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,
R01-DC04331,
R01-DC008780,
T32-DC00038.
National Science Foundation (BCS -0418205),
BCS-0643054,
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. Satrajit Ghosh, Mark Tiede, Dr. Miwako Hisagi, Seth Hall.

Visiting Scientists and Research Affiliates
Dr. Abeer Alwan, Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, California.
Dr. Suzanne E. Boyce, Department of Communication Disorders, University of Cincinnati, Cincinnati, Ohio.
Dr. Margaret Denny, CReSS LLC, Lexington, Massachusetts.
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. Helen Hanson, Department of Electrical and Computer Engineering, Union College, Schenectady, New York.
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. Steven Lulich, Harvard School of Public Health, 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. Anthony Okobi, Brown Medical School, Providence, Rhode Island.
Dr. Rupal Patel, Department of Speech Language Pathology and Audiology, Northeastern University, Boston, Massachusetts.
Dr. Janet Slifka, vlingo Corporation, Cambridge, 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.

Dr. Majid Zandipour, BAE Systems, Burlington, Massachusetts.

Graduate Students
Shanqing Cai, Nancy Chen, Xuemin Chi, Elisabeth Hon-Hunt, Youngsook Jung, Caroline Niziolek, Chi-youn Park, Yoko Saikachi

Undergraduate Students
Erick Fuentes

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 further progress on several major studies.

1.1 Cross-speaker variation in vowel production and its relation to perceptual acuity

We completed a study in which we measured acoustic contrast of produced vowels and acuity of vowel perception in 18 young-adult speakers of American English. The results showed, first, that when vowels are produced in different contexts and speaking conditions, measures of contrast distance (separation in the formant plane) decreased from Clear, to Normal, to Fast, as hypothesized. A predicted trading relation across speakers such that, comparing Clear to Fast speech, some speakers change contrast distance more and duration less while others do the converse was not confirmed. Because of economy of effort, the production of a vowel in Fast speech, compared to Clear, was expected to move toward the center of the vowel space, the more so for point vowels since reaching their outlying goal regions requires greater effort (confirmed).

Speakers who were better able to distinguish between exemplars of vowel sounds with subtle spectral differences produced those sounds with more spectral contrast than speakers with less acute spectral contrast discrimination, as predicted. Finally, speakers who were better able to distinguish between exemplars of vowel sounds with subtle spectral differences were predicted to have smaller vowel goal regions and hence to produce those sounds with less token dispersion than speakers with less acute spectral contrast discrimination (confirmed). The results are interpreted in relation to the neurocomputational model, DIVA.

1.2 Clarity vs. Economy of Effort in Vowel Production

An algorithm was developed to extract kinematic data gathered in the above-described recordings. These data are being extracted to test the hypothesis that there is a trading relation

between clarity and economy of effort in vowel production. Effort is being indexed by peak speed of CV movements and clarity is being indexed by vowel contrast distance.

1.3 Perturbation of vowel-to-vowel formant transitions

We further improved our existing algorithm for real-time perturbation of vowel formant frequencies for use in sensorimotor adaptation experiments. As described previously the improved algorithm now can reliably track and shift multiple formants in changing as well as steady state formant trajectories. In a preliminary study of the motor programming of formant trajectories, subjects pronounced utterances containing diphthongs with formant movement from /ɑ/ to /i/, such as *bik*e and *tight*. The transduced speech signal was processed by this algorithm and fed back to the subject with a delay of 9 ms. The signal processing used LPC analysis and resynthesis to detect and shift the F1 and F2 frequencies during the transition; resynthesis used the LPC residual as the source, so the voiced sounds fed back to the subject sounded reasonably natural.

This approach was used to test two competing hypotheses: 1) entire movement trajectories are planned, vs. 2) only trajectory end-points are planned and trajectory shapes are a result of biomechanical or other low-level movement constraints. The hypothesis was tested by using the apparatus to introduce a formant shift that was in a direction normal to a straight-line F1, F2 trajectory connecting the starting and ending points of the transition; the shift was maximal at the transition mid-point and zero at the starting and ending points. We have now run seven female and four male subjects using this perturbation in an adaptation paradigm.

Results of initial data analyses indicated that the subjects showed significant compensatory and adaptive responses to this type of dynamic perturbation; that is they produced trajectory shifts in directions opposite to the perturbation and maintained those produced shifts for a short while after the perturbation was removed. However, most of the compensatory corrections involved the entire F1-F2 trajectory, including the trajectory end-points that had not been perturbed. Compensatory alterations of the shape of the trajectories were relatively minor, leading to difficulties in interpreting the results and testing the initial hypothesis.

As an effort to improve the dynamic perturbation paradigm, we devised a new type of perturbation which alters the time course of the formant transitions, but does not alter the position or shape of the originally produced trajectory in the F1-F2 plane. This paradigm can be performed in two directions, initial acceleration and initial deceleration, which differ in the direction of the perturbation vectors. We also incorporated a psychophysical test into the experiment protocol in order to study the relationship between perception and production of dynamic formant transitions. Fourteen subjects (7 male, 7 female; 8 acceleration, 6 deceleration) have been run under this paradigm. Data analyses are currently underway.

We are also extending the dynamic perturbation paradigm to tripthongs, exemplified by /iau/ in Mandarin Chinese. The advantages of using triphthong are their longer durations, greater F1-F2 trajectory lengths, and increased trajectory complexity, which permits greater flexibility in design of the perturbation fields. Software design and testing for the triphthong perturbation paradigm are underway.

1.4 fMRI Study of Cross-Category versus Within-Category Perturbations of Vowels

This study builds on a previous fMRI study of auditorily perturbed speech, in which a subject's first formant frequency was shifted in real time during a whole-brain fMRI scan. That study identified the auditory feedback control network underlying speech production. In the current experiment, we provide subjects with two different types of perturbations of the same size in auditory space: one perturbation that moves the subject's speech signal fed back to the subject into another phonetic category (e.g., "bet" sounds like "bit"), and a second which does not move the speech signal across a boundary (e.g., "bet" just sounds like a poorly pronounced "bet"). We have refined the process of determining the direction for these perturbations in the auditory space

formed by the first two formant frequencies of the speech signal. This process is not a simple one due to issues of counterbalancing the perturbation directions across subjects. For example, we need to devise perturbations that will move some subjects' feedback across a category boundary but will not move other subjects' feedback across the same boundary. In addition, we have developed a psychophysical test to determine the location of speech category boundaries in our subjects. Our preliminary results show a great deal of variability in the location of boundaries across subjects for a given vowel pair. This variability enables us to solve the counterbalancing problem by choosing a constant shift amount that will elicit different phonetic percepts depending on the subject.

2. Acoustic Landmarks and Speech Processing

2.1 Consonant landmark detection for speech recognition Introduction

The consonant landmark detector locates the time-points of abrupt acoustic changes, corresponding to closures and releases of consonants. Not only does this algorithm pinpoint the location of abruptness, but it also classifies the detected landmarks according to their characteristics---onsets and offsets of spontaneous glottal vibrations, existence of burst noise, and closures and releases of sonorant consonants. Therefore, this landmark detector can serve as a first step of a knowledge-based automatic speech recognition system, by finding the locations where phonetic information is highly concentrated, and also by determining the type of information that can be found in the vicinity of each detected landmark or between each group of adjacent landmarks.

