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, Dr. Mark Tiede, Dr. Reiner Wilhelms-Tricarico, Dr. Lisa Lavoie, Dr. Chao-Yang Lee, Jennell Vick, Majid Zandipour, 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, Fonix Corporation, Lexington, Massachusetts.
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, Department of Biomedical Engineering, Boston University, Boston, Massachusetts.
Dr. Robert E. Hillman, Mass Eye and Ear Infirmary, Boston, Massachusetts.
Aaron Im, Department of Neuropsychology, 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. Rupal Patel, Department of Bio-behavioral Studies, Teachers College Columbia University, New York.
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
Ying Cao, Lan Chen, Xuemin Chi, Tony Okobi, Heather Gunter, Neira Hajro, Annika Imbrie, Elizabeth Johnson, Chi-Yu Liang, Steven Lulich, Nicole Marrone, Xiaomin Mou, Laura Redi, Ariel Salomon, Jason Smith, Atiwong Suchato, Virgilio Villacorta, Julie Yoo, Sherry Zhao,

Undergraduate Students
Priya Banerjee, Nathan Boy, Bashira Chowdhury, Rena Coen, Kristen Cook, Christina Curry, Megan Foster, Saba Gul, Emily Hanna, Insoo Kim, Mun Yuk Ko, JungEun Lee, Daniel Leeds, Jonathan McEuen, Michael Mejia, Conor Murray, Alvin Carter Powers, Margaret Renwick, Rodrigo Sanchez, Donny Shen, Morgan Sonderegger, Enrique Urena, Kirsten Ware, Sumudu Watugala, Yelena Yasinnik, Man Yin Yee,

Technical and Support Staff
Arlene E. Wint

1. Constraints and Strategies in Speech Production

Introduction

The objective of this research has been to develop and test a theoretical framework in which words in the lexicon are represented as sequences of segments and the segments are represented, in part, as auditory goals. Control mechanisms for the production of speech sound sequences are based on the auditory goals, which, for the present purposes, we equated to acoustic goals. The motor programming to produce sequences of goals utilizes an internal model of relations between articulations and their acoustic consequences. The acoustic goals are determined partly by non-linear quantal relations, called saturation effects, between motor commands and articulatory movements (e.g. with contact between two articulators) and between movements and sound (cf. Stevens, 1972). The relations between motor commands and articulatory movements are influenced by biomechanical constraints, which include characteristics of individual speakers' anatomy 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 acoustic goals that may be due to relations between production and perception mechanisms in individual speakers.

To test hypotheses based on this overview, we have performed a series of psychophysical experiments on as many as 20 subjects (10 females, 10 males), in which we made measures of their speech production (articulator movement and contact and the acoustic signal) and perception. One of these studies included perturbations with bite blocks and masking noise; another included perturbation of auditory feedback by modifying formant frequencies. We have also made advances in vocal-tract and control modeling and have begun to test additional hypotheses – also based on the theoretical overview – by comparing model performance with several different kinds of subject data.

1.1 Cross-subject relations between measures of vowel production and perception.

This study addressed the hypothesis that the more accurately a speaker discriminates a vowel contrast, the more distinctly the speaker produces that contrast. Measures of speech production and perception were collected from 19 young adult speakers of American English. In the production experiment, speakers repeated the words "cod," "cud," "who'd" and "hood" in a carrier phrase at normal and fast rates. Articulatory movements and the associated acoustic signal were recorded, yielding measures of contrast distance between /u/ and /ʊ/ and between /ɑ/ and /ʌ/. In the perception experiment, sets of seven stimuli ranging from "cod" to "cud" and "who'd" to "hood" were synthesized, based on natural productions by one male and one female speaker. The continua were then presented to each of the 19 speakers in labeling and discrimination tasks.

Consistent with the hypothesis, the articulatory measures of produced vowel contrast were correlated across subjects with measures of vowel discrimination.

1.2 The influence of saturation effects and perceptual abilities on production of the /s-ʃ/ contrast.

This study assessed mechanisms underlying production of the sibilant contrast /s-ʃ/. Twenty subjects repeated the words "said," "shed," "sod" and "shod" in carrier phrases in normal, clear, and fast speaking conditions. An hypothesis to be tested was that speakers with more accurate perception of the /s-ʃ/ distinction will produce the two sounds with greater levels of acoustic (and articulatory) contrast than speakers with less accurate perception. To test this hypothesis, acoustic contrast was measured as the difference between spectral means for the two sounds. For measures of perception, the subjects labeled and discriminated between members of a set of seven synthetic stimuli ranging between "said" and "shed". Subjects who demonstrated good perceptual performance had the most distinct sibilant productions. There were five subjects who had difficulty on the discrimination task, and also produced a smaller contrast between the sibilants.

1.3 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, ɪ, ɛ, æ, ɑ/ (in hVd context) from six of our subjects under four possible combinations of perturbation: with a bite block, with masking noise, with both and with no perturbation. A measure of overall vowel contrast, AVS (average spacing among all vowel pairs), in the different conditions showed inter-subject variation in the effect of the perturbations. Generally, overall vowel contrast decreased with increasing amounts of perturbation (none, noise, bite block, both). Values of AVS reduction were strongly related to subjects' average peak vowel discrimination scores from study 1.1, above ($r = .946$; $p < .01$). Thus, articulatory and acoustic perturbation caused the greatest amounts of vowel contrast reduction among the speakers with the most sensitive vowel discrimination ability. Additional analyses of changes in articulatory and acoustic measures are in progress.

