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, Dr. Jae Yung Song, Seth Hall.

Visiting Scientists and Research Affiliates
Dr. Jonathan Barnes, Department of Applied Linguistics, Boston University, Boston, Massachusetts.
Dr. Suzanne E. Boyce, Department of Communication Disorders, University of Cincinnati, Cincinnati, Ohio.
Dr. Jana Brunner, Humboldt University, Berlin, Germany.
Dr. Margaret Denny, CReSS LLC, Lexington, 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. 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, Department of Psychology, Washington University, St. Louis, Missouri.
Dr. Takahide Matsuoka,
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. Janet Slifka, vlingo Corporation, Cambridge, Massachusetts, and Department of Voice Surgery and Rehabilitation, Massachusetts General Hospital, 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. Majid Zandipour, BAE Systems, Burlington, Massachusetts.

Graduate Students
Shanqing Cai, Nancy Chen, Elisabeth Hon-Hunt, Youngsook Jung, Caroline Niziolek, Yoko Saikachi.

Undergraduate Students
Ben Park, Danielle Yuen, PeiLin Ren, and Yu Li

Technical and Support Staff
Arlene E. Wint

1. Constraints and Strategies in Speech Production

1.1 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 the 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 have conducted experiments with speakers and listeners in which we measured articulatory movements, speech acoustics, speech perception, and brain activation. We manipulated speaking condition, phonemic context and speech sound category and we introduced transient and sustained perturbations. We also performed modeling and simulation experiments, in which we adapted the vocal-tract model to the morphologies of individual speakers. We have tested properties of the neurocomputational model by using it to control the individualized vocal tract models in efforts to replicate those speakers' production data.

The project is in a year of unfunded extension, during which we have been completing several studies and conducting several additional ones.

1.2 Feedback perturbation of time-varying formant trajectories: Spectral modifications

Until very recently, almost all experimentation on the nature of phonemic goals in speech and the roles of feedback and feedforward control in achieving them has focused on the production of spectral aspects of steady-state sounds. However, many perceptually important aspects of the spectral structure of speech sounds and sound sequences are time-varying. To investigate time-varying aspects of speech motor programming, we have conducted a sensorimotor adaptation study that examined the responses of speakers of Mandarin Chinese to time-varying perturbations in the auditory feedback of their produced acoustic trajectories for Mandarin Chinese triphthongs (like diphthongs, but containing three vowel sounds instead of two). The triphthong /iau/ was chosen partly because of the large span of its trajectory in F1 x F2 space and its long duration, during which perturbations would presumably be more perceptually salient. Part

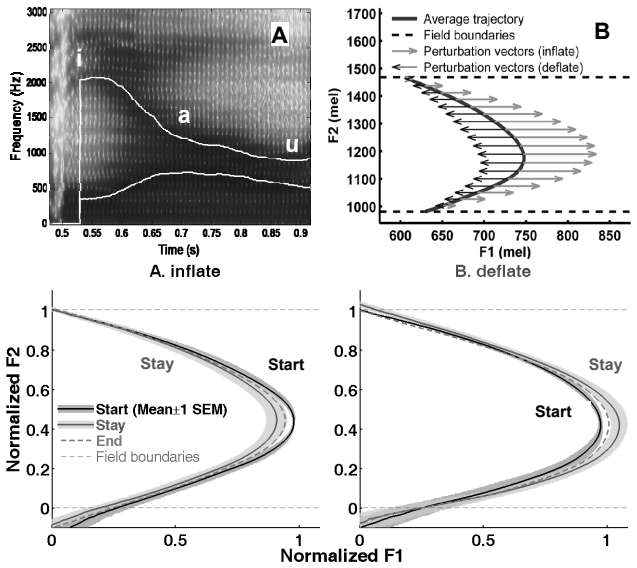

Figure 1. **A**: A spectrogram of a token of /iau/ with white tracks of F1 and F2 vs. time. **B**: A schematic diagram of such a trajectory (Average trajectory) in the F1x F2 plane. The sets of arrows illustrate "inflate" and "deflate" perturbation vector fields. **C**: Resulting trajectories, averaged across subjects, for responses to the inflate perturbation in panel C. **D**: Subject responses to the deflate perturbation.

A of Fig. 1 shows a spectrogram of a token of /iau/ with white tracks of F1 and F2 vs. time. Fig. 1B shows a schematic diagram of such a trajectory (Average trajectory) in the F1x F2 plane. The sets of arrows illustrate "inflate" and "deflate" perturbation vector fields, which had zero magnitudes at the beginning and end of the triphthong and maximum magnitudes at the point of greatest F1 excursion. Twelve subjects experienced inflate perturbations; eleven subjects, deflate perturbations. The experiment was divided into four phases: baseline (called start), ramp, full-perturbation (stay) and no-perturbation (end) – with a large number of repetitions in each phase. Resulting trajectories, averaged across subjects, are shown for responses to the inflate perturbation Fig. 1C and to the deflate perturbation in Fig. 1D. The average trajectories show compensation by both groups: decreased trajectory curvature in response to the inflate perturbation and increased curvature in response to the deflate perturbation. In both cases, the average end trajectories (no perturbation) fall between the start and stay trajectories, indicating some temporary modification of feedforward commands for producing the trajectories, as we have observed previously for steady-state vowels. Even though the differences between the averaged trajectories is relatively small, there were statistically significant effects of phase in both the deflate and inflate groups. The study also found modifications of motor programs for some related but different vowels, despite the fact that the subjects never experienced perturbations to the auditory feedback of those vowels. This finding indicates generalization of auditory-motor adaptation across different types of vowels. We are currently analyzing the detailed patterns of this motor generalization to reach a better understanding of the dynamic auditory-motor organization of the vowel space, which has not been studied extensively before.

1.3 Feedback perturbation of time-varying formant trajectories: temporal modifications

The aim of this pilot experiment was to demonstrate a method for testing the hypothesis that Persons Who Stutter (PWS) show over-reliance on auditory feedback for speech movement timing. We used PC-based real-time signal processing software (Cai et al., 2008) to parse sentences produced by the subject in real-time and perturb the auditory feedback of the diphthong /ai/ in disyllabic words. This perturbation altered the temporal profile of the trajectories in formant (F1 x F2) space without significantly altering the starting point or direction of the trajectory in that space. Specifically, it accelerated or decelerated the diphthong transition in the first syllable of each of three words: *bypassed*, *flytrap* and *pie-crust*. To ensure a context similar to natural speech, the words were embedded in short carrier sentences such as "Say flytrap clearly." Figure 2A shows an instance of the deceleration-type perturbation on the word "flytrap". The microphone was placed at a fixed distance of 10 cm from the subject's mouth. The perturbed auditory feedback was played back to the subject via a pair of insert ear phones with a 12-ms delay. The feedback level was approximately 12 dB louder than the microphone signal, in order to sufficiently mask bone-conducted feedback. Articulatory movement signals were recorded simultaneously with our ElectroMagnetic Midsagittal Articulometer (EMMA) system.

The subjects practiced the stimulus sentences several times before the recording. Then they were trained to produce the sentences at a moderate speaking rate of about 135 syllables/min through visual feedback of speaking rate. After the subject mastered the appropriate speaking rate, he/she was instructed to maintain this rate throughout the rest of the experiment. However, in order to avoid unnatural efforts to stabilize speech timing, feedback regarding speaking rate was no longer given during the recording part of the experiment. The recording consisted of 80 repetitions of each sentence, of which 20 repetitions (25%) were randomly selected and perturbed to decelerate the /a/-/i/ transition in the auditory feedback. In the rest of the trials, the subjects heard unperturbed auditory feedback.

