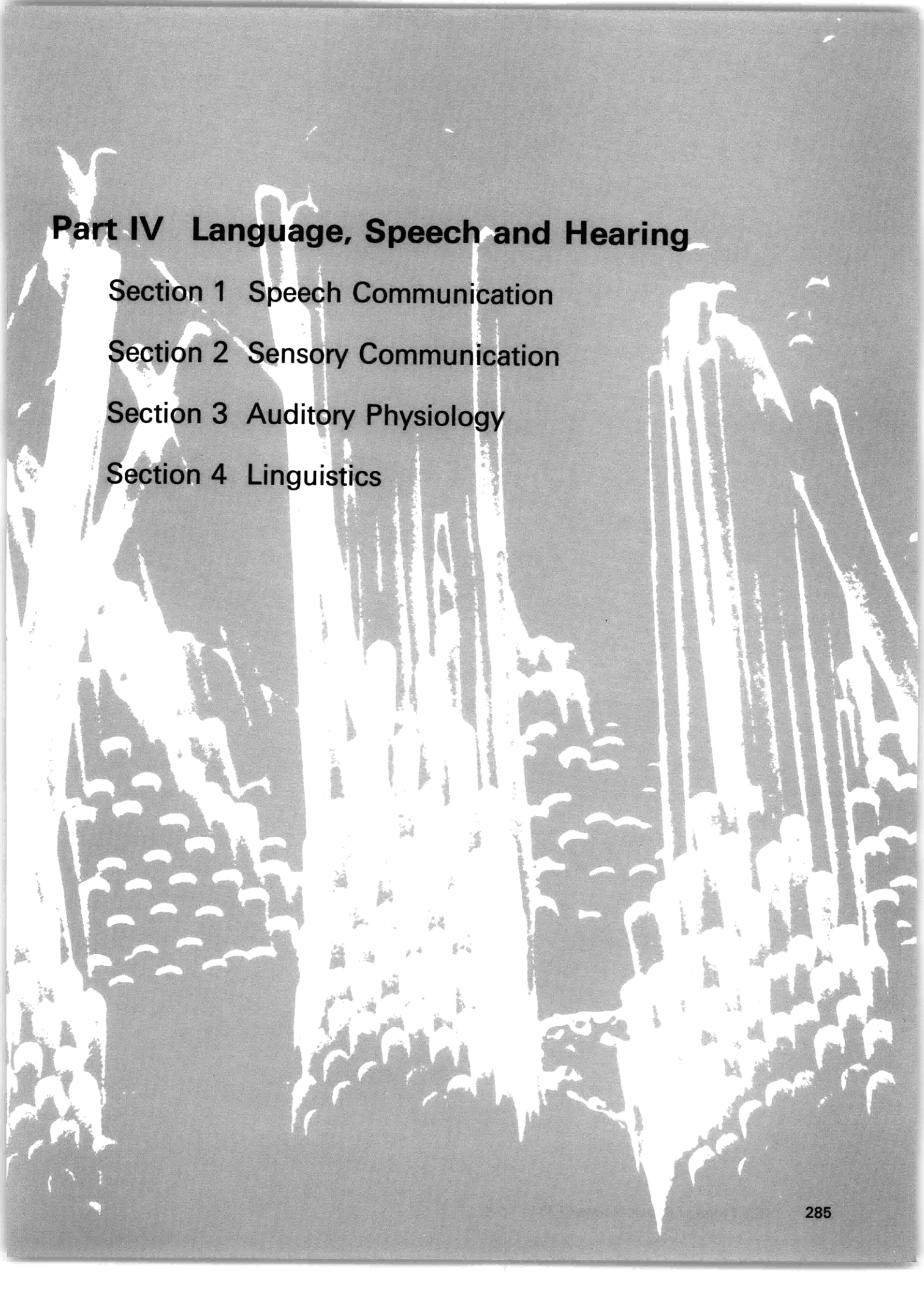

Part IV Language, Speech and Hearing

Section 1 Speech Communication

Chapter 1 Speech Communication

Chapter 1. Speech Communication

Academic and Research Staff

Professor Kenneth N. Stevens, Professor Jonathan Allen, Professor Morris Halle, Professor Samuel J. Keyser, Dr. Corine A. Bickley, Dr. Suzanne E. Boyce, Dr. Carol Y. Espy-Wilson, Dr. Marie K. Huffman, Dr. Michel T. Jackson, Dr. Melanie L. Matthies, Dr. Joseph S. Perkell, Dr. Mark A. Randolph, Dr. Stefanie R. Shattuck-Hufnagel, Dr. Mario A. Svirsky, Dr. Victor W. Zue

Visiting Scientists and Research Affiliates

Giulia Arman-Nassi,[1] Dr. Richard S. Goldhor,[2] Dr. Robert E. Hillman,[3] Dr. Jeannette D. Hoit,[4] Eva B. Holmberg,[5] Dr. Tetsuo Ichikawa,[6] Dr. Harlan L. Lane,[7] Dr. John L. Locke,[8] Dr. John I. Makhoul,[9] Dr. Carol C. Ringo,[10] Dr. Noriko Suzuki,[11] Jane W. Webster[12]

Graduate Students

Abeer A. Alwan, Marilyn Y. Chen, Helen M. Hanson, Andrew W. Howitt, Caroline B. Huang, Lorin F. Wilde

Undergraduate Students

Anita Rajan, Lorraine Sandford, Veena Trehan, Monnica Williams

Technical and Support Staff

Ann F. Forestell, Seth M. Hall, D. Keith North, Arlene Wint

1.1 Introduction

Sponsors

C.J. Lebel Fellowship
Dennis Klatt Memorial Fund
Digital Equipment Corporation

---

[1] Milan Research Consortium, Milan, Italy.

[2] Audiofile, Inc., Lexington, Massachusetts.

[3] Boston University, Boston, Massachusetts.

[4] Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona.

[5] MIT and Department of Speech Disorders, Boston University, Boston, Massachusetts.

[6] Tokushima University, Tokushima, Japan.

[7] Department of Psychology, Northeastern University, Boston, Massachusetts.

[8] Massachusetts General Hospital, Boston, Massachusetts.