1.4 Sensorimotor adaptation: Development of an acoustic feedback modification apparatus and perturbation of the production of /r/.

We have developed an apparatus, including a DSP board and controlling software, to modify the formants of vowels and glides that are fed back to the subject (with a 32 ms delay) as he pronounces simple utterances. The signal processing uses LPC analysis and resynthesis to detect and shift formant frequencies according to a specified algorithm; the resynthesis uses the LPC residual as the source, so the voiced sounds fed back to the subject sound reasonably natural. Insert headphones are used to effectively prevent the subject from hearing his own air-conducted formant structure. During test utterances, the subject hears masking noise instead of his own production through the apparatus. The noise level is high enough to mask the formant structure of vowels and glides without being annoying to the subject. We have used the apparatus to raise the F3 of /r/ in utterances such as /ɑrɑ/ and have observed compensatory F3 lowering.

1.5 Articulatory gestures without acoustic consequences.

This study departs from earlier findings and interpretations of tongue tip gestures that had no audible consequences in the /ktm/ sequence in "perfect memory" (Kiritani et al. 1975; Browman and Goldstein, 1980). To add to the relatively small original data set and investigate underlying mechanisms, we recorded tongue movements and the acoustic signal from 22 speakers of American English as each produced the sequence in "She had a *perfect memory* for details" under three rate conditions: normal, fast and clear. Release bursts were evident for /k/ an

average of 70-91% of the time and for /t/, an average of 11-84% of the time, with the lowest values in the fast condition. These results agree with the original observations: in running speech /t/ bursts are not typically present and in fast production even the /k/ burst can be suppressed. Given this pattern, we examined the movement data to determine whether articulatory gestures associated with /t/ production occur even in the absence of any discernible acoustic effect. All productions, regardless of rate, included the tongue body movement associated with the /k/ articulation. In order to identify tongue tip gestures for /t/ that were independent of tongue body movements for /k/, we examined the time course of the Euclidean distance between tongue tip and tongue blade transducers. We observed an increase, i.e., tongue distension, associated with the /t/ gesture in every case, even those for which gestural overlap masks both /k/ and /t/ acoustic releases. These results indicate that there are articulatory consequences of these stop consonants. We are currently examining the acoustics of the preceding vowel for further possible acoustic consequences of the observed movements.

1.6 Geometric representation of a human tongue for biomechanical modeling.

A quantitative three-dimensional geometrical model of the adult human tongue has been developed. The model represents most muscles of the tongue as a data structure that can be used for finite element simulations. The model includes the tongue body from the apex to the root. It extends downwards to the floor of the mouth, and also represents the tissue inside the mandible, including the geniohyoid muscle and the mylohyoid muscle. All muscles in the model are represented as directional fields at a large number of points within the finite element data structure. The genioglossus, styloglossus, hyoglossus, inferior and superior longitudinal muscles as well as transverse and vertical intrinsic muscles are also represented. The model's morphology was derived from images of the female specimen from the Visible Human data set using custom-built software to make very detailed examinations of sections at arbitrary angles through the reconstructed image volume.

1.7 Control of tongue movements in auditory and articulatory spaces.

This project investigated the planning and control of vowel-to-vowel sequences using a computer model of the vocal tract. The vocal tract model consists of an improved version a 2D biomechanical/physiological tongue model (Payan and Perrier, 1997) that is based on data from a French speaker (FS), with the addition of jaw rotation and translation and lip opening and protrusion movements. To identify values of input parameters for each tongue muscle, we examined x-ray images of FS for the steady state vowels /i/, /ɑ/, /u/, /e/, and /o/. The acoustic transfer functions of the vocal tract were optimized such that the values of formants F1-F4 were within ±10% of the corresponding values from FS's acoustic data. Planning a sequence of phonemes involves two internal models: a) a forward model (FM), which transforms vocal tract configurations into corresponding acoustic parameters and b) an inverse model (IM) responsible for generating the directional mapping from the auditory/acoustic space to the motor command space. A comparison of F1, F2, and F3 from FS to simulation results from the model for /i/-/ɑ/, /i/-/ɛ/, and /i/-/u/ shows that planning a V-V sequence in either acoustic space or motor space produces formant trajectories similar to those of FS's data, and at the same time has a smooth progression of muscle lengths between the two vowels.

1.8 A pneumatically operated device to perturb mandibular closing movements.

We have developed and tested a pneumatically operated device for perturbing mandibular closing movements. The device has been designed for use in experiments involving imaging of vocal-tract configurations and movements. It consists of a small, tubular-shaped inelastic balloon that is held in place between the molar teeth on one side. The balloon is connected via semi-rigid plastic tubing to a small air cylinder driven by a powerful solenoid. Activation of the solenoid causes inflation of the balloon to a diameter of about 1 cm at a pressure of 4-5 psi within 100 ms. Under control of the computer that displays stimuli to the subject and records acoustic and movement signals, the solenoid can be activated for selected utterances at a predetermined delay from the onset of the subject's voicing. Preliminary data show that average trajectories for

the tongue and lips are virtually the same when the mandible is prevented from closing as when it is not. To achieve these lip and tongue blade trajectories, the heights of the tongue and lower lip with respect to the mandible must have increased in the perturbed trials.

