Speech Communication

Sponsors:

C.J. Lebel Fellowship, Dennis Klatt Memorial Fund, Donald North Memorial Fund, National Institutes of Health (Grants R01-DC00075, R01-DC01291, R01-DC01925, R01-DC02125, R01-DC02978, R01-DC03007, 1 R29 DC02525, T32-DC00038), National Science Foundation [Grants INT-9615380 (US-France Cooperative Research), and INT-9821048 (US-Germany Cooperative Research)]

Academic and Research Staff:

Professor Kenneth N. Stevens, Professor Jonathan Allen, Professor Morris Halle, Professor Samuel J. Keyser, Dr. Corine A. Bickley, Dr. Marilyn Chen, Dr. Jeung-Yoon Choi, John Gould, Dr. Jay Moody, Dr. Joseph S. Perkell, Dr. Stefanie Shattuck-Hufnagel, Seth Hall, Dr. Reiner Wilhelms-Tricarico, Jennell Vick, Majid Zandipour

Visiting Scientists and Research Affiliates:

Dr. Suzanne E. Boyce, Department of Communication Disorders, University of Cincinnati, Cincinnati, Ohio; Sherrie Brown, Department of Neuropsychology, Massachusetts General Hospital, Boston, Massachusetts; Dr. Gabriella Di Benedetto, Department of Information and Communication, INFOCOM, University of Rome, Rome, Italy; Dr. Carol Y. Espy-Wilson, Department of Electrical Engineering, Boston University, Boston, Massachusetts; Dr. Krishna Govindarajan, SpeechWorks International, Boston, Massachusettts; Dr. David Gow, Department of Psychology, Salem State College, Salem, Massachusetts, and Department of Neuropsychology, Massachusetts General Hospital; Dr. Frank Guenther, Department of Cognitive and Neural Systems, Boston University, Boston, Massachusetts; Dr. Helen M. Hanson, Sensimetrics Corporation, Somerville, Massachusetts; Dr. Robert E. Hillman, Mass Eye and Ear Infirmary, Boston, Massachusetts; Dr. Caroline Huang, SpeechWorks International, Boston, Massachusettts; Carrie Landa, Department of Neuropsychology, Massachusetts General Hospital, Boston, Massachusetts; Dr. Harlan Lane, Department of Psychology, Northeastern University, Boston, Massachusetts; Dr. John I. Makhoul, Bolt Beranek and Newman, Inc., Cambridge, Massachusetts; Dr. Sharon Y. Manuel, Department of Speech Language Pathology & Audiology, Northeastern University, Boston, Massachusetts; Dr. Melanie Matthies, Department of Communication Disorders, Boston University, Boston, Massachusetts; Dr. Richard McGowan, Sensimetrics Corporation, Somerville, Massachusetts; Dr. Alice Turk, Department of Linguistics, University of Edinburgh, Edinburgh, United Kingdom; Dr. Lorin Wilde, Lernout & Hauspie Speech Products, Burlington, Massachusetts; Jane Wozniak, Speech and Language Pathology, Private Therapy, Massachusetts.

Graduate Students:

Lan Chen, Harold Cheyne, Laura Dilley, Heather Gunter, Michael Harms, Dameon Harrell, Andrew Howitt, Roy Kim, Aaron Maldonado, Hale Ozsoy, Kelly Poort, Janet Slifka, Jason Smith, Felice Sun

Undergraduate Students:

Carolyn Chen, Xixi D'Moon, Shinning Duh, Mengkiat Goh, Emily Hanna, Stefan Hurwitz, Anna Khasin, Shuley Nakamura, Diana Ng, Karen Robinson, Jeremy Vogelmann, Linda Yu, Sophia Yuditskaya

Technical and Support Staff:

Arlene E. Wint

Contents:

1. Studies of Normal Speech Production

1.1 Constraints and strategies in speech production

1.1.1 Development of facilities

Constructing speaker-specific articulatory vocal tract models for testing speech motor control hypotheses. To help lay the foundation for our physiological modeling work, software was written in MATLAB that allows relatively fast construction of speaker-specific articulatory vocal tract models from midsagittal vocal tract profiles, acquired with Magnetic Resonance Imaging (MRI). Data were obtained from two speakers producing a list of approximately 30 speech sounds. MRI images were first binarized to allow differentiation of air and tissue regions. Images were then segmented and vectorized into relevant contour shapes (lips, hard palate, velum, laryngeal region, tongue, and jaw). For each subject, the autocorrelation of these shapes was fitted by regression over the jaw components followed by principal component analysis (PCA) on the remaining terms, thus uniquely defining a basis in the shape space. Pruning this basis for the most significant components lead to the definition of a speaker specific parameter space that allows the representation, up to a residual error, of any vocal tract profile. The resulting parameters approximate the primary articulatory degrees of freedom of the individual speaker. A speaker-specific algorithm for converting midsagittal profiles to area functions was also derived; this algorithm was tuned using acoustic data collected while the subject is producing phonemes in the MRI machine. (Also supported by NIDCD grant 1 R29DC02852 to Prof. Frank Guenther of Boston University.)

