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-DC02978,
R01-DC03007,
RO1-DC04331,
T32-DC00038.
and National Science Foundation (BCS - 0418205)

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. Margaret Denny, Dr. Satrajit Ghosh, Majid Zandipour, Mark Tiede, 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, Nuance Communications Inc., Burlington, Massachusetts.
Dr. David Gow, Department of Psychology, Salem State College, Salem, Massachusetts, and Department of Neuropsychology, Massachusetts General Hospital, Boston, Massachusetts.
Dr. Frank Guenther, Department of Cognitive and Neural Systems, Boston University, Boston, Massachusetts.
Dr. Andrew Howitt, Otolith, visit site at: http://www.otolith.com.
Dr. Robert E. Hillman, Department of Voice Surgery and Rehabilitation, Massachusetts General Hospital, Boston, Massachusetts.
Dr. Harlan Lane, Department of Psychology, Northeastern University, Boston, Massachusetts.
Dr. Sharon Y. Manuel, Department of Speech Language Pathology & Audiology, Northeastern University, Boston, Massachusetts.
Dr. Melanie Matthies, Department of Communication Disorders, Boston University, Boston, Massachusetts.
Dr. Richard McGowan, CReSS LLC, Lexington, Massachusetts.
Dr. Lucie Menard, Department of Linguistics and Language Education, University of Quebec, Montreal, Canada.
Dr. Rupal Patel, Department of Speech Language Pathology and Audiology, Northeastern University, Boston, Massachusetts.
Dr. Alice Turk, Department of Linguistics, University of Edinburgh, Edinburgh, United Kingdom.
Dr. Nanette Veilleux, Department of Computer Science, Simmons College, Boston, Massachusetts.
Dr. Lorin Wilde, STAR, Massachusetts.

Graduate Students
Lan Chen, Nancy Chen, Xuemin Chi, Elisabeth Hon, Youngsook Jung, Steven Lulich, Xiaomin Mou, Tony Okobi, Chi-youn Park, Yoko Saikachi, Kushan Surana, Julie Yoo, Sherry Zhao

Undergraduate Students
Akua Adu-Boahene, Amjad Afanah, Yin Mon Mon Aung, Natalie Cheung, Mingyan Fan, Corinna Hui, Benjamin Levick, Mish Madsen, Olayemi Oyebode, Anunaya Pandey, Alicia Patterson, Rebecca Pomerantz (Harvard University), Rodrigo Sanchez, Morgan Sonderegger, ChengHua Tong, Jessie Wang, Dilini Warnakulasuriyarachchi, Sophie Wong, Yelena Yasinnik

Technical and Support Staff
Arlene E. Wint

1. Constraints and Strategies in Speech Production

Introduction

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

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

During this last year, we have made progress on several major studies and have further developed our facilities.

1.1 Variation in vowel production

Acoustic and articulatory recordings were made of vowel productions by young adult speakers of American English - 10 females and 10 males - to investigate effects of speaker and speaking condition on measures of contrast and dispersion. The vowels in the words *teat, tit, tet, tat, tot* and *toot* were embedded in two-syllable "compound words" consisting of two CVC syllables, in which each of the two syllables comprised a real word, the consonants were /p/, /t/ or /k/, the two adjoining consonants (at the concatenation of the two syllables) were always the same, the first syllable was unstressed and the second, stressed. Variations of phonetic context and stress were used to induce dispersion around each vowel centroid. The compound words were embedded in a carrier phrase and were spoken in normal, clear and fast conditions. Initial analyses of F1 and F2 on 15 speakers have shown significant effects of speaker, speaking condition (and also vowel, stress and context) on vowel contrast and dispersion around means. Generally, dispersions increased and contrasts diminished going from clear to normal to fast conditions. Further analyses are underway. Results will be related to measures of the speakers' perceptual acuity for vowel contrasts.

1.2 Trajectory planning in the concatenation of larger units

We have obtained and are analyzing data bearing on the interplay of sufficient auditory contrast and economy of production effort, in the context of overlapping articulatory gestures for adjacent consonants. /kt/ clusters formed across word boundaries (*pack top*) compared with those originating in the lexicon (*pact op*) show distinct patterns of behavior: Compared to lexical sequences, those formed across word boundaries show in general greater separation in timing between points of maximum constriction at normal speech rates, and greater reduction in these separations under fast production rates. Lexical sequences tend to preserve the rate-scaled relative phasing between the two constrictions. Regardless of utterance type, distinct release bursts are not produced consistently, especially as speaking rate increases. Acoustic analysis is ongoing to determine the extent to which distinct articulatory gestures leave traces in the residual acoustic signal.