2. Effects of Hearing Status on Adult Speech Production

Introduction

This work aims to continue and extend 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 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.

2.1 Audio-visual integration in normal-hearing and hearing-impaired subjects

In listeners with normal hearing, the sight of a speaker's face articulating a syllable can influence the auditory percept, most observably when the auditory and visual stimuli are different from one another. This study investigated differences in audio-visual (AV) integration ("the McGurk effect") between adults with hearing loss who wear hearing aids (HA) and their normal-hearing (NH) counterparts. The following hypothesis was tested: HA users will rely more on visual input and thus be biased more toward the visual stimulus in the mismatch condition. Both the NH and HA groups observed computer-presented audio-visual stimuli from three speakers that paired the consonants /b/ /d/ /g/ with the vowels /ɑ/ /i/ /u/ in three conditions (auditory-only, visual-only, and audio-visual). Participants labeled each stimulus according to the consonant perceived. Responses were coded into four categories: fusion (e.g., a percept of /d/ in response to a visually presented /g/ and an auditorily presented /b/), combination (a percept of both /b/ and /g/), auditory (/b/), or visual (/g/). Data analysis examined the relative strength of visual influences in the two groups. Results showed a fusion and visual bias in HA users and an auditory bias for NH participants. For example, the results illustrated in Fig. 1 indicate that the HA group was more influenced by the visual modality than the NH group in stimuli that would traditionally elicit a fusion response.

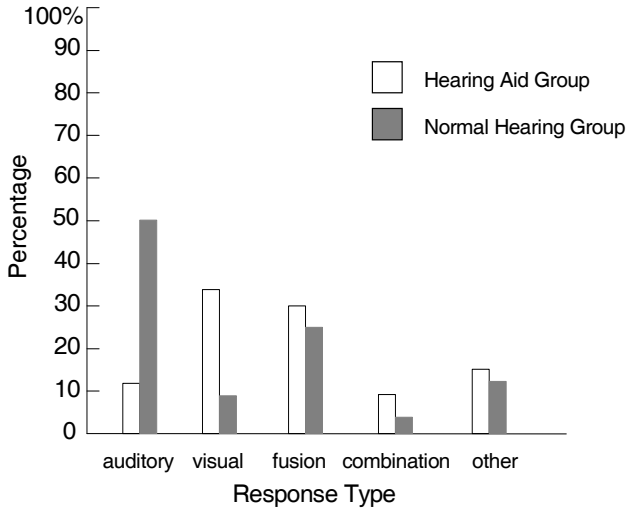

Figure 1: Percentage of response types given by the HA and NH groups to stimuli that traditionally elicit a fusion response.

2.2 Effects of hearing status and perturbation with a bite block on vowel production

This study investigated the effect of hearing status on adaptation to a bite block in vowel productions of normal hearing (NH) adults and adults who use cochlear implants (CI). CI speakers were tested prior to and following experience with the implant. Different sized bite blocks (BB) were used to create unusual degrees of mandibular opening for vowel productions in an /hVd/ context ("had", "head," "heed," "hid," and "hod"). Four conditions were elicited from each NH and CI speaker: (1) no BB with hearing (CI processor on), (2) no BB with no hearing (NH speakers with masking noise and CI speakers with processor off), (3) BB with no hearing and (4) BB with hearing. Prior to fitting with the implant, CI speakers were tested without hearing in two conditions: (1) No BB and (2) BB. Spectra of the vowel productions were analyzed for dispersion of tokens in the F1-F2 plane in the four conditions. Results indicated that while the NH speakers showed no significant changes in vowel formant dispersion with perturbation from a bite block, all of the CI speakers did, supporting the hypothesis that the vowel formants of the CI speakers would be more susceptible to perturbation from a bite block. Furthermore, both groups had greater dispersion of the vowel tokens from their target means in the absence of auditory feedback.

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

3.1 Overview

Aspects of the Lexical Access from Features project have been developed over the years by a relatively wide group of researchers that includes faculty, research staff, and students. An effort to incorporate the results from the disparate studies and theses has been initiated. The scope of this effort is to pilot a software version of the model in a unified platform. The algorithms are being implemented in a MATLAB-based platform with calls to 'c' modules where appropriate. This effort involves setting standards for data storage (specification of uniform file formants), data retrieval (development of a tool library for file read/write), and data display (consistent graphics that incorporate data across modules). Two modules, the consonant landmark detector and the vowel landmark detector, which were developed as part of doctoral thesis work, have been converted to MATLAB with incorporation of these new standards.

