15. 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. T. Carrell, Dr. C. Chapin Ringo, M. Cohen, Dr. B. Delgutte[11], Dr. M. 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], Dr. H. Kawasaki, Dr. D.H. Klatt, K. Krohn, Dr. L.S. Larkey, Dr. B. Lherm[11], Dr. J. Locke[18], Dr. S. Maeda[11], Dr. J.I. Makhoul[17], Dr. L. Menn[19], Dr. P. Menyuk[20], Dr. J.L. Miller[21], Dr. J.M. Pardo Munoz[11], Dr. J.S. Perkell, Dr. P.J. Price, Dr. S. Shattuck-Hufnagel, Dr. J. Vaissiere[11], T. Watanabe[11]

Graduate Students

A. Alwan, A.M. Aull, C. Aoki, C. Bickley, F. Chen, S. Cyphers, C. Espy, J. Glass, R. Goldhor, C. Huang, D. Huttenlocher, R. Kassel, L. Lamel, N. Lauritzen, H. Leung, J. Pitrelli, M. Randolph, S. Seneff, C. Shadle, J. Unverferth

C.J. LeBel Fellowship
Kurzweil Applied Intelligence
National Institutes of Health (Grants 5 T32 NS07040 and 5 R01 NS04332)
National Science Foundation (Grant 1ST 80-17599)
Systems Development Foundation
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, Neurolinguistics Laboratory, Massachusetts General Hospital Institute of Health Professions

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

[20] Professor of Special Education, Boston University

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

15.1 Speech Recognition

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

- to develop techniques for incorporating acoustic-phonetic knowledge, as well as other relevant linguistic 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.

15.1.1 Lexical Stress and its Application in Large Vocabulary Speech Recognition

This research effort is focused on two related issues concerned with the determination of lexical stress for isolated words from the acoustic signal. First, considering prosodic information as a separate source of knowledge, we investigate the amount of lexical constraint provided by stress information. Second, we implement a system that derives the stress information from the acoustic signal. In order to determine the lexical constraints provided by stress information, the polysyllabic words in the Merriam-Webster Pocket Dictionary are mapped into their corresponding stress pattern classes. The results indicate that, from stress and syllable information, the largest class size constitutes 22% of the lexicon. An overall expected class size of 19% illustrates the constraining power of the stress information.

Prior to the actual development of the stress determination system, we studied the acoustic correlates of stress based on a database of 350 words, spoken by seven talkers, three male and four female. We then designed and developed a system that determines the stress pattern of isolated words and performs subsequent lexical access. The system initially segments the speech signal into broad phonetic classes. The initial segmentation provides pointers to regions of the acoustic signal that are sonorant. For some words, such as Massachusetts, the detection of each syllable is relatively straightforward, since each sonorant region, as determined by the initial segmentation, corresponds to a syllable unit. For other words, such as yellow or anxiety, however, further processing focusing on the contextual (temporal) information is necessary. Once the sonorant regions of the syllables are established, known acoustic correlates of stress, such as duration, energy, and fundamental frequency, are extracted from each syllable. A feature vector is then associated with each syllable. A relative comparison of the feature vectors within a word provides a normalization for interspeaker variability, such as speaking rate or pitch. The location of stress is determined by a Euclidean distance measure between the feature vectors and an 'optimum' feature vector, which represents the maximum value for each of the features. Finally, lexical access based on the derived stress pattern provides a list of word candidates. Phonological rules are incorporated to account for variations in the number of syllables from the lexical baseforms.

The system was evaluated on a database of 1600 isolated words, spoken by 11 speakers, six male and five female, with varying degrees of difficulty for syllable detection. System performance can be evaluated according to several criteria. The system correctly labels the stressed syllable for 98% of the words. This performance is measured with no regard for potential errors either in the number of syllables or the distinction between unstressed and reduced syllables. If we further require that the number of syllables be correct, then the performance of the system is reduced to 90%. Stipulating that the entire stress pattern be correctly labelled further reduces the system performance by 3%. Finally, lexical access, based on stress pattern information alone, results in retrieving the correct word in 83% of the cases. Most of this additional error can be attributed to the lack of proper translation of acoustic realizations into well-formed phonological rules.

