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. Marie K. Huffman, Dr. Michel T. Jackson, Dr. Melanie Matthies, Dr. Joseph S. Perkell, Dr. Stefanie Shattuck-Hufnagel, Dr. Mario A. Svirsky, Seth M. Hall

Visiting Scientists and Research Affiliates

Dr. Shyam S. Agrawal,[1] Dr. Tirupattur V. Ananthapadmanabha,[2] Dr. Vladimir M. Barsukov, Dr. Corine A. Bickley, Dr. Suzanne E. Boyce, Dr. Carol Y. Espy-Wilson,[3] Dr. Richard S. Goldhor,[4] Dr. Robert E. Hillman,[3] Eva B. Holmberg,[5] Dr. Caroline Huang,[6] Dr. Harlan Lane,[7] Dr. John Locke,[8] Dr. John I. Makhoul,[9] Dr. Sharon Y. Manuel, Dr. Carol Ringo,[10] Dr. Anthony Traill,[11] Dr. David Williams,[12] Giulia Arman Nassi, Torstein Pedersen,[13] Jane Webster[14]

Graduate Students

Abeer A. Alwan, Hwa-Ping Chang, Marilyn Y. Chen, Helen M. Hanson, Mark A. Johnson, Ronnie Silber, Lorin F. Wilde

Undergraduate Students

Kerry L. Beach, Venkatesh R. Chari, Anna Ison, Sue O. Kim, Laura Mayfield, Bernadette Upshaw, Monnica J. Williams

Technical and Support Staff

D. Keith North, Arlene E. Wint

---

[1] CEERI Centre, CSIR Complex, New Delhi, India.

[2] Voice and Speech Systems, Bangalore, India.

[3] Boston University, Boston, Massachusetts.

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

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

[6] Dragon Systems, Inc., Newton, Massachusetts.

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

[8] Massachusetts General Hospital, Boston, Massachusetts.

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

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

[11] University of Witwatersrand, South Africa.

[12] Sensimetrics, Inc., Cambridge, Massachusetts.

[13] University of Karlsruhe, Germany.

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

1.1 Introduction

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.

Sponsors

C.J. Lebel Fellowship
Dennis Klatt Memorial Fund
National Institutes of Health
- Grant T32-DC00005
- Grant R01-DC00075
- Grant F32-DC00015
- Grant R01-DC00266[15]
- Grant P01-DC00361[16]
- Grant R01-DC00776[17]

National Science Foundation
- Grant IRI 89-10561
- Grant IRI 88-05680[15]
- Grant INT 90-24713[18]

1.2 Studies of the Acoustics, Perception, and Modeling of Speech Sounds

1.2.1 Liquid Consonants

An attribute that distinguishes liquid consonants (such as /l/ and /r/ in English) from other consonants is that the tongue blade and tongue dorsum are shaped in such a way that there is more than one acoustic path from the glottis to the lips. One of these paths passes over the midline of the vocal tract over much of its length whereas the other, shorter, path traverses around the side of the tongue dorsum and blade in the oral portion of the tract. Measurements of the acoustic spectra of sounds produced when the vocal tract is in such a position show some irregularities in the frequency range 1.5 to 3.5 kHz. Extra spectral peaks and valleys not normally seen in nonnasal vowels or in glides are often evident in the spectrum, and some of the peaks have amplitudes that deviate from the amplitudes expected for vowels.

In an effort to understand the acoustic basis for these spectral properties, we have initiated a theoretical study of the behavior of a vocal-tract model in which there is a bifurcation of the acoustic path. Equations for the vocal-tract transfer function (from glottal source to lip output) have been developed and the transfer function has been calculated and displayed for dimensions that are estimated to be within the expected range for liquids. One outcome of this preliminary analysis is that an additional pole and zero is introduced into the transfer function in the expected frequency range (2.0 to 3.0 kHz) when the length of the tract over which there is more than one path is in the range 6-8 cm. This additional pole-zero pair is associated with the acoustic propagation time around the two portions of the split section of the tube. Further theoretical work must be supplemented by additional data on the configurations of the airways for liquids and by improved estimates of the acoustic losses for these types of configurations.

1.2.2 Fricative Consonants

The need to further describe the acoustic properties of fricatives is evident in rule-based speech synthesis, where these sounds continue to present intelligibility problems. In order to provide a baseline measure of intelligibility, identification tests have been performed with natural fricative-vowel tokens and corresponding synthetic tokens, generated using a research version of Klatt's text-to-speech rule-based system. All the strident alveolar (s, z) and palatal ($\int$, ζ) fricatives were identified correctly. The weaker labiodental (f, v) and dental (θ, δ) fricatives were frequently confused with each other. As expected, intelligibility of text-to-speech fricatives was poorer than natural speech.

Acoustic analysis was performed to determine differences between natural and synthetic stimuli that could account for observed differences in intelligibility. Our results confirmed the need for

---

[15] Under subcontract to Boston University.

[16] Under subcontract to Massachusetts Eye and Ear Infirmary.