3.2 Toward modules for identifying place features for vowels: the feature [tense]

In English, the features that determine the identity of vowel segments include the feature [tense] or [ATR] (advanced tongue root). The feature [+tense] has been used to encompass the vowels that are produced on the extreme edges of the acoustic and articulatory spaces. The [-tense] or lax counterparts to these vowels are generally produced closer to the center of those spaces. In English (and in some other languages) this contrast applies only to nonlow vowels. A study was conducted based on the hypothesis that the extreme positioning of vowels in both articulatory and acoustic space evolves during the course of vowel production. Two measures that attempt to track these changes over time were evaluated for a pilot dataset consisting of 48 citation-form vowels in stressed position, from each of three male speakers (144 total). The vowels used were the non-low tense vowels /i/, /e/, /o/, and /u/ and the non-low lax vowel, /ɪ/, /ɛ/, /ɔ/, and /ʊ/ (where /ɔ/ is less uniformly accepted to be a lax vowel.). These two measures, the slope of the first formant (F1) movement across the vowel and the location in time of the energy peak in the F1 region as a percentage of vowel duration, were selected based on the expected movements of the articulators. Tense vowels are produced with a widening in the cross-sectional area of the pharynx. When the pharynx is widened, the tongue body may be displaced forward and possibly upward, leading to a narrowing in the constriction in the oral region. These adjustments in the tongue body lead to a lowering of the first formant frequency for tense vowels as compared to lax vowels. In terms of a Helmholtz resonator approximation, the tense vowels have a greater acoustic compliance in the pharyngeal region and a larger acoustic mass in the oral region

leading to a lower F1 than in the lax configuration. The non-low lax vowels are produced with a wider constriction in the oral region and are frequently accompanied by a lowering of the mandible. These actions, according to perturbation theory, should cause F1 to rise. By the same theory, the narrower constriction for tense vowels should lower F1. The slope of the first formant movement across the vowel distinguished between tense and lax vowels with a 90% accuracy. The location in time of the energy peak in the first formant region as a percentage of vowel duration classified the vowels with an 83% accuracy. These measures will be evaluated with a larger corpus that includes running speech. This is the first step in the development of a module to evaluate cues for the feature [tense]. The module aims to incorporate information from a family of cues that may include the two measures of first formant slope and location of vowel landmark as well as vowel duration, absolute formant values, and some measure of spectral balance.

3.3 Module for consonant place of articulation

We have shown that good classification for stop-consonant place of articulation performance can be obtained using 12 manually-extracted cues based on the acoustic theory of speech production of stop consonants. The classification score for stop-vowel (CV) tokens was close to 100% when the cues were used along with additional information on voicing, vowel frontedness and gender of the speaker of each CV token. As the next step, we have attempted to measure these cues automatically from the recorded acoustic signal, and we are using these automatically determined cues to classify stop consonants. The utterances are consonant-vowel-consonant (CVC) sequences spoken in isolation. Some modification was made to the original set of cues, resulting in 3 cues containing the information on the formant frequencies of the adjacent vowels and 7 cues containing the information on the burst release. The distributions of the values of these cues were found to be generally consistent with what we predicted from the speech production theory. The classification scores were evaluated using a Leave-One-Out cross validation method. When the information on vowel frontedness and voicing were provided, the best classification score was 86% for vowel-stop (VC) context. The back vowel context gave a higher score than the front vowel context.

The classification of CV and VC tokens was also performed using different combinations of the 10 cues. A certain combination of a subset of cues always performed well across different vowel frontedness and voicing context, while using all of the cues available did not always give the best classification score. The results have uncovered the fact that some of the cues contain redundant information and, in some contexts, some of the cues are noisy rather than useful. This encourages us, and we are proceeding to investigate the discriminating ability of each cue and cue combinations in various specific contexts to improve the classification result of CV and VC tokens extracted from both isolated CVC utterances and from read sentences.

4. Speech Prosody

4.1 Models of intonation and timing in speech

Progress has been made on two separate projects relating to the study of the representation of intonational and timing characteristics of speech. With regard to the representation of intonational patterns in speech, we have carried out two experiments which aimed to test the degree to which individuals represent specific kinds of intonational variations. The first examines whether the effects of perceptual categories are observable in a production task in which listeners imitate stimuli in which the F0 maximum (peak) or minimum (valley) has been shifted along a continuum that passes through a syllable boundary. The second experiment tests predictions of two theories of prosody regarding the categorical behavior of subjects in an imitation task for stimuli in which the F0 level of one or more syllables has been raised or lowered relative to the remainder of the utterance. The results of each experiment suggest that categorical effects are observable, and that some revision of current models of speech prosody is needed to account for all the findings. The results have implications for speech synthesis as well as for theories of speech intonation.

The second project involves a study of timing in a corpus of read speech. The aim is to investigate how timing attributes of speech relate to perceived rhythm. A large part of the data analysis for an annotated corpus of speech involving multiple speakers has been completed. Results include the finding that most inter-beat intervals (IBIs) in speech fall in the range 200-1000 ms. Moreover, when a greater number of listeners hear speech as regularly timed, the speech is more acoustically isochronous. This indicates that perceived regularity in speech has an acoustic basis, contrary to earlier claims. The corpus is further expected to yield results relating to the nature of variability in speech timing within and across speakers for the same text, as well as to the degree of agreement among listeners about whether isochrony is heard, both of which have implications for modeling speech timing for synthesis and other purposes.