15.1.2 Detection of Nasalized Vowels in American English

This research is concerned with the analysis of nasalized vowels and also the design, implementation, and evaluation of a set of algorithms to detect nasalized vowels. The study utilizes a database originally collected for a larger study of the properties of nasals in American English. The database consists of over 1200 words, excised from continuous speech and recorded from six speakers, three male and three female. All of the recorded words were digitized, and their phonetic transcriptions were aligned with the speech waveform. Using the SPIRE and SPIREX speech analysis tools, acoustic features common to all nasalized vowels were determined. These features included the presence of a low-frequency peak in the short-time spectra, the amplitude of this peak relative to that of the first formant, and a measure of the broadness of the low-frequency peaks.

The detection algorithms were developed in several steps. First, we established the baseline performance of humans on the determination of nasalization in vowels. This was achieved by constructing a set of stimuli consisting of vowels excised from their natural environments. A balanced set of non-nasalized and nasalized vowels were presented to a panel of 20 listeners in a detection test. Corresponding spectrograms were presented to eight subjects who were experienced in detecting nasalized vowels from visual examination of spectrograms. The results indicated that the feature of nasalization in vowels could be correctly identified 60% of the time from either listening or spectrogram reading.

Thus far, we have attempted several different approaches to the problem of detecting nasalized vowels. These approaches differ both in the feature vectors (e.g., LPC coefficients versus measurements made from the short-time spectra), the distance metrics for measuring similarity, and the decision strategy. Our results, based on the evaluation of data from six talkers, indicate that nasalized vowels can be identified from non-nasalized vowels with an accuracy ranging from 70% to 75%. Knowing whether the talker is male or female, or whether the vowel is high or low, increases the performance significantly.

Efforts are under way to detect nasal consonants in continuous speech.

15.1.3 Lexical Access from Spectra (LAFS)

As a contribution to the conference on Variability and Invariance, held at M.I.T. in October, 1983, a paper that updates the LAFS model of speech recognition/perception has been prepared. The paper points out the advantages of using the LAFS approach as a way to incorporate knowledge of acoustic-phonetic details into a recognition system. Unfortunately, it does not seem wise to attempt to build a LAFS model until a better spectral representation/distance metric has been devised, even though it has been found that a metric operating on the slope of the spectrum has been shown to work better in a recognition context that other popular metrics.

15.2 Speech Synthesis

Several projects concerned with improvement of the Klattalk text-to-speech system have been continued or initiated during the past year. The Klattalk program that runs on the Speech Vax has been modified to put out parameter files that can be read by a hand synthesis program called KLSYN. Several students have used this facility as a quick way to obtain a first cut at the synthesis of a word or phrase. The same capabilities have been used in informal experiments in which several combinations of glottal source control parameters are inserted by hand in order to optimize the voice quality.

Intonation rules have been improved by permitting the input of a 'new paragraph' symbol. This symbol causes a greater pause before speaking and elevates the sentence-initial fundamental frequency by raising the baseline of the hat pattern for the first second of speech. Furthermore, the user can now override the program's choice of where the hat rises and falls occur by placing the input symbols 'hat-rise' and 'hat-fall' before the appropriate vowels. In addition, the amount of hat rise or fall, as well as the amount of impulsive rise-fall on each primary stressed syllable can now by controlled explicitly by attaching appropriate numbers to these symbols and the stress symbol.

It requires a fairly sophisticated user to make optimal use of these control possibilities, but preliminary work suggests that this approach is preferred over the older method in which fundamental frequency targets were attached to each phonemic input symbol, after which the program drew straight lines between the specified targets. For example, a student at the Media Lab is going to try to simulate the Liberman/Pierrehumbert f0 rules using this input representation. The flexibility of the new input format is being used to try out various new rules for pitch control before implementing them as code.