1.3 fMRI study of sensorimotor adaptation in the production of vowels

We are using functional magnetic resonance imaging (fMRI) to investigate brain regions contributing to sensorimotor adaptation – speakers' compensatory responses to perturbations of the first formant (F1) in their auditory feedback. Subjects were asked to pronounce utterances while their brain activity was imaged by an MRI scanner. The transduced speech signal was then processed with custom software running on a DSP board with a delay of 18ms and fed back to the subject via electrostatic headphones. The signal processing used LPC analysis and resynthesis to detect and shift the F1 frequency according to a custom-designed algorithm; the resynthesis uses the LPC residual as the source, so the voiced sounds fed back to the subject sound reasonably natural. We have scanned 14 subjects and are in the process of analyzing the data. Half of the subjects showed significant adaptation to the stimuli as assessed through their speech recordings. Preliminary analysis of the imaging data showed typical brain responses for speech production; however, during the perturbation phase, an unexpected reduction in brain activity was observed relative to the activity during speech without perturbation. Analyses are being performed to investigate reasons for such a reduction in activity.

1.4 Control of tongue movements in acoustic and articulatory spaces

This project is investigating the planning of vowel-to-vowel tongue movements using two different control models: planning in acoustic space using feedback mechanisms, and motor planning through feedforward control of muscle activation. The planning models are used to control an improved model of the vocal tract which has been developed based on anatomical and physiological data. The vocal tract model receives (simulated) muscle-activation motor commands and produces acoustic, EMG, force and kinematic data. The vocal-tract model was adapted to an individual speaker, and simulations were conducted with the vocal-tract model using each of the two control models. Both planning strategies were able to replicate the speaker's data equally well. These results lead to the preliminary inference that neither a motor-space planning scheme nor an acoustic-space planning scheme of speech motor control can be rejected as a model of the control of speech production. Results are in agreement with the DIVA model of speech production (Guenther, et al., 2006, *Brain and Language, 96*, 280-301), which hypothesizes that the normal speech is produced under feedforward control, but when abnormal conditions such as acoustic or somatosensory perturbations arise, auditory and somatosensory feedback mechanisms are engaged.

1.5 Development of facilities – generation of speech continua for perceptual testing.

We have developed speech continua for perceptual testing using the Klatt synthesizer. These include two vowel continua (*pap* to *pep* and *pep* to *pip*), a sibilant continuum (*said* to *shed*), a voicing onset time continuum (*doe* to *toe*) and a vowel duration continuum using the utterance *pup*. These continua will be used to determine the perceptual acuity and phoneme category width of subjects who have already participated in a corresponding speech production experiment (1.1). Software was been developed in Matlab to readily generate the continua and carry out the required perceptual testing. We are currently piloting these tests.

2. Modeling of Human Lexical Access

Progress over the past year in the Lexical Access from Features project (LAFF) has focused on three major areas: (1) the role of voice quality and irregular phonation in cue measurement and estimation of prosodic structure; (2) conversion of landmark estimation to a probabilistic framework and development of initial methods for predicting the most likely sequence of landmarks in an utterance; and (3) further progress in estimation of some distinctive features based on appropriate acoustic cues. In addition, progress on clarification of the theory and the corresponding computationally explicit model has been presented as invited talks at two international conferences.

2.1 Detection and use of irregular phonation

A region of phonation is considered 'irregular' if the speech waveform displays either an unusual difference in time or amplitude over adjacent pitch periods that exceeds the normal small-scale jitter and shimmer differences, or an unusually wide spacing of the glottal pulses compared to their spacing in the local environment, indicating an anomaly with respect to the usual, quasi-periodic behavior of the vocal folds. In the past year, we have made progress in analyzing the range of acoustic and physiological variation associated with irregular phonation, in developing automatic methods for classification of regions of phonation as irregular or regular, and in quantifying a basis for interpreting occurrences of irregular phonation as cues for speech boundaries. The focus was placed on naturally occurring irregular phonation at the ends of utterances. Contrary to previous observations, irregular phonation was commonly observed for vocal folds that are relatively widely spread and are in the process of continuing to spread. These glottal actions are consistent with the speaker's preparations to end the current exhalation and prepare for the next inhalation. The resulting pitch periods generally show a short closed phase and long open phase characterized by increased airflow and increased damping in the latter portion of the pitch period.

Four acoustic cues are used to separate regions of regular phonation from regions of irregular phonation in a speaker-independent and context-independent manner for a large number of speakers. In general, cue distributions are widely separated statistically and classify tokens with accuracy rates greater than 90%. The proposed system uses tokens from multiple speakers in various contexts —114 different speakers for training and 37 different speakers for testing. Given the high inter-speaker variation of irregular phonation, the high accuracy rates using multiple speakers support the robustness of the developed system.