Vocal-tract modeling. We have debugged a new version of a 2-D tongue model developed initially by our collaborators in Grenoble, France. The new version is written for MATLAB 5, and it has been modified to run on a PC under Windows NT. The model has been used in preliminary experiments described below.

Development of a system for automatic phonemic segmentation using a combination of statistical and knowledge-based methods. A prototype system was developed, with the goal of reducing the large amount of time spent in the manual labeling of phoneme boundaries in acoustic signals during the initial stages of data extraction. The system first uses Aligner, a commercial speech alignment software package, to synchronize the speech waveform to a supplied text. Then, a custom built knowledge-based segmentation program is used to locate and label segmentation boundaries. The system has been developed initially and tested in the segmentation of acoustic events associated with voiceless stop consonants.

Upgrading hardware and software for our EMMA movement transducer system. We have taken a number of steps to move our data acquisition and analysis away from our old VAX-VMS systems to PC-Windows systems. We have purchased and installed several new PCs and new data acquisition system, along with software for data acquisition, visualization and analysis from ATR in Kyoto, Japan. The software has been modified to enable stimulus presentation to the subject on an LCD display.

Development of a facility for "sensorimotor adaptation" experiments. In support of a planned series of experiments on "sensorimotor adaptation", in which a speaker pronounces utterances while hearing formant-modified feedback of his own speech, we have purchased computer and special-purpose digitigal signal processing hardware, and we have made significant progress in developing the necessary software.

1.1.2 Studies performed.

Generation of subject specific formant patterns for vowels using a phsiological tongue model. We performed a study in which we compared the outuput of a vocal-tract simulation that includes the above-mantioned tongue model with measured acoustic and articulatory data. To generate acoustic data with the simulation, we used the tongue model and a midsagittal distance-to-area function algorithm to

produce a time-varying vocal-tract area function and an acoustic transfer function. The model was optimized to generate subject-specific formant values (F1, F2, F3, and F4) with ±7.5% error in the vowel space (/i/, /e/, /a/, /o/, and /u/). [We are submitting a manuscript regarding our preliminary results of examination of acoustic (F1-F2) and movement trajectories produced by the model and subjects.] Currently, we are working on the development of a feedforward model of acoustic-to-muscle space mapping based on RBF (radial basis function) and second order Sigma-Pi neural networks.

1.1.3 Theoretical developments

A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. A theory of the segmental component of speech motor control has been presented, followed by supporting data. In the theory, speech movements are programmed to achieve auditory/acoustic goals. The goals are determined partly by "saturation effects", which are basic characteristics of speakers' production systems that make it possible to produce a sound output that has some relatively stable acoustic properties, using a somewhat variable motor input. The programming of articulatory movements to achieve auditory goals utilizes an internal model of relations between articulatory configurations and their acoustic consequences. The internal model is acquired and maintained with the use of auditory feedback. The supporting data for this theory come from experiments on speakers with normal hearing, cochlear implant users and a patient with neurofibromatosis-2. (Also supported by NIDCD grants no. DC01291, DC02525, DC02852 and DC00361)

1.2 Respiration and prosody in speech production

This research has continued to examine the lung volume, subglottal pressure, and airflow used by normal speakers when they produce various kinds of sentence material, together with several acoustic measures. These measurements have been supplemented by estimates of the cross-sectional area of the smallest constriction in the vocal tract (calculated from the pressure and flow) and perceptual ratings of syllable prominence. Attention has been focussed on initiations and terminations of utterances. Present results indicate that (1) utterance initiation is a rapid process where the rise in pressure is correlated to the relaxation curve of the chest wall; (2) voiceless onsets start at a lower pressure than voiced onsets, and for those voiceless onsets, subglottal pressure at onset of phonation is pushed closer to peak pressure; (3) utterance termination is a slower process associated with expiratory muscle actions; (4) the amplitude decrease associated with the end of an utterance as well as the end of phonation is largely controlled by a sharp rise in glottal area for utterances of isolated sentences; and (5) the presence of irregular glottal vibration is correlated with regular changes in both area and pressure within each speaker at utterance termination. The measurements of respiratory parameters have been carried out in collaboration with the Voice Laboratory at the Massachusetts Eye and Ear Infirmary, under the direction of Dr. Robert Hillman.