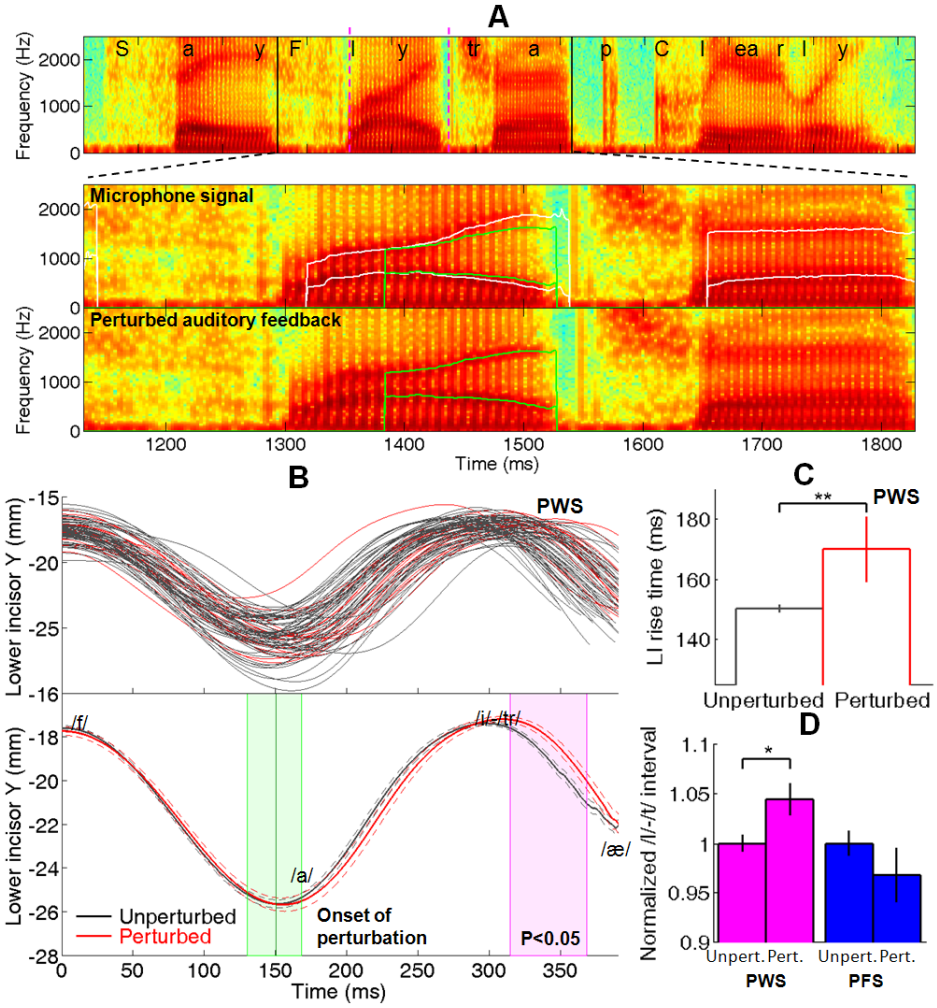

Figure 2. **A.** An example of an auditory perturbation which decelerates the /a/-to-/i/ transition in the diphthong /ai/ of the word "flytrap" embedded in the sentence "Say flytrap clearly" (top). The top and middle panel show the spectrogram of the unperturbed microphone recording. The white traces show the 1st and 2nd formant frequencies (F1 and F2) estimated online using LPC. The bottom panel shows the spectrogram of the perturbed auditory feedback. The green curves in the middle and bottom panels show the perturbed F1 and F2. A short delay (~12 ms) can be seen in the bottom panel. **B**. Top: EMMA recording of the vertical (y) position of the transducer attached to lower incisor (LI) in 60 unperturbed trials (black) and 20 perturbed trials (red) of a PWS. These traces are aligned at the y-position maxima corresponding to the labiodental contact for the fricative /f/. Bottom: average y-position of the LI in the perturbed and unperturbed trials. The dashed curves (mean ± 1 SEM). The green vertical lines and shaded region show the average onset time of the time-varying auditory perturbation (mean ± 1 SD). The magenta shaded region shows the time-interval in which the difference between the perturbed and unperturbed trials reached statistical significance (two-tail t-test, uncorrected P<0.05). **C**. Comparison of the rise times (defined as the time interval between the LI y-position minimum corresponding to the low vowel /a/ and the following LI y-position maximum corresponding to the alveolar stop /t/; error bars show ± 1 SEM) in the perturbed (red) and unperturbed trials. **: p<0.01. D. Normalized /l/-/t/ intervals in the word "flytrap" in perturbed and unperturbed trials: comparison between PWS and matched PFS. See text for the definition of /l/-/t/ interval. *: p<0.02.

Figure 2B shows the effect of a deceleration-type auditory perturbation on the syllabic timing of one PWS. The top part of Fig. 2B shows the vertical (y) position trajectories of the lower-incisor (LI) transducer during 60 unperturbed productions (black) and 20 perturbed productions (red) of the word *flytrap*. These trajectories were time-aligned at the local maximum of the LI y-position that corresponded to the labiodental contact in the fricative /f/. The bottom part of Fig. 2B shows the time average of these trajectories. The auditory perturbation began approximately at the local minimum of LI y-position corresponding to the low vowel /a/ (the green vertical line and shaded area). Significant differences between the perturbed and unperturbed traces were seen approximately 160 ms after the onset of the auditory perturbation (the magenta shaded area; two-tailed t-test; $P<0.05$). This delay is similar to results from other feedback perturbation studies. Of particular interest, the movement of the LI was delayed in the perturbed trials relative to the unperturbed ones for the PWS subject. Figure 2C shows that in the perturbed trials, it took significantly longer for the LI position to reach the local maximum in the alveolar stop /t/ after the local maximum in /a/ (perturbed: 170 ± 54 ms; unperturbed: 150 ± 2.3 ms; $P=0.0037$, two-tail t-test).

The response of the PWS to the deceleration-type auditory perturbation was compared to the response of an age-, sex- and handedness-matched Person with Fluent Speech (PFS) to the same type of auditory perturbation. This between-subject comparison was based on manually labeled landmarks corresponding to /l/ and /t/ in the word “flytrap”. The interval between the onset of the lateral /l/ and the alveolar stop /t/ was significantly longer in the perturbed trials than in the unperturbed ones in this PWS, whereas the matched PFS did not show a significant change in the /l/-/t/ interval (Fig. C.3.2.D). A two-way analysis of variance (ANOVA) indicated a significant {Fluency × Perturbation status} interaction ($P<0.05$). We hypothesize that due to the PWS’s over-reliance on auditory feedback for speech movement timing, the belated auditory feedback of the signal corresponding to the completion of the syllable /flai/ in the perturbed trials caused a later-than-normal initiation of the following syllable /træp/.

Although such findings await confirmation in more subjects, the observed difference is consistent with the hypothesis that a central characteristic of stuttering is overreliance on auditory feedback for speech motor control.

Together, the experiments described in Sections 1.1 and 1.2 demonstrate for the first time that time-varying aspects of speech movements are programmed very precisely and the feedforward component of this programming may be modified when temporal aspects of produced sounds do not match pre-programmed auditory expectations. Such results are compatible with a neurocomputational model in which motor goals for vowels consist partly of regions in multi-dimensional auditory-temporal space and speech is produced with a combination of feedforward and feedback control.

1.4 Pilot study of variability in speech articulatory kinematics

As a methodological basis for future studies, we have explored the use of two measures of speech movement variability: 1) The widely used spatio-temporal index (STI, Smith et al., 1995; Smith and Kleinow, 2000), a measure of the inter-trial variability of articulatory movements which corrects for the variations in speaking rate and overall movement amplitude, and 2) a new measure of variability of movement phasing across articulators that does not include a spatial component. To demonstrate the use of these two measures, we chose a set of movement signals from the pilot study of time-varying auditory perturbations described above (Section 1.2), in which we expected to see such variability differences between movements from unperturbed and perturbed utterances.

Figure 3A shows an example of the movements of the lower incisor (LI, red) and lower lip (LL, blue) during production of the word “flytrap”. From each of the 80 trials, we extracted a segment of the y-trajectory of LI. As the dashed vertical lines show, the segment began at the local maximum of LL y-position which corresponded to the labiodental /f/, and ended at another local

maximum of LL y-position corresponding to the final bilabial stop /p/. Figure 3B shows the extracted LI-y trajectory from the 60 unperturbed and 20 perturbed trials. It can be seen that on average, the onsets of the time-varying perturbations were near the minimum of the LI y-position which corresponded to the low vowel /a/ in the diphthong /ai/. As shown in Fig. 3C, these trajectories were time- and space-normalized. Based on these spatiotemporally normalized trajectories, we divided the normalized time axis into 1000 bins and quantified the inter-trial variability in each bin, in the perturbed and unperturbed sets, respectively. Figure 3D shows the result of this computation, indicating that the perturbed trials showed significantly greater spatiotemporal variability at two time intervals following the onset of the time-varying auditory perturbation. STI was computed as the arithmetic mean of the SDs in all 1000 time bins in the perturbed group and the unperturbed group, respectively. As Fig. 3E shows, the perturbed set exhibited a greater STI than the unperturbed set.

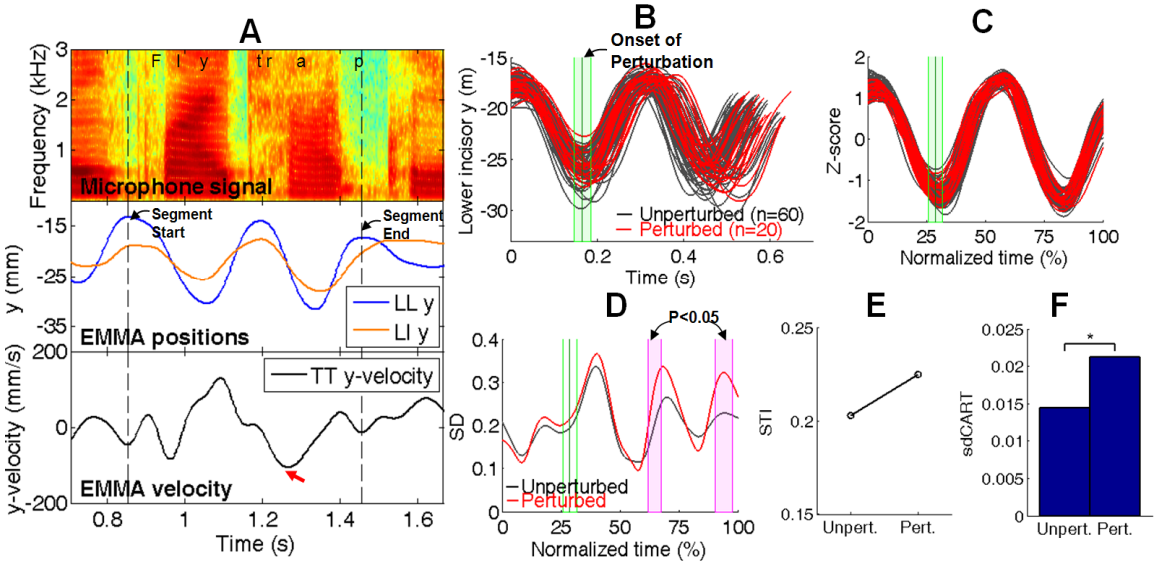

Figure 3. Analysis of spatiotemporal stability in productions of the word "flytrap" under time-varying auditory perturbation. **A.** Extraction of a segment of lower-incisor (LI) movement based on the local maxima in the y-position of lower lip (LL). The two vertical dashed lines show the start and end of a segment. The red arrow in the bottom plot points to the local minimum of the velocity of tongue-tip (TT) movement in the y dimension, which is used in the computation of cross-articulator relative timing (CART). **B.** The LI y-position trajectory segments extracted from 60 unperturbed trials (black) and 20 perturbed trials (red). The green vertical line and shaded region show the mean (±1 SD) onset time of the time-varying perturbation (See C.3.2 for details). **C.** The time- and space-normalized trajectories of LI movements (see text for details). **D.** SD of the normalized LI y-position (z-score) computed on a frame-by-frame basis in normalized time. The magenta shaded regions show time intervals in which the perturbed trials show significantly higher SD than the unperturbed trials (two-tailed F-test, $p<0.05$ uncorrected). **E.** Comparison of the STI from the unperturbed and perturbed trials. **F.** The SD of CART (sdCART) of the TT velocity minimum. * indicates significance at $p<0.02$ (two-tailed F-test).

