22. Speech Communication

Academic and Research Staff

Prof. K.N. Stevens, Prof. J. Allen, Prof. G. Fant,[13] Prof. M. Halle, Prof. S.J. Keyser, Prof. V.W. Zue, Dr. W. Cooper,[14] Dr. F. Grosjean,[15] Dr. S. Hawkins,[16] Dr. R.E. Hillman,[17] Dr. A.W.F. Huggins,[18] Dr. J. Jakimik, Dr. D.H. Klatt, Dr. L.S. Larkey, Dr. B. Lyberg,[19] Dr. J.I. Makhoul,[18] Dr. E. Maxwell, Dr. L. Menn,[20] Dr. P. Menyuk,[21] Dr. J.L. Miller,[22] Dr. J.S. Perkell, Dr. P.J. Price, Dr. S. Shattuck-Hufnagel, M. Hamada,[19] E.B. Holmberg,[23] S.-Q. Wang [19]

Graduate Students

C. Bickley, F. Chen, K. Church, S. Cyphers, C. Espy, R. Goldhor, D. Huttenlocher, L. Lamel, H. Leung, M. Randolph, C. Shadle, S. Steneff

C.J. LeBel Fellowship
Systems Development Foundation
National Institutes of Health (Grants 5 T32 NS 07040-08 and 5 R01 NS 04332-20)
National Science Foundation (Grants 1ST 80-17599 and MCS-8112899)
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

[13] Visiting Professor

[14] Associate Professor, Department of Psychology & Social Relations, Harvard University

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

[16] Assistant Professor of Communicative Disorders, Emerson College

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

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

[19] Visiting Scientist

[20] Aphasia Research Center, Boston University

[21] Professor of Special Education, Boston University

[22] Assistant Professor, Department of Psychology, Northeastern University

[23] Department of Speech Disorders, Boston University

22.1 Speech Recognition

The overall objectives of our research in machine recognition of speech are (1) to develop techniques for incorporating acoustic-phonetic knowledge and knowledge of lexical constraints into speech recognition systems; (2) to carry out research aimed at collecting, quantifying, and organizing this knowledge; and (3) to develop prototype systems based on these principles. During the past year progress has been made on several projects related to those broad objectives.

22.1.1 Phonological Properties of Large Lexicons

A given language is constrained in terms of the possible ways that speech sounds or segments can be combined to form meaningful words. Knowledge about such constraints is implicitly possessed by native speakers of a given language. For example, a native English speaker knows that "vnuk" is not an English word. He/she also knows that if an English word starts with three consonants, then the first consonant must be an /s/, and the second consonant must be either /p/, /t/, or /k/. Such knowledge is presumably utilized by speakers of a given language in the process of speech perception, particularly when the acoustic cues of a speech sound are missing or distorted, and it would clearly be advantageous to incorporate knowledge of this kind into a speech recognition system.

We recently conducted a study of the phonotactic constraints of American English by examining the phonemic distributions in the 20,000–word Merriam Webster's Pocket Dictionary. In one part of this study we mapped the phonemes of each word into one of six broad phonemic categories: vowels, stops, nasals, liquids, and glides, strong fricatives, and weak fricatives. (For example, the world "speak", with a phonemic string given by /spik/, is represented as the pattern: [strong fricative] [stop] [vowel] [stop].) We found that, even at this broad phonetic level, approximately one–third of the words in a 20,000–word lexicon can be uniquely specified. If we define the "cohort" of a particular pattern as the set of words having that pattern, then the size of the cohort is the number of words having that pattern. (For example, the words "speak" and "steep" belong to the same cohort.) The average cohort size for the 20,000–word lexicon was found to be approximately 2, and the maximum cohort size was approximately 200. In other words, in the worst case, a broad phonetic representation of the words in a large lexicon can reduce the possible word candidates to around 1% of the lexicon. Furthermore, over half of the lexical items belong to cohorts of size 5 or less. These results demonstrate that speech recognition systems can use broad acoustic-phonetic classifications of words to reduce the number of possible word candidates to a very small set. Such broad classifications can be performed more reliably than detailed phonetic recognition.