1.3 Coarticulation of rounding in obstruent consonants as estimated from acoustic data

When a consonant such as /s/ is produced between two rounded vowels, the rounding for the vowels tends to spread into the consonant. Previously reported articulatory data have shown that this spreading of rounding into the consonant is more complete in Turkish than it is in English. Turkish has harmony for vowel rounding, and this different use of the feature [round] in the two languages accounts for the differences in rounding observed in the articulatory data. In the present study, a similar behavior for the two languages was observed for the intervocalic consonant /s/ based on acoustic data. Rounding for /s/ can be inferred by measuring the frequency region of the major spectrum prominence: rounding increases the length of the cavity in front of the constriction, and causes the frequency of the front-cavity resonance to be lowered. Acoustic data of this kind were obtained from several speakers of English and Turkish producing the intervocalic consonant /s/ in the environment of rounded and unrounded vowels. The acoustic data were consistent with the previously published articulatory data: the spread of rounding into the consonant was significantly greater for Turkish than for English.

2 Speech Research Relating to Special Populations

2.1 The speech of cochlear implant patients

Development of facilities. We have developed hardware and software for new PC-based perceptual and intelligibility tests. We have further streamlined our setup for stimulus modification experiments, in which implant users' auditory feedback is switched on and off and changes in speech parameters are measured.

Longitudinal studies. This year**,** we have made three baseline pre-implant recordings on each of three additional CI subjects. Four speakers returned for post-implant recordings one week, one month, and three months following processor activation. Three of our 12 implant subjects have returned for their one-year visits and three others have returned for two-year visits. Thirty-six recordings were made this year with CI subjects. Over 85% of our recordings were digitized and more than 75% of the data have been extracted and analyzed.

Perceptual studies. At the time of each speech recording, subjects discriminate eleven consonants and eight vowels from the natural speech of a same-gender normally hearing speaker. In general, the perceptual abilities of all of our CI subjects have improve longitudinally. We have begun to assess whether implant users' evolving abilities to perceive and produce various features covary. Acquired data indicate that improvements in perception are related to improvements in production when production is aberrant in the pre-CI condition. Similarly, decreases in perception have been noted to co-occur with diminished production.

Intelligibility testing. Intelligibility of implant users' speech has been tested before and at intervals after the initial activation of their speech processors. Three repetitions of each of 8 vowels and 11 consonants were extracted from speaker's two pre-implant sessions and last two post-activation sessions and assembled in random order. In order to assess *changes* in intelligibility from pre- to post-implantation and over time in our very intelligible subjects, noise was mixed with the signals before presentation to listeners. The resulting signals were presented in individual listening sessions with each of ten normal-hearing listeners. Analysis of the resulting data is currently underway for a planned test of whether normal-hearing listeners' abilities to perceive various features in implant-users' speech covary with changes in implant-users productions.

Coarticulation. In order to examine the role of hearing status in controlling coarticulation, eight English vowels in /bVt/ and /dVt/ syllables, embedded in a carrier phrase, were elicited from seven postlingually deafened adults in repeated recording sessions both before and up to a year after they received cochlear implants and their speech processors were turned on. Measures were made of F2 at obstruent release and at 25 ms intervals until the final obstruent. With the change in hearing status, four of the speakers significantly reduced the size of their vowel spaces in the F1-F2 plane, while the others increased them. All speakers but one also reduced vowel duration significantly. Four of the speakers had lower dispersion of vowel formant values around vowel midpoint means, but the other three did not. An index of coarticulation, based on the ratio of F2 at vowel onset to F2 at midvowel target, was computed. Changes in the amount of coarticulation after the change in hearing status were small and nonsystematic for the /bVt/ syllables, while those for the /dVt/ syllables averaged a three percent increase – within the range of reliability measures for two hearing control speakers. The slopes of locus equations and ratios of F2 onsets in point vowels tend to confirm that hearing status had little if any effect on coarticulation in the postlingually deafened speakers, leading to the conclusion that hearing does not play a direct role in regulating anticipatory coarticulation in adulthood.

