A COMPUTATIONAL APPROACH TO RECOGNITION OF SPEECH FEATURES USING MODELS OF AUDITORY SIGNAL PROCESSING

Thomas Holton
School of Engineering, San Francisco State University, San Francisco, CA 94132 USA

ABSTRACT

We present a computational approach to the detection of important speech features, such as formants and pitch, based on a model of auditory signal processing. Algorithms have been designed to be computationally simple, physiologically reasonable and to emulate human psychophysical performance.

INTRODUCTION

Most current approaches to computer speech recognition are based on a spectrographic approach to feature extraction[1]. In this approach, the energy of speech is measured as a function of frequency, and parameters derived from the resulting spectrum are compared to a template or rule. Spectrographic techniques include computation of FFTs, extraction of LPC and cepstral coefficients and processing by filter banks.

Spectrographic approaches suffer from well-known problems. Because spectrograms are sensitive to anything that changes the relative magnitude of in-band energies, their performance is often severely degraded in situations of practical interest; for example, in conditions of reduced spectral bandwidth (over the phone) or in the presence of background or line noise.

In our approach, we have sought to understand the fundamental strategy used by the auditory system to process speech signals and apply this understanding to the design of improved algorithms for detection of speech features. We have:

- developed a comprehensive model of signal processing by the peripheral and early central auditory system,
- studied the response of this model to speech and other stimuli, and
- distilled what we believe are important signal processing techniques of the auditory system into practical algorithms for feature extraction that provide noise-immune, speech-specific detection of formants and pitch pulses in sonorant parts of speech.

RESULTS

A model of auditory signal processing

The model of auditory signal processing[2] includes components describing the external and middle ear, a detailed three-dimensional hydro-mechanical model of the cochlea, a biophysical model of mechano-electric transduction by the cochlear hair cells, a description of the time-dependent synaptic chemistry of hair cells and auditory-nerve fibers including models of the hair cell's calcium channel and synapse and a 'micro-neural-net' description of signal processing in the cochlear nucleus. A comparison of the predictions of this model with experimental physiological data in response to both simple stimuli (i.e. tones) and complex stimuli (i.e. speech) suggests that the model adequately describes essential features of auditory signal processing.

The response of the model to /a/

Figure 1 shows the response of the auditory model to a voiced utterance, /a/, spoken by a male speaker. The model response to this utterance comprises two distinct spatio-temporal patterns occurring in alternation. We term these patterns the *impulsive epoch* and the *synchronous epoch*. The impulsive epoch occurs in response to the glottal pulse. In this epoch, most fibers respond at a rate that

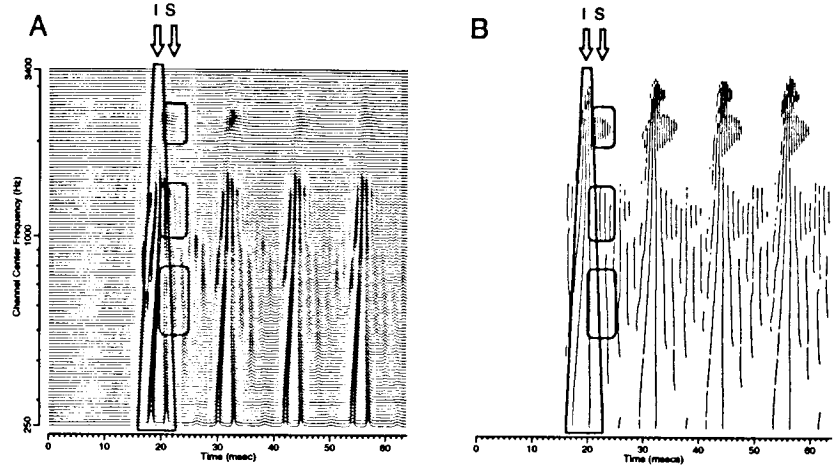

Figure 1. A. The response of 120 model fibers with characteristic frequencies (CFs) spanning the range from 250 Hz (bottom trace) to 3.4 kHz (top trace). Each waveform represents the probability density function of neural discharge for an auditory nerve fiber innervating one location along the cochlea which is maximally sensitive to a particular CF. The responses of fibers have been time aligned to remove the delay that results from the transit time of sound along the basilar membrane and the delay of neural response. The impulsive (I), *and* synchronous (S) *epochs are marked. B. The times at which the nerve fiber ensemble in A is most likely to fire. This plot results from processing the waveforms of A with a threshold-crossing algorithm that places a tick mark at the times at which each fiber is most likely to fire. The plot gives a stylized description of the pattern of timing information that this ensemble of fibers delivers to the brain in response to /a/.*