Phonetic variability is not evenly distributed across words and sentences. Certain segments of a word are relatively phonetically invariant, while others are highly variable. For example, the /n/'s in "international" cannot be deleted, while the /t/ and the reduced vowels can be. A set of experiments we are running indicate that the phonetically variable segments of a word provide much less lexical constraint than phonetically invariant segments. That is, the phonetically variable parts of a word do not aid in differentiating it from other words in the lexicon. In one experiment we used a 6–class broad phonetic representation like the one in our original study. However, in this case, segments in unstressed syllables were ignored unless they were nasals, glides or strong fricatives. So, for instance, "international" would be classified as [vowel] [nasal] [nasal] [vowel] [fricative] [nasal] [glide], which does not depend on phonetically variable parts of the word. This classification scheme partitions the 20,000 word lexicon such that 47% of the words were in cohorts of size 5 or less, which compares favorably with the 54% obtained in the experiment described above. This result demonstrates that speech recognition systems can use broad phonetic classifications to find a small set of possible matching words, even in light of the high phonetic variability in natural speech.

22.1.2 Lexical Access

It is well–known that phonemes have different acoustic realizations depending on the context. Thus, for example, the phoneme /t/ is typically realized with a heavily aspirated strong burst at the beginning of a syllable as in the word "Tom", but without a burst at the end of a syllable in a word like "cat". Variations such as these are often considered to be problematic for speech recognition since they can be viewed as a kind of 'noise' that makes it more difficult to hypothesize lexical candidates given an input phonetic transcription.

In our view, the speech utterance is modeled in terms of two types of acoustic/phonetic cues: those that vary a great deal with context (e.g., aspiration, flapping) and those that are relatively invariant to context (e.g., place, manner, voicing). We have designed a recognizer to exploit variant cues by parsing the input utterance into syllables and other suprasegmental constituents using phrase–structure parsing techniques. Invariant constraints are applied in the usual way to match portions of the utterance with entries from the lexicon.

Only part of the proposed recognizer has been implemented. We have written a program to parse lattices of segments into lattices of syllables and other phonological constituents. Phonological constraints are expressed in terms of phrase–structure rules. For example, we could restrict aspiration to syllable initial position (a reasonable first approximation) with a set of rules of the form:

1. utterance --> syllable

2. syllable --> onset rhyme

3. onset --> aspirated–t | aspirated–k | aspirated–p | ...

4. rhyme --> peak coda

5. coda --> unreleased–t | unreleased–k | unreleased–p | ...

This sort of context-free phrase-structure grammar can be processed with well-known parsers like Earley's Algorithm. Thus, if we completed this grammar in a pure context-free formalism, we could employ a straightforward Earley parser to find syllables, onset, rhymes and so forth. However, it has been our experience that the context-free formalism is not exactly what we want for this task. To serve this purpose, we have implemented a simple language of matrix (lattice) operations. This implementation appears to be easier to work with than many others.

22.1.3 Acoustic Cues for Word Boundaries

Words in continuous speech are rarely delineated by pauses. There are, however, acoustic cues indicating word and syllable boundaries, and knowledge of these cues often helps a listener or a speech recognizer to segment an utterance into words and syllables. For example, the /tr/ sequence in word pairs such as grey train and great rain have very different acoustic realizations although their phonemic transcriptions are essentially the same.

As a pilot study, we have investigated acoustic cues for word boundaries at labial stop-sonorant clusters, similar to the example given above. In this study minimal pair phrases were embedded in carrier sentences in order to enhance any differences between the members of the pairs. The phrases were constrained to provide identical surrounding context for the stop and sonorant under investigation. The format of the phrases was given by:

V(#)labial stop(#)sonorant V,

where V represents a vowel and (#) an optional word boundary. In total there were 15 minimal pair phrases representing the clusters /bl/, /br/, /pr/, and /pl/, such as sheep raid versus she prayed and bay block versus Babe lock. Three male speakers recorded the minimal pairs in the carrier phrase was "I said - - - - not - - - - ." Each member of the minimal pair occurred in both the sentence medial and sentence final positions of the carrier phrase. The study indicated that:

—The sonorant is devoiced when in a cluster with an unvoiced consonant.