Language-specific changes in vowel spaces. In collaboration with the University of Miami Medical School, we measured vowel productions of two groups cochlear implant users when their implant speech processors were on and when they were off. The groups are 5 native speakers of Cuban Spanish (5 vowels) and 6 native speakers of American English (9 vowels used). For each subject, the average inter-vowel distance in the F1-F2 space was calculated and compared between hearing and non-hearing states. For the English speakers, whenever there was a significant difference between hearing and non-hearing states the average inter-vowel distance was *greater* with hearing than without. The average distance was larger for the Spanish speakers than the English speakers. The Spanish hearing-related changes were somewhat more variable than the English ones, but in general, the average Spanish inter-vowel distances were *smaller* with hearing than without. We speculate that since the English vowel space is more crowded, when English implant users can hear, they tend to pronounce vowels more distinctly from one another to help assure intelligibility. With a less crowded and larger vowel space, when Spanish implant users can hear, they use more economical productions, presumably without sacrificing intelligibility. Considered together, the changes in the two languages illustrate a trade-off between distinctiveness of acoustic goals and economy of effort.

Production and perception of liquids /r/ and /l/. Speech production, perceptual testing and speech intelligibility data for /r/ and /l/ were obtained for eight subjects both pre- and post cochlear implant (CI). Formant transition analysis for the CI subjects and two normal-hearing subjects indicated that /r/ and /l/ could be reliably differentiated by the extent of the F3 transition and the distance in Hz between F2 and F3 at the consonant-vowel boundary. Subjects who had a limited contrast between /r/ and /l/ pre-implant and who showed improvement in their perception of these consonants demonstrated greatly improved production of /r/ and /l/ six months post-CI. The speech production changes noted in the acoustic analyses were corroborated by intelligibility improvements as tested with a panel of normal-hearing listeners. The combination of auditory feedback changes and the complex and variable articulation of /r/ supports the hypothesis that the goals of segmental speech movements are acoustic in nature.

Production and perception of vowel pairs. Speech production, perceptual testing and speech intelligibility data for eight vowel pairs were obtained for eight subjects pre- and post-implant. Analyses of Euclidean distance of F1 and F2 between pairs for the CI subjects indicated that pair members could be reliably differentiated. Subjects who had a limited or exaggerated contrast between individual pairs (as compared with normative values) pre-implant and who showed improvement in their perception of these pairs demonstrated greatly improved production of the pairs six months post-CI. The speech production changes noted in the acoustic analyses were corroborated by intelligibility improvements as measured by a panel of normal-hearing listeners. As in the /r/-/l/ study, the combination of auditory feedback changes and the adjustments in production of vowel pairs supports the hypothesis that the goals of speech are acoustic in nature.

2.2 Acoustic characteristics of speech produced by speakers with neuromotor disorders

The aim of this project is to develop acoustic measures that describe some of the properties of utterances produced by speakers with neuromotor disorders. These measures are also made for a group of normal talkers. It is hoped that these quantitative acoustic measures can supplement evaluations of the speech by clinicians. One criterion for the selection of the measures, then, is that they should be correlated with clinical judgments of the speech. The acoustic measurements should also assist clinicians in making inferences about specific deviations in articulatory movements and control for speakers with these types of disorders. During the past year, research in this area has focussed on stop consonants and on vowels.

The utterances from speakers with dysarthria were collected in an earlier study. They consisted of several repetitions of a series of monosyllabic words by four female and four male speakers, five with spastic dysarthria, one with athetosis, one with spastic dysarthria plus athetosis, and one with cerebellar ataxia. Additional recordings of the same words were made by several normal speakers.

For the words beginning with stop consonants, three kinds of data were collected from the recordings: (1) listener judgments of the consonant identity and the quality of the stop consonants; (2) judgments of several attributes of the stop consonants from observations of spectrograms of the words; and (3) detailed acoustic measurements of attributes such as the burst spectra and formant transitions. Data on the intelligibility of the words were available from an earlier study. The data show that a rank ordering of the speakers based on listener judgments of the consonant identity and quality of the stop consonants is very similar to a rank ordering based on overall intelligibility of the speakers. Some aspects of the results based on observations of spectrograms correlate well with the listener judgments, but the data derived from spectrogram observations provide more explicit descriptions of the reasons for the reduced evaluations of intelligibility and quality by listeners. Detailed acoustic measurements of burst spectra often showed significant differences between the spectra for the speakers with disordered production and those for normal speakers. There was not a clear relation between measures of these differences and listener judgments of the consonants. Evidently single measures of attributes of stop consonants at localized regions in time are not very predictive of listener's judgments of the quality or intelligibility of the stop consonants.