Once a region of irregular phonation has been detected, the next step is to appropriately interpret what, if any, information about the spoken utterance can be determined. In Surana and Slifka (2006), the potential use of irregular phonation as a cue for the segmentation of continuous speech was analyzed for two dialect regions of the TIMIT database (read, isolated utterances). For 1331 hand-labeled irregular tokens from 114 speakers, 78% of the irregular tokens occur at word boundaries and 5% occur at syllable boundaries. Of the irregular tokens at syllable boundaries, 72% are either at the junction of a compound-word (e.g. "outcast") or at the junction of a base word and a suffix. Overall, the sentences in our database have a total of about 11000 word boundaries, and about 10% of these are marked with irregular phonation. Of the irregular tokens which do not occur at word or syllable boundaries, 70% occur adjacent to voiceless consonants mostly in utterance-final location. These observations support irregular phonation as an acoustic cue for syntactic boundaries in connected speech.

2.2 Probabilistic formulation of landmark detector for a model of speech perception

Landmarks are the points of time when important events occur in a speech signal. These events provide evidence for the presence of articulator-free distinctive features. For every segment or bundle of distinctive features in an utterance, there is always one articulator-free feature. Knowledge of this feature and the landmarks associated with it provide a starting point for identifying cues for other (articulator-bound) features for the segment. Landmarks are classified broadly into three types --- consonant, vowel and glide, and there are three types of consonant landmarks --- glottis, sonorant and burst. Our current research is focused on these three consonant landmark types.

An algorithm to detect consonant landmarks was developed several years ago, and we are designing and testing a new landmark detector using parts of this earlier algorithm as a starting point. One problem with the earlier algorithm was that it was developed as a deterministic process with fixed thresholds, and therefore it did not retain any information about the filtered out instances, and it did not provide any information to help the detection of false alarms. However, such information may be useful later in the lexical access process, especially given the range of contextual variation in the speech signal.

To overcome this shortcoming, we are reformulating the landmark detection process into a probabilistic process that leads to a graded measure of landmark strength. In this research, thresholds are lowered to include more candidates, and then a probability value is calculated for each candidate to provide information for unclear instances. The implementation uses different sets of cues to detect the three types of consonant landmarks and computes a probability estimate for each landmark based on a set of training data. After training on 28 speakers and testing on another 10 speakers (all from the TIMIT database), two of the three types of landmarks (*glottal* and *burst*) are reliably detected. Greater than 90% of the detected landmarks agree with hand-labeled landmarks and only about 4% of the landmarks are missed. However, the third landmark type, *sonorant*, has room for improvement with roughly 75% accuracy. To compensate for the large number of false alarms, we are constructing an algorithm that selects the sequence of landmarks that are likely to be true landmarks, making use of known constraints that are expected in continuous running speech.

2.3 Classification of fricative consonants

We have developed procedures for classifying the four places of articulation for the [+continuant] consonants in English: /f/, /θ/, /s/, /ʃ/, and their voiced counterparts /v/, /ð/, /z/, /ʒ/. These consonants are first classified as [+strident] and [-strident], based primarily on the high-frequency amplitude of the frication noise in relation to the amplitude of the adjacent vowel in the same high-frequency region. Within each of these classes, the place of articulation for the [+strident] consonants is identified based on the spectrum shape of the frication noise. The distinction between the nonstrident labial and dental consonants is based primarily on the onset of the second formant frequency in the adjacent vowel.

Identification of the nonstrident dental consonant /ð/ presents special problems. In spoken English, this fricative is the most frequently occurring sound in word-initial position, largely because it occurs in this position in a number of function words. In the development of a model for human or machine lexical access in running speech, it is important, then, to be able to recognize this segment.

We have examined the segmental-contextual effects on the continuancy of the voiced dental /ð/ in American English and whether certain acoustic attributes of this consonant are preserved despite possible modification of the continuancy feature of this fricative. Word-initial /ð/ cases, extracted from continuous speech of 146 speakers from the TIMIT database, were frequently (averaging 65%) stoplike when they were in utterance-initial position or when the preceding phoneme was voiceless and/or [-continuant]. This stoplike modification occurred less frequently (averaging 39%) when /ð/ was preceded by a voiced fricative and rarely (averaging 13%) when preceded by a vowel or liquid consonant. A comparison of stoplike /ð/ and /d/ cases under similar contexts showed that the burst peak location, burst spectrum shape, and F2 at vowel onset averages were all statistically different between the two groups. The acoustic data suggested that the dental place of articulation was preserved for the modified /ð/. Preliminary automatic classification experiments involving salient acoustic attributes indicated that F2 at vowel onset may be a reliable cue for the dental place of articulation in /ð/.

3. Quantal Theory, Enhancement, and Speech Production Models

3.1 Quantal/enhancement theory relating phonetics and phonology

During the past year we have been continuing to refine the quantal/enhancement theory that attempts to provide a link between abstract discrete phonological features of language and the surface phonetic realization of these features. The theory attempts to develop physical and physiological bases for the universal inventory of distinctive features that appear to account for the phonemic contrasts that are observed in the languages of the world.