The speech perception group at Indiana University has been conducting formal evaluations of the intelligibility of Klattalk/DECtalk. Results from the modified rhyme test, Harvard sentences, and Haskins semantically anomalous sentences indicate that DECtalk has half the error rate of the best other system tested, which was MITalk-79, about one quarter the error rate of the Prose-2000, and one tenth the error rate of the Votrax Type-n-Talk. On the other hand, the error rate is about 5 times as great as is obtained with natural speech, indicating that further work is needed. Place of articulation for final nasals and nonstrident fricatives are the most serious remaining problems.

The error rate for the female voice of DECtalk is almost twice the rate for the standard male synthetic voice. Current research has focused on examination of the likely source of each error, and the formulation of better rules to correct segmental errors.

A mini-seminar entitled "Using DECtalk as an Aid to the Handicapped" was organized at the 1984 Convention of the American Speech-Language and Hearing Association. The talk, which is currently being written up, describes the history of synthesis by rule and the current technology. David Pisoni gave a paper on comparative evaluation of current text-to-speech systems. Joan Forman described efforts of the Digital Equipment Corporation to provide DECtalk hardware and engineering support to over a dozen efforts to provide new technology to the handicapped. These include a text editor for the blind, a portable speaking aid for a young lady in a wheelchair, a talking phone message system for a non-speaking NASA scientist, several other specially designed speaking aids, automated generation of cassette tapes of technical books, and a speaking video-text service for a blind person with a personal computer and modem.

This mini-seminar will be repeated at a meeting of specialists January 31, 1985 in New Orleans at the request of ASHA.

A preliminary version of a phonemic synthesis by rule program for Japanese has been put together. While based on English Klattalk, the new program is intended to be more general so that researchers attempting to synthesize other languages can use it as a framework.

Extensive computer-assisted spectral analysis of Japanese syllables and sentences has been done, with the goal of formulating rules to synthesize Japanese sentences by rule within the general framework specified in Klattalk. We have recorded syllables and sentences from three speakers, and have made spectrograms of all the recordings. Measurements have been made of formant motions in the CV syllables. Phonological recoding phenomena in sentences have been studied and tentative rules have been formulated.

We have also measured segment durations in a corpus of several paragraphs of Japanese. On the basis of these data, we have written a computer program consisting of a set of duration rules and an evaluation metric for the rule system. As we try new rules, or optimize the constants in the rules already present, we obtain a statistical measure of the fit of the predicted durations to the data. A set of rules having about the same accuracy as the English rules of DECtalk now exists and will be documented in a future publication.

Fundamental frequency rules for Japanese have been formulated on the basis of published papers by Fujisaki and others. The rules are now being tuned to fit the data from our speaker. A new fundamental frequency extraction algorithm (based on the perceptual model of J. Goldstein) that has been programmed on the Vax computer has proven to be useful in producing fundamental frequency plots that can be aligned with the Vax spectrogram plots to facilitate interpretation.

15.3 Physiology of Speech Production

15.3.1 A Pilot Experiment on Coarticulation and Trading Relationships between Tongue Position and Lip Protrusion to Accomplish the /s/-/š/ Contrast

The hypothesis for this experiment is based on acoustic considerations: in order to produce a sufficiently large cavity in front of the constriction, lip protrusion for /š/ would be greater in an environment containing consonants that require tongue fronting (e.g., following /θ/) than in an environment requiring tongue backing (following /k/). Measurements were made in three subjects of maximum lip protrusion during the consonants /s/ and /š/ as they occurred in a variety of environments, including real word combinations and nonsense utterances. Lip protrusion was reliably greater for /š/ following a fronted consonant (/θ/) in only one case for one subject, that of nonsense utterances spoken by the experimenter. The lack of reliable differences for all other cases was interpreted as evidence against the hypothesis.