Method

2.1.1. Individual landmark detection

The landmark detection algorithm is developed in three stages. The first stage provides a probabilistic algorithm that detects the consonant landmarks individually, that is, without considering the relationship among different types of landmarks. This algorithm is largely a direct conversion of Liu's consonant landmark detector into a probabilistic system (Liu, 1996). The new design has the following three advantages: First, the overall system has been redesigned to have a separate knowledge-base and a computation core system. Secondly, the new algorithm allows more candidates to be detected, by increasing the sensitivity to acoustic changes. Finally, the algorithm was developed as a probabilistic system by assigning a probability of each landmark candidate being a true landmark. The probability measure is used in later processing steps to sort out false alarms that have arisen due to increased sensitivity of the landmark candidate detection algorithm.

2.1.2. Landmark sequence determination

The second stage of the landmark detection algorithm makes use of the strict dependency that we observe among true landmark sequences to determine the most likely sequence among the detected landmarks. This process is meaningful not only because it determines a single landmark sequence that can be directly input to the next step of automatic speech recognition, but also because this process models the relationship among different landmark types with a simple bigram representation.

The bigram model that represents the constraints on possible landmark pairs has significance in two ways. One is that among 36 theoretically possible pairs of landmark types, only 16 pairs are articulatorily feasible. The other 20 landmark sequences are physiologically impossible to produce. Therefore, the bigram representation of landmark sequence constraints can be effectively used to filter out unlikely sequences of landmarks. Another advantage is that each landmark pair describes some of the acoustic properties of the signal between the two landmarks.

Therefore, when a landmark sequence that follows the bigram constraints is observed, the articulator-free features of the signal can be directly predicted from the landmark sequence.

2.1.3. Representation of reliable and ambiguous regions

The last stage of the algorithm distinguishes the regions where the landmark sequences can be determined reliably from the regions where more than one possible sequence can be hypothesized. It is expected that where one speaks more carefully, e.g., near word boundaries or lexical stresses, cues to the distinctive features will be produced more clearly and the landmarks will be detected more reliably, and so the reliably detected landmarks can be given more focus in later stages of speech recognition. On the other hand, the ambiguous regions can be given multiple possibilities of landmark sequences, and evidence for additional cues can be sought based on the possible choices, so that the landmark sequence can be determined with more confidence.

2.1.4. Result and conclusion

On the TIMIT test set, 91% of all the consonant landmarks and 95% of obstruent landmarks are located as landmark candidates. The bigram-based process for determining the most likely landmark sequences yields 12% deletion and substitution rates and a 15% insertion rate. An alternative representation that distinguishes reliable and ambiguous regions can detect 92% of the landmarks, while 40% of the true landmarks are judged to be reliable. The deletion rate within reliable regions is as low as 5%.

The landmark detection algorithm has two major improvements: bigram modeling of landmark sequence, and representation of ambiguous and reliable regions. The bigram transition asserts that the resulting set of landmarks should follow the strict rules of landmark sequencing, and it reduces the possibility of contradiction between the sets of distinctive features estimated from adjacent landmarks. Once the sequence of landmarks is detected, it can help reduce the number of possible word candidates in a large vocabulary speech recognition. It is estimated that when the landmark sequence is known, the average number of possible word candidates can be reduced from 20,000 to 5.6 words with the worst-case result equal to 722 words, which is about 3.6% of the total number of candidates.

It has been observed that the reliably detected regions generally correspond to lexical stresses or word boundaries. It has been generally accepted that stressed syllables have robust acoustic cues for phonetic features. Gow and Gordon (1995) also suggest that in continuous speech, word onsets show more robust acoustic realization of phonetic features and are less variable in terms of phonological assimilation than other parts of words. Therefore, it would be worthwhile to investigate the relationship between reliable detection of landmarks and suprasegmental features such as word onsets and lexical or prosodic stress. If it is verified that landmark reliability corresponds to perceptual islands of reliability, the landmarks within the reliable regions can be used as providing valuable information for lexical access.

References

[1] D.W. Gow and P.C. Gordon, “Lexical and Prelexical Influences on Word Segmentation: Evidence from Priming,” *Journal of Experimental Psychology: Human Perception and Performance 21*: 344-359 (1995).

[2] S.A. Liu, “Landmark Detection for Distinctive Feature-Based Speech Recognition,” *Journal of the Acoustical Society of America 100*: 3417-3430 (1996).

2.2 Patterns of landmark change in continuous speech

The model of speech recognition proposed here rests on the assumption that a particular type of

feature cue, (i.e. the abrupt changes termed landmarks that are associated with consonant closures and releases, vowels and glides) is robustly detected by listeners in the first critical stages of the lexical access process. This approach assumes that the landmarks predicted by the feature specifications of words in the speaker's mental lexicon generally survive the rigors of articulatory overlap and weakening in spontaneous speech production. To test this assumption, and to determine the most common patterns of landmark change or loss in continuous speech, a database of 1 hour of spontaneous task-driven speech was elicited from 8 female speakers: the MIT American English Maptask. In a sample of this corpus, each predicted landmark was hand-labeled as either realized, changed or lost. Preliminary results for one conversation (240 secs., 610 words, analysis completed for 1003 of 2750 predicted landmarks) show that 86% of landmarks were realized overall (Shattuck-Hufnagel and Veilleux 2007).

These results suggest that the majority of landmarks are available for detection both by human listeners and by automatic recognition algorithms, even in highly-spontaneous, unselfconscious speech. They also provide a more reasonable benchmark against which to measure the performance of an automatic landmark detection algorithm (Liu 1996, Park 2008), Moreover, missing or changed landmarks occurred in sharply limited circumstances (e.g. many predicted landmarks are not realized for the coronal stops /t/ and /d/, particularly in word-final consonant clusters; voiced /dh/ is often realized as a stop (Zhao 2006); velar /g/ is often realized as a glide-like approximant with no visible closure silence or release burst, etc.) In an accompanying study using classification trees, (Veilleux and Shattuck-Hufnagel, 2008), some of the factors that appear to facilitate these changes were examined, including prosody, word structure, morphosyntactic categories and landmark type. Because patterns of implementation, change and loss for predicted landmarks appear to be systematic and predictable, these results suggest it will be both possible and useful to develop recognition algorithms that take advantage of the limited feature-cue-based character of variation in spontaneous speech signals.

References

[1] S. Shattuck-Hufnagel and N. Veilleux, "Robustness of Acoustic Landmarks in Spontaneously-Spoken American English" *Proceedings 16th International Congress of Phonetic Sciences ICPhS-07,* Saarbrucken, Germany, 925-928, 2007.

[2] S.A. Liu, "Landmark Detection for Distinctive Feature-Based Speech Recognition," *Journal of Acoustical Society of America 100*: 3417-3430 (1996).

[3] C. Park, "Consonant Landmark Detection for Speech Recognition", PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2008.

[4] S.Y. Zhao, "The Stop-Like Modifications of /ð/: A Study in the Analysis and Handling of Speech Variation", PhD thesis, Graduate Program in Speech and Hearing Biosciences and Technology, Harvard-MIT Division of Health Sciences and Technology, MIT, 2007.