STI is a measure that primarily captures the variability of the movements of a single articulator (the LI, or the mandible, in the example above). To quantify token-to-token variability in the coordination (relative timing) of the movements of several articulators, we devised another kinematic measure which we call *cross-articulator relative timing (CART)*. For this example, we identified the timing of the velocity minimum of tongue-tip (TT) movement in the vertical direction (the red arrow in the bottom plot of Fig. 3A) relative to the segment defined by two successive events in the movement of a different articulator (the LL, dashed vertical lines in Fig. 3A). This velocity minimum corresponds to the articulatory transition from the consonant /r/ to the low vowel /æ/. For each trial, CART is a number between 0 and 1, with a greater value corresponding to later occurrence in the segment. The SD of CART (sdCART) across multiple repetitions is a measure of the variability of inter-articulator timing. Figure 3F shows significantly greater sdCART of the TT velocity minimum in the perturbed trials than in the unperturbed ones (SD in perturbed trials: 0.0144; SD in perturbed trials: 0.0212; $P<0.02$, two-tailed F-test).

1.5 Pilot study of white-matter connectivity from structural, functional and diffusion weighted brain images

The aim of this study was to demonstrate that we can identify and quantify specific neuroanatomical markers based on current MRI scanning sequences. The studies were carried out on a Siemens 3-Tesla scanner with a 32-channel head coils.

Structural localization. We collected high-resolution (1 mm isotropic) T1-weighted structural images from a PWS. The structural images were processed using FreeSurfer software (Dale et al., 1999) to generate parametric-mesh representations of cortical surfaces. The cortical surfaces and the volume were then classified automatically into finer cortical and subcortical regions. FreeSurfer also automatically classified WM below cortical regions using a speech-oriented brain atlas developed in Dr. Guenther's lab (see Ghosh et al., 2008). Figures 4[B, C, and E show the automatic parcellations of the cortical surface and underlying WM into smaller regions of interest. Gyrification measures can be estimated from the cortical surfaces (Schaer et al., 2007); gray matter volume can be measured using these cortical regions (Han et al., 2006); and mean values of functional anisotropy can be extracted using the white matter parcellation. The cortical and WM parcellations are also used to determine connectivity with DTI data.

Functional localization. Using a sparse functional scan, we collected brain activity data from a PWS while he: 1) pressed a button with the right index finger, or 2) uttered the sound /m/, or 3) opened and closed the jaw or 4) listened to white noise stimuli. The aim was to localize sensorimotor regions that are often implicated in stuttering. Ten repetitions of each of these behavioral tasks were randomly interleaved between data acquisition periods in a sparse scanning paradigm. Data from a subset of brain regions (See Fig. 4A) were acquired using blood-oxygen level dependent (BOLD) echo planar imaging (EPI) scans and were processed using SPM (motion correction, co-registration, statistical analysis, contrast estimation) and FreeSurfer (cortical surface generation, smoothing, display of results). Figures 4D1-4 show the corresponding statistically significantly activated regions from comparing task related activity to a passive fixation task.

Tractography. Diffusion tensor imaging data (2 mm isotropic resolution, 60 directions) were used to

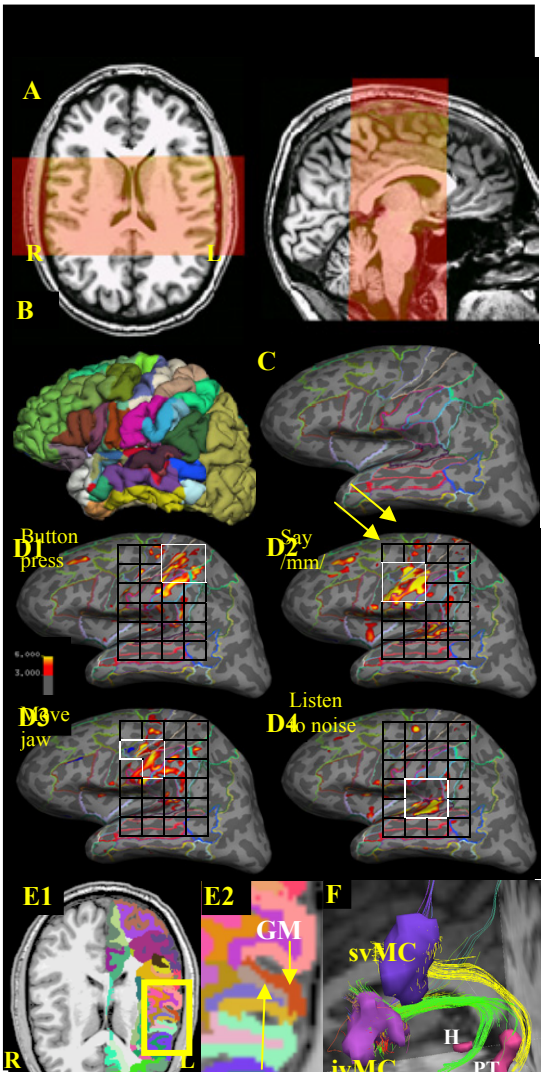

Figure 4 Determination of some neuroanatomical markers from one PWS. **A.** Extent of functional data acquisition overlaid on 1mm isotropic structural scans. **B.** Speech-specific parcellation of cortex on gray matter surface. **C.** Outline of parcellation on inflated brain surface. **D.** Statistical maps from function localizer overlaid on inflated brains surfaces. The grid is overlaid to facilitate comparison between different sources of activity due to the different tasks. (D1: right hand button press; D2: saying /m/; D3: jaw movement; D4: listening to white noise). **E.** Parcellation of structural image into gray matter (GM) regions (see B) and underlying white-matter (WM) regions. E2. Blow-up of boxed region corresponding to left sensorimotor cortex from E1. **F.** Diffusion tracts using the left ventral motor cortex (vMV) activation (see arrows in D2) as seeds. Green tracts from the inferior vMC activity connect to Heschl's gyrus (H) and planum temporale (PT) while the yellow tracts from the superior vMC activity connect to PT and superior temporal gyrus (STG).

estimate tracts and the mean FA values underlying cortical regions. The diffusion data were processed with FSL (eddy current correction, probabilistic tractography) and the Diffusion Toolkit (tractography) and visualized with TrackVis (Wang and Weeden, 2007). The locations of the peaks of left hemisphere motor cortical activity from the utterance /m/ (see arrows in Fig. 4D2) were used as seeds to estimate the two tracts shown in Fig. 4[F. The WM parcellation shown in Fig. 4E was used to estimate mean FA values shown in Fig. 5 These FA values are within the range reported in Chang et al. (in press); the lowest FA value was found in the WM underlying left ventral premotor cortex, which is consistent with observations from Watkins et al. (2008).

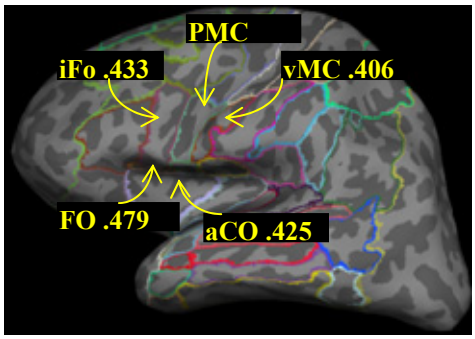

Figure 5. Mean FA values of WM underlying speech regions around motor cortex from a PWS. Premotor cortex (PMC) shows the lowest FA value. (vMC – ventral motor cortex, aCO – anterior central operculum, FO – frontal operculum, iFO – inferior frontal gyrus *pars obicularis*.

The volumetric measurements from GM regions such as planum temporale, gyrification measures from the peri-Sylvian regions, mean FA and strength of connectivity measures from the DTI imaging and the localization of activity from the functional scans will enable us to systematically quantify structural neuroanatomical characteristics of PWS and PFS.

1. 5 The influence of perceptual categories on auditory feedback control during speech

Speakers use auditory feedback to monitor their own speech, ensuring that the observed output matches the intended output. By altering the speech feedback signal before it reaches the ear, we can induce perceived errors and observe both the acoustic and neural consequences. This study investigates the neural mechanisms responsible for the detection and consequent correction of auditory errors, as well as the influence of phonetic categories on this auditory feedback control. Our aim is to compare the neural activation due to a phonetic change — for example, a change from the word “bed” to the word “bad” — with the activation due to a non-linguistic auditory change — for example, a change from a prototypical example of “bed” to an altered version of the same word.

For each subject, vowel production data were collected for six English vowels. Perceptual boundaries between pairs of vowels were assayed using a word categorization task with formant-shifted stimuli. Subjects assigned phonetic labels to vowel tokens generated from their own productions, shifted in the first and second formant frequencies (F1 and F2). Functional magnetic resonance imaging was then employed to measure neural responses to subjects’ speech with and without auditory perturbation. During each trial, subjects were presented with a word to read aloud or with a control stimulus. On one out of every four trials, F1 and F2 were perturbed in real time in one of two directions, one which resulted in a perceptual category change and one which did not. This altered speech was fed back to subjects through headphones, creating a sudden, unexpected mismatch between the vowel target and the perceived realization.