4.2 Prosodic structure and phonetic variation

In line with our interest in developing principles governing the variability of implementation of distinctive features in different contexts, we have been examining the properties of syllable-final /t/. We have collected data from 7 speakers of American English producing word-final /t/ in utterances like *edit us* and *escort us* and syllable-final /t/ in inflections like *escorted*. These utterances were produced with and without contrastive stress on the following word or syllable. The contrastive stress yield utterance pairs like *Please say edit IT*, *don't say edit US*, and *Please say escortED*, *don't say escortING*. The results show that the speakers sometimes treat the constituent [verb + inflection] as a resyllabification context (i.e., aspirate the /t/ in escortED), but not the constituent [verb + pronoun]. This observation suggests that the two sequences form different levels of prosodic constituent with different consequences for the final /t/.

Another observation of interest is that word-final /t/s that are released directly into a following vowel exhibit a wide range of implementations, ranging from a dip in amplitude with continuous voicing throughout, to a stoplike closure with very low amplitude release noise, to a /d/-like stop with a release burst but little or no aspiration. This continuum raises the possibility that this flapping context in American English, i.e. word-final /t/ followed by a vowel, elicits a continuum from /t/-like through /d/-like to flapped (i.e. almost glide-like) renditions. Such a result would suggest the usefulness of rethinking the flap as a category of phonetic variation in American English. There appears to be a continuum of abruptness and narrowness of articulatory constrictions.

4.3 Repetition of F0 patterns in phrases: Influence on word segmentation

We have carried out experiments on the effect of F0 alternation on perceived organization of syllables into words. Results show that repeated high-low alternations on syllable strings such as *boy friend ship side walk way* cause listeners to organize the string into bisyllabic words (*boyfriend shipside walkway*); when the initial and final syllables are removed, creating a new low-high alternation pattern on *friend ship side walk*, listeners reorganize the string into a corresponding new set of bisyllabic words *friendship sidewalk*. In contrast, when F0 is flat, listeners report more one-syllable words, showing that F0 alternation is one of the factors that influences the perceptual organization of incoming streams of syllables, just as has been shown for incoming streams of musical notes or pure tones.

4.4 The phonological shape of function words in American English

A phonological analysis of the Function Words (FWds) vs. Content Words (CWds) of American English has been carried out, based on the Brown Corpus of one million words of text. A first step in this study involved determining which words in the lexicon were Content Words and Function Words, based on the 20 form classes used in the Brown Corpus labels. (Some categories, such as Adverbs and Exclamations, were designated Indeterminate Words and analyzed separately because there is no consensus in the syntactic literature about their CWd/FWd status). The study consisted of (a) determining the frequency of FWds in the corpus and its lexicon; results showed that the 292 lexical FWds make up approximately 50% of the corpus while the 34,000+ CWds make up the other 50%; and b) analyzing the phonological shape of the FWds; results showed that these words are more likely to be monosyllabic and to

begin with a vowel than are CWds. In fact, these asymmetries are amplified in the corpus, so that more than 90% of the monosyllabic vowel-initial words in the corpus are FWds. If similar distributional asymmetries occur in spoken language, it means that if listeners can determine that a given syllable is a monosyllabic vowel-initial word, it has a very high probability of being a FWd; such information could be useful in building an initial parse of the utterance. There is some reason to believe that listeners can do this, i.e., can determine the monosyllabic word status of a syllable. Acoustic cues to monosyllabicity might include its duration (i.e. monosyllabic lengthening makes e.g. *pea* longer than *peo-* in *people*), and cues to its vowel onset structure might include acoustic-phonetic cues to the final position of a preceding consonant (e.g. /k/ produced with less aspiration in *seek us* than in *see cuss*).

4.5 Materials and analysis tools

We have generated a spontaneous speech corpus that will be used for analysis of phonetic variation and the factors that influence it, such as function word/content word status and prosodic structure and prominence. Preparations for acoustic-phonetic analysis include initial course-grained labeling of acoustic landmarks (i.e., abrupt acoustic change) associated with consonants, prosodic labeling of intonational phrase boundaries and pitch accent prominences, and form class labeling.

To test the hypothesis that speech-accompanying gestures of the hands, shoulders, head, eyebrows, etc. are timed with respect to prosodic constituent boundaries and prominences, a corpus of videotaped lectures is being digitized and separately labeled for the prosodic structure of the speech and the gestural organization of the speaker's movements. These two streams of labels will then be correlated to test claims such as a) many gestures begin and end in conjunction with prosodic constituents, and b) many gestures reach their extremum or goal (termed the 'stroke' in gesture analysis literature) on pitch accented syllables. Results will provide insight into the coordination of the planning processes for speech-producing and non-speech-producing movements, and enrich our understanding of the role of prosody in speech production.