[17] Under subcontract to Massachusetts General Hospital.

[18] U.S.-Sweden Cooperative Science Program.

improved modelling of the source changes at the fricative-vowel boundary. Intelligibility of synthetic labiodental and dental fricatives is poorer than natural, even when formant transitions appear to be reproduced accurately. Natural tokens showed subtle noise amplitude variations that were not found in synthetic stimuli. Aspiration noise was found more frequently for [f] than [θ] for natural stimuli, while no text-to-speech tokens included aspiration. Offset of noise and onset of voicing at the fricative-vowel boundary was too abrupt for voiced synthetic stimuli. Preliminary results suggested speaker dependencies in the observed noise variations and time-dependent emergence of high-frequency peaks. We plan to apply our results to modify rules for manipulating speech sources and evaluate the effect of these changes on the intelligibility and naturalness of synthetic fricatives.

1.2.3 Acoustic Properties Contributing to the Classification of Place of Articulation for Stops

The production of stop consonants generates several kinds of acoustic properties: (1) the spectrum of the initial transient and burst indicating the size of the cavity anterior to the constriction; (2) place-dependent articulatory dynamics leading to different time courses of the noise burst, onset of glottal vibrations and formant transitions; (3) formant transitions indicating the changing vocal tract shape from the closed position of the stop to a more open configuration of the following vowel. A study has been initiated to measure the relative contributions of these acoustic properties to the classification of the consonantal place of articulation using a semi-automatic procedure. The acoustic data consisted of a number of repetitions of voiceless unaspirated stops in meaningful words spoken by several female and male speakers. The spectra averaged over the stop release and at the vowel onset were used as the acoustic features. Speaker independent and vowel independent classification was about 80% using either the burst or vowel onset spectrum, and a combined strategy led to higher accuracy. These results, together with further acoustic analysis of the utterances, suggest that additional acoustic properties related to articulatory dynamics, such as the detailed acoustic structure of the burst, the time course of the formant transitions, and the voice onset time, should be included in a model for stop-consonant recognition.

1.2.4 Modeling Speech Perception in Noise: the Stop Consonants as a Case Study

A doctoral thesis on this topic by Abeer A. Alwan was completed during the past year. The abstract for that thesis follows:

This study develops procedures for predicting perceptual confusions of speech sounds in noise by integrating knowledge of the acoustic properties which signal phonetic contrasts of speech sounds with principles of auditory masking theory. The methodology that was followed had three components: (1) quantifying acoustic correlates of some phonological features in naturally-spoken utterances and using the results to generate synthetic utterances, (2) developing a perceptual metric to predict the level and spectrum of the noise which will mask these acoustic correlates, and (3) performing a series of perceptual experiments to evaluate the theoretical predictions.

The focus of the study was the perceptual role of the formant trajectories in signalling the place of articulation for the stop consonants /b, d/ in consonant-vowel syllables, where the vowel was either /a/ or /ε/. Nonsense syllables were chosen for the perceptual study so that lexical effects such as word frequency did not bias subjects' responses. Computer-generated, rather than naturally-spoken, syllables were used to provide better control of the stimuli.

In the analysis/synthesis stage, the acoustic properties of the stop consonants /b, d/ imbedded in naturally-spoken CV syllables were quantified and the results were then used to synthesize these utterances with the formant synthesizer KLSYN88. In the context of the vowel /a/, the two synthetic syllables differed in the $F2$ trajectory: the $F2$ trajectory was falling for /da/ and was relatively flat for /ba/. In the C/ε/. context, both $F2$ and $F3$ trajectories were different for the consonants: $F2$ and $F3$ were flat for /dε/, whereas they were rising for /bε/.

A metric was then developed to predict the level of noise needed to mask a spectral peak corresponding to a formant peak. The metric was based on a combination of theoretical and empirical results. Two types of masking were studied: within-band masking (where the formant was within the bandwidth of the noise masker) and above-band masking (where the formant was above the upper cutoff frequency of the masker). Results of auditory masking theory, which was established primarily for pure tones, were used successfully to predict within-band masking of formant peaks. The predictive measure in this case was the signal-to-noise ratio in a critical band

around the formant frequency. The applicability of the results of masking theory to formant peaks was tested by conducting a series of discrimination and detection experiments with synthetic, steady-state vowels.

In the above-band masking case, it was found that predictions based on the two methods known for predicting aspects of this kind of masking (ANSI standards and Ludvigsen) did not agree with experimental results. An empirical algorithm was developed to account for the experimental data.

In the final stage of the study, a series of identification tests with synthetic CV utterances in noise was conducted. Two noise maskers were used in the experiments: white noise, and band-pass noise centered around the *F*2 region. The spectral prominences associated with *F*2 and *F*3 have a lower amplitude during the transitions from the consonant than in the steady-state vowel, so that it is possible, using a steady-state noise, to mask portions of a formant transition without masking the formant peak in the vowel. Subjects' responses were analyzed with the perceptual metric developed earlier. Results showed that when the *F*2 transition for C/ɑ/ or the *F*2 and *F*3 transitions for C/ɛ/ were masked by noise, listeners interpreted the stimuli as though the formant transitions were flat. That is, /dɑ/ was heard as /bɑ/ and /bɛ/ was heard as /dɛ/.