Our current research on these topics includes: (1) examining the role of the acoustics of the sublaryngeal system in shaping the distinctive features for vowels, particularly the front/back distinction; (2) investigating the role of enhancement in explaining the variability that is observed in production of nasal codas and of certain fricative consonants in Mandarin Chinese; and (3) collecting evidence from other languages that can shed light on the quantal/enhancement theory.

3.2 Physical basis for the front-back distinction for vowels

As a continuation of our research on the role of acoustic coupling between the vocal tract and the subglottal system on speech production, we have been examining the details of the acoustic discontinuity that occurs in the vicinity of the second formant F2 as it passes through the second subglottal resonance in a diphthong like /ai/ in English.

We have developed a theoretical model that incorporates the subglottal impedance and measures of the impedance of the glottal opening during phonation --- an impedance that may be speaker dependent. From this model it is possible to estimate the amount of acoustic amplitude fluctuation (the “attenuation”) and the discontinuity in frequency of the F2 spectrum peak (frequency jump) for various glottal openings.

The range of values of the measured frequency jump for a number of versions of the words “hide” and “hoid”, both in citation form and in carrier sentences, produced by 7 male and 7 female speakers, was 99 Hz to 311 Hz. The average attenuation value at the discontinuity was in the range 1.2 to 4.2 dB. Although there were occasional utterances that showed no significant frequency jump, all utterances showed an amplitude discontinuity. These discontinuities are in a range predicted by the model.

Acoustic data on the second subglottal resonance and on the second formant frequency were also collected for 14 speakers each producing 10 repetitions of 10 words containing monophthongs, such as “hid” and “hawed”. The second subglottal resonance was also measured. For these words, it is expected that the measured F2 frequency should be below the second subglottal resonance for back vowels but above that resonance for front vowels. This prediction was true for all front vowels and for most back vowels. The occasional exceptions were for a few cases the vowels /ʌ/, /U/, and /u/. For these back vowels it appears that the final alveolar consonant in the context /bVt/ had the effect of raising F2.

These observations for diphthongs and for monophthongs vowels provide some support for the hypothesis that speakers adjust the production of vowels so that for front vowels F2 is higher than the second subglottal resonance and for back vowels F2 is below this subglottal resonance.

3.3 The role of lower airway resonances in defining distinctive feature contrasts

In the course of our research on the acoustic and articulatory processes underlying some of the distinctive features for vowels and consonants, we have encountered situations in which acoustic coupling between the resonances of the subglottal respiratory system and those of the vocal-tract proper play a potential role in shaping some of these features. In order to quantify these influences more precisely, we have undertaken a detailed study of the anatomy and the acoustics of the subglottal airways. This study makes use of existing anatomical data on the respiratory system, and involves a detailed computational analysis of the impedance of these airways, particularly the poles and zeros of the impedance up to frequencies of about 3 kHz.

Among the speech-related observations that emerge from this computational study are the role of the second subglottal resonance and its interaction with the vocal tract in providing a basis for the distinctive feature [back], the potential role of the third subglottal resonance in shaping the feature [high] for front vowels, and (more speculatively) the potential role of these two subglottal resonances in providing “landmarks” in formant space that shape the time course of F2 and F3 trajectories that occur at vowel-consonant boundaries.

3.4 The nature of aspiration in stop consonants in English

The releases of aspirated stops in English are typically modeled as having three consecutive phases, which overlap somewhat in time: (1) transient, (2) frication, and (3) aspiration. Close examination reveals that the noise spectrum in what is considered the aspiration phase is sometimes dominated by one spectral prominence, rather than several prominences as expected with a glottal source. In this work we explore the possibility that frication noise generated during the third phase may sometimes dominate the aspiration noise. The nature of the radiated sound during the production of both voiced and unvoiced stop consonants is examined for the three places of articulation in English and with several different following vowels. Data from five subjects have been observed. Results can be summarized as follows. (1) Spectra measured during the early portion of the voice onset time (VOT), assumed to be frication, have similar shapes for both voiced and voiceless stops. Some individual differences are observed, and are more likely to be due to differences in oral cavity geometry than strategy. (2) Individual differences are much more marked in spectra measured during the latter portion of the VOT. For some speakers, the nature of the noise source in this portion of the VOT varies greatly with place of articulation, and there is also some evidence of vowel dependency. Based on these results, we conclude that the first 10-20 ms of noise following the stop release is part of the "interior" portion of the consonant, that is, it is a defining acoustic cue corresponding to the place of articulation feature. The remainder of the VOT is "exterior" to the consonant, that is, it is due to enhancing gestures manifested during the following vowel. In this "exterior" phase, the nature of the noise can vary due to a given speaker's strategy to provide enhancing cues to place of articulation: (1) primarily aspiration noise generated at the spread vocal folds; (2) primarily frication noise generated at a supraglottal constriction; or (3) a mix of aspiration and frication.