—The onset of a word-initial /r/ is more gradual than when the /r/ is in a cluster.

—Good measures to differentiate the minimal pairs include: voice-onset time, stop-gap duration, sonorant duration, and formant frequency measurements such as F_2 and (F_2 - F_1) for /   / and F_2 and F_3 for /r/.

A more extensive study of the acoustic cues for word boundaries is now being conducted. We will be concentrating on the thirty-odd allowable word-initial clusters in English and how the acoustic cues differ when a word boundary divides that cluster.

A supporting study is examining the statistics of allowable consonant clusters in word-initial, word-final, and word-internal positions in words. These data will make it possible to determine the probability that syllable or word boundary occurs when a particular consonant sequence is detected.

22.1.4 Speaker-Independent, Continuous Digit Recognition

An acoustic-phonetically based continuous digit recognizer is being developed to study the utility of phonetic knowledge in a recognition system. The system is also intended to demonstrate a modular design, so that it can be expanded by modification of knowledge sources. The digit vocabulary was chosen because it forms a constrained task under which many coarticulation and prosody effects occur. This allows one to develop a system, in a constrained environment, which can handle many of the effects occurring in continuous speech. The recognizer will embody concepts found to be important when reading spectrograms. These concepts include using robust information, considering context, using multiple cues and checking for consistency, using acoustic knowledge for initial recognition and then using higher level knowledge to constrain the possibilities, and initially listing of all possibilities and then ruling some out based upon acoustic evidence. These ideas will be incorporated by structuring the system to combine robust, bottom-up information with syntactic and lexical top-down constraints.

The system is to be composed of modular components containing various types of knowledge. The types of knowledge may be acoustic, phonetic, phonological, and lexical. Each component is to be a separate unit which may be modified independently. This modularity will enable one to expand the system to larger vocabularies. The modular system under development is to serve as the framework upon which extensions to larger vocabularies can be made.

22.1.5 LAFS Recognition Model

A LAFS (Lexical Access From Spectra) network has been generated for the task of recognizing connected digits. It is hoped that this representation will be appropriate for speaker-independent connected digit recognition, but preliminary experiments indicate that the spectral distance metric initially proposed for LAFS (a weighted slope metric in the critical-band spectral domain) is not good enough for the task. Current research is directed at the discovery of better metrics.

22.1.6 Interactive Speech Research Facilities

As we mentioned earlier, part of our research goal is to describe and quantify the acoustic characteristics of speech sounds in various phonetic environments. Since such studies often involve the examination of a large body of data, it is essential that we provide a graceful, interactive environment for speech research.

Our research effort in speech recognition makes extensive use of an interactive research facility called SPIRE, developed at M.I.T. specifically for acoustic phonetic research. The SPIRE system draws heavily on the capabilities of a unique personal computer, called the Lisp machine, developed at M.I.T.'s Artificial Intelligence Laboratory. The SPIRE system allows users to record, transcribe, store, and retrieve spoken utterances. Users are able to configure the high-resolution display to include spectrograms, spectral slices, transcriptions, LPC parameters, energy measures, and other parameters computed from speech. This is accomplished interactively with a pointer, which can also be used to allow a user to listen to sections of the speech, edit waveforms, examine data values, alter display options, and perform other functions.

In addition to SPIRE, an acoustic phonetic experimental facility is also being implemented. This facility, called SPIREX, provides an environment for speech researchers to formulate and execute complicated acoustic phonetic experiments. With the help of SPIREX, statistically meaningful results can be obtained in a convenient manner.