[5] N. Veilleux, and S. Shattuck-Hufnagel, "Automatic Detection of the Context of Acoustic Landmark Deletion," accepted for presentation at ICSLP 2008, New Zealand.

2.3 Word boundary detection using landmarks Introduction

This report is concerned with the extraction of acoustic cues for word boundaries --- a topic that has been examined in some depth over the years, leading to several proposed candidates for word boundary estimation. In order to examine these and other candidates, a new database was prepared. In the light of new findings of how different prosodic factors such as lexical stress, pitch accent, and phrase-level prosody might affect acoustic correlates for word boundaries, a database that was tightly controlled for all these factors was designed and recorded. The database was focused on consonants in the context of two different vowels /i/ and /o/. Several classes of American English consonants were measured, where the only variable was their word

position --- being either word-initial or word-final position. Based on a preliminary study it was observed that a simple duration measure gave the best separation of word-initial and word-final consonants. This report presents data and discussion for this set of duration measures.

2.3.1 Results – Comparison between word-initial vs. word-final consonants

Comparison of the average durations of the word-initial consonants and the word-final consonants for individual speakers showed that the word-initial consonants were consistently longer than the word-final consonants for all classes of consonants. The absolute values of the durations, however, showed some individual differences, presumably a consequence of different speaking rates, speaking styles, or anatomical differences. There was not, however, a significant difference in these results for the two vowel environments [i] and [o]. In addition to being significantly shorter than the word-initial consonants, the word-final consonants frequently showed the presence of two other attributes; glottalization of the following vowel and insertion of a pause after the consonant. The extent to which these acoustic modifications were present was somewhat speaker-dependent; about one-half of the speakers showed these attributes. For a given category of word-initial consonant (e.g., voiced or voiceless stops), the duration was about the same for all places of articulation (e.g., the duration was the same for b, d, and g). For the final consonants, on the other hand, there were significant differences in the durations for these places of articulation.

2.3.2 Some properties of word-initial consonants

For the word-initial consonants, the differences in durations, averaged over speakers and consonants within a class, are summarized in Table 1. For the stop consonants there was an increase in duration from nasals to voiced stops to voiceless stops. The voiceless fricatives were slightly shorter than the voiceless stops, and the affricates were somewhat longer than their stop counterparts. Within each class the variation across the different articulators was small, with the exception of voiceless fricatives, for which the fricative [sh] had a duration somewhat greater than [f] and [s] (not shown in Table 1).

nasals	**voiced stops**	**voiceless stops**	**voiceless fricatives**	**affricates**
m, n	b, d, g	p, t, k	f, s, sh	jh, ch
80	95	150	138	118 155

Table 1. Average durations of various word-initial consonant classes in milliseconds. Averages are over 6 speakers and over 5 consonant classes and 2 vowel contexts.

The differences in duration between consonant classes can be explained, at least in part, by reviewing what is known about the articulatory movements and the influence of the intraoral pressure on these movements. For the nasal consonants, there is essentially no change in the intraoral pressure as the vocal tract shape passes from the end of the preceding vowel to the beginning of the following vowel. Therefore the consonant closures and openings were not influenced by the intraoral pressure. For the voiced stops, the active articulator is moved close to the opposing articulator (e.g., the tongue blade is close to the hard palate), but leaving a narrow constriction so that airflow continues to maintain glottal vibration. Some intraoral pressure builds up, but this pressure is substantially less than the transglottal pressure. Production of the voiceless stops requires a complete closure of the constriction formed by the active articulator, relative to voiced stops. Creation of this closure and the resulting buildup of intraoral pressure and the subsequent release of this pressure require a substantial increase in duration of about 55 milliseconds on average. In the case of voiceless fricatives the average duration is smaller than that for voiceless stops. The voiceless affricates are only slightly longer than the voiceless stop consonants. This increased duration is presumably a consequence of the sequence of two events that characterize the affricates. The increased duration for the voiced affricate relative to the voiced stop is much greater since that stop is not aspirated.

In summary, then, the duration data are consistent with a view that the combination of the required articulatory movements for the various consonant categories, together with the influence of airflow through the constricted vocal tract and its interaction with the impedance of the vocal tract walls, cause consistent changes in the duration of consonant production. The greatest of these influences is the contrast between voiced and voiceless obstruents.

2.3.3 Discussion

This experimental study of acoustic cues to word boundaries has provided data for relatively simple two-syllable utterances with well-controlled prosody. The data show that the word-initial consonants can be predicted from simple acoustic measures of consonant duration if the articulator-free features of the consonants can be determined. These results are essentially independent of the place of articulation for obstruent consonants and for nasal consonants, although other sonorant consonants have not been tested. Consonants in word-final position can have a number of acoustic properties that differ in several ways from those for the word-initial consonants. For utterances that are not well controlled for prosody, it is expected that the consonant durations will be modified somewhat, particularly for consonants in utterance-final or phrase-final positions.

The data on durational differences as cues for word boundaries, together with the influence of articulator-free features, fit well into a landmark-feature model of human speech recognition (Park, 2008). Word boundaries are based on estimates of just a few cues, such as cues for the voicing feature, the feature continuant, the feature sonorant, and the time between landmarks for the boundaries. Identification of place of articulation and some other features is not a requirement at this stage of the model. However, the identification of these features (primarily articulator-bound features) is enhanced by knowledge of word boundaries and of the articulator-free features derived at the landmark stage. For example, knowledge of word boundaries can help to reduce the cohort of possible utterances. Experimental work suggesting that knowledge of articulator-free features (similar to so-called manner features) without clear estimation of articulator-bound features can lead to good decoding of sentence material, whether by human listeners (Shannon et al. 1995) or by automatic speech recognition (Huttenlocher and Zue, 1984).

References

[1] D.P. Huttenlocher and V.W. Zue, “A Model of Lexical Access from Partial Phonetic Information,” *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-84, Vol. 9*, 391-394, 1984.

[2] C. Park, “Consonant Landmark Detection for Speech Recognition”, PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2008.

[3] R.V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech Recognition with Primarily Temporal Cues,” *Science 270*: 303-304 (1995).

3. Characteristics of Consonants and Vowels

3.1 Vowel nasalization in American English

This study quantifies acoustic variation of vowel nasalization that arises from consonant context in American English. While qualitative articulatory trajectories and phonetic descriptions suggest a vowel is nasalized in carryover contexts, few acoustic studies have examined this issue in detail. We have carried out a quantitative investigation to determine the degree and extent of nasalization in vowels (V) in several different consonant contexts (nonnasal C or nasal N). In particular the contexts were 0-CVC (no adjacent nasal consonants), 1-NVC-CVN (1 adjacent nasal), and 2-NVN (nasals preceding and following).

A total of 900 tokens containing the vowel /i/ were collected from six male speakers of American English. Examples of the target words are (1) NVN: *mean*; (2) CVN: *team*; (3) NVC: *neat*. The acoustic measure of nasalization in the vowels was A1-P1, a measure proposed by Chen (1997), where A1 is the amplitude of the largest harmonic in the low frequency range, and P1 is the largest harmonic in the frequency range 700 to 1500 Hz. P1 has been shown to be in the frequency range of the first nasal resonance in a nasal vowel. Since A1 decreases and P1 increases during vowel nasalization, A1-P1 is expected to be smaller when a vowel is more nasalized. Measurements of A1-P1 were taken by applying 20 ms Hamming windows and computing 512-point Fourier transforms at 10, 20, 30, 40, and 50 ms from the closures or releases of the consonants. The durations of the vowels ranged from 100-250 ms.