3.5 Codas in Standard Chinese: The role of enhancement

In Standard Chinese (SC), there are only two consonants that can occur in coda position --- the consonants /n/ and ŋ/ --- while the nasal consonants /m/ and /n/ can occur in prevocalic position (as in English). In English, almost all consonants can appear in syllable coda position, including the three nasal consonants /m/, /n/, and /ŋ/. English and SC also differ in the inventory of vowel contrasts: SC has just 3 or 4 contrasting vowels, where English has a much larger inventory. We have examined the acoustic characteristics of the nasal consonants in the two languages, and we have observed this combination of a relatively sparse vowel inventory in SC as well as the special role of nasal consonants in syllable-coda position raises the possibility that these consonants may exhibit more variability in their acoustic and articulatory attributes, particularly their influence on the attributes observed in preceding vowels.

Several acoustic and perceptual studies of these nasal consonants in various vowel contexts have been carried out, including comparison with nasal consonants in similar contexts in English. The following observations have emerged from these studies.

(1) When SC listeners are asked to identify the four vowel-nasal sequences /an/, /aŋ/, /æn/ and /æŋ/ in terms of their own language, which has only the two codas /An/ and /Aŋ/ following the

low-vowel /A/, they identify /æn/ as /An/ and /aŋ/ as /Aŋ/, that is, the coda with the front vowel in English is identified as /n/ and the coda with the back vowel is identified as /ŋ/.

(2) Acoustic analysis of the sequence /An/ in SC shows that the vowel is nasalized throughout much of its length, and through this interval the second formant frequency is relatively high, indicating a fronted tongue body. Likewise the vowel in /Aŋ/ is also nasalized but shows a lower F2, indicating a backed tongue body configuration. The nasal murmur at the end of these utterances tends to be short, and sometime lacks evidence of complete oral closure. Similar observations are made with the utterances with the mid vowel /e/, i.e., /en/ and /eŋ/, again with the vowel being fronted before /n/ and more backed before /ŋ/.

(3) When the nasal coda /n/ and /ŋ/ is preceded by a high vowel /i/ or /u/, there is essentially no shift in vowel for the two different codas.

(4) The prevocalic nasal consonants /m/ and /n/ in SC exhibit acoustic properties similar to those for the same prevocalic consonants in English.

One can interpret these data for the nasal codas in terms of an enhancement theory which postulates that the perceptual saliency of these codas in SC is enhanced by introducing a fronting or backing of the preceding vowel. When the vowel is a low or mid vowel this fronting or backing can be introduced without compromising the perceptual saliency of the vowels which do not have a front-back distinction. For high vowels, for which there is a front-back contrast, however, this kind of enhancement cannot be used.

4. Studies of Speech Development and Speech Disorders

4.1 Improvement of Electrolarynx speech

In collaboration with Dr. Robert Hillman at the Voice Surgery and Rehabilitation laboratory of the Massachusetts General Hospital, we have continued our participation in a project whose aim is to improve the intelligibility and naturalness of speech produced with an Electrolarynx device (EL). This device is used as a substitute for the larynx for speakers whose larynx has been removed. As a first step in evaluating the influence of various modifications of the EL device, we have used the speech synthesizer KLsyn to hand-synthesize utterances which are very similar to EL speech. The synthesized versions of EL speech are judged to be very similar to the original EL-produced utterances.

As an initial experimental step, the F0 contour was manipulated in the synthesizer to be similar to the F0 contour produced by a normal speaker with sentence-length utterances produced in a normal way. This produced a significant improvement in the quality relative to the fixed F0 speech produced by the EL device. As a next step, the F0 contour in the synthesized EL speech was manipulated by a signal derived from the RMS amplitude of the original EL speech. Indeed, these amplitude fluctuations in EL speech were similar to those in original spoken utterances. These observations suggest the possibility of implementing a modification of the EL device that automatically calculates an F0 contour based on measurements of the time-varying signal that is generated from the EL speech. Further perceptual evaluation of this method of generating F0 contours based on amplitude fluctuations have been planned for various types of utterances that include those for which various types of F0 contours are expected, such as questions and manipulations of emphasis.

5. Effects of Hearing Status on Adult Speech Production

This work has continued and extended our program of research on postlingual deafness and the role of hearing in speech production. We have been 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 have described changes in speech communication that take place in these deaf adults when they receive cochlear implants. We have aimed 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, a number of completed studies (described in previous RLE reports) were submitted, accepted for publication or published.

