22. Speech Communication

Academic and Research Staff

Prof. K.N. Stevens, Prof. J. Allen, Prof. M. Halle, Prof. S.J. Keyser, Prof. V.W. Zue, A. Andrade[11], Dr. D. Bradley[11], Dr. Martha Danly[12], Dr. F. Grosjean[13], Dr. S. Hawkins[14], Dr. R.E. Hillman[15], E.B. Holmberg[16], Dr. A.W.F. Huggins[17], C. Hume, Dr. H. Kawasaki, Dr. D.H. Klatt, Dr. L.S. Larkey, Dr. John Locke[18], Dr. B. Lyberg[11], Dr. J.I. Makhoul[17], Dr. E. Maxwell, Dr. L. Menn[19], Dr. P. Menyuk[20], Dr. J.L. Miller[13], Dr. J.M. Pardo Munoz[11], Dr. J.S. Perkell, Dr. P.J. Price, Dr. S. Shattuck-Hufnagel, S.-Q. Wang[11], T. Watanabe[11]

Graduate Students

A.M. Aull, C. Aoki, C. Bickley, F. Chen, K. Church, S. Cyphers, C. Espy, J. Glass, R. Goldhor, D. Huttenlocher, L. Lamel, H. Leung, M. Randolph, S. Seneff, C. Shadle

C.J. LeBel Fellowship

Systems Development Foundation

National Institutes of Health (Grants 5 T32 NS07040 and 5 R01 NS04332)

National Science Foundation (Grant 1ST 80-17599)

U.S. Navy - Office of Naval Research (Contract N00014-82-K-0727)

Kenneth N. Stevens, Dennis H. Klatt, Joseph S. Perkell, Stefanie Shattuck-Hufnagel, Victor W. Zue

[11] Visiting Scientist

[12] Staff Member, Wang Laboratories, Inc.

[13] Associate Professor, Department of Psychology, Northeastern University

[14] Postdoctoral Fellow, Haskins Laboratories

[15] Assistant Professor, Department of Speech Disorders, Boston University

[16] Research Scientist, Department of Speech Disorders, Boston University

[17] Staff Member, Bolt, Beranek and Newman, Inc.

[18] Director, Graduate Program in Speech-Language Pathology, MGH Institute of Health Professions, Massachusetts General Hospital

[19] Assistant Professor, Department of Neurology, Boston University School of Medicine

[20] Professor of Special Education, Boston University

22.1 Studies of Acoustics and Perception of Speech Sounds

We have continued our investigations of the acoustic and perceptual correlates of the features that distinguish between utterances in different languages. The aim of these studies is to establish the inventory of acoustic properties that are used to make these distinctions, and to attempt to uncover the principles governing the selection of this inventory of properties that appear to be used across all languages. The techniques we are using in this research include acoustic analysis of utterances representative of the distinctions, and human perception of synthetic utterances in which the relevant properties are systematically manipulated. The phonetic features that we have been studying over the past year include: the nonnasal-nasal distinction for vowels in certain languages of India (as well as the influence of nasal consonants on vowels in English); the use of duration of a preceding vowel or nasal consonant to signal voicing for an obstruent consonant in Portuguese; the distinction between tense and lax vowels in English, including differences in duration, formant trajectories, and relative prominence of spectral peaks; the acoustic properties that distinguish between the liquids and glides /ℓ, r, w, y/ in English; the relative roles of formant transitions and bursts in stop-consonant perception; the distinction between velars and uvulars in Arabic; and the distinctions between dental, alveolar, and palatal consonants in certain Australian languages.

In addition to our studies of the segmental aspects of speech production and perception, we are continuing our investigation of the influence of fundamental frequency on the comprehension of sentences. We have found apparent differences between speakers in the contribution of the F0 contour in a reaction-time task involving sentence comprehension, and we are investigating the source of these individual differences.