15.3.2 An Alternating Magnetic Field System for Tracking Articulatory Movements

Construction and evaluation of a prototype system for tracking multiple articulatory movements in the midsagittal plane has been completed. The measurement area is 150 mm x 150 mm, allowing for coverage from the rear wall of the pharynx to the lips and the bridge of the nose to the base of the tongue. The system uses alternating magnetic fields generated by two transmitter coils which are mounted in a magnetically-transparent assembly. A device has been constructed that mounts the transmitter assembly on a subject's head and allows positioning of the measurement area to suit individual subjects. Receiver-transducers consist of small (4 mm x 4 mm x 2 mm) biaxial inductors which are attached to articulatory structures and have fine lead wires that are connected to demodulating circuits. The biaxial transducer design provides measurements that are independent of transducer tilt. Output voltages are low-pass filtered, digitized and converted to Cartesian coordinates in software. Output can be displayed as X and Y coordinates vs time or X vs Y trajectories.

Testing of the system has shown that: (1) individual transducers have slightly different characteristics, requiring transducer-specific field calibration, gain, and tilt-correction factors; (2) accuracy, repeatability and resolution of measurements are better than 0.25 mm; (3) measurements are not affected by amalgam dental fillings, off-midsagittal plane transducer placement of up to 5 mm and proximity of two transducers of as little as 8 mm; (4) circles produced with a radius as large as 75 mm at transducer tilt angles of 0, + and - 30 degrees are concurrent; and (5) absolute distance measurements can be affected by the precision with which a transducer is oriented when being attached to an articulator. A pilot vowel-repetition experiment has demonstrated the usefulness of the system.

PC boards for transmitter and receiver electronics have been manufactured and are being tested in preparation for construction of a full, multichannel system capable of tracking the movements of up to seven transducers, with two additional fixed transducers to be used for a maxillary frame of reference. Receiver circuits are designed so they can readily be adapted to take advantage of a possibly superior scheme for correcting for transducer tilt.

15.3.3 Acoustic Characteristics of Fricatives and Fricative Models

In a continuation of our studies of the characteristics of turbulence noise sources in speech production, we have compared spectra of spoken fricatives and of mechanical models designed to mimic fricatives, in order to test the applicability of such models to speech. Power spectra were computed from recordings of five speakers (two male, three female) uttering the sustained fricatives [φ, f, θ, s, š, x]. Mechanical models were assembled using a variety of articulatory and aerodynamic data and previous work with models. The models fell into three groups according to source location and type. To produce [s, š] an obstacle is required at the approximate location of the teeth. [φ, f, θ] have a very short front cavity; the jet of air must be located near the tube wall or other surface to mimic the role of the lips. For [x, ç] the palate and possibly alveolar ridge act as a surface anterior to the constriction at which turbulence is generated. Although perfect spectral matches are not attained in all cases, the mechanical models can be used to test the adequacy of source-filter representations of the fricatives.

15.3.4 Objective Assessment of Vocal Hyperfunction

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. In a pilot study, we have completed recording, data processing and data extraction on 30 normal and dysphonic subjects, and we are currently performing statistical analyses of these data. We have also begun work on an expanded project and we are currently performing several studies designed to improve experimental techniques.

15.3.5 Laryngeal Modelling

A two-mass model of the vocal folds has been developed, and the behavior of this model has been examined when it is coupled to vocal-tract configurations with various degrees of constriction. More extreme constrictions during vowellike articulations result in a less peaked and somewhat longer waveform of the glottal volume-velocity pulse, leading to a spectrum with more low-frequency energy. This effect is due to the increased acoustic mass and resistance of the supraglottal airways, as discussed by M. Rothenberg [in K.N. Stevens and M. Hirano (Eds.) Vocal Fold Physiology, (University of Tokyo Press 1981)], who assumed a fixed glottal area function. The acoustic behavior of the constricted supraglottal tract causes large fluctuations in the pressure in the glottis within a cycle of vibration, and these pressure changes can influence the waveform of vocal-fold displacement (and of glottal area function). Data from the model have implications with regard to the phonetic classification of sonorant segments according to degree of vocal-tract constriction.