[9] Bolt Beranek and Newman, Inc., Cambridge, Massachusetts.

[10] University of New Hampshire, Durham, New Hampshire.

[11] First Department of Oral Surgery, School of Dentistry, Showa University, Tokyo, Japan.

[12] Massachusetts Eye and Ear Infirmary, Boston, Massachusetts.

National Institutes of Health
Grants T32 DC00005, 5-R01 DC00075, F32 DC00015, S15 NS28048, R01 NS21183,[13] P01 NS23734,[14] and 1-R01 DC00776
National Science Foundation
Grants IRI 88-05680[13] and IRI 89-10561

The overall objective of our research in speech communication is to gain an understanding of the processes whereby (1) a speaker transforms a discrete linguistic representation of an utterance into an acoustic signal, and (2) a listener decodes the acoustic signal to retrieve the linguistic representation. The research includes development of models for speech production, speech perception, and lexical access, as well as studies of impaired speech communication.

1.2 Models, Theory, and Data in Speech Physiology

1.2.1 Comments on Speech Production Models

A paper on speech production models has been prepared for the 12th International Congress of Phonetic Sciences (Aix-en-Provence, France, August 1991). The paper discusses modeling which covers the entire transformation from a linguistic-like input to a sound output. Such modeling can make two kinds of contribution to our understanding of speech production: (1) forcing theories of speech production to be stated explicitly and (2) serving as an organizing framework for a focused program of experimentation. As an example, a "task dynamic" production model is cited. The model incorporates underlying phonological primitives that consist of "abstract articulatory gestures." Initial efforts have been made to use the model in interpreting experimental data. Several issues that arise in such modeling are discussed: the validity of the underlying theory, incorporating realistic simulations of peripheral constraints, accounting for timing, difficulties in using models to evaluate experimental data, and the use of neural networks. Suggestions for an alternative modeling approach center around the need to include the influence of speech acoustics and perceptual mechanisms on strategies of speech production.

1.2.2 Models for Articulatory Instantiation of Feature Values

Data from previous studies of speech movements have been used as a basis for several theoretical papers focusing on models for the articulatory instantiation of particular distinctive feature values. Two of the papers (prepared in collaboration with colleagues from Haskins Laboratories) address models of coarticulation for the features +/- ROUND and +/- NASAL. In both papers, we review conflicting reports in the literature on the extent of anticipatory coarticulation and present evidence to suggest that these conflicts can be resolved by assuming that segments not normally associated with a rounding or nasalization feature show stable articulatory patterns associated with those features. In one of the papers, we point out that these patterns may have been previously interpreted as anticipatory coarticulation of rounding or nasality. We further argue that timing and suprasegmental variables can be used to show that this is the case and that the coproduction model of coarticulation can best explain the data. In the other paper, we use similar data to argue against the "targets and (interpolated) connections" model proposed by Keating. In a third paper on this general topic, coarticulatory patterns of rounding in Turkish and English are compared and evidence is given to show that their modes of articulatory organization may be different.

1.2.3 Modeling Midsagittal Tongue Shapes

Two projects on articulatory modeling are in progress. The first is concerned with the construction of a quantitative, cross-linguistic articulatory model of vowel production, based in cineradiographic data from Akan, Arabic, Chinese, and French. Programs for displaying, processing, and partially automating the process of taking measurements from cineradiographic data have been developed for the purpose of gathering data for this work. The second project is concerned with quantitatively testing the generalizability of this model to similar data gathered from speakers of English, Icelandic, Spanish, and Swedish.

Future work will focus on the degree to which this model generalizes from vowel production to consonant production. Velar and uvular consonants, which, like vowels, are articulated primarily with

[13] Under subcontract to Boston University.

[14] Under subcontract to Massachusetts General Hospital.

the tongue body, should be able to be generated by the model with no additional parameters; it is an open question whether or not the same will be true of other consonants such as pharyngeals and palatals, it almost certainly will not be true of coronals.

1.3 Speech Synthesis

Research on speech synthesis has concentrated for the most part on developing improved procedures for the synthesis of consonant-vowel and vowel-consonant syllables. With the expanded capabilities of the KLSYN88 synthesizer developed by Dennis Klatt, it is possible to synthesize utterances that simulate in a more natural way the processes involved in human speech production. The details of these procedures are based on theoretical and acoustical studies of the production of various classes of speech sounds, so that the manipulation of the synthesis parameters is in accord with the mechanical, aerodynamic, and acoustic constraints inherent in natural speech production.

For example, using the improved glottal source, including appropriately shaped aspiration noise, procedures for synthesizing the transitions in source characteristics between vowels and voiceless consonants and for synthesizing voicing in obstruent consonants have been specified. Methods for manipulating the frequencies of a pole-zero pair in the cascade branch of the synthesizer have led to the synthesis of nasal, lateral, and retroflex consonants with spectral and temporal characteristics that match closely the characteristics of naturally-produced syllables. The synthesis work is leading to an inventory of synthetic consonant-vowel and vowel-consonant syllables with precisely specified acoustic characteristics. These syllables will be used for testing the phonetic discrimination and identification capabilities of an aphasic population that is being studied by colleagues at Massachusetts General Hospital.

An additional facet of the synthesis research is the development of ways of characterizing different male and female voices by manipulating the synthesizer parameters, particularly the waveform of the glottal source, including the noise component of this waveform. We have found informally that it is possible to reproduce the spectral properties of a variety of different male and female voices by adjustment of these parameters.

1.4 Speech Production of Cochlear Implant Patients

1.4.1 Longitudinal Changes in Vowel Acoustics in Cochlear Implant Patients

Acoustic parameters were measured for vowels spoken in /hVd/ context by three postlingually-deafened cochlear implant recipients. Two female subjects became deaf in adulthood; the male subject, in childhood. Longitudinal recordings were made before and at intervals following processor activation. The measured parameters included formant frequencies, *F*0, *SPL*, duration and amplitude difference between the first two harmonic peaks in the log magnitude spectrum (*H*1 - *H*2). A number of changes were observed from pre- to post-implant with inter-subject differences. The male subject showed changes in *F*1, vowel durations and *F*0. These changes were consistent with one another; however, they were not necessarily in the direction of normalcy. On the other hand, the female subjects showed changes in *F*2, vowel durations, *F*0 and *SPL* which were in the direction of normal values and, for some parameters, tended to enhance phonemic contrasts. For all three subjects, *H*1 - *H*2 changed in a direction which was consistent with previously-made flow measurements. Similar data from additional subjects are currently being analyzed.