A program of research aimed at an acoustic analysis of Japanese syllables and sentences had been initiated. The purpose is to see if it might be feasible to synthesize Japanese sentences by rule within the general framework specified in Klattalk, and further, to determine the minimal data set that would permit such an effort to succeed. We have recorded syllables and sentences from three speakers, and have made measurements of formant motions in the CV syllables and durations in sentences. Phonological recoding phenomena in sentences are being studied and tentative rules have been formulated. We are in the process of analyzing these data.

A digitized data base consisting of 129 CVC English words spoken by 12 men, 12 women, and 12 children has been acquired from a cooperating group. The words include many different CV and VC syllables. This data base will be analyzed in the coming year to determine the extent to which the synthesis-by-rule framework described elsewhere in this report can be used to characterize the data, produce meaningful averages, and perhaps suggest better ways to approach the problem of phonetic recognition by machine.

As we proceed with these phonetic studies, we are continuing to revise and to elaborate on our theoretical views concerning invariant acoustic correlates of phonetic features and the role of redundant features in enhancing phonetic distinctions.

22.2 Speech Production Planning

Spontaneous speech error patterns suggest at least two conclusions about normal speech planning:

(1) it involves the formulation of representations in terms of sublexical elements like phonemic segments (C,V), syllable onsets (CC), and syllable rhymes (VC); and

(2) it includes processing mechanisms that manipulate these units, e.g., a serial ordering process. We are examining the nature of the sublexical planning processes by analyzing constraints on and patterns in the manipulation of these units, in speech errors (both spontaneous and experimentally elicited) and pig-latin-like play languages. Recent findings provide information about the following questions:

What are the planning units? Although spontaneous speech errors often involve syllabic units like single C's and V's, consonant clusters, and -VC's, they seldom move or change whole syllables (unless the syllable also corresponds to a morpheme). This suggests that at some point in production planning, syllable-like frames provide the organizing schema for syllabic processing units. Using tongue twisters like "play cat plot coy," we are pursuing the question of how and when onset clusters like /pl/ come apart in errors, and when they function as whole units. In a second line of investigation we are examining the spontaneous and elicited output of speakers who use play languages that manipulate sublexical fragments in various ways, in order to gain insight into the nature of the representation of syllabic and subsyllabic units.

What are the representational dimensions? On the assumption that shared representational dimensions make two elements more likely to interact in an error, we have examined spontaneous errors to see which dimensions are shared by pairs of target and intrusion elements. One obvious finding is position similarity: Two target segments can interact when they appear in similar positions (like initial /t/ and /d/ in "top dog"). Elicitation experiments have shown that shared position in the word is more conducive to errors than is shared position in the stressed syllable (e.g., /ǰ/ and /d/ are more likely to interact in "july dog" than in "legit dog"). This indicates that word-sized units play a role in the representation of an utterance at the point where sublexical ordering errors occur.

A second finding is phonemic context similarity: Interacting consonant segments are often followed by identical vowels, as in /f/ and /p/ in the error "pat fack" for "fat pack." Experimental elicitation errors show this same contextual similarity constraint.

We are now using an elicitation technique to ask whether context similarity affects some error types more than others. Such a finding would support a model in which different sublexical error types (exchanges, substitutions, shifts, etc.) are associated with different aspects of phonemic processing (see next section).

What are the sublexical processing stages or mechanisms? Is there a single sublexical planning mechanism or are there several? If there are several, then it is a reasonable assumption that each is susceptible to its own particular type and pattern of errors. Against the simplest assumption of one mechanism are several findings in the elicited error data: (1) Words behave differently from nonwords. (2) Content words and function words participate differently in sublexical errors in spontaneous speech: many more sublexical errors occur in content words than in function words, even though function words are more frequent. (3) Stressed-syllable-position similarity does not interact with word-position similarity in facilitating sublexical interaction errors.

Taken together, these results suggest the existence of at least two planning mechanisms in normal speech processing that make reference to sublexical units: one that involves the organization of real words and morphemes of English into grammatically well-formed utterances, and another that controls the organization of strings of phonemic elements (whatever their grammatical status) into utterable sequences. Experiments now in progress test this hypothesis in the following ways: by comparing sublexical errors in content words vs. function words, in the recitation of prespecified lists and phrases vs. the generation of fresh new utterances, and between segments in similar phonemic contexts vs. dissimilar ones.