15.4 Speech Production Planning

Building on the model of speech planning proposed earlier on the basis of constraints observed in sublexical speech error patterns, current research has two goals. The first is to distinguish among different sublexical processing mechanisms by differences in their error patterns, and to provide evidence for their ordering during the planning process. The second is to take the first steps toward integrating a proposed slots-and-fillers model with prosodic theory, particularly its claims about the association of abstract segments with previously obtained 'timing slots.' Experiments and error analyses directed toward these ends have been focused on three main questions:

(1) What larger units are actively represented at the point where segmental interaction of errors occur?

Evidence from error elicitation experiments shows that position in the word is a powerful determinant of segmental errors, and is more significant than either (a) position of the segment in the syllable or (b) the lexical stress of the syllable. This suggests that at the point in production planning when single-segment errors arise, the word is still a significant element in the representation. This finding is compatible with the model proposed earlier, in which segmental errors arise during processing of items stored together in a buffer arranged as tables of words or morphemes.

Similarity of position in the syllable is also a powerful constraint on error interactions, since elicited errors almost never occur between initial and final syllable position. However, syllable position similarity by itself is apparently not adequate to provoke interaction errors in the absence of other similarities such as stress. One interpretation of this set of results is that while both the word positions and the syllable positions of segments are effectively represented at the point when segmental errors occur, it is the word frame rather than the syllable frame that is actively involved in the process whose malfunction results in segmental errors.

(2) Which kinds of errors are differentially affected by variations in such speech dimensions as the grammatical shape of the utterance or its real-word versus nonword status?

Earlier experiments showed that lists and phrases elicit errors in different patterns: identical stimulus words show substantially fewer errors in final position when they are uttered in the phrase condition than when they are uttered in the list condition. This 'final-position protection effect' raises the possibility that segmental or sublexical errors arise by two separate error mechanisms: one which is subject to errors primarily in word-initial position and is insensitive to the list/phrase distinction, and another which is subject to errors in non-initial positions and is highly sensitive to the requirements of integrating words into grammatical phrases.

The hypothesis that there are two segmental error mechanisms draws further support from another difference in elicited errors. This one arises for real word versus nonwords stimuli. If the same target consonants are spoken in nonwords (versus real words), the number of initial-position errors goes up significantly, while the number of non-initial errors remains unchanged. Again, this suggests the possibility that the planning process includes two distinct loci at which segmental or sublexical errors can occur: one sensitive to the word status of the stimulus and affecting initial segments only, and the other insensitive to word status and affecting segments at other locations.

(3) Finally, what is the acoustic-phonetic evidence to support the claim, often made on the basis of perceptual intuition alone, that segments introduced by errors undergo the same phonetic adjustments to their immediate context as correct target segments?

This claim is an important part of the argument that segmental errors occur when the segments in the utterance being planned are represented abstractly, and that later processing integrates all segments into their context appropriately. Pairs of monosyllables like 'pack pig' occasionally exchange their final segments in error elicitation experiments; preliminary evidence for ten such exchange errors suggests strongly that the adjustment for duration of voicing appropriate to the voiced or voiceless character of the final segment is made for error-introduced segments as well as for correct target segments. This observation supports the hypothesis that segment interaction errors occur before at least some phonetic adjustment processes during speech production planning.

Current experiments address the two major goals of differentiating and ordering planning mechanisms that are subject to different kinds of errors, and integrating the model with prosodic theory, in several further ways. These include: examining the effect of similarity of vowel context on consonantal errors; following up hints of a position similarity constraint for a unit larger than the word; and determining the constraints on possible voluntary manipulation of sublexical elements, using language-game-like experimental paradigms. Other areas under investigation include the acoustic-phonetic correlates of lexical stress in languages other than English, and the use of lexical stress information in the perceptual placement of boundaries in non-lexical stimuli.

15.5 Studies of the Acoustics and Perception of Speech Sounds

We have continued a series of studies aimed at specifying the acoustic and perceptual correlates of the phonetic features that are used to signal phonetic distinctions in language. Recent progress on some of these studies is summarized below.