1.4.2 Effects on Vowel Production and Speech Breathing of Interrupting Stimulation from the Speech Processor of a Cochlear Prosthesis

We have completed a study on the short-term changes in speech breathing and speech acoustics caused by 24 hours of processor deactivation and by brief periods of processor activation and deactivation during a single experimental session in cochlear implant users. We measured the same acoustic parameters employed in the above-described longitudinal study of vowel production and the same physiological parameters used in a longitudinal study of changes in speech breathing.

We found significant and asymmetric effects of turning the processor on and off for a number of speech parameters. Parameter changes induced by short-term changes in the state of the processor were consistent with longitudinal changes for two of the three subjects. Most changes were in the direction of enhancing phonetic contrasts when

the processor was on. Although each subject behaved differently, within a subject the pattern of changes across vowels was generally consistent. Certain parameters, such as breathiness and airflow proved to be closely coupled in their changes, as in the longitudinal studies. (This coupling occurred even in the subject whose responses to processor activation where inconsistent with the long-term changes obtained in the longitudinal study.) The greater responsiveness of speech parameters to the onset of processor activation than to its offset supports the hypothesis that auditory stimulation has a "calibrational" role in speech production, retuning pre-existing articulatory routines. Many of the findings thus far implicate control of the larynx as a major focus of recalibration, but other speech production parameters need to be analyzed. As suggested by certain findings about parameter relations in the longitudinal study, we expect that as we progress in our understanding of those relations, we will be able to give a principled account of many of the apparent differences we observed among these initial subjects.

1.4.3 Methods for Longitudinal Measures of Speech Nasality in Cochlear Implant Patients

We have been examining the validity of longitudinal measures of nasality in cochlear implant patients, based on acoustic spectra, sound levels and outputs of nasal and throat accelerometers. Speech materials consist of isolated utterances and reading passages. Preliminary observations indicate the following: The ratio of RMS values of nasal and throat accelerometer outputs (a technique used in published experiments) may be influenced by: (1) variation in the relative levels of the two signals during the recommended calibration maneuver, i.e., production of a sustained /m/; and (2) substantial changes in SPL that accompany onset of "auditory" stimulation from a cochlear prosthesis. These observations raise uncertainty about using the throat accelerometer output as a reference and the sensitivity of this kind of measure to longitudinal changes in nasality across experimental sessions. In addition, measures of harmonic and formant amplitudes from acoustic spectra may be confounded by changes in coupling to tracheal resonances that also accompany the activation of the prosthesis. These observations and additional measures and calibration strategies are being explored further.

1.5 Phonatory Function Associated with Misuse of the Vocal Mechanism

1.5.1 Vocal Nodules in Females

In this study, we used our methods for obtaining measurements from the inverse filtered glottal waveform, electroglottographic signal (EGG), and average transglottal pressure and glottal airflow to compare phonatory function for a group of twelve female bilateral vocal-fold nodules patients with a control group of age-matched normal females. Statistical analysis of the data involved analysis of variance (ANOVA) and analysis of covariance (ANCOVA with SPL the covariate) procedures to test for significant differences between the groups. Standard and regressed Z scores were also used to compare measurements for individual nodules subjects with the normal group.

The largest number of significant differences between the nodules and normal group were found for the loud voice condition, with normal voice having the second largest number of significant differences and soft voice the smallest number of differences. A majority of the significant differences between nodules and normal subjects were associated with higher SPL used by the nodules group. Increased SPL in the nodules group was accompanied, most notably, by increases in parameters. These are believed to reflect greater potential for vocal-fold trauma (i.e., increased transglottal pressure, AC (modulated) glottal flow, peak glottal flow and maximum rate of glottal flow declination) due to high vocal-fold closure velocities and collision forces. In addition, an examination of individual Z-score profiles revealed that three of the twelve nodules patients displayed instances in which SPL was within normal limits, but AC flow and/or maximum flow declination rate was abnormally high. In addition, there were four nodules patients who displayed instances in which SPL, AC flow and maximum flow declination rate were all abnormally high, but AC flow and/or maximum flow declination rate was proportionally higher than the SPL.

In summary, the results suggest that for most female nodules patients, abnormalities in vocal function are primarily associated with use of excessive loudness. Thus, for patients who fit this profile, simply reducing loudness through direct facilitation and/or vocal hygiene training may represent appropriate treatment. However, there appears to be a smaller proportion of female nodules patients whose abnormalities in vocal function are not completely accounted for by excessive loudness. These patients may be dis-

playing an underlying "pattern" of vocal hyperfunction which could require additional therapeutic procedures to ameliorate.

1.5.2 Speech Respiration Associated with Vocal Hyperfunction: a Pilot Study

This pilot study compared respiratory function in an adult female nodules patient with a sex, age and body-type matched normal control. Inductance plethysmography (a Respitrace) was used to obtain a variety of speech and non-speech respiratory measures at normal and "loud" vocal intensity.

The non-speech respiratory performance of both subjects was within normal limits. The matched control also displayed normal values for measures of speech respiration. In contrast, the subject with vocal-fold nodules showed clear evidence of deviant speech respiration during oral reading such as abnormally low inspiratory volumes, with the majority of expiratory limbs for speech being initiated at levels within the quiet tidal volume range. During loud reading, the nodules patient initiated speech at lung volumes that were equal to or less than those used at normal loudness. This result is quite unlike normal speakers who tend to inhale to higher lung volumes in preparation for louder utterances. The nodules patient also displayed consistent, marked encroachment into the expiratory reserve volume during reading at normal loudness, with even further encroachment during loud reading. This pattern is deviant from normal speakers who tend to terminate utterances close to the functional residual capacity. Taken together, the results for the nodules subject reflect reduced efficiency of speech respiratory function. This patient produced alveolar pressures required for speech by expending greater expiratory muscular effort, instead of inhaling like normal speakers to larger lung volumes, thereby taking advantage of greater natural recoil forces to assist in generating required pressures. The nodules patient also tended to stop at inappropriate (non-juncture) points within sentences to inhale and evidenced loss of substantial volumes of air prior to the initiation of phonation.