It was also found that when only the *F*2 trajectory is masked, achieved by selectively masking *F*2 with a band-pass noise masker, then amplitude differences in the *F*3 and *F*4 regions could be used as cues for place information in the C/ɑ/ case even though the trajectories of these higher formants did not differ for the two consonants.

1.2.5 Unstressed Syllables

We are continuing acoustic and perceptual studies of the vowels and consonants in unstressed syllables in utterances produced with different speech styles. Our aim is to determine what attributes of these syllables are retained and what attributes are modified or deleted when reduction occurs. We hope that the results of these studies can contribute to the development of models for lexical access that are valid for various speech styles.

In one study we have examined the acoustic manifestation of /ð/ in the definite article **the** when this article occurs in different phonetic environments. In particular we have looked at the sequence **in the** in a large number of sentences produced by several speakers. In almost all of these utterances, acoustic analysis indicates that the underlying consonant /ð/ is produced as a nasal consonant. That is, the consonant is produced with a dental closure, and the velopharyngeal opening from the preceding nasal consonant spreads through the dental closure. Production of this underlying voiced fricative as a [-continuant] consonant is a common occurrence (e.g., in absolute initial position, or following an obstruent, as in **at the** or **of the**), and consequently use of a stop-like closure following **in** is not unexpected. Comparison of the duration of the nasal murmur for the sequences **in a** and **in the** show that the duration for the latter sequence is consistently longer, with the dividing point being at about 50 ms.

In a second group of studies, we are examining the acoustic evidence for the presence of unstressed vowels in bisyllabic utterances like **below** and **support**, so that they are distinguished from the monosyllabic words **blow** and **sport**. Acoustic measurements for several speakers producing these utterances show that the speakers' attempts to implement the unstressed vowels can take several forms. These forms are consistent with a view that there are two basic requirements for an unstressed vowel. One is that the airway above the glottis should be less constricted during the vowel than in the adjacent consonants, and the other is that the glottis assumes a configuration that is narrower than that in the adjacent consonants if those consonants are obstruents. Both of these adjustments contribute to the generation of a maximum amplitude in some part of the frequency range to signal the presence of the vowel. One manifestation of these intended movements in our data is an increase in the time from the release of the first consonant to the release of the second consonant for the two-syllable utterance relative to the one-syllable one. Other acoustic measurements (such as aspiration in /p/ of **support**, and changes in spectrum amplitudes and formant frequencies between the two consonant releases in **below**) also provide evidence for the unstressed vowel. Preliminary listening tests with synthetic versions of one of these word pairs (**support-sport**) have demonstrated the perceptual importance of these various acoustic attributes.

A third set of experiments has compared the temporal and amplitude characteristics of stop consonants in post- and prestressed positions in an utterance. The data show consistent reduction in closure duration and in voice onset time for consonants in post-stressed positions. Burst amplitudes (in relation to the following vowel) for post-stressed stops are less than those for prestressed stops when the consonants are voiceless but not when they are voiced. Poststressed velar stop consonants show considerable variability in

duration and in the degree to which they exhibit imprecise releases with multiple bursts.

These studies of unstressed syllables are consistent with the notion that the intention to produce the components of these syllables can take many different acoustic forms. Models of lexical access must be sensitive to these different acoustic attributes.

1.2.6 Reductions in Casual Speech in German

For each language, there are modifications or reductions that occur in the production and acoustic patterns of utterances when they are spoken casually compared to when they are spoken carefully. Many of these reductions are language-specific, but it is expected that some of the principles governing the reductions might be universal across languages. In a preliminary effort to uncover some of these principles, a study of reductions in casual speech in German has been undertaken.

Several sentences were spoken by three speakers of German, at four speech rates ranging from careful to rapid. Acoustic characteristics of a number of components of these sentences were measured. These included measurements of the durations of several types of units, and measurements of formant frequencies and glottal vibration characteristics for selected vowels and consonants. For the more casual or rapid speech modes, the sentence durations were shorter, with function words and unstressed vowels contributing more to the duration reduction than content words and stressed vowels. The durations of the nasal consonant regions of the sentences were reduced less than other portions of the sentences. The presence of a nasal consonant (or a sequence with nasal consonants) was always preserved in the rapid utterances, whereas the distinction between voiced and voiceless obstruent consonants was often neutralized. Vowel formant frequencies for stressed vowels in content words showed greater stability with changes in rate than did the formants for unstressed vowels or vowels in function words. Some of the reductions that occur in function words in German appear to be more extreme than reductions in function words in English.