Psychophysical results showed a compensatory shift of the first two formants during the perturbed conditions. fMRI data for two subjects showed increased activation of bilateral inferior frontal gyrus, superior temporal gyrus, and supplementary motor areas for across-category shifts compared with within-category shifts. In accordance with previous results, shifted conditions showed more cortical activation in superior temporal gyrus and right inferior frontal gyrus than unshifted conditions. In general, cortical activation was greater in extent for shifts that crossed a category boundary than for those that did not, even though these shifts were of the same magnitude.

Vowel category boundaries, as assayed by a perceptual categorization test, are often asymmetric within subjects. This asymmetry enables a direct contrast between across- and within-category perturbations in a single subject. That is, a constant shift magnitude can elicit different phonetic percepts, depending on the direction of the shift and the location of category boundaries in formant space. A within-category shift was found to activate bilateral superior temporal gyrus and right inferior frontal gyrus, while a cross-category shift evoked greater activation in superior temporal gyrus and inferior frontal gyrus bilaterally.

1.6 Evidence of an articulatory saturation effect in the production of 's.'

Previous work [e.g., JSLHR, 47, 1259-1269, 2004] demonstrated that speakers of American English who consistently activated a contact sensor on the lower alveolar ridge with the tongue tip during 's' but not 'sh' tended to produce greater sibilant contrast than speakers who didn't consistently show this contact difference. This study used data from articulatory movements (EMMA), from measures of contact of the tongue tip with the alveolar ridge and from analysis of the acoustic signal to investigate a hypothesized articulatory saturation effect: abrupt articulatory and acoustic changes result from continuous forward movement of the tongue, which does not have to be controlled precisely in stopping when contact occurs. In tentative support of this hypothesis, preliminary observations from productions of continuous 'sh'-to-'s' transitions show movements of tongue-blade points which slow but do not stop immediately following articulatory contact and/or a rapid rise in sibilant spectral mean.

1.7 Development of facilities

Our software for data acquisition, extraction and analysis was enhanced in several ways: These enhancements included:1) modification of our EMMA control software to support real-time tracking and presentation of subject jaw opening amplitude for use in experiments with manipulated feedback, extended support; 2) provision of support for F0 and formant measurement–values are now tracked continuously through each token of interest using two distinct approaches for F0 and multiple LPC orders for formants; 3) development of an interactive tool for reviewing results of algorithmic data extraction in problematic cases; and 4) development of several approaches for acoustic analysis of sibilant consonants to facilitate comparison of produced acoustics with corresponding articulatory movements in a continuing study of saturation effects in sibilant production.

2. Effects of Hearing Status on Adult Speech Production

This project aims to advance knowledge of the roles of hearing in speech and to evaluate new hypotheses based on a model of those roles. We are using acoustic recordings of speech, perceptual tests, and several types of intervention in experiments with individuals with normal hearing and individuals who have postlingual deafness and receive cochlear implants. According to our model, the goals of speech movements are in sensory domains. Motor commands to achieve auditory goals are determined by the combined operation of feedback and feedforward control mechanisms. Feedforward control is almost entirely responsible for generating articulatory movements in mature speech production. However, if auditory feedback detects a mismatch between auditory goals and the heard consequences of ongoing speech movements, the detected error leads to the generation of feedback-based corrective motor commands, which in turn serve to update feedforward commands for subsequent movements.

In our experiments, the usefulness of a speaker's auditory feedback is reflected in that speaker's acuity for the speech parameters under study, and the size and spacing of the speaker's phonemic goal regions are indexed by measures of the speaker's produced phonemic contrasts. The relative contributions of feedback and feedforward control are assessed by blocking and unblocking feedback, and by introducing interventions and measuring speakers' compensatory responses. The main objective of the proposed research, then, is to extend our understanding of

relations among speakers' auditory acuity, the phonemic contrasts they produce, and the roles of feedback and feedforward control. To reach this objective we are engaged in a series of experiments involving: 1) *Auditory acuity and produced phonemic contrast,* 2) *Auditory acuity and produced lexical stress,* 3) *Digital perturbation of vowel spectra,* 4) *Mechanical perturbation of sibilant spectra* and 5) *Vowel imitation and auditory feedback*. The resulting comprehensive picture should provide significant new insights into the roles of speaker acuity in feedback and feedforward control of speech motor planning to achieve auditory goals.

We have run a complete set of procedures on 9 pilot cochlear implant users and 6 pilot speakers with normal hearing. These studies include a battery of perceptual tests and experiments on: 1) Auditory acuity and produced phonemic contrast, 2) Produced lexical stress, 3) Digital perturbation of vowel spectra, and 4) Vowel imitation and auditory feedback. Enough data have been gathered, processed and analyzed in some of these pilot studies to report selected preliminary results, as described below.

2.1. Perceptual testing

Perceptual testing comprises measures of residual hearing in the unimplanted ear of implant users, perception of vowels and consonants in CVC syllables and perception of monosyllabic words.

Preliminary finding: Two cochlear-implant users and six participants with normal-hearing were tested with both vowel and consonant perception tasks using pre-recorded CVC stimuli spoken by one male and one female. The 8 vowel stimuli in /pVt/ context were "pat, pet, pete, pit, poot, pot, put, putt"; the 11 consonants in /Cɑt/ context were "bot, cot, dot, got, lot, pot, rot, shot, sot, tot, zot." Listeners heard 9 repetitions of each syllable at a comfortable level and indicated the word that they heard by choosing from a closed set list of options displayed on a computer monitor. The male and female stimulus sets were tested separately for a total of 4 tasks. Listener responses were organized into confusion matrices and analyzed with a MATLAB-based implementation of the sequential information analysis (SINFA) procedure. Two sets of input features were tested; one set included vowel height, vowel position in the front–back dimension, and the tense–lax distinction and consonant features included place, manner, and voicing. A second set provided a larger number of features which included voicing, place, continuant, sonorant and lateral descriptors for consonants as well as height, front/back, rounding and tense/lax descriptors for vowels.

As expected, the cochlear implant users had more difficulty with the task; overall they scored 43.2% (sd=11.4) on the consonants and 63.2% (sd=8.8) on the vowels. Listeners with normal hearing scored 98.1% (sd=2.3) and 97.1% (sd=4.1) on the overall consonant and vowel perceptual tasks respectively. The stimuli produced by the female speaker were somewhat more difficult for all listeners. The compact set of features generally accounted for more transmitted information than the larger set. For both vowels and consonants, all three input features contributed to the total transmitted information. Further analyses are planned to optimize the feature sets.

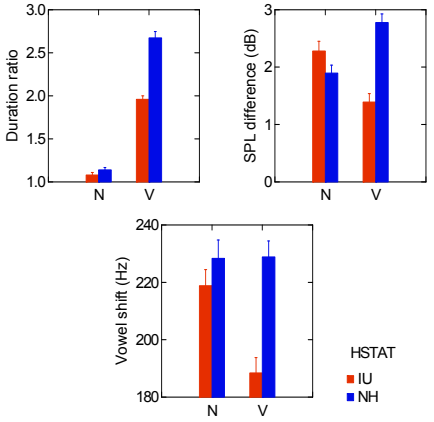

Figure 6. Duration ratio, SPL difference and vowel shift (Hz) as a function of word type (Noun, Verb) and Hearing Status (Implant User, Normal Hearing (NH).

2.2. Production of lexical stress

Prosodic variables such as stress may rely more on closed-loop feedback control than segmental productions since these variables change more slowly with time and thus are not as negatively impacted by the delays inherent in auditory feedback. To investigate the control of lexical stress, we are measuring, in hearing controls and implant users, the extent of the speakers' use of four acoustic parameters in producing lexical stress contrasts, each one normalized by its value in the unstressed syllable. Subjects are prompted to produce repetitions of words differing by the location of lexical stress (e.g., exTRACT-EXtract). Each is embedded in a short sentence that provides meaningful context.

Preliminary findings: As shown in Fig. 6, data from five implant users (IU) and three normal-hearing (NH) subjects allow the following observations. With duration measures, the contrast between stressed and unstressed syllables is greater in the verbs than the nouns for both groups, but the contrast is less for the implant users. Normal-hearing speakers use vowel quality (measured as locus in the formant plane) to signal stress for both nouns and verbs; implant users do the same but to a lesser degree. Finally, both groups signal stress with sound level but implant users less so in the case of the verbs. In brief, both groups appear to use the same cues to lexical stress but the implant users' contrasts are less marked. Fundamental frequency data are being extracted. Such data will be used to compare the production of lexical stress contrasts by normal-hearing speakers and hearing-impaired speakers and to track changes that occur with experience with a cochlear implant.

2.3 Mapping the vowel perceptual space

We have continued to refine an interactive procedure for eliciting subject judgments of vowel quality, in order to map their vowel spaces. Following pilot testing to determine the most effective procedure, we have now finalized a task that combines both a search through the F1xF2 vowel plane for a given target (e.g. "heed") with subject rating information on a five point scale. The results of these tests will be used to determine: the degree of correspondence (mapping) between speakers' production and perceptual vowel spaces, differences in this mapping between hearing-impaired and normal-hearing speakers, and changes in the mapping as a function of experience with a cochlear implant.

2.4. Auditory Feedback and Vowel Imitation

This study was designed to investigate the roles of feedforward and feedback control in speech motor planning by using a mimicry paradigm that makes it possible to evaluate the relative influences of those two subsystems. According to our theoretical framework, the effects of feedforward commands can be seen in speakers' "canonical" productions of speech sounds – viz., sustained vowels spoken in isolation. A mimicry paradigm that gives the speaker a synthetic vowel target at some distance from his or her canonical form, is predicted to create competition between feedforward commands that would generate the canonical form and auditory feedback that would yield a spectral match to the presented target. The effects of this competition can be seen in the speaker's imitation, which lies somewhere between the speaker's own canonical form and accurate reproduction of the target.