22.3 Auditory Models and Speech Processing

In an attempt to gain insight into the mechanisms used by human listeners in the processing of speech, we are developing models of the peripheral auditory system and we are examining the response of these models to selected classes of speech sounds. These models contain sets of bandpass filters with critical bandwidths, and process the outputs in different ways. One of the models incorporates a special procedure for detecting synchrony in the waveform at the filter outputs, and provides a representation of vowellike sounds with enhanced spectral prominences. This model is being tested with a variety of nonnasal and nasal vowels produced by speakers with a range of fundamental frequencies, and its behavior is being compared with the results of more conventional types of spectral analysis. Another model is oriented toward simulating the auditory response to speech events characterized by rapid frequency and amplitude changes, and incorporates adaptation properties similar to those observed in the peripheral auditory system. The behavior of this model is being tested with speech and speechlike sounds that simulate stop gaps of various durations and degrees of abruptness, and sounds in which various rates of spectrum change occur near an onset.

This work on implementation of auditory models has been paralleled by the continuing development of theoretical ideas concerning the potential utility of these models, and the directions for future research on auditory modeling. Ideas that have been produced in published papers include: (1) the necessity for reexamining and possibly broadening the bandwidths that are appropriate for speech analysis at low frequencies; and (2) the types of "central filtering" that are appropriate, beyond the stage of peripheral processing.

22.4 Physiology and Acoustics of Speech Production

A pilot study of jaw movements produced during selected speech and nonspeech tasks performed at different rates has been completed. Several hypotheses have been developed as a result of this study, and these have been concerned with the relation between articulatory effort and the articulatory targets selected by a subject when producing alternating jaw movements at different rates. In preparation for a more detailed study of these questions, algorithms have been developed to automatically extract movement distance, time, peak velocity and peak acceleration from displacement, velocity and acceleration signal stream data. These algorithms will be used to increase the efficiency of data processing, thereby allowing us to explore constraints on jaw movements with larger numbers of subjects and more varied experimental conditions. Experiments will be designed to test the hypotheses that were addressed in or have grown out of the pilot study.

Progress has been made in evaluating a full prototype of an alternating magnetic field transducer system for simultaneous tracking of several midsagittal-plane points inside and outside the vocal tract. The prototype electronics were completed and are being tested in conjunction with a new, small transducer. For this testing, several precision mechanical components have been designed and built. These components include a plastic jig to hold the transmitter coils and eventually mount them on a subject's head, a plastic testing device which enables us to move a transducer in circles of different radii and through straight lines with different degrees of transducer tilt, and an apparatus that enables us to precisely translate and tilt a transducer with respect to the transmitters for calibrating the field. A procedure has been devised to record data directly into the computer. Data processing algorithms have been developed to convert the raw data into Cartesian coordinates in the midsagittal plane, and an algorithm has been developed to plot midsagittal-plane trajectories of transducer displacement. The data processing and plotting algorithms have been used to evaluate the performance of the system in tests which include the input of circles and straight lines using the testing device. Several problems with the electronics and calibration techniques have been encountered and overcome, and testing is near completion. Work on the finished laboratory apparatus has begun.

In collaboration with Dr. Robert Hillman and Ms. Eva Holmberg at Boston University, we have continued work on a project on the use of non-invasive aerodynamic and acoustic measures during vowel production to study hyperfunctional and other voice disorders. Recording and data processing techniques have been further refined, and data collection and processing is proceeding on a number of normal and pathological subjects. This work is exploring the utility of measures of glottal resistance to airflow, vocal efficiency and various measures derived from glottal airflow waveforms as predictors of voice dysfunction. Preliminary work on devising test procedures has pointed out the importance of carefully controlling speaking mode and rate for obtaining oral measurements which can be used for the reliable estimation of glottal function.