22.2 Auditory Models and Analysis Techniques

Our research on speech recognition and on perceptual correlates of phonetic features has led us to ask whether the relevant acoustic properties associated with phonetic categories might be represented in a more prominent fashion if the speech signal were first processed by a model of the peripheral auditory system. We are developing two such models that are designed to represent different temporal and spectral attributes of speech sounds. Both models process the speech through a set of bandpass filters with critical bandwidths. In one model, special attention is given to the response of the auditory system to signals exhibiting rapid frequency and amplitude changes. In the other model, the processing takes advantage of synchrony in simulations of nerve firing patterns to enhance formant peaks and to extract fundamental frequency.

An invited paper has been written reviewing the status of work on auditory modeling as it relates to models of speech perception. Conclusions were that critical-band filtering, or filtering with wider bandwidth low-frequency filters was a very good idea, but that the field is unsettled in that we do not know the importance of the time-locked aspect of neural encoding. This paper was presented at a conference in Stockholm.

22.3 Speech Synthesis

The formant synthesizer that is used for stimulus preparation in a number of perceptual studies has been rewritten in the language C in anticipation that the PDP-9 computer will be replaced by a newer machine. The synthesizer voicing source has been improved in flexibility by adding control parameters that govern the ratio of the open phase to the total period, the degree of rounding of the waveform corner at glottal closure, the amount of breathiness, and the amount of alternating jitter.

A 20,000-word phonemic dictionary has been hand-edited and converted to speech using a synthesis-by-rule program in order to verify the correctness of the pronunciations. Then an automatic pattern discovery program was used to try to find letter-to-phoneme rules in English.

Results suggest that simple pattern discovery procedures are not sufficiently powerful, and syllable parsing strategies are needed to discover appropriate rules for vowels.

Student projects have begun in the use of a text-to-speech system in a text editor for the blind and on synthesis of Japanese segments by rule.

22.4 Physiology of Speech Production

An overall aim of our work on the physiology of speech production is to develop models for the control of the various articulatory structures as they are manipulated to achieve a desired sequence of acoustic properties for an utterance. In the course of this research we are developing new techniques for the measurement of articulatory movements and for the analysis and processing of multi-channel articulatory data.

During the past year we have completed a cross-language pilot study of sources of variation in strategies for anticipatory coarticulation of lip rounding. A motivation for this work is to determine the extent to which different speakers use different strategies for producing anticipatory movement patterns for lip rounding depending on the language. Results of an experiment on lip retraction for "neutral" consonants occurring between rounded vowels show that variation in anticipatory movement patterns may be related to language-specific patterns of vowel diphthongization.

In collaboration with Dr. Winston Nelson of Bell Laboratories and Dr. John Westbury of the University of North Carolina we have completed a study on three subjects of repetitive jaw movements for speechlike and non-speech movement tasks. Results vary widely across subjects. One subject has movement patterns which appear to be governed by physical constraints at high repetition rates and factors related to speech-motor control skill (such as the possible use of auditory feedback) at lower rates. For this subject there appears to be a region in which both types of influences could be interacting with one another, at about 4 Hz. Movement patterns for the other two subjects suggested that they were less capable of approaching these boundary conditions, perhaps because they were less "skilled" at performing the rather unusual task. Techniques are being developed to explore these findings further.

In collaboration with Dr. Robert Hillman and Ms. Eva Holmberg at Boston University, we have begun work on a project on the use of non-invasive aerodynamic and acoustic measures to study hyperfunctional and other types of voice disorders. Techniques have been worked out to record intra-oral air pressure, oral air flow and the acoustic signal in a way which enables us to derive a number of indirect measures of vocal function. Data processing algorithms have been developed for determining glottal resistance to air flow, vocal efficiency, and parameters derived from the glottal air flow waveform (with the use of inverse filtering). Initial results on normal subjects show substantial agreement with the work of others and with aerodynamically and acoustically based theory. Manipulation of pitch is reflected in changes in parameters describing the shape of glottal air flow waveforms, and manipulation of loudness is reflected in parameters related to pulmonic driving forces and force of vocal-fold adduction. There were also results that were inconsistent with previous work, and these will be explored further.

In a related study, an inverse filtering technique has been used to observe differences in glottal waveforms of normal, creaky, and breathy vowels as spoken at different pitches by two female and two male talkers.