5. Detailed Acoustic/Articulatory Studies of Speech-Sound Production

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

An obstruent consonant is produced by forming a narrow constriction in the vocal tract with one of three articulators --- the lips, the tongue blade, and the tongue body. Pressure is built up in the airway behind this constriction, and turbulence noise is generated in the vicinity of the constriction. This noise is continuous for fricative consonants, and occurs at the release of the closure for stop consonants. Measurements and approximate calculations of the movements of these structures show that the displacements in the vicinity of the region of constriction are influenced in part by forces due to pressures upstream from the constriction and in the constriction itself. These displacements can have a significant influence on the amplitude and time course of the turbulence noise that is generated in the vicinity of the constriction, both for stop and for fricative consonants.

We have begun a more detailed computational modeling of the production of these classes of consonants produced with the tongue blade. The tongue blade is represented by an array of finite elements, and the pressures and flows through the model of the airway are based on standard aerodynamic theory. The mechanical properties of the tongue blade surface are estimated based on published studies, but it is hoped that these properties can be fine-tuned based on comparison of the model behavior with observations of the acoustics, the pressures, the airflows, and the movements observed in natural speech production. Preliminary observations of the flows and movements of the model simulations for a stop consonant are consistent with published data on these parameters for stop consonant releases.

5.2 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. For example, aspiration noise may be generated at the vocal folds during the second phase, but this phase is dominated by frication. 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. Results suggest some subjects follow the classical model, but other subjects produce a mix of frication and aspiration during the third phase. 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.

5.3 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 consonant-vowel syllables were inserted into a carrier phrase. 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). Three female subjects and two male subjects have been analyzed. The F0 contours following voiceless and voiced obstruent consonants were compared with a baseline contour obtained for syllables in which the initial segment was a nasal consonant.

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. Analysis of data from other subjects will allow us to get a clearer picture of the effect of prosodic environment on obstruent F0.

5.4 Possible physical basis for some vowel categories

The acoustic characteristics of vowels are traditionally described in terms of formant frequencies, usually F1 and F2. Different vowels are represented as points in a two-dimensional plot of F2 versus F1, and this two-dimensional space is assumed to be continuous. However, when there is a small average opening through the glottis to the trachea, acoustic coupling between the vocal tract and the trachea can introduce additional peaks in the spectrum due to resonances of the subglottal system. When a subglottal resonance is close to one of the formants, the spectrum may not show a well-defined prominence at the frequency of this formant. This observation is consistent with a theory of coupled resonators. The first two resonances of the subglottal system are in the range 500-700 Hz and 1400-1800 Hz, respectively, for adult speakers. An attempt to locate F1 or F2 in these frequency regions can lead to a spectrum for which the frequency of the

spectrum prominence in this region is somewhat unstable due to the acoustic coupling between the supraglottal and subglottal system. These frequency regions are roughly the boundaries in F1-F2 space between vowels that have been traditionally classified as [+low] and [-low] in one case and [-back] and [+back] in the other.

We are collecting acoustic data on the first two or three formant frequencies of vowels and diphthongs in words produced by several male and female talkers. Simultaneous recordings are also made from an accelerometer attached to the talker's neck just above the sternal notch, but below the larynx. These recordings are expected to show the frequencies of the subglottal resonances. The aims of these experiments are to determine whether, for a number of speakers, this instability in spectral prominences in these frequency ranges can be consistently observed, and to examine whether, for each speaker, these regions for boundaries between vowel categories defined by the features [low] and [back]. Preliminary data have shown this type of instability for some speakers, and we are now collecting a larger set of data to determine the relation of these measured ranges of instability to the vowel categories.

5.5 An acoustic study of strident fricatives: Mandarin Chinese

Mandarin distinguishes between flat and palatalized post-alveolar fricatives in addition to the alveolar-palatoalveolar distinction that has been shown to have well-defined acoustic correlates. The goal of this study is to quantitatively explicate the acoustic characteristics of these three strident fricatives and to evaluate the mapping from the acoustic properties onto phonetic features. Acoustic analyses were conducted of 216 fricative-vowel syllables produced by six speakers in three vowel and four tone contexts. Measurements were taken of the spectral properties of the frication noise, amplitude comparisons between fricative and the following vowel, and spectral properties of the vowel. The results indicate the lower frequency limit of the frication noise is associated with distinct formant regions: F4 or higher for the alveolar [s], F3 for the palatalized postalveolar [ɕ], and F2 for the flat post-alveolar [ʃ]. Several measures statistically distinguish all three fricatives, including maximal spectral peak frequency, spectral mean, variance and skewness, amplitude difference between fricative and vowel in the F3 and F4 regions, and F2 frequency at vowel onset. In terms of a universal inventory of distinctive features, it is proposed that one distinction between these three fricatives is captured by the feature [distributed], with /ɕ/ being classified as [+distributed] and /s/ and /ʃ/ as [-distributed]. The feature [+distributed] implies that the tongue blade is shaped against the palate to form a relatively long constriction, in contrast to a shorter constriction for the [-distributed] cognate. The distinctive feature [anterior] contrasts the [+anterior] /s/ and the [-anterior] /ʃ/, with this feature being unspecified for /ɕ/.