1.3 Speech Synthesis

1.3.1 Synthesis of Syllables for Testing with Different Populations of Listeners

Ongoing improvements in procedures for synthesizing consonant-vowel and vowel-consonant syllables have led to the creation of a wide inventory of synthesized syllables. One use of these syllables has been testing the phonetic discrimination and identification capabilities of an aphasic population that is being studied by colleagues at Massachusetts General Hospital. Pairs of syllables have been synthesized which differ in one phonetic feature, such as [bα-vα] (stop-continuant), [wα-vα] (sonorant-obstruent), [nα-lα] (nasal-nonnasal). Contrasts of these sorts have been studied infrequently in the past. More commonly used contrasts such as place ([bα-dα-gα],[αm-αn]) have also been synthesized. We attempted to create highly natural-sounding stimuli by manipulating every acoustic characteristic which has been observed to contribute to the phonetic contrast in question. The other acoustic characteristics of the pair of stimuli were identical. This synthesis approach facilitated the creation of continua between pairs of stimuli. Several continua have been generated by varying each acoustic parameter which differs between the endpoints in nine steps. These continua are being presented informally to normal-hearing listeners in order to assess their naturalness and will be used in studies at Massachusetts General Hospital. An initial goal of the research with aphasic listeners is to determine whether individuals or groups of individuals show a pattern in phonetic discrimination indicating that ability to discriminate some features is a prerequisite for discriminating others.

1.3.2 Synthesis of Non-English Consonants

The flexibility of the glottal source in the KLSYN88 synthesizer should make it capable of generating obstruent consonants with a variety of laryngeal characteristics. Hindi (as well as a number of other languages) contrasts stop consonants on the dimensions of both voicing and aspiration. These stop consonants can be produced with four different places of articulation. All of these consonants have been generated (in consonant-vowel syllables) with the synthesizer. Of particular interest are the voiced aspirated stops. Synthesis of these consonants requires that vocal-fold vibration continue through much of the closure interval, with glottal spreading initiated just prior to the consonantal release. The glottal parameters

are adjusted to give breathy voicing in the few tens of milliseconds after release, followed by a return to modal voicing. All syllables were judged to be adequate by native listeners.

The synthesizer has also been used to generate click sounds for perception tests by speakers of a Khoisan language that uses an inventory of clicks. Generation of the abrupt clicks /!/ and /ǂ/ required that a transient source be used to excite the parallel resonant circuits of the synthesizer, with appropriate adjustment of the gains of the circuits to simulate the resonances of the oral cavity that are excited by the release of the tongue tip or tongue blade.

1.4 Studies of Speech Production

1.4.1 Degradation of Speech and Hearing with Bilateral Acoustic Neuromas

We have begun a project in which we are studying the relation between speech and hearing in people who become deaf from bilateral acoustic neuromas (NF2). The primary goal of this study is to add to the understanding of the role of hearing in the control of adult speech production. The rationale and approach to this work is similar to our ongoing work on the speech production of cochlear implant patients. Speech acoustic and physiological parameters will be recorded and speech perception will be tested in a group of (still hearing) NF2 patients who are at risk of losing their hearing. The same production parameters will be recorded at intervals for the subset of patients from the initially-recorded group who suffer further hearing loss. Thus far, we have begun the recruitment of patients from across the country, we have ordered several special-purpose transducers, and we have nearly completed the design of a comprehensive test protocol.

1.4.2 Trading Relations Between Tongue-body Raising and Lip Rounding in Production of the Vowel /u/

Articulatory and acoustic data are being used to explore the following hypothesis: the goals of articulatory movements are relatively invariant acoustic targets, which may be achieved with varying and reciprocal contributions of different articulators. Previous articulatory studies of similar hypotheses, expressed entirely in articulatory terms, have been confounded by interdependencies of the variables being studied (lip and mandible or tongue body and mandible displacements). One case in which this complication may be minimized is that of lip rounding and tongue-body raising (formation of a velo-palatal constriction) for the vowel /u/. Lip rounding and tongue-body raising should have similar acoustic effects for /u/, mainly on the second formant frequency, and could show reciprocal contributions to its production; thus, we are looking for negative correlations in measures of these two parameters. We are using an Electro-Magnetic Midsagittal Articulometer (EMMA) to track movements of midsagittal points on the tongue body, upper and lower lips and mandible for large numbers of repetitions of utterances containing /u/ in controlled phonetic environments. Initial analyses from four subjects of articulatory displacements at times of minima in absolute velocity for the tongue body during the /u/ (i.e., "articulatory targets") show evidence against the hypothesis for one subject and weakly in favor of the hypothesis for the other three.

We are currently examining: (1) trading relations between the two parameters as a function of time over the voiced interval of each vowel, and (2) relations among the transduced measures of tongue raising and lip protrusion, the resulting vocal-tract area-function changes (using measurements from dental casts and video recordings of facial movements), and changes in the vowel formants (using an articulatory synthesizer).

1.4.3 Refinements to Articulatory Movement Transduction

We have concluded that our EMMA system almost always produces sufficiently accurate measurements of the positions of articulatory structures, but we have observed in experiments with subjects (see 1.4.2 above) that there can be circumstances under which measurement error may approach or exceed acceptable limits. We are currently running bench tests to explore this issue, with the goal of developing procedures and criteria for the objective assessment of measurement error.