Studies of the characteristics of turbulence noise sources in speech production have been continued with further measurements of the amplitude and spectrum of sound that is generated when air is passed through constrictions in mechanical models and impinges on obstacles (simulating the teeth or lips). Evidence has been gathered to show that the source of sound produced in this way has the characteristics of a dipole source, and can be modeled in one dimension by a series source of sound pressure. Further experiments and theoretical analysis is proceeding with various configurations of the constrictions and obstacles in mechanical systems that simulate human production of fricative sounds.

22.5 Speech Recognition

The overall objectives of our research in machine recognition of speech are:

— To develop techniques for incorporating acoustic-phonetic and phonological knowledge into speech recognition systems;

— To carry out research aimed at collecting, quantifying, and organizing such knowledge; and

— To develop prototype systems based on these principles.

During the past year progress has been made on several projects related to these broad objectives.

22.5.1 An Isolated-Word Recognition Model Based on Broad Phonetic Information

In 1982, we conducted a study on the structural constraints imposed by a language on the allowable sound patterns. Words were indexed into a lexicon based on broad phonetic representations, and it was found that the number of words sharing a common representation in this form is very small. As a result, we proposed an approach to isolated word recognition for large vocabularies. In this proposal, the speech signal is first classified into a string of broadly-defined phonetic elements. The broad phonetic representation is then used for lexical access, resulting in a small set of word candidates. Finally, fine phonetic distinctions determine which of these word candidates were actually spoken.

We have extended this model in several directions. First, the potential role of word-level prosodic information in aiding lexical access was examined. Again using the Merriam Websters Pocket dictionary as our database, we found that knowledge of the three-level (i.e., stressed, unstressed, and reduced) stress pattern of a given word, or simply the position of the most stressed syllable, significantly reduces the number of word candidates in lexical access. A project is underway to determine the stress patterns of isolated words from the acoustic signal while utilizing the constraints on allowable stress patterns.

Second, refinements were made to our original model in order to minimize the effect of acoustic variability and front-end errors. Specifically, lexical access is based on prosodic information, and segmental information derived from stressed syllables, where the acoustic information is known to be robust. We are developing a preliminary implementation of a word hypothesizer based on partial phonetic information.

22.5.2 Speaker-Independent Continuous Digit Recognition

We also explored for continuous speech the use of constraints like those described in the previous section. Specifically, we focused on the task of continuous digit recognition as a feasibility demonstration. In our proposed model, the first step is the broad classification of the segmentable portions of the acoustic signal. Next, the segment lattice is scanned for all possible digit candidates spanning the utterance. Finally, the best digit string is selected as the answer based on further detailed acoustic comparison of the digit candidates. We have completed a preliminary implementation of the system up to the point of lexical access. Formal evaluation indicated that the "fatal" error rate is about 1%. In other words, 1% of the time the correct digit cannot be recovered. The number of digit candidates has correspondingly been reduced by about 60%.

22.5.3 Properties of Consonant Sequences within Words and Across Word Boundaries

It is well known that, for a given language, there are powerful constraints limiting the permissible sound sequences within words. Sound sequences across word boundaries, on the other hand, are only limited by the allowable combination of words. In some cases, knowledge of the phoneme sequence in continuous speech uniquely specifies the location of the word boundary, while in other cases, phonotactic knowledge is not sufficient. For example, the word boundary can be uniquely placed in the sequence /... m g l ... /, as in the word pair "same glass", whereas the word boundary location is ambiguous in the phoneme sequence /... s t r ... / without further acoustic information. The /... s t r ... / may have a word boundary in one of three places as in "last rain", "mouse trap", and "may stretch."

As a step towards a better understanding of the acoustic phonetic properties of consonant sequences within and across word boundaries, we have studied the distributional properties of these sequences. We focused our inquiry on whether there exist structural constraints limiting the potential consonant sequences across word boundaries in English. The database consists of several text files from different discourses, ranging in size from 200 to 38,000 words. In each case we tabulated the number of distinct word-initial, word-medial, word-final, and word-boundary consonant sequences and their frequency of occurrence. It was found that, on the average, only 20% of the observed word-boundary sequences also occur in a word-medial position. Of those word-boundary sequences that can only occur across word-boundaries, approximately 80% have a unique boundary location.