15.5.1 Vowel Perception, the "Center of Gravity" Effect, and Synchrony Models

Chistovich and her colleagues have observed that when the frequency spacing between two formants is within about three critical bands, the auditory percept is like that of a single equivalent formant, in that a change in relative formant amplitude has about the same perceptual effect as a change in formant frequencies (shifting the center of gravity of the collective energy concentration, and thus changing the identity of the vowel). In work reported in 1982, Klatt was unable to replicate these findings — a change in formant amplitude produced no measurable change in phonetic quality of a vowel. A new perceptual experiment has been designed in order to reconcile these two results by examining the various differences in stimuli and experimental design in the two studies. Results indicate that large changes in the relative levels of F1 and F2 for [*a*] can change the perceived phonetic quality of the vowel, but the change is not the same as that induced by moving the formant frequencies, because opposing changes in these two variables do not tend to cancel one another in the percept.

In a further study of the vowel perception process, we have been examining the properties of a synchrony model with respect to formant frequency information. Of particular interest is the case where the fundamental frequency F0 is high and the first formant frequency F1 is low. Perceptual data have been collected to show that, even in these cases, the listener behaves as if he is capable of extracting the true formant value rather than the nearest harmonic or an energy-weighted average of prominent harmonics.

Preliminary results of an effort to duplicate this behavior in a synchrony model of peripheral processing are not encouraging. The bandpass filters that are the first processing step of any model of the periphery have an adverse effect on inter-axis-crossing intervals, biasing them to be closer to harmonic intervals than would be expected from perceptual data. Work will continue on this problem.

15.5.2 The Nasal-Nonnasal Distinction for Vowels and Consonants

We have completed a perceptual study of the feature nasal for vowels. Further work on the perception of the nasal feature across languages is showing that, while speakers of different languages agree on the stimulus characteristics leading to the nasal-nonnasal distinction (spreading of spectral energy in the first formant region), they show considerable differences in the timing of nasalization (in a CVC syllable ending in a nasal consonant) that is accepted as natural in their languages. When a nasal consonant is in prevocalic position, perception experiments suggest that it is the nasalization of the vowel in the initial few tens of msec that contributes primarily to the identification of the nasal feature for the consonant.

15.5.3 Tense-lax Distinction for Vowels

A series of perceptual tests has been carried out in which American English-speaking subjects were asked to identify tense and lax vowels. Several vowel parameters were manipulated in synthetic nonsense words of the form [dVs]: formant frequency, duration, breathiness, first-formant bandwidth, and formant trajectories. The tense/lax pairs included [iɪ], [æ ɛ], [u U], and [a ʌ]. Results so far indicate that shortening the vowel shifts judgements toward the lax vowel, and that this effect is greater in the low vowel pairs than in the high pairs. The effect of breathiness shows differences between subjects, the effect of first-formant bandwidth is negligible, and the effect of formant trajectory suggests that listeners are responding to a weighted average of the first-formant frequency F1 over the vowel rather than to the maximum F1 or to an extrapolated F1.

15.5.4 Other Acoustic and Perceptual Studies

Other ongoing research that is examining the acoustic and perceptual correlates of phonetic features includes: (1) the features that distinguish between uvular and pharyngeal consonants, and that distinguish these consonants from other classes; (2) the acoustic manifestation of aspiration in the voiceless stop consonants in English and in the consonant [h]; (3) the acoustic properties that distinguish the compact stop consonants [k g] from the diffuse consonants [p t b d].

15.5.5 Auditory Models and Speech Processing

During the past year, two theses concerned with speech processing in auditory models have been completed. One of these models focuses on the detection and enhancement of periodicities and spectral prominences in the speech signal, and the other is oriented toward the processing of onset, offset, and durational properties of the speech signal.