6. Speech Prosody

6.1 Intonational phonology

In collaborative work with Professor Jonathan Barnes (Boston University), Prof. Nanette Veilleux (Simmons College), and Alejna Brujos (Boston University), we are investigating the domain of realization of the low tonal target just before the final rise in contours that express shock and disbelief, as in “Elizabeth?!?” realized with a high F0 on –liz-, followed by a low before the final rise during –beth. Our F0 measures for 9 speakers suggest that this low is aligned at its left edge with a metrically strong syllable, and at its right edge with a final syllable, rather than being aligned with a single syllable.

In a separate study, we are following up on our earlier finding of temporal alignment between spoken pitch accents and gestural ‘hits’ (i.e. movements of hands, head, eyebrows characterized by sudden sharp stops that can be aligned with the speech signal) by expanding our database of prosodically, segmentally, and gesturally labeled lecture samples. In this expanded database, we are also comparing the alignments of hits produced by different articulators, since preliminary results for one speaker suggested that hand hits align with the accented syllable, but head hits tend to occur more often with the syllable following the spoken accent.

6.2 Patterns of spoken errors

We have begun a collaboration with Drs. Marianne Pouplier and Louis Goldstein (Yale University/Haskins Laboratories), who have recently reported articulatory measures showing that when speakers produce certain types of repetitive tongue twisters (e.g. top cop top cop etc.), they make errors that involve the intrusion of individual articulatory gestures (e.g tkop tkop) rather than the segmental substitutions (e.g. t → k) commonly reported in error corpora that are collected by simply listening to spontaneous communicative speech. Their hypothesis that such gestural error patterns can account for all sound-level errors is in direct competition with the hypothesis that has been advanced at MIT and elsewhere, that most such errors reflect the substitution, exchange, addition or omission of whole-segment planning units. Together we are designing a set of elicitation experiments (using both acoustic and articulatory analysis) to test the alternative hypothesis that tongue twisters, with their alternating patterns, are particularly likely to elicit single-gesture errors that regularize the articulatory rhythms (i.e. one /t/ and one /k/ for each –op), while speech elicited by increasingly more natural speaking conditions, i.e. those that invoke the full complement of planning processes including lexical access, the generation of syntactic and prosodic structure and phonetic variation, will induce an increasingly higher proportion of whole-segment errors.

6.3 Spectral correlates of lexical prosody

We have initiated a new series of experiments that are designed to evaluate the acoustic cues for lexical stress in English. In particular, the experiments are examining the acoustic production and perceptual correlates of lexical stress in two-syllable CV words embedded in a carrier phrase in which the pitch accent is placed on different syllables within the carrier phrase, either in the target word or in a word preceding the target word.

In the production experiment, real-word CVCV utterances are elicited from several speakers, with pitch accents placed in different positions in the carrier phrase. The recorded utterances are subjected to acoustic analysis in order to quantify various temporal and spectral attributes of the utterances.

In the perceptual experiment, target stimuli with different timing patterns and different attributes of the glottal source were synthesized using the Klatt formant synthesizer KLsyn. These synthesized CVCV "utterances" were always of the form /da da/. These target sequences were inserted in a spoken carrier phrase containing a pitch accent two syllables preceding the target. Several hundred target utterances with different combinations of duration, spectrum tilt, and aspiration noise were synthesized and presented to a group of listeners. The listeners were instructed to state which of the syllables in the target stimulus was more "prominent", and to assign a confidence to this judgment.

One outcome of the preliminary production experiments is that there was an influence of the position of the pitch accent on the response to the target word, the influence being smaller for more leftward pitch accent placements relative to the target words. From these data it was possible to make a correction for the influence of the focal accent on the spectral characteristics of the target, yielding results that provided better estimates of the influence of lexical stress alone. These corrected data showed that the spectral tilt, given by H1*-A3*, was greater for the syllable with less stress. Other correlates of lexical stress observed in this preliminary study include the presence of noise at high frequencies, and syllable duration, with syllable duration being the most salient. Further measurements with more subjects are in progress.

6.4 Subglottal pressure contours of speech utterances

In order to model P_s variation for speech synthesis we are studying the relationship between characteristics of the P_s contour and prosodic events. Subglottal pressure (P_s) contours for speech are described as having three phases: initiation phase, constant or declining working phase, and termination phase. In past work, it was found that the initiation phase is relatively easy to identify, but the transition from the working phase to the termination phase is less clear. Confounding issues could include segmental impedances, pitch accents, and phrase and boundary tones, all of which can have local effects on P_s. We have attempted to control tones and segments at the ends of utterances in order to better identify final fall. Lung pressure is estimated from esophageal pressure (corrected for lung volume). Several preliminary findings (based on a subset of all data) have been made. First, the distribution of pitch accents, phrase tones, and boundary tones affects the clarity of the transition from the working phase to the termination phase, but details of this effect appear to vary by speaker. Second, the location of the nuclear pitch accent of an utterance does not define the start of the termination phase. Thirdly, utterance offset generally occurs after the termination phase begins, but the time between these two events is highly variable. This latter result suggests that the production factors controlling phonation offset are somewhat independent of those controlling the start of the termination phase. Therefore, phonation offset can not be used to pinpoint the onset of the termination phase. Finally, line-fitting techniques provide reasonable estimates of the onset of the termination phase, but we continue to explore alternative methods for identifying this important landmark in P_s contour.