In order to make the original EMMA electronics function with a revised transmitter configuration, it was necessary to design and implement a phase correction circuit. This circuit has now been hard wired and installed in a box which plugs into the main unit.

1.4.4 Modeling of the Lateral Dimension in Vocal-Fold Vibration

The forces which drive vocal-fold vibration can be simulated using the Ishizaka-Flanagan two-mass model, but the model misses some of the subtleties of actual fold vibration. In particular, acoustic studies by Klatt and Klatt indicate that during breathy voicing, the vocal folds close gradually along their length, rather than all at once as in the two-mass model. We have developed a model, based on the two-mass model, which treats the fold surface as a pair of beam elements which are acted upon by the glottal pressure, a compressive stress which is proportional to the fold displacement, and a shear stress which is proportional to the deformation of the fold's inner surface. Rather than assuming a single displacement for the entire length of the fold, the displacement is smoothly interpolated from the arytenoid cartilage, through the middle of the fold, to the anterior commissure. With two degrees of freedom, the model is capable of imitating the two-mass model for tightly adducted arytenoid cartilages, and of simulating gradual fold closure during breathy voicing.

1.5 Speech Production Planning

Our studies of the speech production planning process have focussed on developing and testing models of two aspects of phonological processing: segmental (primarily the serial ordering of sublexical elements) and prosodic (particularly the phenomenon of apparent stress shift).

1.5.1 Segmental Planning

Earlier analyses of segmental speech errors like "parade fad" → "farade pad" showed that word-position similarity is an important factor in determining which segments of an utterance will interact with each other in errors, even when word stress is controlled. This result supports the view that word structure constrains the serial ordering process (i.e., the process which ensures that the sub-parts of words occur in their appropriate locations in an utterance). In fact, for some kinds of speech errors, up to 80 percent occur in word-onset position. The predeliction for word-onset errors, at least in English, may arise from a difference in processing for these segments. We are testing this hypothesis with a series of error elicitation experiments to determine whether the asymmetry is found only for the planning of syntactically-structured utterances.

1.5.2 Prosodic Planning

Recent theories of phonology have addressed the question of the location of prominences and boundaries in spoken language. These two aspects of utterances are not fully determined by syntactic and lexical structure, but instead suggest the necessity of a separate hierarchy of prosodic constituents, with intonational phrases at the top and individual syllables at the bottom. One aspect of this theoretical development has been the positing of a rhythmic grid for individual lexical items in which the columns correspond to the syllables in the utterance and the rows to degrees of rhythmic prominence. It has been argued that this grid representation should be extended to entire utterances, partly on the basis of the occurrence of apparent "stress shift." Stress shift occurs when the major perceptual prominence of a target word is heard not on its main-stess syllable (as in "massaCHUsetts") but on an earlier syllable (as in "MAssachusetts MIRacle") in certain rhythmic configurations.

Our acoustic and perceptual analyses of this phenomenon have shown that in many cases the perception of apparent stress shift is the result of the disappearance of a pitch marker from the main-stress syllable ("CHU" in the above example) when the target word appears in the longer phrase. This disappearance of the pitch accent leaves an impression of relative prominence on the earlier syllable "MA-," even though its duration does not systematically increase. Although the rise in fundamental frequency on the early syllable may be greater when the target word occurs in the longer phrase, just as a stress shift account would predict, this increased rise is also found for phrases like "MONetary POLicy," where no shift can occur because the first syllable of the target word is the main stress syllable. This result suggests that the increased *F*0 movement is associated with the length of the phrase, rather than with the migration of rhythmic prominence to the left in the target word.

In summary, it appears that apparent stress shift is at least in part a matter of where the pitch accents in an utterance are located, raising the possibility that speakers tend to place pitch accents near the beginnings of prosodic constituents. We are currently testing this hypothesis with paragraphs in which the target word occurs early vs. late in its intonational phrase. We are testing this possibility in both elicited laboratory speech read aloud, and in a corpus of FM radio news speech. If we can demonstrate that pitch accent placement is at least partially determined by the beginning of a new prosodic constituent, and that pitch accents can be reliably detected in the signal, we will have a basis for proposing that listeners make use of the occur-

rence of pitch accents to help them determine which portion of an utterance to process as a constituent.

1.5.3 Integration of Segmental and Prosodic Planning

We have begun a series of experiments to explore the possibility that the segmental and prosodic aspects of production planning interact. Initial results suggest that metrically regular utterances (with repeated use of the same stress pattern) provoke fewer segmental errors than metrically irregular utterances. This preliminary finding is compatible with Shattuck-Hufnagel's hypothesis that the serial ordering process for sublexical elements occurs during the integration of syntactic/morphological information with the prosodic framework of the planned utterance. It may be that, when the prosody of an utterance is partially predictable, with stresses occurring at regular intervals, the planning process is less demanding and serial ordering operates more reliably.