We have conducted a vowel mimicry experiment based on a previous study by Viechnicki (2002). We have elaborated Viechnicki's paradigm by: 1) using two groups of subjects, cochlear implant users and matched controls with normal hearing, 2) using two auditory feedback conditions, *normal* and *blocked,* to help separate the effects of feedback and feedforward control and 3) including measures of speaker auditory acuity.

Figure 7 shows hypothetical vowel imitations produced by a high- and a low-acuity subject. Synthesized targets correspond to the tick marks on the dotted lines. A single canonical vowel is indicated by the large filled circle in each panel. Hypothetical mean values of the speakers' imitations are indicated by small filled circles, with dispersions shown by dashed ellipses. The numbers indicate the order of the synthesized targets along the continuum, and the thin solid lines connect the imitation means with their targets.

Methods. Each subject made a preliminary recording of 10 "clear" pronunciations of each of three target words, *hid, head* and *had*. The resulting formant values of these canonical vowels were used to anchor the subject's synthetic vowel continuum, which consisted of algorithmically-generated /hVd/ stimuli. The stimuli contained vowel nuclei 250 ms long and F1 and F2 values that were at eight evenly spaced intervals between the canonical /ɪ/ and /ɛ/ and between the canonical /ɛ/ and /ʌ/ (plus two stimuli that extended beyond each end of the continuum to assist in delineating the prototypes of each category).

For the imitation task, the subject was presented aurally with a target stimulus, under instructions to imitate as accurately as possible. The subject's productions were recorded and vowel formants extracted. The subjects experienced a set of trials without hearing blocked for the imitations (20 repetitions of each stimulus, random order), then a set with hearing blocked for the imitations by masking noise (10 repetitions, random order).

Speakers' acuity was tested by measuring their JNDs (just-noticeable differences) for the synthetic continua.

We used the following formula to predict F1 and F2 formant values of speakers' imitation responses:
$F_i = k_o + k_1T + k_2C$, where T is the formant value of the imitation target and C is the formant value from the speaker's canonical vowel (prolonged pronunciation without phonetic context).

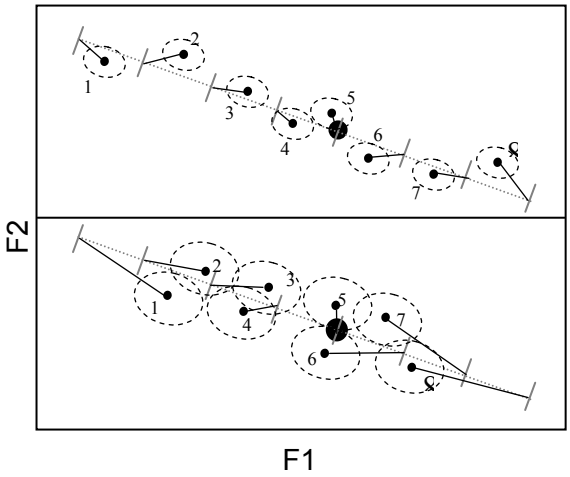

Figure 7. See text for details

Hypotheses and Preliminary Results. Thus far, the following hypotheses have been tested across nine speakers with normal hearing.

Hypothesis 1: Imitation formants can be predicted from target and canonical vowel formants. Canonical formants will have more weight without feedback. *Results:* Imitation formants were predicted reliably, especially F2. Canonical vowels received substantial weight but not more weight without feedback.

To account for overall shifts of imitation responses with respect to the targets, the remaining hypotheses were tested after shifting each speaker's responses by the ratios of mean F1 and F2 of his or her imitations and targets in mel space.

Hypothesis 2: Without auditory feedback, imitation responses will be significantly further from the target. *Result:* This hypothesis was supported.

Hypothesis 3: Imitation responses from high acuity (small JND) speakers will be closer to imitation targets than those from low acuity speakers. *Result:* There was a significant correlation ($p < .05$; 1-tailed) of speaker acuity and imitation accuracy for the *hid-head* continuum, in the with-feedback condition. (Data are currently unavailable for the *head-had* continuum.)

Hypothesis 4: With auditory feedback unavailable, imitation responses will be nearer canonical values. *Result:* This hypothesis, currently tested only for the *hid-head* continuum, was not supported.

Hypothesis 5: Without feedback, imitation responses will cluster more around canonical values. *Result:* There was a non-significant trend for more clustering without feedback.

3. Acoustic Articulatory Evidence for Quantal Vowel Categories: The Features [low] and [back]

In recent years, research in human speech communication has suggested that the inventory of sound units that are observed in vowels across languages is strongly influenced by the acoustic properties of the human subglottal system (Stevens, 1998; Chi and Sonderegger, 2004, 2007; Lulich et al., 2007; Lulich 2009). That is, there is a discrete or quantal set of possible vowel features that are constrained by the interaction of the acoustic/articulatory properties of the vowels and a small set of attributes that are observed in the subglottal region below the vocal tract. The hypothesis that subglottal resonances govern feature boundaries was tested, by exploring the relations among subglottal resonances, vowel formants and features for three populations: adult speakers of American English; adult speakers of Korean; and children learning American English.

3.1 The first subglottal resonance (SubF1) and the vowel feature [low] for adult speakers of American English

Previous research has suggested that the second subglottal resonance SubF2 may play a role in defining F2 boundary between [+back] and [-back] vowels in American English (Stevens, 1998; Chi and Sonderegger, 2004, 2007; Lulich et al., 2007; Lulich 2009). This observation raised a possibility that the first subglottal resonance SubF1 can play a role in defining another feature contrast for vowels. That is the feature [low]. This possibility was tested using the database of Chi and Sonderegger (2004, 2007) which was originally recorded and collected for study of SubF2 effects on speech. The relationship between first formant frequencies (F1) of vowels and the first subglottal resonances (SubF1) was explored for both diphthongs and monophthongs. The values of SubF1 for three adult male and six adult female speakers were in the range of 540 – 690 Hz. For English adult speakers producing the diphthong [ai], the time courses of F1 frequency and amplitude were obtained, using a window of one pitch period length. The F1 peak for vowels showed irregularity such as frequency discontinuity and amplitude attenuation as it passed through the frequency of the speaker's SubF1. The amplitude attenuation was found for all subjects in the range of 3 – 7 dB. The frequency discontinuity was often found depending on speakers, in the range of 220 – 300 Hz. The location of frequency discontinuities near SubF1 was 40 Hz from the value of SubF1 obtained by an accelerometer, providing a useful estimate of the individual speaker's SubF1. From monophthong vowels, analysis of *F*1 frequency distributions shows a boundary between [+low] and [-low] vowels at the speakers' SubF1: 99% of tokens for low vowels were above the speaker's SubF1, while 96% of tokens for non-low vowels were below the speaker's SubF1.

From a variety of reports in the literature, the first formant frequencies of vowels were collected from 10 languages in the world for adult male speakers and for adult female speakers, separately. The collected data of F1 were divided into two groups: a [+low] vowel group and a [–low] vowel group. The result shows that the boundary between these two groups agreed with the average value of SubF1 from the laboratory study with English, and the reported values in the literature. There were dips in the *F*1 distributions near the average value of *SubF*1, which were 600 Hz for males and 700 Hz for female speakers.

3.2 The second subglottal resonance (SubF2) and the vowel feature [back] for adult speakers of Korean

Previous research has suggested that the second subglottal resonance may play a role in defining the F2 boundary between [+back] and [-back] vowels in American English (Stevens, 1998; Chi and Sonderegger, 2004, 2007; Lulich et al., 2007; Lulich 2009). This raised a possibility that the second subglottal resonance can play a role in defining the boundary of the feature [back] in other languages as well. Subglottal coupling effects on speech are expected to be universal and independent of language because the coupling arises from physical properties of human speech production system. However, few acoustic studies have been carried out in other languages far from English (Madsack et al, 2008; Wang at al. 2008). We studied the relationship between F2 of Korean vowels and SubF2, to test whether the relation between this subglottal resonance and the boundary for a different set of feature contrasts (i.e. [back] vs. [front] vowels), demonstrated earlier for American English, might also be observed for this cross-linguistic study.

Speech signals and subglottal signals were obtained from four adult male and four adult female speakers of Korean, including Korean monophthong and diphthong vowels. To obtain subglottal signals, an accelerometer was attached to the neck below the larynx of a speaker. Nine Korean monophthong vowels were produced in the context of /hVd/. The result of diphthong analysis showed that there were frequency discontinuities and amplitude attenuations in the trajectory of F2 near the SubF2 values of speakers, as in English. From monophthong analysis, results showed that the F2 boundary between [back] vowels and [front] vowels was placed near SubF2 in Korean adult speech, as in American English. For central vowels, F2 values were placed either side of SubF2, but F2 values were always avoided SubF2 values.

Especially for the vowel /a/ which has no contrast of the feature [back], similarities and differences in implementing the backness of vowels in relation to different vowel inventories were further examined, in the different environments of adjacent consonants. From several adult speakers of Korean producing /a/ in context of /CaC/, speech signals and subglottal signals were simultaneously obtained. The result showed that F2 values of /a/ were always avoided the speaker's SubF2. The direction of F2 shift was related to the SubF2 values depending on adjacent consonants. If the adjacent consonants were labial or velar, F2 of the low vowel was below SubF2, whereas if the consonants were alveolar, F2 of the vowel was above SubF2. The amount of F2 shift due to adjacent consonants was different for each individual, in the range of 100 – 260 Hz. These findings suggest that SubF2 may have a role in assimilation of the feature [back] in Korean, as opposed to leading to a contrast for the feature in English. Similar observations are expected for the allophone in Mandarin or Japanese.