Thus, the results indicate that abnormalities in speech respiration can occur in a patient with a hyperfunctionally-related voice disorder (nodules). This finding will be pursued in future studies of additional patients.

1.5.3 Comparisons Between Inter- and Intra-speaker Variation in Aerodynamic and Acoustic Parameters of Voice Production

In this study, intraspeaker variation of non-invasive aerodynamic and acoustic measurements of voice production was examined across three recordings for three normal female and three normal male speakers. Data from one recording for each of fifteen females and fifteen males served as normative references. The following measurements were made for productions of sequences of the syllable /pæ/ in soft, normal and loud voice and low and high pitch: the inverse filtered airflow waveform average transglottal air pressure and glottal airflow, and the amplitude difference between the two first harmonics of the acoustic spectrum. Linear regression analyses between individual values of SPL and each of the other parameters were performed for data pooled across the three recordings. In addition, comparisons between the individual subjects and the reference groups were performed using Z-score analysis.

Preliminary results showed that a small amount of SPL variation could be accompanied by large variation in other parameters. For some parameters (e.g., maximum flow declination rate and AC flow), both inter- and intra-speaker variation were significantly related to SPL. For such parameters the effects of SPL can be removed statistically, and comparisons between individual subjects and group data can readily be made. For other parameters, which showed large inter-speaker variation not significantly related SPL, intra-speaker variation for individual subjects could be orderly related to SPL (e.g., for DC flow offset). For such parameters, data from several baseline recordings can be useful to establish to what extent parameters—for which both inter- and intra-speaker variation was large and not significantly related to SPL (e.g., average flow)—were considered less useful as indices of abnormal vocal function. These data suggest that in order to quantify normal variation, both inter- and intra-speaker variation should be considered.

1.6 Studies of Acoustics and Perception of Speech Sounds

1.6.1 Vowels

Previous reports have described an acoustic study of vowels in which the effects of consonantal context, lexical stress, and speech style were compared using the same database of utterances. That study, which had been extended, was completed during the past year and included perceptual experiments in which listeners identified vowels excised from the utterances in the database. The completed data compilation verifies that consonant context affects the vowel midpoints more than lexical stress and speech style. The direction and magnitude of the formant frequency shifts were consistent with findings of previous studies. The liquid and glide contexts, /w/, /r/, and /l/, lowered the *F*2 frequency of front vowels, especially lax front vowels, on the order of one Bark relative to the *F*2 frequencies when the same vowels are adjacent to stop consonants. Shifts for *F*1 tended to be smaller than shifts in *F*2, even on a Bark scale, and were less consistent across speakers.

The formant frequency midpoints and durations of vowels carrying primary stress were shown to differ only slightly on average from those vowels carrying secondary stress, if the other factors were held constant. Vowels in speech read continuously also differed only slightly on the average from vowels in spontaneous speech.

In general, the data show that variations in vowel midpoint formant frequencies, durations, and trajectory shapes are correlated with the perception of the vowel by human listeners. For example, /ɛ/ tokens which have *F*1 - *F*2 midpoint values typical of /ʌ/ tend to be identified as /ʌ/, and /e/ tokens which are short and lack a /y/ offglide, typical characteristics of lax vowels, tend to be misidentified as lax vowels.

Aspects of the trajectories which are important for characterizing the vowel were sought. The trajectory was used to derive a representation of the vowel by one point per formant, a modified "midpoint." Performance by a Gaussian classifier was the criterion used to evaluate different representations of the vowels. If the effect of perceptual overshoot for *F*2 and perceptual averaging for *F*1 was simulated, and the resulting modified midpoint was used as an input for the classifier, performance was somewhat better than if the durational midpoints were used as input. However, the best performance was achieved if the raw data—the quarter-point, midpoint, and three-quarter point of the trajectory and the duration—were used as input to the classifier. The improved performance with the raw data over the modified midpoints shows that not all of the significant aspects of the trajectory have been captured in a one-point representation. It may be that a new one-point representation could be found which would result in as high performance as the raw data. Alternatively, it may be necessary to use more than a modified midpoint to fully characterize a vowel.

Of all the representations used as input to the statistical classifier, the raw data also result in the best agreement of the classifier with the human performance. If the classifier is also allowed to train and test on vowels in stop and liquid-glide contexts separately, agreement with the listeners' responses (and performance in the conventional sense, i.e., agreement with the transcriber's phonemic labels) improves further. The improvement due to separating the contexts suggests that humans perform vowel identification in a context-dependent manner.

1.6.2 Analysis and Modeling of Stop and Affricate Consonants

We have been developing models for the production of various types of stop consonants including voiceless aspirated stops, voiced stops, affricates, and prenasal stops. In all of these models, estimates are made of the changes in cross-sectional area of the airways as the various articulators (glottis, velopharyngeal opening, supraglottal constriction, pharyngeal expansion or contraction) are manipulated to produce the different classes of consonants. From these estimates, aerodynamic parameters such as flows and pressures are calculated, and the time course of different acoustic sources at the glottis and near the supraglottal constriction are then determined. The modification of the sources by the vocal-tract transfer function is included to obtain predictions of the spectral changes in the radiated sound for different phases of the consonant production. The predictions are then compared with acoustic measurements of utterances produced by several speakers. In some cases where there are discrepancies between the acoustic data and the predictions, revised estimates of the model parameters are made. This process leads to reasonable agreement between the acoustic measurements and the predictions and provides a way of estimating the time course of the articulatory changes from the acoustic data.

In the case of voiceless aspirated stop consonants, the model includes a glottal spreading maneuver and manipulation of the pharyngeal walls to inhibit expansion of the vocal-tract volume during the

consonantal closure interval. The model generates a quantitative description of the sequence of acoustic events following the consonantal release: an initial transient, an interval of frication noise (at the supraglottal constriction), an interval of aspiration noise (at the glottis), a time in which there is breathy vocal-fold vibration, and, finally, modal vocal-fold vibration. For a voiced stop consonant, active pharyngeal volume expansion is included in the model, and the increase in intraoral pressure is thereby delayed.