All subjects showed a significant influence of the adjacent nasal consonants on the vowel except for one subject who failed to show an effect of the consonant on a preceding vowel. Otherwise the influence of the nasal consonant as measured by A1-P1 ranged from about 5-15 dB depending on the speaker and on the position of the consonant in relation to the vowel. The difference varied only slightly over a time interval of 40 ms adjacent to the vowel. It is expected that this "carryover" effect will be different in different languages, depending on the phonemic inventory in the language.

References

[1] M.Y. Chen, "Acoustic Correlates of English and French Nasalized Vowels," *Journal of the Acoustical Society of America 102*: 2360-2370 (1997).

3.2 Acoustic characteristics of glides /j/ and /w/

As part of a larger study of acoustic characteristics of glides, patterns in the movements of the fundamental frequency of phonation (F0) have been measured during utterances of the glide segments /j/ and /w/. Two male and two female American English speakers produced intervocalic glides in six different vowel contexts and five different prosodic contexts. Target utterances were **VCV** nonsense words embedded in carrier phrases; in each utterance, **C** was one of the glides /j, w/, and **V** was one of the vowels /i, u, e, o, ae, a/. Carrier phrases were constructed such that there was either a high pitch accent (H*) on the first **V**, a low pitch accent (L*) on the first **V**, H* on the second **V**, L* on the second **V**, or no pitch accent during the target utterance.

The fundamental frequency (F0) was measured from the waveform at each pitch period, creating a contour from the steady-state portion of the first vowel through the glide segment to the steady-state portion of the second vowel. These F0 contours show pronounced valleys during the glide segments, in which the frequency of phonation is decreased relative to the surrounding prosodic contour, even though no prosodic pitch target exists during the glide. Fig. 1 shows an example of an F0 contour from a female utterance of the sequence /uju/, plotted together with the first formant frequency (F1) contour from the same utterance. In this example, the F0 contour reaches a minimum in the middle of the glide segment, approximately 50 Hz lower than its value during the preceding vowel. The only prosodic target in this utterance is a H* on the following vowel; thus the presence of the F0 valley during the glide segment cannot be explained by currently accepted prosodic theories alone.

These data can be interpreted instead in terms of acoustic loading on the glottal source, the effects of which are most pronounced when a narrow vocal tract constriction (as in a glide segment) lowers the first formant frequency (F1) into the range of F0. As explained recently by Titze (2008), increased inertive loading on the glottal source when F1 and F0 are brought into proximity tends to decrease F0 from its theoretical frequency in the absence of such interaction. In the current study, the F0 valleys occur more frequently for higher average fundamental frequencies, as in female speech. In addition, the valleys occur more often when surrounded by

high vowels (with low F1) than low vowels. These results support the acoustic loading hypothesis, as the interaction effects are more often observed when F0 and F1 are in closer proximity to each other. Effects of surrounding prosodic contexts are also observed, with F0 valleys more often occurring during glides when the following vowel is pitch accented; this may be an effect of consonant strengthening before pitch accented vowels.

Work currently in progress seeks to quantify the interactions between F0 and F1 during glide segments, accounting for surrounding vowel and prosodic contexts. These interactions may provide an enhancing acoustic cue to the presence of glides in the speech stream, perhaps particularly for female speakers or high vowel contexts. Future work may test the perceptual value of such cues through experiments with listening subjects and synthetic glide/non-glide utterances.

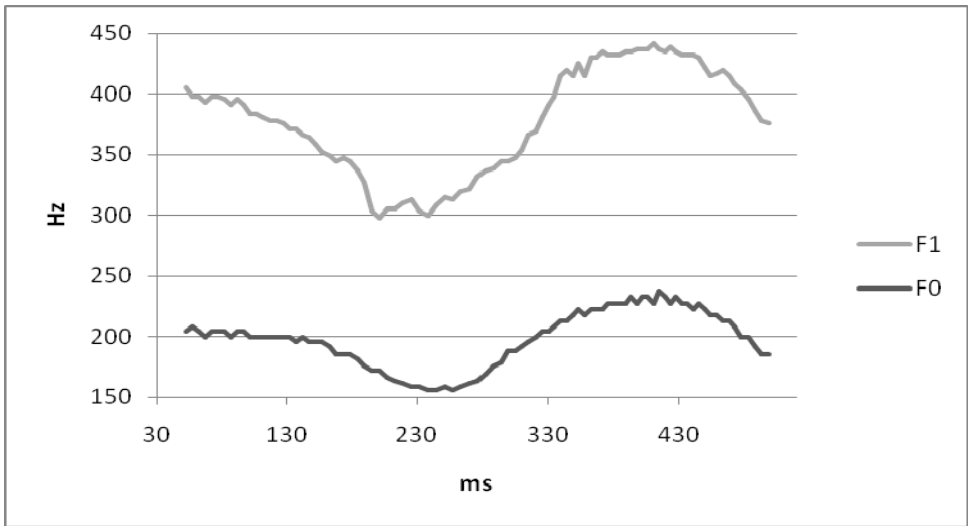

Fig. 1: Fundamental frequency (F0) and first formant frequency (F1) contours during a female utterance of the sequence /uju/ with high pitch accent on the second vowel. The midpoint of the glide segment /j/ occurs around time index 250 ms. At this point the F0 contour displays a pronounced valley, although there is no low pitch accent prosodic target in this sequence. Rather, the F0 valley coincides with the minimum in F1, which is consistent with the effects of acoustic loading on the glottal source.

Reference

[1] I.R. Titze, "Nonlinear Source-Filter Coupling in Phonation: Theory. *Journal of the Acoustical Society of America 123 (5)*: 2733-2749 (2008).

3.3 Relation between the subglottal resonances and the vowel categories across languages

Previous research has suggested that the subglottal resonances define the vowel features [back] and [low] for English. In testing this hypothesis, we sought to determine whether the vowel feature boundaries are independent of language. As preliminary research, we made recordings of speech and subglottal signals simultaneously for several adult Korean speakers using an accelerometer. We found amplitude attenuation or frequency discontinuity in the first and second formant frequency near the subglottal resonances in Korean vowels as in English vowels. The boundary between [+low] and [-low] vowels agrees with a speaker's first subglottal resonance, while the boundary between [+back] and [-back] vowels agrees with a speaker's second subglottal resonance. These findings provide evidence that the subglottal resonances define the vowel

features independent of language, as well as insights into why there exist natural vowel categories common to the languages.

4. Quantal Theory and Enhancement

4.1 Some possible new insights

In recent years it has been proposed that the acoustic and articulatory parameters for a number of phonemic contrasts or distinctive features in language are based on quantal relations between acoustics and articulation (Stevens, 1989; Keyser and Stevens, 2006). For example, the articulatory attribute that gives rise to a particular distinctive feature tends to create a stable acoustic property that is relatively insensitive to deviations in the articulatory parameter. From such observations it can be hypothesized that there are particular states of the speech production system from which the basic building blocks of phonological units are derived.