The study of vowels includes measurements of formant frequencies and glottal characteristics, including fluctuations of these attributes within a word and consistency across repetitions of a word. Listeners made judgements of the identity and breathiness of the vowels excised from the words, and these judgments were related to the acoustic measures. Errors in vowel identification were correlated with tongue-body position, as inferred from deviations in formant frequencies relative to those for normal speakers. Perception of voice quality or breathiness was correlated with measures of noise in the vowel spectra and waveforms and with measures of spectrum tilt and of low-frequency spectrum shape, although the strength of the correlation depended on the gender of the speaker. The degree to which the vowels of the dysarthric speakers differed from normal depended to some extent on the features of the adjacent consonants and on the height and frontness of the intended vowel.

3 Speech production planning and its relation to prosody and speech errors

Work on speech production planning focussed on two questions: what can we learn about the phonological planning process from speech errors and their corrections, and what can we learn from duration adjustments in spoken phrases. Speech error correction patterns distinguish between word-level errors, like "left" for "right" and sound-level errors like "tar cowed" for "car towed", in the sense that corrections of word-level errors are more likely to be prosodically marked than corrections of sound-level errors. This suggests that the two types of errors may occur at different levels in the production planning process. However, a substantial number of errors are ambiguous between word-level and sound-level errors, e.g. "keep Tar Talk on the air" for "keep Car Talk on the air". These ambiguous errors have sometimes been interpreted as evidence that word-level and sound-level processing levels interact, causing errors that reflect both types of constraints simultaneously. However, our analysis of 90 corrected errors in the MIT Digitized Speech Error Corpus showed that the correction patterns for word-sound-ambiguous errors are like those for sound-level errors, i.e. ambiguous errors are less likely to be prosodically marked. As measured by both prosodic labelling (+/- Pitch Accented correction) and f0 values, ambiguous errors are indistinguishable from unambiguous sound errors, and significantly different from unambiguous word errors, providing additional support for the view that these two levels of processing are separate in production planning.

Duration lengthening at the ends of spoken phrases has long been noted, especially in the final syllable, and research over the past 10 years has shown that this pre-boundary lengthening is a) governed by prosodic constituents, and b) hierarchically organized, reflecting the structure of the Prosodic Hierarchy. That is, duration lengthening at the ends of higher-level constituents, such as the Utterance and the Intonational Phrase, is greater than lengthening at the ends of lower-level constituents, such as the Prosodic Word. In our own experiments, we found that, as in other languages, most of the lengthening is located in the final rhyme of the Intonational Phrase, and increases monotonically from the syllable onset to the nuclear vowel to the final coda consonants. Moreover, we noted some additional duration lengthening extending leftward from the final syllable; preliminary analysis suggests that this additional lengthening is blocked by the onset of the main-stress syllable. Further analyses are under way to

determine whether this evidence for an interaction between lexical stress patterns and phrase-level prosodic phonetics is
reliable.

In contrast, we also studied duration shortening as a cue to constituent structure in phrases like “tuna choir” vs. “tune acquire” vs. “tune a choir”. It is well established that in e.g. “tuna choir”, the “tun-“is shorter and the schwa is longer than in “tune acquire”. In a comparison of 11 such triads, we determined that a) this difference reflects Polysyllabic Shortening, showing that “tuna” forms a more coherent constituent than “tune a-“, b) duration patterns do not provide any evidence of word-final lengthening, and c) durations for tokens like “tune a choir”, although more variable, generally have intermediate values between those for “tuna choir” and for “tune acquire”. This is compatible with the hypothesis that boundaries between a content word and a function word are not as strong as boundaries between two content words, a claim consistent with current definitions of structures in the Prosodic Hierarchy.

Finally, we carried out a perceptual study of the effects of repeated F0 contours on the perceptual organization of syllables into words. The hypothesis we tested was that repeated F0 contours provide an organizational frame for the syllables of an utterance, so that speakers will hear a string of syllables like "boyfriendshipsidewalkway", produced with an alternating high - low F0 contour, as two-syllable words i.e. boyfriend + shipside + walkway, because the repeated H-L pattern defines the domain of a perceptual constituent. Results supported the hypothesis: few listeners transcribed single-syllable words. Preliminary results from an experiment with flat F0 suggested that the alternating pattern was necessary to elicit the two-syllable-word transcriptions; follow-up work in progress using synthetically-flattened F0 will determine whether speakers produced any additional cues to word structure.