The model for the production of the affricate /č/ in English specifies four components: (1) an initial release phase for the tongue tip, (2) a relatively slow downward movement of the tip (30-50 ms), (3) a steady palato-alveolar fricative portion, and (4) a release of the fricative portion into the vowel. From the model it is possible to calculate the amplitude, spectrum, and time course of the components of the sound resulting from these movements, as well as the airflow. The acoustic and aerodynamic predictions are on the range observed in actual productions of affricates. In the initial 30-50 ms there is a brief transient (1-2 ms), followed by a gradually rising noise amplitude, with a spectral prominence corresponding to the acoustic cavity in front of the tongue tip. During the later fricative phase, there is acoustic coupling to the palatal channel, leading to a second major prominence in the spectrum. The single prominence in the initial phases and the two prominences in the later frication spectrum achieve amplitudes that are comparable to or greater than the corresponding spectral prominences in the adjacent vowel. The sequence of acoustic events cannot be described simply a /t/ followed by /š/ or as a /š/ with an abrupt onset, but is unique to the affricate.

1.6.3 Fricative Consonants: Modeling, Acoustics, and Perception

We have continued our modeling studies of the production of fricative consonants with an analysis of the aerodynamic and acoustic processes involved in the production of these consonants and have completed an investigation of the voicing distinction for fricatives. The fricative studies reported previously have been extended to include quantitative estimates of the amplitudes and locations of turbulence noise sources at the glottis and at the supraglottal constriction for various places of articulation. Reasonable agreement has been obtained between acoustic characteristics of fricatives in intervocalic position. The acoustic and perceptual studies of voicing in fricatives (in collaboration with Sheila Blumstein of Brown University) have led to a delineation of the roles of the extent of formant transitions and the presence of glottal vibration at the edges of the obstruent interval in determining the voicing status of intervocalic fricatives. Data have also been obtained for fricative sequences with mixed voicing (e.g., the /zf/ sequence in **his form**), and have shown the predominant role of regressive voicing assimilation relative to progressive assimilation.

1.6.4 Acoustic Properties of Devoiced Semivowels

When the semivowels /w,y,r,l/ occur in clusters with unvoiced consonants, they are sometimes at least partially devoiced so that much of the information about their features is in the devoiced region. While the acoustic characteristics of the semivowels which are manifest as sonorants are fairly well understood, little research has been conducted on those semivowels which are fricated. In this acoustic study, we recorded groups of words such as "keen," "clean," "queen," and "cream" at a slow and fast rate by two speakers, one male and one female. We are investigating the spectral characteristics of the word-initial consonants and consonant clusters to determine the attributes in this region which signal the presence of a semivowel and the attributes which distinguish among the devoiced semivowels. So far, our findings suggest that the spectral characteristics of the consonant clusters can change considerably over time due to the presence of the semivowel. Data are being collected to quantify changes in the spectral characteristics within the consonant cluster from an initial frication burst to noise generated at the constriction for the semivowel to voicing as the semivowel constriction is released. For example, one notable time-varying property of the [sw] cluster in "sweet" is the change from a rising spectral shape where the main concentration of energy is between 4-5 kHz (at the beginning of the frication noise) to a falling spectral shape where the main concentration of energy is between 1-2 kHz (towards the end of the frication noise), as the place of noise generation shifts from a postdental to a velar location.

1.6.5 Spreading of Retroflexion in American English

In this acoustic study we investigated the factors which influence the spreading of retroflexion from a postvocalic /r/ into the preceding vowel, /ɑ/, in American English. We consider the acoustic correlate for the feature "retroflex" to be a low third formant (*F*3) which is close in frequency to the second formant (*F*2). Minimal pair words with initial and final obstruent consonants were recorded by six speakers at a slow and a fast

speaking rate. Several basic *F*3 trajectories could be observed during the /*ar*/ region, depending upon the timing of the minimum in the *F*3 trajectory in relation to the beginning and ending of the sonorant region. Given the same word said by different speakers at different speaking rates, the minimum in the *F*3 trajectory can occur at the beginning, in the middle and at the end of the sonorant region. In addition, *F*3 can remain fairly flat at a low frequency throughout the sonorant region. We believe that this variability in the *F*3 trajectory can be explained by several factors. First, there is a certain basic *F*3 trajectory needed to articulate a postvocalic /r/: a downward movement from the *F*3 frequency in the previous sound and an upward movement to either a neutral position or the *F*3 position for the following sound. The *F*3 slopes on either side of the minimum will depend upon the context. Second, only a portion of this full *F*3 trajectory may be observable during the sonorant region. How much and what portion of the *F*3 trajectory can be observed will depend upon the duration of the sonorant region and on how early the speaker starts to produce the /r/. The anticipation of /r/ as well as the duration of the sonorant region appear to depend on speaking rate, consonantal context and speaker differences. Further analysis is planned to determine more accurately the minimum time needed to execute the articulation of /r/.

1.6.6 Studies of Unstressed Syllables

We have begun an investigation of the acoustic properties of the vowels and consonants in unstressed syllables in utterances produced with different speech styles. The aim of this study is to determine the kinds of reductions that can occur in unstressed syllables and to attempt to establish what aspects of the syllables are retained in spite of the reductions. As a first step in this project, we are examining vowels and consonants in unstressed syllables in which the consonants preceding and following the vowel are both voiceless (as in **sup**port, cla**ssical, pot**ato). These are situations in which the vowel can become devoiced, but in which some remnant of the vowel remains in voiceless form, or in which temporal properties of the surrounding segments signal the presence of the syllable.

When there is severe reduction of the vowel, the presence of the unstressed syllable appears to be signaled in various ways: (1) the vowel is devoiced, but there is an interval in which the presence of a more open vocal-tract configuration between the consonants is signaled by aspiration noise; (2) in a word like **support,** when the vowel is devoiced, the frication (and possibly aspiration) noise within the initial /s/ shows an increased amplitude toward the end of the consonant, indicating that the consonant is being released into a more open vocal-tract configuration or that the consonant itself is more open toward the end of its duration; (3) properties of the adjacent consonant (such as aspiration in /p/ in **support** or in /k/ in **classical**) indicate that the consonant sequence could not have arisen unless an unstressed vowel were inserted between the consonants. These properties have been observed in a preliminary study, and a larger database of utterances is being prepared to conduct a more detailed investigation.