The approach of the first model is to process the incoming speech signal through a system which simulates what is known about peripheral auditory processing, and then to apply a synchrony measure to accentuate the spectral attributes that are known to be important for the identification of the phonetic content of speech. The design of the synchrony measure is motivated in large part by a preconceived notion of what represents a 'good' result. The main criteria are that peaks in the original speech spectrum should be preserved, but amplitude information, particularly general spectral tilt factors or overall gain, should be deemphasized. A spectral representation that is smooth in frequency and time, without sacrificing resolution, is considered desirable.

The peripheral model includes a bank of critical-band filters, followed by a dynamic range compression scheme and a nonlinear half-wave rectifier. The model derives outputs for a set of 30 filters, covering the frequency range from 230 to 2700 Hz, thus representing the region of importance for sonorant segments of speech. The outputs of the peripheral model are then further processed through a hypothesized 'central' processor, which consists of a spectral analyzer and a parallel pitch estimator. These two independent systems both make use of a Generalized Synchrony Detector [GSD], which detects specific prominent periodicities present in the input waveform.

The GSD algorithm, the major novel idea of the model, is a ratio of the envelope amplitude of a sum waveform to the envelope amplitude of a difference waveform, where the sum and difference are obtained by adding or subtracting, respectively, the waveform with itself delayed by a specified period. Since a ratio is computed, the algorithm achieves a normalization with respect to signal level, an important aspect that reduces the effect of overall signal level and eliminates fluctuations in amplitude over time due to random placements of the window relative to glottal pulses. For spectral analysis, the output of each peripheral level channel is processed through a GSD which is tuned to the center frequency of the peripheral filter, in order to measure the extent of dominance of the center frequency in the wave shape. Channels centered on formant frequencies thus respond strongly, whereas the response for channels near formants is reduced, because synchrony is to the formant frequency rather than to the filter frequency. Thus the spectral contrast between peaks and valleys is enhanced.

The fundamental frequency is extracted from a waveform which is obtained by summing the outputs of the peripheral model across the 'place' dimension. This waveform exhibits strong periodicity at the fundamental, but the periodicities at formant frequencies have in general been reduced, relative to the original waveform. The GSD algorithm is applied to the pitch waveform for the range of periods appropriate for fundamental frequency of voicing. The pitch estimate is obtained from the first prominent peak in the resulting 'pseudo autocorrelation,' a plot of the GSD outputs as a function of the delay period. The method can extract the pitch of an isolated tone, as well as the fundamental frequency of a sequence of higher harmonics. It obtains a result for inharmonic sequences that is in line with perceptual data as well.

The performance of the model has been evaluated by processing a number of different synthetic and natural speech tokens through the system. In some cases, results were compared with available perceptual data.

The second model, which is oriented toward temporal properties of speech, consists of three stages, each of which effects a transformation on the acoustic input signal. The spectral analysis stage consists of a bank of filters whose characteristics reflect the frequency response of the basilar membrane. The transduction stage models the cochlea's transformation of signal intensity to expected neural firing rate. The adaptation stage produces a temporal transformation that mimics neural adaptation to sustained stimulation. The output of the model is the expected firing rate of a representative set of auditory fibers in response to the input signal.

The hypotheses regarding auditory neural response has been tested by studying the model's response to a variety of signals. A number of simple signals exhibiting speechlike spectral structure, onset and offset characteristics, and durational properties, were used to demonstrate the model's behavior and to contrast that behavior with the representation of these signals by more conventional analysis techniques. The response patterns of the model were examined for two series of speech stimuli, one of which spans the perceptual boundary between stop consonants and glides, and the other of which spans the boundary between voiced and unvoiced intervocalic stop consonants. These data provide support for an hypothesis that particular properties of the neural response to speech sounds tend to shape the representation in a way that simplifies the detection of phonetic features.

The results demonstrate that acoustic patterns of speechlike signals are transformed by the peripheral auditory model into auditory patterns with quite different properties. They also show that certain perceptual judgements of the tested speech signals by human subjects conform closely with measurable properties of the models' response. These measurable properties correspond to changes in the overall expected firing rate in the mammalian auditory nerve in response to the speech signals.