Additional activities included 1) the continuing expansion of the MIT Digitized Speech Error Corpus; 2) successful completion of a subcontract from Johns Hopkins University to label the prosody of 7+ hours of spontaneous spoken dialogues, material which is now available to us for close analysis of the phonetics of spontaneous communicative speech in our laboratory. For example, this database is the object of 3) our study of severe phonetic modifications of function words, as in "gonna" and "whydja". We will evaluate the hypothesis derived in part from the model of lexical access developed in our group, that a) many of the articulatory landmarks predicted by the underlying representation of these words (i.e. consonant closures and releases, glide opening minima and vowel opening maxima) will be preserved even in cases of severe modification; b) landmarks which are missing from the signal can be predicted by models of inter-gesture timing; and c) landmarks will be best preserved at the edges of prosodic constituents and more modifiable in constituent-medial regions.

4. Models for Lexical Access

4.1 Overview

The development and implementation of a knowledge-based model for accessing words from continuous speech is being carried out in collaboration with Professor Carol Espy-Wilson and her students at Boston University. During the past year we have refined our model of the lexical access process in human listeners. In this model, the input is the acoustic signal in running speech, and the output is the sequence of words. The principal steps in the model are: (1) initial signal processing to identify the locations of landmarks in the signal indicating acoustic prominences for vowels, certain minima in amplitude for glides, and acoustic discontinuities for consonants; (2) signal processing in the vicinity of the landmarks to determine cues that help to estimate the articulatory states and movements in the vicinity of the landmarks; (3) combining of the cues, together with information about the context in which the landmarks occur, to estimate the segments and distinctive features underlying the utterance; (4) from the estimates of segments and features in (3), match against the lexicon to hypothesize possible word sequences; and (5) from each of these hypothesized word sequences, calculate the sequence of landmarks and acoustic attributes that could result from the sequence, compare with the initial estimates in (1), (2), and (3), and accept or reject the word sequence. The components of the lexical-access model that have undergone revision in the past year are those relating to the stages in which graded acoustic cues related to articulatory states and movements are combined to estimate discretely specified segments and features

that are then matched against the lexicon, which is specified in terms of these segments and features. This revised view attempts to make a clear distinction between context-dependent acoustic cues for articulatory events and the discrete (and more abstract) features that underlie these events.

4.2 Identification of vowel landmarks

Landmark-based speech processing is a component of lexical access from features that is being developed in the Speech Communication Group. Detection and classification of landmarks is a crucial first step in this model. This work implements a vowel landmark detector based in part on a syllabic segmentation algorithm developed by P. Mermelstein in 1975. That algorithm was based on detection of peaks in amplitude of the speech signal. The performance of the detector is scored against the TIMIT database, using a novel algorithm to convert the segmental transcriptions to a landmark representation for scoring. The results show that substantial improvement in performance can be gained by modifying the frequency range for peak detection. An additional advantage of this modification is that post processing to remove fricative peaks is no longer necessary, and the algorithm is therefore substantially simpler. Statistical studies of the vowels in the TIMIT database have been done, to provide further insight into which characteristics of vowels and their immediate phonetic environment tend to inhibit the detection of the vowels, and which environments result in reliable detection of vowel-internal amplitude peaks. Results of the statistical study will be used to guide further enhancements to the vowel landmark detector.

4.3 Detection of consonant voicing

Our current model of the lexical access process includes several modules that estimate various features of vowels and consonants when the locations of the vowel and consonant landmarks are known. The feature estimates are based in part on the selection and extraction of appropriate cues measured in the vicinity of the landmarks. We have carried out the preliminary design and testing of a module for detecting voicing for consonants. Consonant production and conditions for phonation are first examined, to determine acoustic properties that may be used to infer consonant voicing. The acoustic measurements are then examined in different environments to determine a set of reliable acoustic cues. These acoustic cues are measured in the vicinity of the consonant closure and release landmarks, and include fundamental frequency, difference in amplitudes of the first two harmonics, cutoff first formant frequency, and residual amplitude of the first harmonic around consonant landmarks. Hand measurements of these acoustic cues result in error rates of voicing detection around 10% for isolated speech, and 20% for continuous speech. Combining closure/release landmarks reduces error rates by about 5%. Comparison with perceived voicing yield similar results. When modifications are discounted, most errors occur adjacent to weak vowels. Automatic measurements increase error rates by about 3%. Training on isolated utterances produces error rates for continuous speech comparable to training on continuous speech. These results show that a small set of acoustic cues based on speech production may provide reliable criteria for determining the values of features. The contexts in which errors occur correspond to those where human listeners make errors, and expressing acoustic information using features provides a compact method of describing these environments.