1.6 Speech Research Relating to Special Populations

1.6.1 A Psychophysically-based Vowel Perception Model for Users of Pulsatile Cochlear Implants

Pulsatile multichannel cochlear implants that encode second formant frequency as electrode position (*F*0-*F*2 strategy) represent different vowels by stimulating different electrodes across the electrode array. Thus, the ability of a subject to identify different vowels is limited by his ability to discriminate stimulation delivered to different electrodes. A vowel perception model was developed for users of such devices. This model is an extension of the Durlach-Braida model of intensity resolution. It has two parts: an "internal noise" model that accounts for vowel confusions made as a consequence of poor perceptual sensitivity, and a "decision" model that accounts for vowel identification errors due to response bias. The decision model postulates that each stimulus is associated with a "response center" which represents the sensation expected by the subject in response to that stimulus. To test the vowel perception model we used d' scores in an electrode identification experiment with a male subject and predicted this subject's vowel confusion matrix employing three sets of response centers. The "natural" set was determined by places that would be maximally stimulated by each vowel in a normal cochlea; the "optimal" response center was determined by places actually stimulated by each vowel during the vowel identification experiment; and the "good fit" set used response centers that were manually adjusted to improve the model's prediction of the real matrix. Model predictions were tested against a vowel confusion matrix obtained from the same subject in a separate study.

Results using the optimal set of response centers are encouraging: the model predicted successfully some characteristics of the confusion matrix, in spite of a five-year lapse between the experiment that provided sensitivity data and the experiment where the confusion matrix was obtained. Nevertheless, there were discrepancies between predicted and real confusion matrices. These discrepancies are interestingly biased—the subject frequently mistook some vowels for other vowels with higher *F*2. These errors are consistent with a set of response centers that are close to the optimal set, but somewhat shifted in the direction of the natural set. One implication of these preliminary results is that cochlear implant users make use of the central nervous system's plasticity but at the same time this plasticity may not be unlimited.

1.6.2 The Effect of Fixed Electrode Stimulation on Perception of Spectral Information

This study was carried out in collaboration with Margaret W. Skinner and Timothy A. Holden of the Department of Otolaryngology-Head and Neck Surgery, Washington University School of Medicine, Saint Louis. The study attempts to find the underlying reasons for improvements in speech perception by users of the Nucleus cochlear implant when they employ the new Multipeak stimulation strategy instead of the older *F*0-*F*1-*F*2 strategy. For voiced sounds, the *F*0-*F*1-*F*2 strategy stimulates two channels every pitch period: these electrodes are chosen based on instantaneous estimates of *F*1 and *F*2. The Multipeak strategy stimulates two additional channels in the basal region of the cochlea. Channels 4 and 7 are typically used for voiced sounds, with delivery of pulses whose energy is proportional to the acoustic energy found in the 2.0-2.8 kHz and 2.8-4.0 kHz ranges.

We designed an experiment that employed a Multipeak-like condition, a *F*0-*F*1-*F*2 condition and a control condition, where fixed-amplitude pulses were sent to channels 4 and 7. Two subjects were tested using vowel and consonant identification tasks. The control condition provided vowel identification performance at least as good as the Multipeak condition, and significantly better than the *F*0-*F*1-*F*2 condition. Consonant identifi-

cation results for the control condition were better than for the *F*0-*F*1-*F*2 condition, but not as good as for the Multipeak condition. These results suggest that part of the improvement observed with the Multipeak strategy is due to stimulation of the fixed basal channels, which act as perceptual "anchors" or references, allowing subjects to identify the position of electrode pairs that vary in position along the electrode array.

1.6.3 Acoustic Parameters of Nasal Utterance in Hearing-Impaired and Normal-Hearing Speakers

One of the more prevalent abnormalities in the speech of the hearing impaired that contributes to reduced intelligibility is inadvertent nasality. Theoretically, a wider first formant should reflect a greater loss of sound energy within the nasal cavity, which has a relatively large surface area compared to the oral cavity of vowels. A more prominent spectral peak in the vicinity of 1 kHz should reflect a larger velopharyngeal opening, according to a theoretical analysis of a nasal configuration based on admittance curves. From acoustic analysis of the speech of hearing-impaired children, reduced first formant prominence and the presence of an extra pole-zero pair between the first and second formants characterize spectra of nasalized vowels. The difference between the amplitude of the first formant and the amplitude of the extra peak, *A*1-*P*1, was a measure that correlated with listener judgments of the degree of vowel nasalization. To obtain further validation of these parameters as measures of nasalization, *A*1 and the pole-zero spacing were systematically manipulated in synthetic utterances. Perceptual experiments with synthesized, isolated static vowels [i],[I],[o],[u] showed that both parameters contributed to the perception of nasality. Another perceptual experiment with a number of synthesized words (of the form bVt) gave results indicating somewhat different relative importance of the two parameters. Correlation of *A*1-*P*1 with the average nasality perception judgments of ten listeners was found for both groups of stimuli. Another acoustic characteristic of speech of the hearing-impaired children is the change of the extra peak amplitude over time within the vowel due to their lack of coordination of movement of the velum with movement of other articulators. From a pilot test, time variation of the first formant bandwidth and frequency of the nasal zero had a greater effect on nasality judgments when the average *A*1-*P*1 was small. This finding suggests that a somewhat nasalized vowel is perceived to be even more nasal when there is large velum movement during production of the vowel.