5.6 Gestural overlap of stop-consonant sequences

This study uses an analysis-by-synthesis approach to discover possible principles governing the coordination of oral and laryngeal articulators in the production of English stop-consonant sequences. Individual recordings were made of two male and two female native American-English speakers reading phrases which include voiced and voiceless stop consonants in word-initial (V#CV) and word-final (VC#V) positions, as well as in VC#CV stop-stop consonant sequences. Acoustic data including formant movements, closure durations, release bursts, and spectrum shape at low frequencies were analyzed. Results agree with earlier findings of more overlapping of oral gestures in sequences with front-to-back order of place of articulation than those with back-to-front order [I. Chitoran, L. Goldstein, and D. Byrd, Laboratory Phonology 7, 419-448 (2002)]. For example, F2 movements into the C1 closure are more affected by C2 in VC1#C2V sequences with front-to-back order of place of articulation than in those with back-to-front order. In addition, the closure durations of C1 and C2 in sequence are shortened with respect to the aggregate closure durations of singleton C1 and C2, to a greater extent for sequences with front-to-back order of place of articulation than for those with back-to-front order. However, the presence of the C1 burst is more-likely preserved in sequences with front-to-back order of place of articulation than in those with back-to-front order; this perhaps serves to preserve acoustic cues for the features of C1 when they are likely to be obscured by C2 in front-

to-back sequences. Analysis of spectrum shape at low frequencies indicate that on average, the open quotient and the acoustical loss at the glottis in V1 are only slightly affected by the voicing characteristic of C2; similarly, voicing characteristics of V2 are slightly affected by that of C1. Evidence for the overlapping of laryngeal gestures is stronger and varies on an individual basis. For example, one speaker may have much smaller acoustical loss at the glottis in vowels after voiced-voiceless sequences than in those following voiceless-voiceless sequences, whereas for another speaker, the acoustical loss is relatively the same in vowels following both types of consonant sequences.

Articulatory timing estimates will be made from the acoustic data analysis. Based on the gestural estimates, the same consonant sequences uttered by the speakers will be generated using HLsyn, a quasi-articulatory synthesizer. The synthetic utterances will be acoustically and perceptually compared to the actual utterances in order to verify and refine the articulatory timing estimates from which possible principles could be derived.

6. Acoustic Characteristics of Sounds Produced by Children in the Age Range 2-4 Years: Stop Consonants

One part of our research is attempting to track the acquisition of speech skills of children in the age range 2-4 years by eliciting selected utterances from a group of these children and performing detailed acoustic analysis of some components of these utterances. In interpreting the results of this analysis, we are particularly interested in making inferences about how the children are coordinating the various articulatory, respiratory, and laryngeal movements involved in the production of the sounds in various contexts. Our initial emphasis is on the production of stop consonants: the details of how the primary articulator for a stop consonant is moved toward closure, the time course of the release, the manipulation of the glottal opening and the state of the vocal folds, the control of respiration, and the changes in these attributes over this age range of 2-4 years.

Toward this end, we have recruited about 15 children with ages near 2; 6 years, and we have developed a protocol for eliciting and recording desired utterances in an informal setting. Initial acoustic analyses of these utterances are yielding certain baseline information on each child, such as the average spacing of the formants for vowels (related to the vocal tract length for the child) and the average spectrum shape of the vowels (related to the phonation source for the child). Average data from several repetitions of each of the six word-initial stop consonants for one child are: (1) the duration of the initial frication noise burst at consonant release is less than that observed for adults, suggesting an initial faster consonant release in the first few milliseconds; (2) the spectrum of the noise burst is in a frequency range consistent with the child's short vocal tract, and indicate a roughly correct placement of the articulator; (3) a slow rise in the first formant frequency following the release for alveolars and velars suggests a slower movement of the jaw and tongue body toward the vowel target, relative to that for adults.

Journal Articles, Published

Hanson, H.M. and Stevens, K.N. "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn". J. Acoust. Soc. Am., 112: 1158-1182 (2002).

Keating, P. and Shattuck-Hufnagel, S., A prosodic view of word form encoding for speech production. UCLA Working Papers in Phonetics 101: 112-156 (2002).

Lavoie, L., "Subphonemic and Suballophonic consonant variation: The role of phoneme inventory". Zentrum fur Allgemeine Sprachwissenschaft Berlin Papers in Linguistics 28: 39-54 (2002).

Perkell, J., Zandipour, M., Matthies, M. and Lane, H. "Clarity versus economy of effort in speech production: A preliminary study of inter-subject differences and modeling issues". J Acoust Soc Am., 112: 1627-1641 (2002).

Perkell, J., and Zandipour, M. "Clarity versus economy of effort in speech production: Kinematic performance spaces for cyclical and speech movements". J Acoust Soc Am., 112: 1642-1651 (2002)

Stevens, K.N. "Toward a model for lexical access based on acoustic landmarks and distinctive features". J. Acoust. Soc. Am., 111: 1872-1891 (2002).

Journal Articles, Accepted for Publication

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

Journal Articles, Submitted for Publication

Chen, M.Y. Nasal detection module for a knowledge-based speech recognition system, submitted to J. Phonetics.

Perrier, P., Payan, Y., Zandipour, M. and Perkell, J.S. Factors that shape articulatory movements: A modeling study of the biomechanical properties of the tongue during the production of velar stop consonants, submitted to J. Acoust. Soc. Am.