4.4 Identification of place features for vowels

As with other modules in our proposed lexical access model, the identification of the place features for vowels is achieved by measuring certain acoustic cues in the vicinity of landmarks and combining these cues in a way that is influenced by context. As a step toward the design of such a module for vowels, we have made a series of measurements of vowels in a database of 100 sentences produced by two female and two male speakers. A subset of five vowels was selected for this initial study. Measurements of the first three formant frequencies and the amplitudes of the spectral prominences for these formants were made at 10-millisecond intervals throughout each vowel. In agreement with other studies, identification of vowels based only on formant frequencies is subject to errors. However, the data show that the front-back distinction for vowels can be achieved with a low error rate based only on the frequency F2 of the second formant, independent of gender of the speaker. Acoustic theory, together with anatomical data on vocal-tract dimensions for female and male speakers can help to explain this observation. There are, however, some contexts and levels of stress where consonant context must be taken into account in

estimating the feature [back]. Estimation of the features [high] and [low], however, shows considerable overlap. The data show that the context influencing vowel height includes both the adjacent consonants and the height of the vowels in the adjacent syllables. Further measurements with a larger database of utterances are needed to develop the parameters of a module that estimates vowel height.

4.5 Identification of place features for stop and nasal consonants

This study reports on measurements of transitions of the second formant (F2) for syllable-initial and syllable-final nasal and stop consonants in English, as cues to place of articulation. F2 values and the amount and direction of F2 changes in a 20 ms interval were determined in vowels adjacent to consonant "implosion" and release. The corpus included consonants in sentences and isolated nonsense syllables. The F2 transitions for a given place of articulation are roughly similar for nasals and stops for different syllable positions, as expected, but some interesting differences exist. Since the stop-consonant bursts influence the spectrum sampling point relative to release, there are some shifts in F2 onset frequencies for stops, as compared to nasals. Syllable-initial alveolars show additional manner differences, which can be attributed to differences in tongue configurations for alveolar nasals relative to stops in back-vowel contexts. These latter shifts are tentatively ascribed to differential coarticulatory effects for /n/ versus /d/ and /t/, differences which might be related to the lack of a distinctive syllable-initial velar nasal in English. There are also systematic shifts in F2 transitions as a function of syllable position, for both nasals and stops.

4.6 Detecting nasals

Detection of certain discontinuities in the speech signal indicates the presence of a consonant closure or release. In the current lexical access model, these landmarks are further classified as being nonsonorant or sonorant. Sonorant consonants include nasal and lateral consonants. In certain phonetic environments, a nasal murmur may not be present, and consequently a simple sonorant landmark detector for consonants may not find such a nasal. Based on acoustic data from a variety of utterances containing nasal consonants, revised methods for detecting nasals have been developed. In English, nasalization in a vowel is a cue that there is an underlying nasal consonant adjacent to the vowel. The revised nasal-detecting methods utilize not only identification of sonorant landmarks but also the detection of nasalization in vowels, based on measures of first-formant prominence and the presence of additional nasal resonances in the spectrum.

4.7 Lexical access from a partial segment/feature representation

One step in the proposed lexical access model is to determine a sequence of words that is consistent with the estimated segments and features that are derived from analysis of the signal. This step involves searching for a match between the estimated segment/feature representation and the segment/feature specifications for words and word sequences in the lexicon. Such a matching procedure has been developed. In its present implementation, each underlying segment in the utterance must be detected, although the labeled features for each segment may be only a partial list (including a minimal specification of a segment as [+consonant] or [+vowel] with no other features). Studies are being carried out to determine the relation between the number and location of the unspecified features, the size of the lexicon, and the number of words, on average, that result from the matching process.

5 Publications

Papers published

Choi, J.-Y. (1999). Detection of consonant voicing: A module for a hierarchical speech recognition system. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge MA.

Dilley, L.C. and S. Shattuck-Hufnagel (1999). Effects of Repeated Intonation Patterns on Perceived Word-Level Organization. In **Proc. 14th International Congress of Phonetics Sciences, (ICPhS'99), Vol. 2,** San Francisco, 1487-1490.

Howitt, A.W. (1999). Vowel Landmark Detection. In **Proc. 6th European Conference on Speech Communication and Technology, Vol. 6**, 2777-2780, Budapest, Hungary.

Manuel, S.Y. and G.C. Wyrick (1999). "Casual speech: A rich source of intriguing puzzles." In **Proc. 14th International Congress of Phonetics Sciences, (ICPhS'99**), **Vol. 1**, San Francisco, 679-682.