This interpretation of quantal relations between articulatory and acoustic parameters implies that the relation defining a distinctive feature makes reference to a particular target value of an acoustic parameter that is observed over a range of values for an articulatory parameter. Thus the defining acoustic attribute is not expected to be a description of a slowly time-varying parameter such as the movement of a spectrum peak for a vowel.

Examination of the defining properties of a number of distinctive features suggests that the features can be classified into two distinct groups. In one group, the quantal properties arise from changes that occur in the vocal tract shape when acoustic coupling between resonators within the system creates abrupt movements of zeros in the transfer function. This group is called ARC, or Acoustic Resonator Coupling. For the other group, the quantal properties are a consequence of abrupt changes in acoustic sources in the system due to interaction between pressures within the vocal tract and the yielding walls of the vocal tract. This group of features is called AMI, for Aeromechanical Interaction. Examples of ARC are (1) the feature [back] which involves acoustic coupling between the second subglottal resonance and the second formant frequency for vowels; and (2) the feature [anterior], which involves changes in affiliation between front and back cavities for obstruent consonants. Examples of AMI include the features [sonorant], [continuant], and [strident]. This classification of features is similar (with some exceptions) to terms that have been used in the past. For example, ARC features are similar to place features, and AMI to manner features. Or the articulator-free and articulator-bound features of Halle (1990) are roughly the same as AMI and ARC features in the proposed classification based on acoustic and articulatory interactions. The proposed classification may help to provide a more quantitative description of the types of features.

While the defining acoustic attribute that underlies a feature provides a potential cue that a listener might use to identify the feature, there may be other cues that are available to a listener. These cues, called enhancing cues, may be present at locations in the signal that are adjacent to the location of the target, and may be influenced by acoustic information that also contains cues for a nearby segment. It may often occur that in running speech there is overlap in the cues for features of one segment with cues for adjacent segments, so that some of the cues, including the defining attributes, may be obliterated. The listener may still be able to identify the feature based on evidence from the remaining cues.

Identification of the various cues for features in running speech, particularly for consonants, is facilitated by the detection of landmarks in the signal (Park, 2008). These landmarks indicate the presence of abrupt changes or transients in the sound. Through identification of three different types of landmarks (representing cessations or onsets of voicing, abruptnesses due to stop consonants, and sudden changes for sonorant consonants), and through constraining the allowable sequences of these landmarks, reasonably accurate identification of these landmarks can be made.

References

[1] M. Halle, "Phonological Features". In W. Bright (ed.) *Oxford International Encyclopedia of Linguistics*, New York: Oxford University Press, *Vol. 3*, pp. 207-212 (1990).

[2] S.J. Keyser and K.N. Stevens, "Enhancement and Overlap in the Speech Chain", *Language 82*: 33-63 (2006).

[3] C.Y. Park, "*Consonant Landmark Detection for Speech Recognition"*, PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2008.

[4] K.N. Stevens, "On the Quantal Nature of Speech", *Journal of Phonetics 17*: 3-46 (1989).

4.2 Quantal theory and sparse vowels

There are languages that have only one low vowel, so that there is no contrast of [+back] and [-back] for the low vowel. The question arises whether in such languages this vowel is made in a way that avoids the second subglottal resonance F2sub. Possible views are: (1) the low vowel is always [+back] or (2) F2sub is avoided but the low vowel is front or back depending on the adjacent consonant. We tested which of these hypotheses is correct for the low vowels in Korean which has only one low vowel. F2 and F2sub were measured in the context /CaC/, where C is a consonant. Consonants are /n/, /t/, /s/, /č/, /d/, /ǰ/, /m/, /p/, /g/, /b/. F2sub was always avoided for the low vowel. The vowel was front or back depending on the adjacent consonant: if the adjacent consonants were labial or velar ([+back]), F2 of the low vowel was lower than F2sub, whereas if the consonants were alveolar ([-back]), F2 of the vowel was higher than F2sub. These results show that the low vowel is shifted by the effect of adjacent consonants, and the contrast in the backness of the vowel is likely related to speakers' F2sub.

A similar influence on the vowel feature [back] can be observed in English for the reduced mid vowel. For example, in an utterance like "pass a dip", in which the reduced vowel is surrounded by alveolar consonants, F2 for the vowel is relatively fronted (above F2sub). On the other hand, when the reduced vowel in surrounded by consonants that are more backed, as in "rub a book", F2 is relatively low and below F2sub.

4.3 Acoustic influences of the lower airways on speech

We have been studying various aspects of the influence of lower airway acoustics on speech. First, we are carrying out computer simulations of the extra-thoracic trachea to determine the effect of its geometry on the lower airway impedance and its influence both on vocal fold vibration via nonlinear source-filter interaction, and also on the resonances (poles and zeros) in the speech signal. Initial results suggest that individual anatomical variations may have a significant effect on the lower airway impedance. Second, we are exploring the possibility that subglottal coupling in speech may occur during the closed phase of vocal fold vibration. Simple models of the vocal folds allow us to infer their mechanical properties from measures of subglottal resonances (pole-zero pairs) in the speech signal, and perhaps also from accelerometer signals of the vibration of the skin of the neck just below the larynx. Third, in collaboration with colleagues at Harvard and MIT, we are working with a simple physical model of vocal fold vibration to determine the effects of vocal fold geometry, tension, and stiffness on voice production. This model differs from other physical models currently being studied by other groups in its simplicity, flexibility, and the ability to apply tension to the vocal folds. Fourth, partly in collaboration with colleagues at the University of Stuttgart, we are expanding our study of the separation of vowel and consonant categories by the subglottal resonances in adult speakers of English and two dialects of German (Madsack et al, 2008; Lulich, to appear). This work follows up on previous work done in the Speech Group over the last several years (Lulich et al, 2007). Fifth, in collaboration with colleagues at UCLA, we are conducting experiments to test whether the subglottal resonances can be detected

automatically from speech in children (speakers of English and Spanish) and adults, and whether these can lead to improved speaker normalization and speech recognition technology. Results thus far indicate that the subglottal resonances can be accurately detected automatically, and that speaker normalization based on these resonances is more robust and efficient than current state-of-the-art techniques (Wang et al, 2008; Wang et al, to appear).

References

[1] Steven M. Lulich. (to appear) Subglottal resonances and distinctive features. Journal of Phonetics.

[2] Shizhen Wang, Steven M. Lulich, Abeer Alwan. (to appear) A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation. Proceedings of Interspeech.

[3] Shizhen Wang, Abeer Alwan, Steven M. Lulich. (2008) Speaker normalization based on subglottal resonances. IEEE International Conference on Acoustics, Speech, and Signal Processing 2008, pp. 4277-4280.

[4] Andreas Madsack, Steven M. Lulich, Wolfgang Wokurek, Grzegorz Dogil. (2008) Subglottal resonances and vowel formant variability: A case study of High German monophthongs and Swabian diphthongs. Proceedings of LabPhon11, pp. 91-92.

[5] Steven M. Lulich, Asaf Bachrach, Nicolas Malyska. (2007) A role for the second subglottal resonance in lexical access. Journal of the Acoustical Society of America 122(4):2320-2327.