7. Speech Corpora

During the past year we have made progress in the labeling and interpretation of acoustic landmarks in the spontaneous task-elicited speech of the Maptask corpus. The observed landmarks are compared with the landmarks predicted from the underlying lexical representation to determine and classify the prosodic, lexical, and segmental locations where landmarks are changed or lost.

In another labeling-related task, we have continued our development of an online tutorial for training labelers for the TOBI system for prosodic labeling, and we have trained several new TOBI labelers. The tutorial will be published under the auspices of the OpenCourseWare program during the summer of 2006.

These and other speech databases from our group are available as part of the MIT Libraries DSpace Initiative, as described in RLE Progress Report No.147 in the section on the Speech Communication Group.

Publications

Journal Articles, Published

S.J. Keyser and K.N. Stevens, “Enhancement and Overlap in the Speech Chain,” *Language 82, no. 1:* 33-63 (2006).

H. Lane, M. Denny, F.H. Guenther, M. Matthies, L. Ménard, J. Perkell, E. Stockmann, M. Tiede, J. Vick, and M. Zandipour, “Effects of Bite Blocks and Hearing Status on Vowel Production,” *J. Acoust. Soc. Am*. *118*: 1636-46 (2005).

H. Lane and J.S. Perkell, “Control of Voice-Onset Time in the Absence of Hearing: A Review,” *J. Speech Lang. Hear. Res*. 48: 1334-43 (2005).

K.N. Stevens, “The Acoustic/Articulatory Interface,” Invited review for *Acoust. Sci & Tech. 26(5): 410-17 (2005)*.

J. Slifka, “Some Physiological Correlates to the End of Phonation at the End of an Utterance,” *J. Voice 20(2)*: 171-86 (2006).

J. Slifka, “Some Challenges in the Detection of Landmarks in Models of Speech Recognition,” Invited paper, *IEICE Technical Report 105*(685): 9-14 (2006).

Journal Articles, Accepted for Publication

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

A.E. Turk and S. Shattuck-Hufnagel, “Phrase-Final Lengthening in American English,” *J.Phonetics*, forthcoming.

Journal Articles, Submitted for Publication

M. Denny, J.S. Perkell, H. Lane, M.L. Matthies, M. Tiede, M. Zandipour, J. Vick, and E. Burton, “Timing of SPL and Contrast Changes in Response to a Modification of Auditory Feedback,” submitted to *J. Acoust. Soc. Am*.

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

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

S. Lulich, A. Bachrach, and N. Malyska, "A Role for the Second Subglottal Resonance in Lexical Access," submitted to *J. Acoust. Soc. Am*.
L. Ménard, M. Polak, M. Denny, H. Lane, M.L. Matthies, J.S. Perkell, E. Burton, N. Marrone, M. Tiede and J. Vick, "Effects of Speaking Condition and Hearing Status on Vowel Production in Postlingually Deaf Adults with Cochlear Implants," submitted to *J. Acoust. Soc. Am*.

J.S. Perkell, M. Denny, H. Lane, M.L. Matthies, E, Stockmann, M. Tiede, J. Vick, and M. Zandipour, "Effects of Masking Noise on Vowel and Sibilant Contrasts in Normal-Hearing Speakers and Postlingually Deafened Cochlear Implant Users," submitted to *J. Acoust. Soc. Am*.

Book/Chapters in Books

J.S. Perkell, F.H. Guenther, H. Lane, N. Marrone, M.L. Matthies, E. Stockmann, M. Tiede and M. Zandipour, "Production and Perception of Phoneme Contrasts Covary across Speakers," in *Speech Production: Models, Phonetic Processes, and Techniques*, eds. J. Harrington & M. Tabain (Psychology Press, 69-84, 2006).

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

Meeting Papers, Presented

A. Brugos, J. Barnes, S. Shattuck-Hufnagel, and N. Veilleux, "A Range of Intonation Patterns Produced in an Elicitation Task," J*ournal of Acoustical Society of America, 119* (*5*), 3301, 2006.

H. Hanson and K.Stevens, "The Nature of Aspiration in Stop Consonants in English", J*ournal of Acoustical Society of America, 119* (*5*), 3393-4, 2006.

S. Lulich, "Modeling the Effects of the Lower Airway on Vowel Spectra," J*ournal of Acoustical Society of America, 119* (*5*), 3303, 2006.