3.3 The development of F1 and F2 frequencies for young children

Previous research and the results of study for the feature [low] have suggested that the first and second subglottal resonances may play a role in defining the boundaries of the features [back] and [low] for adult speakers of American English (Stevens, 1998; Chi and Sonderegger, 2004, 2007; Lulich et al., 2007; Jung and Stevens, 2007; Lulich 2009). This raised a question of when adult-like relations among these vowel features, vowel formant values and subglottal resonances begin to appear for children.

We explored the development of the placement of vowel formants in relation to subglottal resonances for 10 children in the age range of 2;6–3;9 years using the database previously collected for other purposes by Imbrie (2005). To determine when adult-like relations among vowel features, vowel formants and subglottal resonances emerge, for the ten children, the first and second subglottal resonances were estimated from speech data. The mean values of *SubF*1 were in the range of 900 – 1100 Hz, while those of *SubF*2 were between 2200 – 2700 Hz. In this time period, the first and second subglottal resonances decreased with age. The subglottal resonances were different for each individual. Results of F1 and F2 measurements during the time period showed that at the earlier ages, the F1 and F2 values deviated from the expected

relations for some children (6/10 children), but during the six month period in which the measurements were made, there was considerable movement toward the expected values of F1 and F2 in relation to the subglottal resonances. For children who already had achieved an adult-like pattern at the earlier ages, little further changes occurred in F1 and F2 frequencies during the time period. Furthermore, we tested two opposing hypotheses, the quantal hypothesis and the vocal tract growth hypothesis, as accounts of how children's productions changed over time. The result showed that this developmental pattern of F1 and F2 values was inconsistent with an account in terms of simple anatomical increase in the size of the child's vocal tract. The significant changes in F1 or F2 during the 6 month period could be explained by the quantal hypothesis. The transition to the expected relations appeared to occur by the age of 3 years for most of these children. Although an age-related development pattern was found, the speed of development was different for each individual.

Conclusion and discussion

We hypothesized that the first two subglottal resonances may play a role in defining the boundaries for the vowel features [low] and [back]. We tested the hypothesis on three different populations: adult speakers of American English, adult speakers of Korean, and young children learning American English. These three sets of observations provide evidence that subglottal resonances play a role in defining vowel feature boundaries, as predicted by Stevens' (1972) hypothesis that contrastive phonological features in human languages have arisen from quantal discontinuities in articulatory-acoustic space. This research focused on vowel features [low] and [back]. For other features, theses kind of relationships between physical properties of the speech production system and phonological contrasts should be further explored in future. Also, the mechanism of how formant frequencies of young children change according to their subglottal resonances requires further study.

References

Chi, X., and Sonderegger, M. (2004) Subglottal coupling and vowel space, J. Acoust. Soc. Am. 115:2540–2540.

Chi, X., and Sonderegger,M. (2007) Subglottal coupling and its influence on vowel formants. J. Acoust. Soc. Am. 122:1735-1745.

Imbrie, A. (2005) Acoustical Study of Stop Consonants in Children, PhD. Thesis, MIT.

Jung, Y., Stevens, K. N. (2007) Acoustic articulatory evidence for quantal vowel categories: The feature [low]. J. Acoust. Soc. Am. 122:3029.

Lulich, S. M., Bachrach, A., and Malyska, N. (2007) A role for the second subglottal resonance in lexical access. J. Acoust. Soc. Am. 122:2320-2327.

Lulich, S. M. (2009) Subglottal resonances and distinctive features. J Phonetics.

Madsack A., S. M. Lulich, W. Wokurek and G. Dogil. (2008) Subglottal resonances and vowel formant variability: A case study of High German monophthongs and Swabian diphthongs. In Proceedings of LabPhon11, 91-92.

Stevens, K.N. (1972) The Quantal nature of speech; Evidence from articulatory-acoustic data. In P.B. Denes and E.E. David Jr. (Eds.) Human Communication: A Unified View, New York: McGraw-Hill, pp 51-66.

Stevens, K. N. (1998) *Acoustic Phonetics*. MIT Press, Cambridge, Massachusetts.

Wang, S., Lulich, S. M., and A. Alwan (2008) A reliable technique for detecting the second subglottal resonance and its use in cross-language speaker adaptation. In Proc. Interspeech.

4. Acoustic Characterization of the Glides /j/ and /w/ in American English Theory and Enhancement

This study was conducted to identify the acoustic characteristics that differentiate the glides /j,w/ from adjacent vowels in human speech production and perception. Previous acoustic studies comparing glides to their cognate high vowels /i,u/ have largely been limited to measurements of formant frequencies and their rates of change (e.g., Lehiste & Peterson, 1961; Maddieson & Emmorey, 1985; Chitoran, 2002). Given the preponderance of mainly durational data, it has been assumed by many phonologists (e.g., Selkirk, 1984; Catford, 1988) that glides are featurally identical to high vowels, although shorter and in non-syllabic positions. This study readdresses the distinction between glides and high vowels by focusing on acoustic cues to the articulatory targets of glides, which are reached at specific moments in time and need not be related to durational parameters. In particular, this study considers the possibility that the glottal voicing source, previously assumed to be independent of the vocal tract filter, may be affected by the vocal tract constriction in glides, and that these effects may contribute to the defining distinction between the articulator-free features of glides and vowels.

4.1. Acoustic analyses

Acoustic analyses were performed on a recorded database of intervocalic glides, produced naturally by two male and two female speakers in controlled vocalic and prosodic contexts. Each utterance was a vowel-glide-vowel (/VGV/) nonsense sequence in which one of the glides /j,w/ was flanked by one of six American English tense vowels /i,u,o,e,æ,ɑ/. The /VGV/ tokens were embedded in carrier phrases constructed in such a way as to vary the location and type of pitch accent on the flanking vowels. For each of 240 tokens, pitch-synchronous DFT analysis was carried out using analysis windows set to the length of each individual pitch period. Measurements of RMS amplitude (A_{RMS}), fundamental frequency (F0), and first formant frequency (F1) were made from each pitch period from the first vowel through the glide to the second vowel. In addition, measurements of open quotient (OQ) were made at the glide and vowel landmarks, computed as the difference in amplitude of the first two harmonics of the source spectrum, following inverse filtering of the speech signal to remove the effects of the first formant. Measurements of harmonics-to-noise ratio (HNR) were also made at the glide and vowel landmarks, using a pitch-scaled harmonic filter algorithm (Mehta, 2006).

Glides were found to differ significantly from adjacent high vowels /i,u/ in terms of the acoustic cues signaling their articulatory target configurations. Glide landmarks showed significantly reduced A_{RMS}, reduced HNR, increased OQ, and often decreased F0 compared to adjacent high vowel landmarks, with relatively little reduction in F1. The combined acoustic data suggest that glides differ from their cognate high vowels in that the glides are produced with a greater degree of constriction in the vocal tract. The relatively small reduction in F1 in the face of this increased constriction narrowing can be explained by the F1 saturation effect conditioned by the yielding vocal tract walls (Fant, 1972; Stevens, 1998). The narrower constriction does, however, cause an increase in oral pressure, which produces aerodynamic effects on the glottal voicing source, as described by Bickley & Stevens (1986) for artificial constrictions (see Figure 1). This interaction between the vocal tract filter and its excitation source results in skewing of the glottal waveform (see Figure 2), notably increasing the open quotient and fundamental period and decreasing the amplitude of voicing in glides. The voicing amplitude reduction produces a concomitant reduction in HNR, although isolation of the turbulence noise component does not indicate any significant increase in frication.

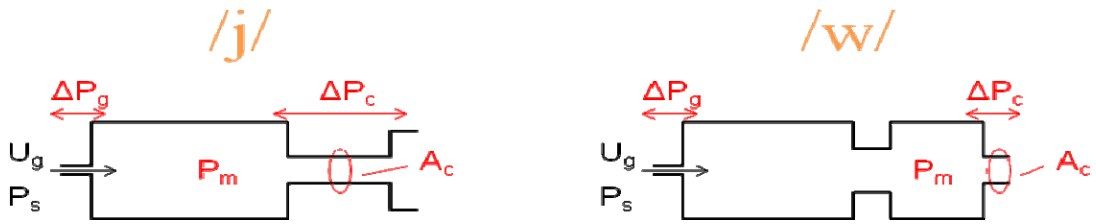

Fig. 1: Vocal tract tube approximations for the glides /j/ (left) and /w/ (right), with the glottal volume velocity source U_g entering the tube filter on the left end, and the mouth opening on the right end. When the vocal tract constriction is narrowed such that its cross-sectional area A_c is smaller than that of a high vowel, a pressure drop ΔP_c may form across the constriction, with a corresponding oral pressure build-up P_m. Since the sum of the pressure drops across the entire vocal tract length must equal the static subglottal pressure P_s, the build-up of oral pressure is balanced by a decrease in the transglottal pressure drop ΔP_g. This decrease in transglottal pressure, coupled with the increase in oral pressure, has weakening and skewing effects on the glottal source waveform.

In glides in high vowel contexts, the relative lack of F1 movement indicates that the major factor causing significant reduction in amplitude is the vocal tract constriction's effect on the glottal source; this effect is directly confirmed by the significant increase in open quotient for all speakers. In glides in other vowel height contexts, larger movements in F1 contribute more to the amplitude reduction; however, the source and filter contributions balance such that the amount of amplitude reduction in glides is relatively invariant to differences in the height of the surrounding vowels. Statistical analysis showed almost no significant effects of surrounding vowel height on ΔA_{RMS} across speakers in this database; the average ΔA_{RMS} for all tokens was 14.6 dB. Some effects of prosodic context were also observed, offering evidence of articulatory strengthening (i.e., increased constriction narrowing) in glides preceding pitch-accented vowels.