1.6.7 Nasalization in English

Nasalization is one of the sound properties, or features, languages use to distinguish words. Understanding the acoustic consequences of nasalization is therefore necessary for understanding how nasal sounds pattern in natural languages, and also for developing models of speech acoustics for use in speech synthesis and recognition. We have been investigating the important spectral consequences of nasalization using analysis of natural speech, and we have carried out perception experiments using both natural and synthetic speech. One set of perceptual experiments investigated how readily listeners can use nasalization on a (natural) vowel to guess the nasality of a following consonant (bead vs. bean). Our results indicate that English speakers notice nasality on a vowel more readily when the experimental task focuses their attention on it. This observation may be because they are not expecting nasalization on vowels, since in English vowels are nasalized only incidentally, before a nasal consonant, rather than being nasalized contrastively, as in French. One of the most consistent spectral effects of nasalization is damping of the first formant. Our analyses of natural speech items indicate that *F*1 prominence decreases with increasing opening of the velopharyngeal port. Perceptual experiments with synthetic speech indicate that even in the absence of other spectral cues to nasalization, a decrease in *F*1 prominence increases the likelihood that a vowel will be heard as nasal. This effect is stronger when *F*1 prominence decreases over the course of the vowel, rather than remaining static, suggesting that a decrease in *F*1 prominence over time is an essential part of the spectral pattern which is heard as nasal. However, it could also be simply due to the fact that time-varying synthetic items sound more natural to the listener than do static items. To choose between these alternative explanations, we plan to run similar experiments with listeners whose native language has contrastively nasalized vowels, in which the spectral cues to nasalization change little over the

course of the vowel. A somewhat less consistent spectral effect of nasalization is the presence of a visible zero in the spectral valley between *F*1 and *F*2. Our perceptual studies with natural speech items suggest that the zero can be an important factor in determining when a vowel is heard as nasal. One may infer from this result that a measure of relative spectral balance in the region between *F*1 and *F*2 may correlate well with independent indications of perceived nasality. This measure is currently under development. The results of this research, along with work reported previously, should lead to an improved understanding of the use of nasalization in natural speech.

1.6.8 Modeling Speech Perception in Noise: A Case Study of Prevocalic Stop Consonants

The present study is part of an attempt to model speech perception in noise by integrating knowledge of the acoustic cues for phonetic distinctions with theories of auditory masking. The methodology in the study is twofold: (1) modeling speech perception in noise based on results of auditory-masking theory and (2) evaluating the theoretical predictions by conducting a series of perceptual experiments in which particular acoustic attributes that contribute to several specific phonetic distinctions are selectively masked. The model predicts the level and spectrum of the masker needed to mask out important acoustic attributes in the speech signal and predicts listeners' responses to noisy speech signals.

We have been focusing on the importance of the *F*2 trajectory in signaling the place of articulation for the consonants /b,d/ in /CV/ syllables with the vowels /*a*/ and /ɛ/ In the /C*a*/ context, the *F*2 trajectory from /d/ to the vowel falls, whereas the *F*2 trajectory in /b*a*/ syllables rises into the vowel and is less extensive than the /d*a*/ trajectory. The reverse is true for /Cɛ/ syllables: the *F*2 trajectory is less extensive for /dɛ/ than it is for /bɛ/, i.e., there is almost no transition in *F*2 from the consonant to the vowel in /dɛ/.

Two perceptual experiments using synthetic /CV/ syllables were conducted. /C/ was either /b/ or /d/; the first set of experiments used the vowel /*a*/, and the second set used the vowel /bɛ/. The synthetic pairs (/b*a*, d*a*/) and (/bɛ/, /dɛ/). were identical except for differences in the *F*2 trajectories (as described above). The utterances were degraded by adding various levels of white noise, and were presented to subjects in identification tests. Results show that for /C*a*/ stimuli with most of the *F*2 transition masked—yielding to the perception of an almost "flat" trajectory— the listeners label the /d*a*/ stimuli as /b*a*/. The reverse is true for /Cɛ/ syllables: when most of the *F*2 transition is masked leaving a flat trajectory appropriate for /dɛ/, the listeners label the /bɛ/ stimuli as /dɛ/. These results indicate that the shape of the *F*2 trajectory is crucial in signaling the place of articulation for these consonants in noise. Results also show that the thresholds where confusion in place of articulation occur can be estimated successfully from masking theory.

In future experiments we will examine listeners' responses to speech signals degraded by shaped noise, rather than white noise, in addition to using a larger set of consonants.

1.7 Speech Production Planning

Our work in speech planning continued the analysis of patterns in segmental speech errors, which provide evidence for models of the phonological planning process. For example, when two target segments in an utterance exchange, as in "pat fig" for "fat pig," they usually share position in both their words and stressed syllables. Since this constraint is widespread, it suggests that one or both of these larger units plays a role in the segmental planning process. In a series of elicitation experiments, we contrasted stimuli like "parade fad" (where /p/ and /f/ share word position but not stress) and "repeat fad" (where /p/ and /f/ share stress but not word position) with "ripple fad" (where /p/ and /f/ share neither word position nor stress). Results show that word position induces the largest number of errors, stress the next-largest, and pairs of segments that share neither positional factor participate in almost no errors. We conclude that both word structure and syllable stress must be incorporated into models of the representation that speakers use for phonological planning.

We are further pursuing the issue of position constraints on segmental errors and their implications for processing representations by examining an earlier finding that word-final segments participate in more errors when speakers produce lists of words (leap note nap lute), and fewer errors when those words occur in grammatically well-formed phrases (From the leap of the note to the nap of the lute) or in spontaneous speech. One possible account of this observation is that the planning of a more complex syntactic structure somehow protects word-final segments; another is that the presence of intervening words provides the protection, and a third possibility is that the more complex prosodic planning of longer grammatical phrases is

responsible. We are comparing the error patterns in short word lists (e.g., lame run ram Len) with sentences of similar length (Len's ram runs lame), and with longer sentences (But if Len has a ram it can run if it's lame) similar to the phrase condition in the earlier experiments. If syntactic structure alone is the factor that changes the proportion of final-segment errors, then short sentences should be as effective as longer sentences in eliciting the difference.