5. Studies of Individual Differences, Speech Development and Disorders

5.1 Effects of hearing status on adult speech production Introduction

The goals of this research are: to deepen significantly our understanding of how a speaker's auditory acuity influences his or her speech motor planning; to describe the speech perception and production of hearing-impaired adults and the effects of cochlear prostheses; and to evaluate and help refine a quantitative model of the role of hearing in speech. We are conducting experiments with normal-hearing speakers, and with postlingually deafened adults who receive cochlear implants. The experiments measure the effects of the implants on speech in recordings made before implantation and up to two years after as speakers' auditory acuity evolves. According to our model of the role of hearing in adult speech motor control, many of the goals of speech movements are in the auditory domain. Consequently, a central theme of this research is the role of auditory perception in the feedback and feedforward control systems that are used to achieve auditory goals during speech. Feedforward control is almost entirely responsible for generating articulatory movements in adults. However, when there is a mismatch between the speaker's intention and the resulting auditory feedback during production of a speech sound, that error leads to corrective motor commands that serve to update feedforward commands for subsequent movements. The speaker's ability to detect such a mismatch depends on his or her auditory acuity.

The proposed research examines the role of auditory acuity when producing phoneme and lexical stress contrasts; when compensating for feedback perturbation of vowel formants; and when imitating synthesized vowels. To measure acuity we present synthetic speech continua for discrimination testing. To assess relations between production and acuity, we measure the degree of separation of produced contrastive phonemes; dispersion of productions around their phoneme means; compensation for introduced vowel formant shifts; and imitation accuracy for vowels. In most of these experiments we also block auditory feedback temporarily in order to

reveal the state of feedforward commands. Analyses take demographic variables such as age at hearing loss and duration of implant use into account.

5.2 Development of facilities

During this period we have completed an extensive rewrite of our approach to data acquisition, now Matlab-based to support fast prototyping and hardware independence. We have also enhanced our procedures for F0 and formant tracking and integrated these into our tools for data analysis. New procedures have been developed for testing auditory perception on both a standard (CNC) word test and for SINFA analysis of confusable features. A previously developed interactive procedure used to elicit judgments of vowel quality for the purpose of mapping subjects' vowel perceptual spaces has been modified in response to extensive pilot testing.

5.3 Pilot studies and subject recruitment

All of the experimental paradigms and procedures have now been extensively piloted and refined. In the pilot studies, we have run all the experimental procedures on eight implant users (4 male, 4 female) and nine age-matched normal-hearing speakers (4 male, 4 female). Complete analyses of these data are in progress, and efforts are underway to recruit new implant candidates for the full set of planned experiments.

5.4 Development of vowel production

The objective of this project is to study the roles of subglottal resonances F1sub and F2sub in defining the vowel features for children. There are huge differences in the physical dimension of the speech production system between adult and infant speakers. With a focus on children of 2;6 to 3;7 years, the characteristics of formant frequencies and amplitudes in the vowels [a] and [ae] were analyzed to examine how the changes in formant frequencies relate to their subglottal resonances. Acoustic data were obtained primarily from a database collected and reported by Imbrie (2005). The hypothesis is that the vowel boundaries are determined by children's subglottal resonances, as in adult speakers. Measures of subglottal resonances for each child were obtained from the locations of discontinuities in the F1 and F2 trajectories in diphthongs and in vowel-consonant transitions. Preliminary results from ten children indicate that F1 frequencies for these vowels are more likely to be higher than F1sub frequencies as the children age. Similarly, F2 frequencies of the two vowels are more likely to fall on opposite sides of F2sub as the children age. The transition from the expected quantal relation appears to occur in the age range between 2 and 3 years. The data show that, even for the small amount of data, it seems that the children do better at the age of 3 than at the younger age 2.5 years. These findings provide insights into how the physiological characteristics of the speech production system affect acoustics when they learn how to produce sounds.

5.5 Consonant codas

We are continuing our investigations of speech development, using the Imbrie Corpus (Imbrie 2005) of recordings from 10 children ages 2;6-3;7 for the study of coda development in CVC target words. Preliminary analysis of 3 speakers shows that these children produce non-adult-like cues to the voicing contrast in coda stops, including a region of noise excitation at the end of the vowel/beginning of closure for voiceless /p,k/, e.g. in 'cup', 'duck', that is not observed for voiced /b,g/ , e.g. in 'tub', 'bug' (Shattuck-Hufnagel et al. 2007). Although this cue is not observed in the adult-directed speech of these children's parents, it has been observed as a contrastive cue ('pre-aspiration') in other languages. Ongoing analysis of the acoustics of this noise excitation (Hanson and Shattuck-Hufnagel, 2008) is aimed at determining whether it arises in the laryngeal region, as aspiration, at the oral closure for the consonant, as frication, or both. These analyses will shed light on the question of whether this noise at the end of the vowel results from immature motor control leading to non-adult-like timing and amplitude of articulatory gestures, or from the child's decision to provide an enhancing cue to the [-voice] feature of the coda

consonant. The interesting observation that CVCs with voiceless onset and coda consonants (e.g. 'cup') are sometimes produced with a vowel that is whispered (i.e. unvoiced) throughout, lends some support to the former hypothesis, since it may occur because the child cannot switch quickly enough from [-voice] to [+voice] to [-voice] cues.

References

[1] H. Hanson, and S. Shattuck-Hufnagel, “Acoustic Cues to the Voicing Contrast in Coda Stops in the Speech of 2-year- olds Learning American English”, *Journal of the Acoustical Society of America 123 (5)*: p. 3320 (2008).

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

[3] S. Shattuck-Hufnagel, K. Demuth, H. Hanson, and K. Stevens, “Acoustic Cues to Voicing Contrasts in Coda Stop Consonants in 2-3-year old Speakers of American English”, presented at the Conference, *Where do Features Come From* held in Sorbonne, Paris, October 2007.

5.6 F0 Control in electrolarynx speech

An electrolarynx (EL) is a battery-powered device that produces a sound that can be used to acoustically excite the vocal tract as a substitute for laryngeal voice production. Although ELs provide laryngectomy patients with the basic capability to communicate, EL speech contains persistent acoustic deficits that result in reduced intelligibility and contribute to the “mechanical” sound quality of EL speech. Major acoustic deficits in EL speech include lack of normal fundamental frequency (F0) variation/control, spectral characteristics (reduced energy below 500Hz), and directly radiated sounds from the EL that interferes with the speech signal. In order to improve EL speech communication system, the current study aimed to develop and evaluate a procedure for controlling the F0 automatically. The idea was to co-vary the F0 of EL utterances based on the variation in the RMS amplitude of the EL speech signal output. In our previous study of acoustic analysis of speech produced both before and after the laryngectomy operations (pre-laryngectomy speech vs. EL speech) by the same speakers, we found significant fluctuations in amplitude in EL speech as a function of time. In particular there was a gradual decrease of amplitude during vowels at the end of declarative utterances, which was similar to what we observed in the corresponding pre-laryngectomy speech. Furthermore, there were generally positive correlations between F0 and amplitude in the pre-laryngectomy speech. Based on these observations, we hypothesized that the variation in amplitude in EL speech could be used as a predictor for estimating F0 in EL speech.