The results of this study suggest that, given a detailed phonetic transcription, it may be possible to propose word boundary locations even if words are not delineated by pauses. When the consonant sequence cannot uniquely specify the location of a word boundary, different placement of word boundaries often results in allophones with substantially different acoustic characteristics, as described in the our last report.

22.5.4 Automatic Phonetic Alignment of Phonetic Transcription with Continuous Speech

The alignment of a speech signal with its corresponding phonetic transcription is an essential process in speech research, since the time-aligned transcription can provide pointers to specific phonetic events in the waveform. Traditionally, the alignment is done manually by a trained acoustic phonetician, who listens to the speech signal and visually examines various displays of the signal. This is a very time-consuming task requiring the expertise of acoustic phoneticians, of whom there are very few. Furthermore, there is the problem of the lack of consistency and reproducibility of the results, as well as human errors associated with tedious tasks.

During the past year, we have developed a system to automatically align a phonetic transcription with the speech signal. The speech signal is first segmented into broad classes using a non-parametric pattern classifier. A knowledge-based dynamic programming algorithm then aligns the broad classes with the phonetic transcriptions. These broad classes provide "islands of reliability" for more detailed segmentation and refinement of boundaries. Acoustic phonetic knowledge is utilized extensively in the feature extraction for pattern classification, the specification of constraints for time-aligned paths, and the subsequent segmentation/labeling and refinement of boundaries. Doing alignment at the phonetic level permits the system to tolerate some degree of inter and intra-speaker variability.

The system was recently evaluated on sixty sentences spoken by three speakers — two male and one female. Ninety-three percent of the segments are mapped into only one phoneme, and 70% of the time the offset between the boundary found by the automatic alignment system and a hand transcriber is less than 10 ms.

We expect that such a system will enable us to collect and label a large speech database, which will directly contribute to our enhanced knowledge in acoustic phonetics. Furthermore, the automatic time alignment system can also serve as a testbed for specific recognition algorithms in that the success of the time alignment procedure depends critically on our ability to describe and identify phonetic events in the speech signal.

22.6 Speech Synthesis

A review of speech synthesis technology has been updated. The review includes a discussion of various methods for synthesizing speech from small building blocks and a description of the techniques used to convert text to speech.

Rules for synthesis of consonant-vowel syllables using a formant synthesizer have been updated and described in a publication. It is argued that consonant-vowel coarticulation has systematic acoustic influences that are described fairly elegantly by hypothesizing up to three distinct allophones for each consonant — one before front vowels, one before back unrounded vowels, and one before back rounded vowels. Within each category, consonant spectra are remarkably invariant for one speaker, and consonant-vowel formant transitions obey a modified locus theory. It remains to be seen whether these results (quite satisfactory for synthesis purposes, as shown by CV intelligibility testing) apply to all English talkers (see analysis research described elsewhere in this report).

A text editor system for the blind has been assembled using a Rainbow personal computer and a Dectalk synthesizer to speak information that a sighted user obtains from a terminal screen during creation and editing of a text file. The software is currently undergoing evaluation and use by two blind professionals. Additional efforts are underway to get Dectalk used in other applications involving the handicapped.

In a theoretical paper, relations between synthesis by rule and linguistic theory have been discussed. Some of the solutions chosen in the search for a suitable abstract linguistic description of sentences in Dectalk, as well as some of the detailed rules for allophone selection and synthesis, may have implications for phonological theory. Synthesis activities not only provide a unique opportunity to test specific postulated rules, but may also guide thinking on the nature of universal phonological frameworks and descriptions. Examples of current descriptive controversies include a determination of the number of lexical stress levels in English. Experience with Klattalk seems to indicate that secondary lexical stress, if it is to be distinguished from unstressed/unreduced, has very restricted uses, no matter what acoustic cues are postulated to distinguish it. An example of Klattalk rules having implications for universal phonetic frameworks is the system of duration rules. Even if the detailed durational system of a language is considered a part of the language-specific feature implementation rules, ways must still be developed to represent the input to these rules adequately.