1.7 Models for Lexical Representation and Lexical Access

As we continue to develop models for lexical access from acoustic patterns of speech, we have been attempting to organize and refine the representation of words, segments, and features in the lexicon. The present version of the model of the lexicon organizes phonological segments into four classes according to the way the features are represented in the sound.

Two of these classes are (1) the vowels and (2) the glides and syllabic consonants. A distinguishing attribute of both these classes is that the spectrum of the sound that is produced when the segments are implemented is either relatively fixed or changes only slowly. There are no acoustic discontinuities, and all of the features of the segments are derived from the spectra sampled over a relatively long time interval and how these spectra change with time. Both classes are produced with a source at the glottis. Within this group of segments, the vowels are considered to form a class, in that the vocal tract does not have a narrow constriction that gives a distinctive acoustic attribute. They are characterized entirely by their formant pattern, and no articulator can be regarded as primary. The glides and syllabic consonants, on the other hand, have spectral characteristics that are imposed by forming a constriction with a particular articulator.

A third class of segments (traditionally classified as [+consonantal]) is produced by making a narrow constriction with a specific articulator. Formation or release of this constriction creates a discontinuity in the sound. The landmark generated by this discontinuity forms a nucleus around which acoustic information about the various features of the consonant is located, possibly over a time interval of $\pm$ 50 milliseconds or more. The fourth class consists of the clicks. The sound is of high intensity and relatively brief. The features identifying a click are contained in the spectral characteristics within a few milliseconds of the click release. These clicks are not directly relevant to English, except that they highlight the contrast between sounds that are of high intensity and concentrated in time (the clicks) and sounds of lower intensity where evidence for the features is spread over a longer time interval around an acoustic discontinuity.

In the proposed lexical representation, the classes of segments and features within these classes are organized in a hierarchical fashion. In the case of

consonants, a distinction is made between features designating the primary articulator that forms the constriction and features designating the activity of other articulators.

As a part of a project in which automatic procedures for lexical access are being developed, we are building a computer-based lexicon of words in which the representation of features for the segments in the words reflects the classification noted above. There are currently about 150 words in the lexicon. The marking of feature values within the lexicon includes the possibility of labeling a feature as being subject to modification by the context. For example, in the word **bat**, the features designating place of articulation for the final consonant are subject to modification (as in **bat man**) but the feature [-continuant] is not. A second component of the implementation of the lexical access model is to develop automatic procedures for identifying landmarks and determining the features whose acoustic correlates appear in the vicinity of the landmarks. Initial efforts in this direction are in progress.

1.8 Speech Analysis and Synthesis Facilities

Several modifications have been made to the Klattools that are used for analysis and synthesis of speech. The KLSYN88 synthesizer has been augmented to include the possibility of selecting a new glottal source (designed by Dr. Tirupattur Ananthapadmanabha) that has potential advantages and flexibility relative to the current inventory of glottal sources in the synthesizer. An inverse filtering capability has also been incorporated in the synthesizer program to allow for the possibility of generating a glottal source derived from a natural utterance.

The KLSPEC analysis program has been modified to include the possibility of obtaining a spectrum that is an average of a number of DFT spectra sampled at one-ms intervals over a specified time span. This type of averaging is especially useful when examining the spectra of noise bursts and fricative consonants. It also has application when using a short time window for measuring the spectrum of a voiced sound to avoid the necessity of careful placement of the window within each glottal period.

1.9 Publications

1.9.1 Published Papers

Alwan, A. "Modelling Speech Perception in Noise: A Case Study of the Place of Articulation Feature." *Proceedings of the 12th International Congress of Phonetic Sciences* 2: 78-81 (1991).

Bickley, C.A. "Vocal-fold Vibration in a Computer Model of a Larynx." In *Vocal Fold Physiology.* Eds. J. Gauffin and B. Hammarberg. San Diego: Singular, 1991, pp. 37-46.

Boyce, S.E., R.A. Krakow, and F. Bell-Berti. "Phonological Underspecification and Speech Motor Organisation." *Phonology* 8: 219-236 (1991).

Burton, M.W., S.E. Blumstein, and K.N. Stevens. "A Phonetic Analysis of Prenasalized Stops in Moru." *J. Phonetics* 20: 127-142 (1992).

Halle, M., and K.N. Stevens. "Knowledge of Language and the Sounds of Speech." In *Music, Language, Speech and Brain.* Eds. J. Sundberg, L. Nord, and R. Carlson. Basingstoke, Hampshire: Macmillan Press, 1991, pp. 1-19.