Slifka, J. (submitted). Respiratory constraints on speech production: Initiations and pauses, submitted to J. Acoust. Soc. Am.

Slifka, J. (submitted). Respiratory constraints on speech production: Ending an utterance, submitted to J. Acoust. Soc. Am.

Wilhelms-Tricarico, R. Geometric representation of a human tongue for biomechanical modeling, submitted to J. Speech, Language and Hearing Res.

Book/Chapters in Books

Beckman, M., Hirschberg, J. and Shattuck-Hufnagel, S., The original ToBI system and the evolution of the ToBI framework for developing prosody transcription systems, to appear in Jun, S. A.., (ed.), *Prosodic Typology and Transcription: A Unified Approach*. Blackwell, (2002).

Stevens, K.N. and H.M. Hanson (**accepted**) Voice acoustics. In R. Kent (ed.), *MIT Encyclopedia of Communication Disorders (MITECD*), Cambridge MA: MIT Press.

Meeting Papers, Presented

Lavoie, L., Realization of underlying stops in different speech styles, Poster presented at 8th Conference in Laboratory Phonology, Yale, CT, June 2002.

Manuel, S. and R. Krakow (2002). Why are final consonants so difficult? Insights from normal speech. Paper presented at ASHA Convention, Atlanta, Georgia, 20 November 2002. Abstract in **ASHA Leader, Vol. 7**, No. 15, Aug. 27, 2002.

Marrone, N., Stockmann, F.H. Guenther, J. Vick, J. Perkell, H. Lane, Audio-visual integration in normal-hearing and hearing-impaired subjects, Paper presented to the 144th meeting of the Acoustical Society of America, Cancun, Mexico, 2-6 December 2002.

Perkell, J., Polak, M., Vick, J., Lane, H., Balkany, T., Stockmann, E., Tiede, M. Zandipour, M., Language-specific, hearing-related changes in vowel spaces: A study of English-and Spanish-speaking cochlear implant users, Paper presented to the 144th meeting of the Acoustical Society of America, Cancun, Mexico, 2-6 December 2002.

Redi, L., Continuous temporal alignment of an F0 peak produces categorical behavior, Paper presented to the 144th meeting of the Acoustical Society of America, Cancun, Mexico, 2-6 December 2002.

Slifka, J., Respiratory system dynamics at prosodic boundaries: utterance onset and pause, Poster at the NATO Advanced Study Institute on Speech Dynamics, Italy, July 2002.

Tiede, M., Perkell, J., Zandipour, M. and Matthies, M. Gestural timing effects in the "perfect memory" sequence observed under three rates by electromagnetometry, Paper presented to the 144th meeting of the Acoustical Society of America, Cancun, Mexico, 2-6 December 2002.

Vick, J. Perkell, J., Lane, H., Matthies, M. Zandipour, M., Stockmann, E., Guenther, F. and Tiede, M., Effects of hearing status and perturbation with a bite block on vowel production, Paper presented to the 144th meeting of the Acoustical Society of America, Cancun, Mexico, 2-6 December 2002.

Vick, J. Perkell, J., Lane, H., Matthies, M. Zandipour, M., Stockmann, E., Guenther, F. Tiede, M. and Hanson, H., Biomechanical saturation and hearing status effects in determining goals for vowels: preliminary results from a bite block study, Paper presented at the Eleventh Biennial Conference on Motor Speech: Motor Speech Disorders and Speech Motor Control. Williamsburg, VA, March 14-17, 2002.

Meeting Papers, Published

Guenther, F.H., and Perkell, J.S. (in press). A neural model of speech production and its application to studies of the role of auditory feedback in speech. To appear in: B. Maassen, R. Kent, H. Peters, P. Van Lieshout, and W. Hulstijn (eds.), Speech Motor Control in Normal and Disordered Speech. Oxford: Oxford University Press.

Shattuck-Hufnagel, S., Prosodic structure and surface phonetic variation: Implications for models of speech production planning. To appear in L. Goldstein (ed.), *Proceedings of Laboratory Phonology 8*.

Stevens, K.N. Toward formant synthesis with articulatory controls. To appear in Proc. TTS IEEE Workshop on Speech Synthesis, Santa Monica, CA.

Theses

Cao, Ying A. *Analysis of Acoustic Cues for Identifying the Consonant /ð/ in Continuous Speech*. M.Eng. thesis, Department of Electrical Engineering and Computer Science, MIT, 2002.

Cheyne II, H. A. *Estimating Voicing Source Characteristics by Measuring and Modeling the Acceleration of the Skin on the Neck*. PhD thesis, Department of Massachusetts Institute of Technology, 2002.

Chi, Xuemin, *Analysis and Synthesis of Acoustic Cues in Mandarin Tones*. M.Eng. thesis, Department of Electrical Engineering and Computer Science, MIT, 2002.

Karlsson-Imbrie, A. K., *Production of Liquids by Speakers with Dysarthria*. M.S. thesis, Department of Electrical Engineering and Computer Science, MIT, 2002.

Liang, Chi-yu, *Acoustic Analysis and Synthesis of Nasal Codas in Mandarin Chinese.* M.Eng. thesis, Department of Electrical Engineering and Computer Science, MIT, 2002.