Progress has been made in the development of an alternating magnetic field transducer system for simultaneous tracking of movements of several midsagittal-plane points inside and outside the vocal tract. A new, considerably smaller (4 mm x 5 mm x 2 mm) biaxial, tilt-compensating transducer-receiver has been designed and built and is being tested. The design of high-quality transmitter and receiver electronics has been worked out and a full prototype is under construction.

Physiological data processing software has been enhanced considerably. The functionality of blocked-data type of processing has been increased to include ensemble averaging, sequence plotting and interactive extraction of time and frequency-domain data. Increased programmability now makes the blocked-data signal processor an extremely flexible and useful tool for data processing and extraction. Software for the processing of simultaneous acoustic and palatographic data has been developed and is near completion.

22.5 Acoustics of Speech Production

Acoustical theories of speech production are generally based on a source-filter model. In the simplest version of the theory the source is either the quasi-periodic glottal air flow or is an aperiodic source resulting from turbulence in the flow in the vicinity of a constriction or obstacle. Recent theoretical and experimental studies have been leading to a more adequate description of these sources and of the interaction between the sources and the acoustics of the vocal tract. As a part of his activities during his visit to M.I.T., Gunnar Fant carried out detailed theoretical studies of the glottal waveform, its interaction with the supraglottal tract, and its variation with frequency. This theoretical work was based in part on experimental data collected at Dr. Fant's laboratory in Sweden.

Studies of the characteristics of turbulence noise sources in speech production have been initiated with a series of measurements of the spectrum of sound that is generated when air is passed through mechanical models containing constrictions. The measured sound spectrum is compared with the spectrum that is calculated for an ideal model with an idealized pressure or velocity source located at various points in the vicinity of the constriction. The aim of these studies is to determine how to represent the source characteristics for turbulence noise in speech, including the distribution of the sources in the vocal tract, the spectrum of the sources as a function of position, and the interaction between the sources and vocal-tract acoustics.

22.6 Speech Production Planning

We continued to test a speech production planning model against predicted patterns in both spontaneous and experimentally-elicited speech errors, asking two questions:

1. What are the planning units (as reflected in error units) and how are they represented (as reflected in consistent similarity constraints linking target elements with the intrusions that replace them)?

2. Are there several separate steps in the planning process (as reflected in differences in the constraints governing the interaction of two elements)?

Observations so far suggest the following points:

1. At least one serial-ordering mechanism operates over individual phonemic elements; errors like "lip yoke" -> "yip loke," involving single phonemic segments, are substantially more common than errors involving individual features or syllabic components like - VC or CC - .

2. Phonemically similar elements (like p/f) are more likely to interact than are phonemically dissimilar ones (like g/m), suggesting that the planning representation captures those facts at the point where segmental errors occur.

3. While one planning process is sensitive to errors at all word locations, a second process is apparently protected against errors in word-final position. This second mechanism comes into play when well-formed phrases (rather than lists) are planned. Thus, errors like "leap note -> [nip lot] and -> [lit nop] are both common, but when these words are embedded in phrases ("From the leap of the note"), word-final errors are rare.

4. The segment-serial-ordering mechanism is significantly more likely to confuse two word-initial segments (as in "July dog") than two stressed-syllable-initial segments (as in "largesse dog"). Moreover, these two similarity factors do not interact, again suggesting two separate error-prone mechanisms.

Ongoing experiments and analyses are addressed to three further questions: (1) Does the similarity of adjacent contextual elements affect the likelihood that two segments will interact in an error, e.g. will there be more f/p interactions in "fad pan" than in "fad pin"? (2) Do different error types, like exchanges and substitutions, occur under different similarity constraints, suggesting that they reflect two separate planning mechanisms? (3) Do function words (prepositions, pronouns, wh-words, etc.) participate in segmental errors with content words, or are they processed separately?

In addition, a range of error elicitation methods have been developed, from sentence generation based on triplets of specified words, through recitation of nonsense syllable strings from memory.