Relations among Production, Perception and the Use of an Articulatory Saturation Effect," *Proceedings of the 6th International Seminar on Speech Productio*n, pp. 219-224, Sydney, December 7-10, 2003.

S. Shattuck-Hufnagel, L. Dilley, N. Veilleux, A. Brugos, & R. Speer, "F0 Peaks and Valleys Aligned with Non-Prominent Syllables can Influence Perceived Prominence in Adjacent Syllables." *Proceedings of International Conference on Speech Prosody*, Nara, Japan, pp. 705-708, March 23-26, 2004.

J. Slifka, K.N. Stevens, S.Y. Manuel, & S. Shattuck-Hufnagel, "A Landmark-Based Model of Speech Perception: History and Recent Developments." *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communicatio*n, MIT, Cambridge Massachusetts, pp. C85-C90, June 11-13, 2004.

K.N. Stevens, "Acoustic and Perceptual Evidence for Universal Phonological Features," *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, pp. 33-38, August 3-9, 2003.

K.N. Stevens, and H.M. Hanson, "Production of Consonants with a Quasi-Articulatory Synthesizer, " *Proceedings of the 15th International Congress of Phonetic Sciences*, pp. 199-202, Barcelona, August 3-9, 2003.

K.N. Stevens, “Invariance and Variability in Speech: Interpreting Acoustic Evidence,” *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge Massachusetts, pp. B77-B85, June 11-13, 2004

A. Suchato, “Classification of Stop Consonant Place of Articulation: Combining Acoustic Attributes,” *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge Massachusetts, pp. C197-C202, June 11-13, 2004.

M. Tiede, J. Perkell, M. Zandipour, M. Matthies, & E. Stockmann, “A New Approach to Pressure-Sensitive Palatography using a Capacitive Sensing Device,” *Proceedings of the XVth International Congress of Phonetics Sciences*, Barcelona, pp. 3149-3152, August 3-9, 2003.

Y. Yasinnik, M. Renwick, & S. Shattuck-Hufnagel, “The Timing of Speech-Accompanying Gestures with respect to Prosody.” *Proceedings from From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge Massachusetts, pp. C97-C102, June 11-13, 2004.

M. Zandipour, F. Guenther, J. Perkell, P. Perrier, Y. Payan, & P. Badin, “Vowel-Vowel Planning in Acoustic and Muscle Space, *Proceedings From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge Massachusetts, pp. C103-C108, June 11-13, 2004.

Theses

A. Suchato, *Classification of Stop Consonant Place of Articulation*, PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2004.

N. Hajro, *Automated Nasal Feature Detection for the Lexical Access from Features Project*, M.Eng. thesis, Department of Electrical Engineering and Computer Science, MIT, 2004.

C-Y. Park, *Recognition of English Vowels using Top-down Method*, S.M. thesis, Department of Electrical Engineering and Computer Science, MIT, 2004.

S.Y. Zhao, *Gestural Overlap of Stop-Consonant Sequences*, M.Eng. thesis, Department of Electrical Engineering and Computer Science, MIT, 2003.