Kuhl, P.K., K. Williams, F. Lacerda, K.N. Stevens, and B. Lindblom. "Linguistic Experience Alters Phonetic Perception in Infants by Six Months of Age." *Sci.* 255: 606-608 (1992).

Lane, H., J.S. Perkell, M. Svirsky, and J. Webster. "Changes in Speech Breathing Following Cochlear Implant in Postlingually Deafened Adults." *J. Speech Hear. Res.* 34: 526-533 (1991).

Perkell, J.S. "Models, Theory and Data in Speech Production." *Proceedings of the 12th International Congress of Phonetic Sciences* 1: 182-191 (1991).

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." *J. Acoust. Soc. Am.* 89: 1777-1781 (1991).

Price, P.J., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. "The Use of Prosody in Syntactic Disambiguation." *J. Acoust. Soc. Am.* 90: 1956-2970 (1991).

Price, P.J., M. Ostendorf and S. Shattuck-Hufnagel. "Disambiguating Sentences Using Prosody." *Proceedings of the 12th International Congress of Phonetic Sciences* 2: 418-421 (1991).

Shattuck-Hufnagel, S. "Acoustic Correlates of Stress Shift." *Proceedings of the 12th International Congress of Phonetic Sciences* 4: 266-269 (1991).

Stevens, K.N. "Some Factors Influencing the Precision Required for Articulatory Targets: Comments on Keating's Paper." In *Papers in Laboratory Phonology I.* Eds. J.C. Kingston and M.E. Beckman. Cambridge: Cambridge University Press, 1991, pp. 471-475.

Stevens, K.N. "Vocal-fold Vibration for Obstruent Consonants." In *Vocal Fold Physiology.* Eds. J. Gauffin and B. Hammarberg. San Diego: Singular, 1991, pp. 29-36.

Stevens, K.N. "The Contribution of Speech Synthesis to Phonetics: Dennis Klatt's Legacy." *Proceedings of the 12th International Congress of Phonetic Sciences* 1: 28-37 (1991).

Stevens, K.N., and C.A. Bickley. "Constraints Among Parameters Simplify Control of Klatt Formant Synthesizer." *J. Phonetics* 19: 161-174 (1991).

Svirsky, M.A., and E.A. Tobey. "Effect of Different Types of Auditory Stimulation on Vowel Formant Frequencies in Multichannel Cochlear Implant Users." *J. Acoust. Soc. Am.* 89: 2895-2904 (1991).

Wilde, L.F., and C.B. Huang. "Acoustic Properties at Fricative-vowel Boundaries in American English." *Proceedings of the 12th International Congress of Phonetic Sciences* 5: 398-401 (1991).

1.9.2 Papers Submitted for Publication

Holmberg, E., R. Hillman, J. Perkell, and C. Gress. "Relationships Between SPL and Aerodynamic and Acoustic Measures of Voice Production: Inter- and Intra-speaker Variation." *J. Speech Hear. Res.*

Perkell, J., M. Svirsky, M. Matthies, and M. Jordan. "Trading Relations Between Tongue-body Raising and Lip Rounding in Production of the Vowel /u/." *PERILUS,* the working papers of the Department of Phonetics, Institute of Linguistics, Stockholm. Forthcoming.

Perkell, J., and M. Matthies. "Temporal Measures of Anticipatory Labial Coarticulation for the Vowel /u/: Within- and Cross-Subject Variability." *J. Acoust. Soc. Am.*

Perkell, J., H. Lane, M. Svirsky, and J. Webster. "Speech of Cochlear Implant Patients: A Longitudinal Study of Vowel Production." *J. Acoust. Soc. Am.*

Shattuck-Hufnagel, S. "The Role of Word and Syllable Structure in Phonological Encoding in English." *Cognition.* Forthcoming.

Stevens, K.N. "Lexical Access from Features." Workshop on Speech Technology for Man-Machine Interaction, Tata Institute of Fundamental Research, Bombay, India, 1990.

Stevens, K.N. "Speech Synthesis Methods: Homage to Dennis Klatt." In *Talking Machines: Theories, Models, and Applications.* Eds. G. Bailly and C. Benoit. New York: Elsevier. Forthcoming.

Stevens, K.N. "Models of Speech Production." In *Handbook of Acoustics.* Ed. M. Crocker. New York: Wiley. Forthcoming.

Stevens, K.N., S.E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski. "Acoustic and Perceptual Characteristics of Voicing in Fricatives and Fricative Clusters." *J. Acoust. Soc. Am.* Forthcoming.

Stevens, K.N. "Phonetic Evidence for Hierarchies of Features." In *LabPhon3.* Ed. P. Keating. Los Angeles: University of California. Forthcoming.

Svirsky, M., H. Lane, J. Perkell, and J. Webster. "Effects of Short-Term Auditory Deprivation on Speech Production in Adult Cochlear Implant Users." *J. Acoust. Soc. Am.* Forthcoming.

Wightman, C., S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries." *J. Acoust. Soc. Am.* Forthcoming.