In a second line of investigation, we are asking how complex prosody affects the process of phonological planning. Our initial experiments determined that speakers using reiterant speech to imitate target sentences (i.e., using repeated instances of the syllable /ma/ to reproduce the prosody of the original) showed that it is more difficult to imitate metrically irregular sentences like "He told us it was too hot to work" than metrically regular sentences like "It was hot so we swam in the pool." Difficulty is measured by whether the number of syllables in the reiterant imitation matches that of the target sentence and by the speaker's estimate of whether the imitation was hard, medium or easy). If prosodic planning and segmental planning interact, then prosodically irregular utterances with normal segment structure should elicit more segmental errors in normally-produced utterances as well. We are currently investigating this possibility.

Finally, in collaboration with Mari Ostendorf of Boston University and Patti Price of the Stanford Research Institute, we have investigated a third issue in production planning: the role of prosodic information (prominence, duration, etc.) in signaling aspects of both the structure and the meaning of utterances. In a study of seven types of syntactic ambiguity, using trained FM-radio speakers, we confirmed earlier (but not unanimous) findings showing that listeners can select the meaning intended by the speaker reliably for some but not all types of bracketing ambiguity. Phonological and acoustic analyses showed that speakers signal these distinctions in part by placement of the boundaries of prosodic constituents and that these boundaries are indicated by such prosodic cues as boundary tones and duration lengthening in the pre-boundary syllable.

1.8 Models Relating Phonetics, Phonology, and Lexical Access

In recent years phonologists have proposed that the features that characterize the segments in the representation of an utterance in memory should be organized in a hierarchical fashion. A geometrical tree-like arrangement for depicting phonological segments has emerged from this proposal. We have been examining whether phonetic considerations might play a role in the development of these new approaches to lexical representation of utterances. Our efforts in this direction are following two different paths: (1) a theoretical path that is attempting to modify or restructure the feature hierarchy derived from phonological considerations, and (2) a more practical path in which we are examining how a lexicon based on these principles might be implemented and how it might be accessed from acoustic data.

Theoretical considerations have led to a separation of features into two classes: (1) articulator-free features that specify the kind of constriction that is formed in the vocal tract, and (2) articulator-bound features indicating what articulators are to be manipulated. The articulator-bound features for a given segment can, in turn, be placed into two categories: (1) features specifying which articulator is implementing the articulator-free feature, and (2) features specifying additional secondary articulators that create a distinctive modulation of the sound pattern. The articulator-free features are manifested in the sound stream as landmarks with distinctive acoustic properties. The articulator-bound features for a given segment are represented in the sound in the vicinity of these landmarks. Details of the feature hierarchy that emerges from this view will appear in papers that are in press or in preparation.

Implementation of a lexicon based on this hierarchical arrangement is beginning, and we are planning the development of a system for automatic extraction of acoustic properties that identify some of the articulator-free and articulator-bound features in running speech.

1.9 Other Research Relating to Special Populations

1.9.1 Breathiness in the Speech of Hearing-impaired Children

One of the more prevalent speech abnormalities in the speech of the hearing-impaired that contributes to reduced intelligibility is excessive breathiness. According to a past study of normal speakers done by D. H. Klatt and L. C. Klatt, *H*1 - *H*2 and aspiration noise are highly correlated to breathiness judgments. (*H*1 and *H*2 are the amplitudes of the first two harmonics, in dB, of the spectrum of a vowel.) A study was done to analyze quantitatively the middle of the vowel of hearing-impaired children and a few normal-hearing children. Acoustic parameters *F*0 (fundamental frequency), *H*1 - *H*2, aspiration noise, and spectral tilt were examined in relation to the vowel perceptual judgments on breathiness done by two experienced phoneticians exposed to the vowel in a word or phrase context. The correlation coefficients were unexpectedly low for all of the acoustic parameters. Also, according to the judges, nasality often accompanied breathiness in the vowels of the hearing-impaired and it was difficult to distinguish the two. To lessen the effect of nasality on breathiness judgments, only vowels with low nasality judgments were analyzed. Although the correlation improved, the change was not large.

The main difference between the study and that of Klatt and Klatt was that the judgments in the Klatt study were done on isolated vowels. Another study was conducted to examine the effect of adjacent segments on breathiness perceptual judgments by presenting to the listeners vowels that were both isolated and imbedded in a word. According to the preliminary study, a large difference was observed in the judgments of imbedded vowels in certain words: more breathiness was detected in the vowel when it was in a word context than when it was isolated. However, more experiments and examination of acoustic events near consonant-vowel boundaries must be done.

Another characteristic of breathiness is the introduction of extra pole-zero pairs due to tracheal coupling. Analysis-by-synthesis was done using KLSYN88 developed by Dennis Klatt. An extra pole-zero pair was often observed around 2 kHz for hearing-impaired speakers but not for normal speakers. However, there was low correlation between prominence of the extra peak and breathiness judgments, suggesting that it may not be possible to use prominence of extra peaks to quantify breathiness. Synthesis was also done in which the extra pole-zero pair was removed. In some cases, the breathiness perceptual judgments were lower than the vowels synthesized with the pole-zero pair, although in most cases there was not much difference.

1.9.2 Speech Planning Capabilities in Children

A project at Emerson College is studying the processes of speech encoding in children both with and without language impairment. One component of that project involves collaboration with our Speech Communications group at RLE. The purpose of this research is to examine the two populations for differences in ability to encode phonological information and translate it into a motor program. We have designed a test of children's word repetition skills that will give insight into their speech planning capabilities, and this test is being run on the two populations of children.

1.9.3 Training the /r/-/l/ Distinction for Japanese Speakers

For adult second-language learners, certain phoneme contrasts are particularly difficult to master. A classic example is the difficulty native Japanese speakers have with the English /r/-/l/ distinction. Even when these speakers learn to produce /r/ and /l/ correctly, their ability to perceive the distinction is sometimes at chance.