Nieto-Castanon, A., and F.H. Guenther (1999). "Constructing speaker-specific articulatory vocal tract models for testing speech motor control hypotheses." In **Proc. 14th International Congress of Phonetics Sciences, (ICPhS'99), Vol. 3,** San Francisco, 2271-2274.

Perkell, J.S., M. Zandipour, M. Matthies, and H. Lane (1999). "Articulatory kinematics: Preliminary data on the effects of speaking condition, articulator and movement type." In **Proc. 14th International Congress of Phonetics Sciences, (ICPhS'99), Vol. 3,** San Francisco, 1773-1776.

Perkell, J., F. Guenther, H. Lane, M. Matthies, P.Perrier, J.Vick, R. Wilhelms-Tricarico, and M. Zandipour, (in press). "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss." To appear in **J. Phonetics**.

Shattuck-Hufnagel, S. and A. Cutler (1999). "The Prosody of Speech Error Corrections Revisited." In **Proc. 14th International Congress of Phonetics Sciences, (ICPhS'99), Vol. 2,** San Francisco, 1483-1486.

Shattuck-Hufnagel, S. (forthcoming). "Phrase-level Phonology in Speech Production Planning: Evidence for the Role of Prosodic Structure." In Merle Horne, (Ed.), **Prosody: Theory and Experiment: Studies presented to Gosta Bruce**, Stockholm:Kluwer.

Stevens, K.N. (accepted). Diverse Acoustic Cues at Consonantal Landmarks. To appear in a Special Issue of **Phonetica.**

Stevens, K.N., S.Y. Manuel, and M.L. Matthies (1999). Place of articulation for consonants: Comparing nasals and stops, and syllable position. **J. Acoust. Soc. Am., Vol. 106**, No. 4, Pt. 2, 2244, (Abstract).

Stevens, K.N., S.Y. Manuel, and M. Matthies (1999). Revisiting Place of Articulation Measures for Stop Consonants: Implications for Models of Consonant Production. In **Proc. 14th International Congress of Phonetic Sciences, (ICPhS'99) Vol. 2**, San Francisco, 1117-1120.

Turk, A.E. (1999). Structural Influences on Boundary-Related Lengthening in English. In **Proc. 14th International Congress of Phonetics Sciences, (ICPhS'99), Vol. 1,** San Francisco, 237-240.

Turk, A.E., and Shattuck-Hufnagel, S. (to appear). "Duration as a Cue to Syllable Affiliation." In **Proc. of Conference on the Phonological Word**, Berlin, October 1997.

Turk, A.E., and Shattuck-Hufnagel, S. (in press). "Word-boundary-related duration patterns in English." To appear in **J. Phonetics**.

Papers to be published

House, A.S. and K.N. Stevens (submitted). "A Longitudinal Study of Speech Production, I: General Findings." J. Acoust. Soc. Am.

House, A.S. and K.N. Stevens (submitted). "A Longitudinal Study of Speech Production, II: Vocalic Characteristics." J. Acoust. Soc. Am.

House, A.S. and K.N. Stevens (submitted). "A Longitudinal Study of Speech Production, III: Stop, Fricative and Nasal Consonants." J. Acoust. Soc. Am.

Kwong, K.W. and K.N. Stevens (submitted). "On the Voiced-Voiceless Distinction for Writer/Rider." J. Phonetics.

Lane, H., Matthies, M.L., Perkell, J.S., Vick, J., and Zandipour, M. (submitted). "The effects of changes in hearing status in cochlear implant users on the acoustic vowel space and coarticulation." J. Speech. Lang. Hear. Res.

Massey, N. and K.N. Stevens (submitted). "Transients at Stop-Consonant Releases." J. Acoust. Soc. Am.

Matthies, M., Perrier, P., Perkell, J. and Zandipour, M. (submitted). "Variation in Speech Movement Kinematics and Temporal Patterns of Coarticulation with Changes in Clarity and Rate." J. Speech, Lang. Hear. Res.

Perkell, J., Zandipour, M., Matthies, M. and Lane, H. (submitted). "Clarity versus Economy of Effort in Speech Production: A Preliminary Study of Inter-Subject Differences and Modeling Issues." J. Acoust. Soc. Am.

Perkell, J., and Zandipour, M. (submitted). "Clarity versus Economy of Effort in Speech Production: Kinematic Performance Spaces for Cyclical and Speech Movements." J. Acoust. Soc. Am.

Perkell, J.S., and F.H. Guenther (submitted). "Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models." J. Phonetics.