In the current study, two declarative sentences ending with vowels produced by two male laryngectomy patients before and after the laryngectomy were used in order to develop procedures for controlling F0. Specifically, the linear regression coefficients between F0 and RMS amplitude in the pre-laryngectomy speech were first determined for each speaker and sentence, and applied to the amplitude variation in EL speech to compute the F0 contour for the EL sentences. An analysis-by-synthesis approach using the Klatt formant synthesizer was employed to modify the F0 contour of the EL speech as computed. Perceptual evaluation was conducted in order to determine whether the proposed approach could significantly improve the naturalness of EL speech. The experiment consists of two tasks: the Paired comparison task and Visual analog scaling task. The F0 modified EL speech based on the amplitude was compared to the EL speech with constant F0 as well as to the EL speech with F0 modifications based on the corresponding pre-laryngectomy F0 contour. The results showed that the addition of amplitude-based F0 modulation resulted in EL speech that was judged to be more natural sounding than EL speech having constant F0, supporting the idea of using a simple linear relationship between amplitude and frequency to compute an F0 contour. It must be noted, however, that there are still other important acoustic factors for improving the quality of the EL speech, including deficient acoustic characteristics of the voicing source, its location provided by the EL transducers, and

modifications in vocal tract transfer functions. Further efforts will be made to examine the perceptual importance of correcting these factors in order to make EL speech more closely approximate the sound quality of normal speech.

6. Prosody

6.1 Production and perception of prosody

Research on the production and perception of prosody includes a) a series of studies developing and testing a method for transforming modal phonation to phonation with irregular pitch periods, and vice versa (Bohm and Shattuck-Hufnagel 2007, Bohm et al. 2008), and using this method to show that listeners make use of a speaker's habitual production of phrase-final irregular pitch periods to recognize the speaker's voice (Bohm and Shattuck-Hufnagel, under review); b) a series of studies of the alignment of F0 peaks and valleys associated with pitch accents in American English, showing that these turning points occur earlier in a syllable that also includes an intonational phrase boundary tone, due to tonal crowding (Shue et al. 2007, 2008), and c) a series of studies developing the idea that the alignment of F0 peaks, valleys and elbows is just one of a set of strategies that speakers use to accomplish a deeper goal, i.e. the realization of the bulk of the high or low pitch accent within the target syllable, providing a robust perceptual cue to contrasting accent types even in unvoiced regions where F0 turning points are missing (Barnes et al. 2006, Barnes et al. 2008, Brugos et al. 2008a).

We continue to develop and refine the ToBI method for prosodic transcription, adding a ALT tier for alternative transcriptions which improves transcriber agreement and satisfaction (Brugos et al. 2008b). This tier also captures significant information about the fact that most ambiguities are between just two possible transcriptions; by recording the distribution of these ambiguities, the new tier has the potential to increase our understanding of the ways in which intonational targets are realized phonetically by the speaker, and successfully recovered (or not) by the listener. Another line of work concerns the nature of intonational phrase-final lengthening: by examining the acoustic durations of phrase-final words with various stress patterns (e.g. Jamaica, Michigan) we determined that final lengthening affects not only the phrase-final syllable (e.g. -ca and -gan), but also the main-stress syllable (e.g. -mai- and Mich-). This main-stress-syllable lengthening occurs even when a non-lengthened syllable (e.g. -chi- in 'Michigan') intervenes between the two lengthening sites, suggesting two separate mechanisms for phrase final lengthening (Turk and Shattuck-Hufnagel 2007). Ongoing work tests the hypothesis that final-syllable lengthening, in contrast to main-stress-syllable lengthening, is a universal of planned movement, i.e. is motorically rather than structurally governed.

References

[1] J. Barnes, A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux, “Turning Points, Tonal Targets, and the English L- Phrase Accent”, paper presented at *Experimental and Theoretical Advances in Prosody*, Cornell University, April 2008.

[2] J. Barnes, S. Shattuck-Hufnagel, A, Brugos, and N. Veilleux, “The Domain of Realization of the L- Phrase Tone in American English”, *Proceedings of Speech Prosody 2006*, Dresden.

[3] T. Bohm, N. Audibert, S, Shattuck-Hufnagel, G.N. Nemeth, and V. Auberge, “Transforming Modal Voice into Irregular Voice by Amplitude Scaling of Individual Glottal Cycles”, *Journal of the Acoustical Society of America 123)*: p. 3886 (2008).

[4] T. Bohm, and S. Shattuck-Hufnagel, “Listeners Recognize Speakers’ Habitual Utterance-Final Voice Quality,” Presented at the Paralinguistic Speech – Between Models and Data Satellite Meeting, ICPhS 07, Saarbrucken, Germany, August 3, 2007.

[5] A. Brugos, J. Barnes, S. Shattuck-Hufnagel, and N. Veilleux, "(At Least) Two Members of the Rise-Fall-Rise Family", poster presented at *Experimental and Theoretical Advances in Prosody*, Cornell University, April 2008a.

[6] A. Brugos, N. Veilleux, M. Breen, and S. Shattuck-Hufnagel, "The Alternatives (Alt) Tier for ToBI: Advantages of Capturing Prosodic Ambiguity", Proceedings of *Speech Prosody 2008*, Brazil, June 2008b.

[7] Y.L. Shue, M. Iseli, N. Veilleux, and A. Alwan, "Pitch Accent Versus Lexical Stress: Quantifying Acoustic Measures Related to the Voice Source", Proceedings of the ICSLP, Antwerp, July 2007.

[8] Y.L. Shue, M. Iseli, S. Shattuck-Hufnagel, N. Veilleux, S.A. Jun, and A. Alwan, "Effects of Boundary Tones on Accent-Related F0 Peak Alignment", *Journal of the Acoustical Society of America 123*: p. 3460 (2008).

[9] A. Turk, and S. Shattuck-Hufnagel, "Multiple Targets of Phrase-Final Lengthening in American English Words," *J. Phonetics 35*: 445-472 (2007).

6.2 Tone distribution and its effect on subglottal pressure during speech

This research is part of a project to characterize the subglottal pressure (P_s) contour associated with a spoken utterance in terms of the distribution of pitch accents and of phrase and boundary tones. More specifically, we would like to know (1) how to place the transitions between the initiation phase, the working phase, and the termination phase, and (2) how to set the slope of the working phase, depending on the types of pitch accents and phrase and boundary tones that occur in an utterance. In this project, we focus on the working and termination phases. We attempt to relate (1) the transition point between working and termination phases, and (2) the slop of the working phase, to the distribution of pitch accents, phrase tones, and boundary tones.
It is found that the nuclear pitch accent does not define the start of the termination phase; the utterance offset is a better marker. Declination rate of the working phase and its relation to the phrase and boundary tones at utterance offset are found to vary among speakers. The results have implications for models of speech production, and for applications such as computer speech synthesis and recognition.

References

[1] H. Hanson, J. Slifka, S. Shattuck-Hufnagel, and J. Kobler, "Tone Distribution and Its Effect on Subglottal Pressure during Speech," *Proceedings 16th International Congress of Phonetic Science ICPhS-07,* Saarbrucken, Germany, 545-548, 2007.

Publications

Journal Articles, Published

X. Chi and M. Sonderegger, "Subglottal Coupling and Its Influence on Vowel Formants," *J. Acoustical Society of America*, *122*: 1735-1745 (2007).