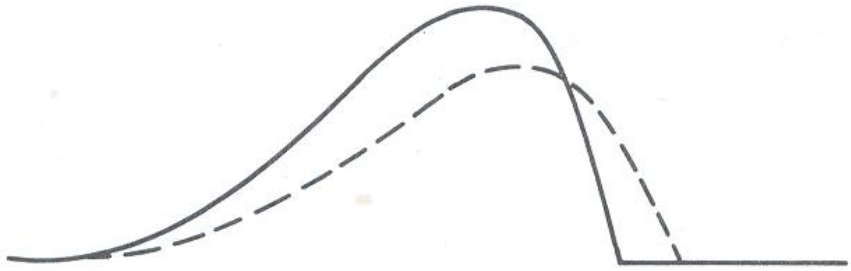

Fig. 2: The solid line shows the typical shape of a pulse of glottal volume velocity for an open vowel. The dashed line indicates schematically the modification of this pulse for a glide which is produced with a relatively narrow constriction in the vocal tract. In this skewed pulse, the peak amplitude of the waveform is decreased, and the open quotient is increased. The increase in open time contributes to an increase in the fundamental period (decrease in fundamental frequency), and may also contribute to some increase in aspiration noise.

4.2. Perceptual study

In the acoustic analyses of natural speech, it was found that the amplitude reduction characteristics of glides are relatively invariant to segmental context, in contrast with the formant characteristics. This suggests that reduction in voicing amplitude may be a particularly useful perceptual cue to the presence of glides in the speech stream. In addition, it was found through measurements of A_{RMS}, OQ, HNR, and F0 that oral pressure build-up during glides has a

weakening effect on the glottal voicing source, contrary to previous assumptions. To test whether these glottal source effects play a major role in the distinction between glides and high vowels, a listening experiment was performed with synthetic tokens designed to isolate and compare the perceptual salience of acoustic cues to the glottal source effects with cues to the vocal tract configuration itself.

Voicing amplitude (representing source effects) and first formant frequency (representing filter configuration) were manipulated in cooperating and conflicting patterns to create percepts of /V#V/ or /V#GV/ sequences, where Vs were high vowels and Gs were their cognate glides. In the responses of eight naïve subjects, voicing amplitude (AV) had a greater effect on the detection of glides than first formant frequency (F1), suggesting that glottal source effects are more important to the distinction between glides and high vowels. In particular, AV reduction was sufficient to cue the presence of a glide in the absence of F1 movement; by contrast, F1 reduction was not sufficient to cue the presence of a glide in the absence of AV reduction. The average phonetic category boundary was 8 dB of amplitude reduction to begin to cue the presence of a glide.

4.3. Discussion

The results of the acoustic and perceptual studies provide evidence for an articulatory-acoustic mapping defining the glide category. It is suggested that glides are differentiated from high vowels and fricatives by articulatory-acoustic boundaries related to the aerodynamic consequences of different degrees of vocal tract constriction. The supraglottal constriction target for glides is sufficiently narrow to produce a non-vocalic pressure drop, but not sufficiently narrow to produce a significant frication noise source. This mapping is consistent with the theory that articulator-free features are defined by aero-mechanical interactions (Stevens & Hanson, in press). According to the results of this study, the articulator-free feature differentiating glides from vowels ([-vocalic], perhaps) may be defined by the aerodynamic effects on the glottal source caused by the unique constriction degree of canonical glides, along with the acoustic cues produced by those effects. Our newly detailed understanding of these acoustic characteristics and their sources provides the potential for numerous improvements to phonological classification systems and speech technology applications.

References

Bickley, C., & Stevens, K. (1986). Effects of a vocal-tract constriction on the glottal source: experimental and modelling studies. *Journal of Phonetics*, **14**, 373-382.

Catford, J. (1988). *A Practical Introduction to Phonetics*. Oxford: Clarendon Press.

Chitoran, I. (2002). A perception-production study of Romanian diphthongs and glide-vowel sequences. *Journal of the International Phonetic Association*, **32** (2), 203-222.

Fant, G. (1972). *Vocal tract wall effects, losses, and resonance bandwidths*. Quarterly Progress and Status Report, Royal Institute of Technology (KTH), Speech Transmission Laboratory, Stockholm.

Lehiste, I., & Peterson, G. E. (1961). Transitions, Glides, and Diphthongs. *Journal of the Acoustical Society of America*, **33** (3), 268-277.

Maddieson, I., & Emmorey, K. (1985). Relationship between Semivowels and Vowels: Cross-Linguistic Investigations of Acoustic Difference and Coarticulation. *Phonetica*, **42**, 163-174.

Mehta, D. (2006). *Aspiration noise during phonation: synthesis, analysis, and pitch-scale modification.* SM Thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, Massachusetts.

Selkirk, E. O. (1984). On the Major Class Features and Syllable Theory. In M. Aronoff, & R. T. Oehrle (Eds.), *Language Sound Structure* (pp. 107-136). Cambridge, Massachusetts: MIT Press.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, Massachusetts: MIT Press.

Stevens, K. N., & Hanson, H. M. (in press). Articulatory-acoustic relations as the basis of distinctive contrasts. In W. Hardcastle, & J. Laver (Eds.), *Handbook of Phonetic Sciences* (2nd ed.). Malden, Massachusetts: Wiley-Blackwell.

5. Development, Perceptual Evaluation, and Acoustic Analysis of 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. ELs provide laryngectomy patients with the basic capability to communicate, but current EL devices produce a mechanical speech quality which has been largely attributed to the lack of natural fundamental frequency (F0) variation. In order to improve the quality of EL speech, we have been aiming to develop and evaluate an automatic F0 control scheme, in which F0 was modulated based on variations in the root-mean-squared (RMS) amplitude of the EL speech signal (Saikachi, Stevens, & Hillman, 2009). The results of perceptual experiments showed that modulating the F0 of EL speech using a linear relationship between amplitude and frequency made it significantly more natural sounding than EL speech with constant F0, providing a preliminary support for the proposed control scheme.

Based on these results, the current study more fully evaluated the prosodic control ability of amplitude-based F0 control scheme by conducting a set of perceptual identification experiments. The vocal tasks consisted of sentence quadruplets containing two declarative sentences and two interrogative sentences, with the location of contrastive stress differing within the statement/question pairs. Four speech-language pathologists recorded speech tasks under three different conditions. In the first condition, subjects used their natural voices. The second condition involved producing the speech material using an EL set at a constant F0. For the third condition, subjects produced the speech tasks using the manual control feature of the Tru-Tone EL to vary F0. The amplitude-based F0 control stimuli were generated from the EL speech with constant F0 condition by linearly covarying the F0 based on the RMS amplitude variation. Each sentence from the entire stimulus set was presented individually to 10 normal-hearing listeners in a sound-isolated booth. The listeners were instructed to categorize each vocalization as either a question or a statement in the first session and were asked if the stimulus had sentence-initial or sentence-final stress in the second session.

The results showed a slight improvement for perception of contrastive stress and slight degradation in perception of intonation in amplitude-based F0 control compared to constant F0 condition, revealing inherent limitations in communicating linguistic contrasts in amplitude-based F0 control scheme. The F0 based on the amplitude of EL speech wave appeared to depend on segmental contexts rather than on the suprasegmental aspects. Further efforts will be made to examine the potential source of amplitude fluctuation in EL speech in order to improve algorithms for real-time enhancement of EL speech based on processing of the EL speech output and to examine the possibility of training an EL user to manipulate the device to produce more natural prosody.

6. Speech prosody

One of the most pressing questions about intonational prosody is how the F0 contour of an utterance is aligned with its words and syllables. Two aspects of this question were addressed: 1) what aspects of the F0 contour does the speaker align, and 2) how is F0 alignment influenced by the number of tonal targets that must be realized on a particular word.

6.1 What aspect of the F0 contour are aligned? It has been widely assumed that the tonal targets in an F0 contour are its turning points, i.e. its peaks (High targets), valleys (Low targets) and elbows (High or Low targets depending on whether the contour is rising or falling), and that these target turning points are what the speaker aligns with the words and syllables of an utterance, to produce pitch accents and boundary-related tones. Although research based on this assumption has resulted in a number of important insights, several significant problems have emerged, including a) indeterminate location of peaks and valleys in non-voiced regions of the speech signal, and b) significant effects of the global shape of a rise-fall on perception even when the

turning points remain the same. In two studies of American English intonational contours, we tested a new view of tonal alignment, i.e. that speakers align the Tonal Center of Gravity of a high F0 region (a global measure for characterizing F0 contours) rather than the peak or the valley. In one study, we showed that TCoG, distinguishes single-accent contours (H* L-) from contrasting double-accent contours (H* L*) more reliably than a turning point-based approach focused on the location/scaling of an F0 elbow. This new measure is sensitive to difficult-to-quantify changes in contour shape, while avoiding the pitfalls of purely shape-based theories. In a second study, using logistic regression, we showed that this new measure also distinguishes two contrasting pitch accents (L+H* vs. L*+H) better than the alignment of a single F0 turning point (i.e. the valley or the peak). Moreover, while the alignment of both turning points does about as well as TCoG in distinguishing these two pitch accents, TCoG results were significantly more robust to noise in the location of critical aspects of the F0 curve. Thus, the hypothesis that speakers align the TCoG of a contour rather than the F0 turning points is potentially transformative in the field, because it avoids the difficulties with locating specific F0 points when the F0 curve is interrupted by voiceless consonants or is ambiguous F0 shapes (such as plateaus), and provides an account of the growing evidence that some aspects of global contour shape, such as the domedness or scoopiness of the rise or fall, have a measureable effect on listeners' perceptions of intonation categories. Perceptual experiments are currently underway to test the hypothesis that when speakers change the alignment of F0 turning points, they do so in the service of changing the TCoG.