In conjunction with the Athena second language learning project, we developed a computer-aided training program to assist in teaching the /r/-/l/ perceptual distinction. The basis of the training is the use of multiple natural tokens and the ability to randomly select tokens and to obtain feedback after presentation. While training was slow, it was effective for a majority of subjects. Many reached 100 percent correct identification when tested after 30-50 sessions.

As another component of this project, a fast spectrograph display was developed as an aid to teaching production of /r/ and /l/.

1.10 Facilities

1.10.1 Articulatory Movement Transduction

Development and testing of our system for Electro-Magnetic Midsagittal Articulometry has been completed. The final stages of this effort included: revising the electronics to eliminate two sources of transduction error, development of a more efficient transducer calibration method and construction of an apparatus to implement the method, improvement of transducer connecting wire to overcome insulation failures, updating and elaboration of signal processing software, development of a new computer program to display and extract synchronous data from articulatory and acoustic signals, and running a variety of tests to validate system performance. These tests indicate that performance is now acceptably close to its theoretical limit. Two trial experiments have been run with subjects. Those experiments have resulted in several important improvements to experimental methods, including techniques for attaching transducers to subjects and presenting utterance materials. Preparations have been made to run experiments on "motor equivalence" and anticipatory coarticulation.

1.10.2 Computer Facilities

Work is in progress on a program for displaying spectrograms on our workstation screens, to increase the efficiency of certain kinds of acoustic analyses. A RISC-based compute server has been integrated into our network of workstations. Among other functions, it will perform compute-intensive calculations (such as for spectrogram generation) at ten times the rate currently possible on our VAXStations. The WAVES acoustic analysis package has been installed on the RISC workstation for future use in work on lexical access.

1.11 Publications

Journal Articles and Published Papers

Butzburger, J., M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. "Isolated Word Intonation Recognition Using Hidden Markov Models." *Proc. ICASSP* 90, pp. 773-776 (1990).

Boyce, S.E. "Coarticulatory Organization for Lip Rounding in Turkish and English." *J. Acoust. Soc. Am.* 88: 2584-2595 (1990).

Boyce, S.E., R.A. Krakow, F. Bell-Berti, and C.E. Gelfer. "Converging Sources of Evidence for Dissecting Articulatory Movements into Core Gestures." *J. Phonetics* 18: 173-188 (1990).

Hillman, R.E., E.B. Holmberg, J.S. Perkell, M. Walsh, and C. Vaughn. "Phonatory Function Associated with Hyperfunctionally Related Vocal Fold Lesions." *J. Voice* 4: 52-63 (1990).

Klatt, D.H, and L.C. Klatt. "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers." *J. Acoustic. Soc. Am.* 87: 820-857 (1990).

Manuel, S.Y. "The Role of Contrast in Limiting Vowel-to-Vowel Coarticulation in Different Languages." *J. Acoustic. Soc. Am.* 83: 1286-1298 (1990).

Perkell, J.S. "Testing Theories of Speech Production, Implications of Some Detailed Analyses of Variable Articulatory Data." In *Speech Production and Speech Modeling.* Eds. W.J. Hardcastle and A. Marchal. Boston: Kluwer Academic Publishers, 1990, pp. 263-288.

Stevens, K.N., and C.A. Bickley. "Higher-level Control Parameters for a Formant Synthesizer." *Proceedings of the First International Conference on Speech Synthesis,* Autrans, France, 1990, pp. 63-66.

Veilleux N., M. Ostendorf, S. Shattuck-Hufnagel, and P.J. Price. "Markov Modeling of Prosodic Phrases." *Proc. ICASSP* 90, pp. 777-780 (1990).

Wodicka, G.R., K.N. Stevens, H.L. Golub, and D.C. Shannon. "Spectral Characteristics of Sound Transmission in the Human Respiratory System." *IEEE Trans. Biomed. Eng.* 37(12): 1130-1134 (1990).

Papers Submitted for Publication

Bickley, C.A. "Vocal-fold Vibration in a Computer Model of a Larynx." Submitted to *Proceedings of the Vocal Fold Physiology Conference,* Stockholm, Sweden, 1989. New York: Raven Press.

Halle, M., and K.N. Stevens. "The Postalveolar Fricatives of Polish." Submitted to *Osamu Fujimura Festschrift.*

Halle, M., and K.N. Stevens. "Knowledge of Language and the Sounds of Speech." Submitted to *Proceedings of the Symposium on Music, Language, Speech and Brain,* Stockholm, Sweden.

Huang, C.B. "Effects of Context, Stress, and Speech Style on American Vowels." Submitted to *ICSLP'90,* Kobe, Japan.

Lane, H., J.S. Perkell, M. Svirsky, and J. Webster. "Changes in Speech Breathing Following Cochlear Implant in Postlingually Deafened Adults." Submitted to *J. Speech Hear. Res.*.

Perkell, J.S., and M.H. Cohen. "Token-to-token Variation of Tongue-body Vowel Targets: the Effect of Context." Submitted to *Osamu Fujimura Festschrift.*

Perkell, J.S., E.B. Holmberg, and R.E. Hillman. "A System for Signal Processing and Data Extraction from Aerodynamic, Acoustic and Electroglottographic Signals in the Study of Voice Production." Submitted to *J. Acoust. Soc. Am.*

Perkell, J.S., and M.L. Matthies. "Temporal Measures of Labial Coarticulation for the Vowel /u/." Submitted to *J. Acoust. Soc. Am.*

Shattuck-Hufnagel, S. "The Role of Word and Syllable Structure in Phonological Encoding in English." Submitted to *Cognition* (special issue).

Stevens, K.N. "Some Factors Influencing the Precision Required for Articulatory Targets: Comments on Keating's Paper." *Papers in Laboratory Phonology I.* Eds. J.C. Kingston and M.E. Beckman. Cambridge: Cambridge University Press.

Stevens, K.N. "Vocal-fold Vibration for Obstruent Consonants." *Proc. of the Vocal Fold Physiology Conference,* Stockholm, Sweden. New York: Raven Press.

Stevens, K.N., and C.A. Bickley. "Constraints among Parameters Simplify Control of Klatt Formant Synthesizer." Submitted to *J. Phonetics*.