22.7 Issues of Variability and Invariance in Speech

National Institutes of Health (Grant 1 R13 NS 20005)
National Science Foundation (Grant BNS 83-11063)
Research Laboratory of Electronics

A symposium on "Invariance and Variability of Speech Processes" was held on October 8–10, 1983. The symposium was organized by members of the Speech Communication Group in collaboration with Prof. Gunnar Fant of the Royal Institute of Technology, Stockholm and Prof. Bjorn Lindblom of Stockholm University. About seventy researchers from the fields of speech production, perception, acoustics, pathology, psychology, linguistics, language acquisition, synthesis and recognition were invited. Twenty-four participants submitted drafts of papers which were distributed in advance of the meeting to serve as foci of discussion for all participants. Final manuscripts of focus and comment papers are being prepared for inclusion in a proceedings book which is scheduled for publication late in 1984.

Among the contributions to the symposium were two papers from the Speech Communication Group. One of these papers reviews currently popular speech recognition techniques in light of the variability normally seen in speech. The paper presents arguments to show why techniques such as dynamic programming, automatic clustering, vector quantization, and LPC distance metrics are limited in their success. In each case, more knowledge of the processes actually contributing to irrelevant variability, as well as knowledge of that portion of the signal contributing useful context-dependent information, would help improve recognition systems. The same paper includes a description of the IBM speech recognition system in terms that permit comparisons between it and other approaches. The "learning" technique employed by IBM is very powerful, but other design choices in the IBM system were not so fortunate. Suggestions for combining the best ideas from LAFS (Lexical Access From Spectra) and IBM are included in the discussion. The positive features of LAFS as a way to incorporate knowledge of acoustic-phonetic details into a recognition system are discussed.

Another contribution to the symposium gives a review of the concepts of distinctive and redundant features, and presents a number of examples of the enhancing properties of redundant features and their role in phonology.

22.8 Computer-Based Speech Research Facilities

A new VAX-750 computer system to replace the old PDP-9 system has been installed. Software from the PDP-9 has been rewritten or modified in order to be useful on the VAX.

KLSYN Synthesizer. A formant synthesizer has been rewritten in floating point and augmented with a new more natural voicing source. Parameter specification from terminal commands has been implemented, and a VT-125 display package has been written to permit visualization of parameter tracks. Interactive graphics will be written when a VS-100 display and graphic input device arrive.

MAKETAPES Stimulus Generation. A program has been written to generate perceptual tests and to produce both answer sheets and response sheets automatically. Test types initially available include (1) identification tests, (2) 4IAX tests, and (3) fixed standard paired comparison tests.

KLSYN Spectral Analysis. A program to replace FBMAIN has been written. It can compare spectra for several waveforms at one time, and has four different spectral representations to choose from: (1) dft, (2) spectrogram-like, (3) critical-band, and (4) linear prediction. Hard copies of spectra and waveforms can be plotted.

SPECTO Crude Spectrogram. A program has been written to produce a crude spectrogram-like printout on the line printer, and a listing of f0 and formant estimates every 10 msec. The program is useful when trying to synthesize from a natural model, and when using KLSPEC with a long utterance so as to "find your way".

KLATTALK. The Klattalk text-to-speech software has been made accessible to users. Stimuli can be prepared, or listings of parameter values can be obtained as a starting point for hand synthesis using KLSYN.

The development of an interactive speech research facility on our Lisp machines has also continued over the past year. In addition to refinements and additions to the capability of the SPIRE and SPIREX systems described in our last report, we have also developed a facility that enables researchers to study the distributional constraints imposed by a language. Researchers can ask questions such as: "How many three-syllable words in the database have primary stress on the first syllable?", or "How often is the homorganic rule violated in the database?", and obtain statistically meaningful results. The software package, including SPIRE, SPIREX, and ALEXIS, are available to interested parties through the MIT patent office.