6.2. How is F0 alignment influenced by the number of tonal targets? In earlier work we found that a High pitch accent had an earlier F0 peak if it occurred on a phrase-final syllable than on a phrase-medial syllable. This result was in accord with earlier studies of effects of tonal crowding, which showed that an F0 peak occurs earlier if other tonal targets (such as boundary tones) are realized on the same syllable. However, this earlier study included results from only 5 speakers, and had several potentially confounding factors, such as the number of syllables, lexical stress pattern and vowel content in the word. In a more thorough study of this phenomenon in 20 speakers, we found that most speakers show the effects of tonal crowding for High pitch accents, but not for Low, adding to the growing body of evidence that Low accents are governed by different principles.

7. Speech Development in Children

Developing a model of how children acquire the phonological feature contrasts and phonetic cues of their adult speech community requires a detailed understanding of how production changes during development. Earlier studies based on listening provided an initial estimate of these patterns; recent instrumental studies of the acoustic and articulatory details of production during development have revealed some surprises, including covert contrasts between distinctive feature categories, and of incomplete acquisition of adult feature-cue patterns (Imbrie, 2005), which can be difficult for adults to hear. Few instrumental studies, however, have focused on the child's cue patterns for coda consonants. Our study of cues to the feature [voice] for coda stops used quantitative acoustic analyses of tokens from the Imbrie corpus of 2-year-old speech to compare child productions with those of their adult caretakers. Results show that a) many children produce a noisy region at the end of the vowel for voiceless codas (in duck, cup) but a long strong voice bar during closure for voiced codas (in bug, tub), and b) these cues may be exaggerated versions of the feature cues of their adult caretakers.

8. Voice quality

Voice quality varies significantly in typical speakers, from 'modal' to breathy to irregular or glottalized pitch periods, and it is gradually becoming clear that this variation is governed, at least in part, by prosodic structure, including the location of intonational phrase boundaries and intonational prominences (pitch accents). This observation raises the question of how listeners

might make use of this information in perceiving speech and recognizing speakers. Earlier studies report systematic differences across speakers in the occurrence of utterance-final irregular phonation; we extended this work to investigate whether human listeners remember this speaker-specific information and can access it when necessary (a prerequisite for using this cue in speaker recognition). Listeners personally familiar with the voices of the speakers were presented with pairs of speech samples: one with the original and the other with transformed final phonation type. Asked to select the member of the pair that was closer to the talker's voice, most listeners tended to choose the unmanipulated token (even though they judged them to sound essentially equally natural). This suggests that utterance-final pitch period irregularity is part of the mental representation of individual speaker voices, although this may depend on the individual speaker and listener to some extent.

Publications

Journal Articles, Published

Bohm, T. and Shattuck-Hufnagel, S. (2009), Do Listeners Store in Memory a Speaker's Habitual Utterance-Final Phonation Type? *Phonetica* 66, 150-166.

Ghosh, S.S., Tourville, J.A. and Guenther, F.H. (2009) "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).

Saikachi, Y., Stevens, K. and Hillman, R. (2009), Development and Perceptual Evaluation of Amplitude Based F0 Control in Electrolarynx Speech. *Journal of Speech, Language and Hearing Research* 52, 1360.

Matthies, M.L, Guenther, F. H., Denny, M., Perkell, J. S., Burton, E., Vick, J., Lane, H., Tiede, M. and Zandipour, M. (2008) Perception and Production of /r/ Allophones Improve with Hearing from a Cochlear Implant, *Journal of the Acoustical Society of America,* 124, 3191-3202.

Journal Articles, Accepted for Publication

Barnes, J., Brugos, A., Veilleux, N. and Shattuck-Hufnagel, S., Turning Points, Tonal Targets, and the English L- Phrase Accent. Language and Cognitive Processes.

Ghosh, S.S., Matthies, M.L., Mass, E., Hanson, A., Tiede, M., Menard, L., Guenther, F.H., Lane, H. and Perkell, J.S., An Investigation of the Relation between Sibilant Production and Somatosensory and Auditory Acuity. Journal of the Acoustical Society of America.

Perkell, J.S., Movement Goals and Feedback and Feedforward Control Mechanisms in Speech Production. Journal of Neurolinguistics (invited submission for a special issue, Neural Theory of Language).

Shue, Y., Shattuck-Hufnagel, S., Iseli, M., Jun, S.-A., Veilleux, N. and Alwan, A., On the Acoustic Correlates of High and Low Nuclear Pitch Accents in American English. Speech Communication.

Stevens, K. and Keyser, S., "Quantal Theory, Enhancement and Overlap," J. Phonetics, forthcoming.

Book/Chapters in Books

Shattuck-Hufnagel, S. (in press), The Role of the Syllable in Speech Production Planning. In Cairns, C. and Rainey, E. (Eds.), *Proceedings of the Workshop on the Syllable*, City University of New York, January 2008.

Shattuck-Hufnagel, S., Demuth, K., Hanson, H. and Stevens, K. (in press), Acoustic Cues to Stop-Coda Voicing Contrasts in the Speech of American English 2-3-year-olds. In Clements, N. and Ridouane, R. (Eds), *Where do features come from?* Paris: The Sorbonne.

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

Meeting Papers, Published

Demuth, K., Shattuck-Hufnagel, S., Song, J.Y., Evans, K., Kuhn, J. and Sinnott-Armstrong, M. (2009), Acoustic cues to stop coda voicing contrasts in 1-2-year-olds' American English. *Journal of the Acoustical Society of America* 125, 2570A.

Jung, Y. (2009), Subglottal effects on the vowels across language: Preliminary study on Korean. *Journal of the Acoustical Society of America* 125, 2638A.

Perkell, J.S. (2009). Movement goals and feedback and feedforward control mechanisms in speech production, Proceedings of the Third International Symposium on Biomechanics, Human Function and Information Science, Kanazawa, Japan, Feb. 20-22, 2009, Vol. II, pp. 21-45.

Perkell, J.S., Matthies, M.L., and Tiede, M.K. (2009), Evidence of an articulatory saturation effect in the production of /s/ in American English. *Journal of the Acoustical Society of America* 125, 2569(A).

Shue, Y-L., Shattuck-Hufnagel, S., Iseli, M., Jun. S.-A., Veilleux, N. and Alwan, A. (2009), Effects of intonational phrase boundaries on pitch-accented syllables in American English. *Proceedings of Interspeech* (Brisbane, Australia, Sept 2008) (won prize for best student paper) p. 873-876.

Tiede, M., Goldstein, L., Shattuck-Hufnagel, S., Perkell, J., and Matthies, M. (2009), Optimization of articulator trajectories in producing learned nonsense words. *Journal of the Acoustical Society of America* 125, 2499A.

Bohm, T., Audibert, N., Shattuck-Hufnagel, S., Nemeth, G. and Auberge, V. (2008), Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles. *Journal of the Acoustical Society* 123, 3886A

S. Cai, M. Boucek, S.S. Ghosh, F.H. Guenther, and J.S. Perkell, "A System for Online Dynamic Perturbation of Formant Frequencies." Proceedings of the 8th International Seminar on Speech Production, Dec. 8-12, Strasbourg, 2008, p. 65-68.

Hanson, H. and Shattuck-Hufnagel, S. (2008), Acoustic cues to the voicing contrast in coda stops in the speech of 2-year-olds learning American English. *Journal of the Acoustical Society* 123, 3320A.

Hunt, E.H. (2008), Acoustic characteristics of glides /j/ and /w/: interactions with phonation frequency. *Journal of the Acoustical Society of America* 124, 2519A (Winner of Best student paper award).

Jung, Y., Lulich, S.M and Stevens, K.N. (2008), Development of subglottal quantal effects in young children. *Journal of the Acoustical Society of America* 124, 2519A.

J.S. Perkell, H, Lane, S.S. Ghosh, F.H. M.L. Matthies M.K. Tiede Guenther, and, L. Ménard, "Mechanisms of vowel production: auditory goals and speaker acuity," Proceedings of the 8th International Seminar on Speech Production, Dec. 8-12, Strasbourg, 2008, p. 29-32.

Veilleux, N., Barnes, J., Shattuck-Hufnagel, S. and Brugos, A., (2008), Tones and turning points: Intonational primitives in phonetics and phonology. *Journal of the Acoustical Society of America* 124, 2497A.

Meeting Papers, Presented

Shattuck-Hufnagel, S. and Turk, A. (2009), An experimental investigation of Abercrombian Feet in American English. Presented at the Workshop on the Foot, City University of New York, January 2009.

Song, Y.Y., Demuth, K. and Shattuck-Hufnagel, S. (2009), The acoustic realization of velar stop codas in utterance-final vs. utterance medial position. Presented at the Workshop on Child Phonology, Austin, Texas, June 2009.

Theses

E. Hon Hunt, Acoustic Characterization of the Glides /j/ and /w/ in American English, PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2009.

Y. Jung, *Acoustic Articulatory Evidence for Quantal Vowel Categories: The Features [low] and [back],* PhD thesis, Graduate Program in Speech and Hearing Biosciences and Technology, Harvard-MIT Division of Health Sciences and Technology, MIT, 2009.

Y. Saikachi, *Development, Perceptual Evaluation, and Acoustic Analysis of F0 Control in Electrolarynx Speech,* PhD thesis, Graduate Program in Speech and Hearing Biosciences and Technology, Harvard-MIT Division of Health Sciences and Technology, MIT, 2009.

A. Nti, Dialects Will Illuminate Words, Master's thesis, Department of Electrical Engineering and Computer Science, MIT, 2009.