S. Lulich, A. Bachrach, and N. Malyska, "A Role for the Second Subglottal Resonance in Lexical Access," *J. Acoustical Society of America*, *122*: 2320-2327 (2007).

A. Turk, and S. Shattuck-Hufnagel, "Multiple Targets of Phrase-Final Lengthening in American English Words," *J. Phonetics 35*: 445-472 (2007).

V. Villacorta, J.S. Perkell, and F.H. Guenther, "Sensorimotor Adaptation to Feedback Perturbations of Vowel Acoustics and its Relation to Perception," *J. Acoustical Society of America 122*: 2306-2319 (2007).

Journal Articles, Accepted for Publication

S.S. Ghosh, J.A. Tourville, and F.H. Guenther, "A Neuroimaging Study of Premotor Lateralization and Cerebellar Involvement in the Production of Phonemes and Syllables," Journal of Speech, Language, and Hearing Research (PubMed ID: 18664692).
K. Stevens and S. Keyser, "Quantal Theory, Enhancement and Overlap," *J. Phonetics*, forthcoming.

Journal Articles, Submitted for Publication

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

M.L. Matthies, F.H. Guenther, M. Denny, J.S. Perkell, E. Burton, J. Vick, H. Lane, M. Tiede, and M. Zandipour, "Perception and Production of /r/ Allophones Improve with Hearing from a Cochlear Implant," submitted to Journal of the Acoustical Society of America.

Book/Chapters in Books

K.N. Stevens, and H.M. Hanson, "Articulatory-Acoustic Relations as the Basis of Distinctive Contrasts", accepted for publication in *Handbook of Phonetic Sciences*, ed. W.J. Hardcastle, (Amsterdam: IOS Press).

Meeting Papers, Presented

J. Barnes, A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux, "Turning Points, Tonal Targets, and the English L- Phrase Accent", paper presented at *Experimental and Theoretical Advances in Prosody*, Cornell University, April 2008.

T. Bohm, N. Audibert, S, Shattuck-Hufnagel, G.N. Nemeth, and V. Auberge, "Transforming Modal Voice into Irregular Voice by Amplitude Scaling of Individual Glottal Cycles", *Journal of the Acoustical Society of America 123)*: p. 3886 (2008).

A. Brugos, J. Barnes, S. Shattuck-Hufnagel, and N. Veilleux, "(At Least) Two Members of the Rise-Fall-Rise Family", poster presented at *Experimental and Theoretical Advances in Prosody*, Cornell University, April 2008.

A. Brugos, N. Veilleux, M. Breen, and S. Shattuck-Hufnagel, "The Alternatives (Alt) Tier for ToBI: Advantages of Capturing Prosodic Ambiguity", Proceedings of *Speech Prosody 2008*, Brazil, June 2008.

H. Hanson, and S. Shattuck-Hufnagel, "Acoustic Cues to the Voicing Contrast in Coda Stops in the Speech of 2-year- olds Learning American English", *Journal of the Acoustical Society of America 123)*: p. 3320 (2008).

Y.L. Shue, M. Iseli, S. Shattuck-Hufnagel, N. Veilleux, S.A. Jun, and A. Alwan, "Effects of Boundary Tones on Accent-Related F0 Peak Alignment", *Journal of the Acoustical Society of America 123)*: p. 3460 (2008).

D.A. Robin, F.H. Guenther, S. Narayana , A. Jacks, J. Tourville, A.E. Ramage, J.L. Lancaster, C. Franklin, S. Ghosh, P.T. Fox, "A Transcranial Magnetic Stimulation Virtual Lesion Study of Speech," Proceedings of Conference on Speech Motor Control. Monterey, California, USA, 2008.

J.S. Perkell, "Movement Goals and Feedback and Feedforward Mechanisms in Speech Production," Invited talk at an International Workshop, *Is a neural theory of language possible? Development of Unified Representations in Natural and artificial Systems*, Centro di Ricerca Interdisciplinare sul Linguaggio and the CONTACT Project Learning and Development of Contextual Action, Lecce, Italy, June 28-30, 2007.

S. Shattuck-Hufnagel, K. Demuth, H. Hanson, and K. Stevens, "Acoustic Cues to Voicing Contrasts in Coda Stop Consonants in 2-3-year old Speakers of American English", presented at the Conference, *Where do Features Come From*, held in Sorbonne, Paris, October 2007.

Y.L. Shue, M. Iseli, N. Veilleux, and A. Alwan, "Pitch Accent Versus Lexical Stress: Quantifying Acoustic Measures Related to the Voice Source", Proceedings of the ICSLP, Antwerp, July 2007.

Meeting Papers, Published

S.S. Ghosh, M. Hamm, K. Jahns, C. Triantafyllou, "Using High Resolution fMRI to Identify Individual-Specific Speech Motor Regions," Proceedings of the XVIth Conference of the International Society for Magnetic Resonance in Imaging, 2008.

S. Wang, A. Alwan, and S.M. Lulich, "Speaker Normalization Based on Subglottal Resonances," *Proceedings International Conference on Acoustics, Speech, and Signal Processing ICASSP-08*, Las Vegas, Nevada, USA, 2008.

C. Park, N.F. Chen, and Y. Jung, "Determining and Interpreting Acoustic Landmark Sequences in American English," J. Acoustical Society of America 122: 2972 (2007).

J.S. Perkell, "Sensory Goals and Control Mechanisms for Phonemic Articulations," Invited paper for the special session, *Of Mouths, Ears, Eyes and Brains: The Sensory Motor Foundations of Spoken Language*, Proceedings of the XVIth International Congress of Phonetic Sciences. Saarbrücken, Germany: 16th ICPhS Organizing Committee, 2007.

M. Tiede, S. Shattuck-Hufnagel, B. Johnson, S. Ghosh, M. Matthies, M. Zandipour and J. Perkell, "Gestural Phasing in /kt/ Sequences Contrasting Within and Cross Word Contexts," Proceedings of the XVIth International Congress of Phonetic Sciences. Saarbrücken, Germany: 16th ICPhS Organizing Committee, 2007.

Meeting Papers, Submitted or under Review

S. Cai, M. Boucek, S.S. Ghosh, F.H. Guenther, and J.S. Perkell, "A System for Online Dynamic Perturbation of Formant Frequencies." To be presented at 8th International Seminar on Speech Production, Strasbourg, France, Dec. 8-12, 2008.

J.S. Perkell, S.S. Ghosh, F.H. Guenther, H, Lane, M.L. Matthies, L. Ménard, and M.K. Tiede, "Mechanisms of Vowel Production: Auditory Goals and Speaker Acuity," To be presented at 8th International Seminar on Speech Production, Strasbourg, France, Dec. 8-12, 2008.

Theses

X. Chi, *Word Boundary Detection Using Landmarks: A Survey of Consonants*, PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2008.

C. Park, *Consonant Landmark Detection for Speech Recognition,* PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2008.

S.Y. Zhao, *The Stop-Like Modifications of /ð/: A Study in the Analysis and Handling of Speech Variation*, PhD thesis, Graduate Program in Speech and Hearing Biosciences and Technology, Harvard-MIT Division of Health Sciences and Technology, MIT, 2007.