M.L. Matthies, F.H. Guenther, M. Denny, J.S. Perkell, E. Burton, J. Vick, M. Tiede, and H. Lane "Perception and Production of /r/ Allophones Improve with Hearing from a Cochlear Implant, *J. Acoust. Soc. Am. 118,* 1964 2005.

L. Ménard, M. Denny, H. Lane, M.L. Matthies, J.S. Perkell, E. Stockmann, J. Vick, M. Zandipour, T. Balkany, M. Polak, and M. Tiede. "Effects of Speaking Condition and Hearing Status on Vowel Production in Postlingually Deaf Adults with Cochlear Implant," *J. Acoust. Soc. Am. 118*, 1964 2005.

C-Y. Park and J. Slifka. "Towards a Probabilistic System for Estimation of Acoustic Landmarks for Speech Recognition," J*ournal of Acoustical Society of America, 119* (*5*), 3339, 2006.

J.S. Perkell, M. Denny, H. Lane, M.L. Matthies, E. Stockmann, M. Tiede, and J. Vick, "Effects of Masking Noise on Vowel and Sibilant Contrasts in Normal-Hearing Speakers and Postlingually Deafened Cochlear Implant Users," *J. Acoust. Soc. Am. 118*, 1963 2005.

M. Pouplier, M. Goldstein, S. Shattuck-Hufnagel, and M. Tiede, “The Effect of Prosodic Phrasing on Speech Errors,” *Laboratory Phonology 10*, Paris, June 2006.

S. Shattuck-Hufnagel (to appear), “Prosody First or Prosody Last? Evidence from the Phonetics of Word-Final /t/ in American English,” *Proc. Laboratory Phonology 8*, New Haven, June 2004.

S. Shattuck-Hufnagel, and N. Veilleux, “Factors Affecting Phonetic Variation of American English /k/,” J*ournal of Acoustical Society of America, 119* (*5*), 3301-2, 2006.

J. Slifka, “Acoustic Cues to Vowel-Schwa Sequences for High, Front Vowels,” *Journal of the Acoustical Society of America, 118*, 2037, 2005.

J. Slifka, and K. Surana, “Evaluation of Token-Based Acoustic Measures for Classification of Irregular Phonation in Normal Speech,” J*ournal of Acoustical Society of America, 119* (*5*), 3339, 2006.

K. Stevens, “Defining and Enhancing Attributes for Features,” (Invited) “Special Session” J*ournal of Acoustical Society of America, 119* (*5*), 3268, 2006.

S. Zhao, “Contextual Effects on the Continuancy of /ð/,” J*ournal of Acoustical Society of America, 119* (*5*), 3300, 2006.

Meeting Papers, Published

A. Andrade and J. Slifka “A Phonetic Study of Sibilants Produced by 2 Speakers of a Northern Portuguese Dialect,” *Proceedings of XXI Conference of the Association of Portuguese Linguistics*, Portugal, 2005.

J. Barnes, S. Shattuck-Hufnagel, N. Veilleux, and A. Brugos, “The Domain of Realization of the L-Phrase Tone in American English,” *Speech Prosody* Dresden, Germany, PS3-11-163, 2006.

J. Slifka, “Acoustic Cues, Landmarks, and Distinctive Features: a Model of Human Speech Processing,” *6th International Symposium on Natural Language Processing*, Chang Rai, Thailand, 29-36, 2005.

A. Suchato and P. Punyabukkana, "Factors in Classification of Stop Consonant Place of Articulation", In *INTERSPEECH-2005*, Lisbon, Portugal, 2969-2972, 2005.

K. Surana and J. Slifka, (accepted) “Is Irregular Phonation a Reliable Cue Towards the Segmentation of Continuous Speech in American English?” *Speech Prosody* Dresden, Germany, PS7-11-177, 2006.

Y. Yasinnik, S. Shattuck-Hufnagel, and N. Veilleux, “Gesture Marking of Disfluencies in Spontaneous Speech.” *Proceedings of DiSS’05, Disfluency in Spontaneous Speech Workshop*, Aix-en-Provence, France, 173-178, 2005.

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

Theses

E. Hon, *An Acoustic Analysis of Labialization of Coronal Nasal Consonants in American English*, SM Thesis, Department of Electrical Engineering and Computer Science, MIT, 2005.

X. Mou, *Nasal Codas in Standard Chinese – A Study in the Framework of the Distinctive Theory*, PhD thesis, Harvard-MIT Division of Health Sciences and Technology, 2006.

K. Surana, *Classification of Vocal Fold Vibration as Regular or Irregular in Normal, Voiced Speech*, M.Eng. thesis, Department of Electrical Engineering and Computer Science, MIT, 2006.

V. Villacorta. *Speech Sensorimotor Adaptation to Acoustic Perturbations in the First Formant of Vowels and its Relation to Perception,* PhD Thesis, Harvard-MIT Division of Health Sciences and Technology, MIT, 2005.