corresponds with their best or characteristic frequency (CF), giving the pattern of response of the ensemble of fibers a splayed appearance. In the synchronous epoch which follows, several groups of fibers respond distinctly at a rate that corresponds to the frequency of a proximal formant. We poetically term each group of fibers entrained to one formant an "island of synchrony". There appear to be at least three sharply delineated islands of synchrony: fibers with CFs between approximately 500 and 800 Hz are synchronized to F1; fibers with CFs between 1000 and 1400 Hz are synchronized to F2; fibers with CFs above 2000 Hz are synchronized to F3.

The alternation of an impulse-like pattern with a synchronous pattern is highly characteristic of the response to voiced speech. These observations suggest that the alternately impulsive and synchronous nature of the model's response could be used to locate and track linguistically interesting quantities such as the times of occurrence of pitch pulses and the frequencies of the formants. Our approach has been to build separate physiologically motivated "detectors" for the impulse-like first epoch and the synchronous second epoch and then use these detectors to identify formants and pitch pulses.

The response of the auditory model to an impulse

Figure 2 shows the response of the

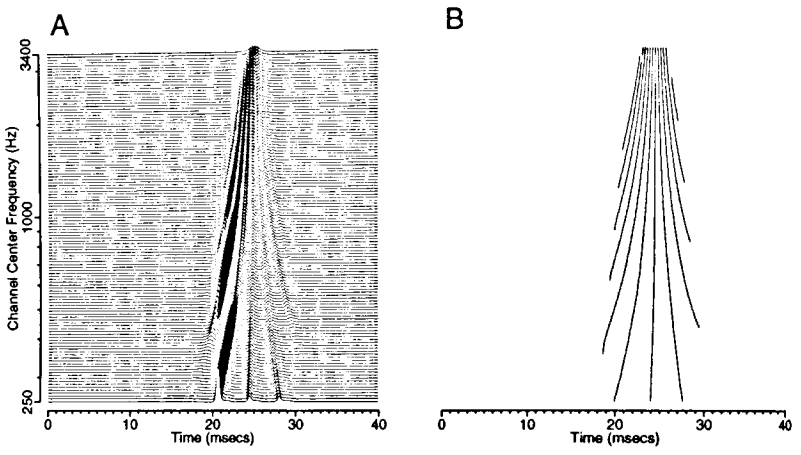

Figure 2. A. The response of the auditory model to an impulse. B. The result of processing the waveforms in A through a threshold-crossing algorithm that puts a tick mark at the times each fiber is most likely to fire.

auditory model to an impulse. Examined individually, every fiber tends to respond at a rate equivalent to its own CF. An engineering approach to designing a detector of this impulsive epoch might be to assert that an impulse is detected if, at any moment, enough fibers respond at a rate equivalent to their own CF. Algorithmically, one might implement this by computing an interval histogram of the time between firings for each fiber, taking the inverse to get a distribution of firing rate and extracting the dominant frequency component by a transform method[5]. However, there is no evidence that the brain has any processes analogous to those of forming or inverting histograms, or performing transforms to extract frequency components.

What the brain most likely can do is to detect patterns occurring in the response of a large number of simultaneously active parallel channels. We suggest that what is interesting about this picture is not an *individual* fiber's response, but the pattern of response of the *ensemble* of fibers. Specifically, the cochlea's response to an impulse is characterized by a "splayed" pattern of firing: before the peak of the impulse, fibers of lower CF respond before those of higher CF; after the peak of the impulse, fibers of lower CF respond after those of higher CF. In order to detect this pattern, we propose an array of cells, each of which correlates the response from a small number of adjacent channels and produces an output when this sequential, tonotopically organized pattern of firing is seen in the input for a period of time. The signal processing operations involved here are simple, physiologically reasonable time-correlation pattern detections; this approach does not require the computation of non-physical quantities like histograms and transforms.

While it is possible to build a detector that finds impulsive features in the stimulus using the approach just outlined, there are two practical problems with this idea: 1) *Computational complexity:* generating waveform plots such as those in Figure 1 requires the solution of a system of nonlinear, time-varying differential equations that specify the cochlear-mechanical, hair-cell and neural components of the model. The solution of these equations is highly computationally intensive; 2) *Temporal granularity:*

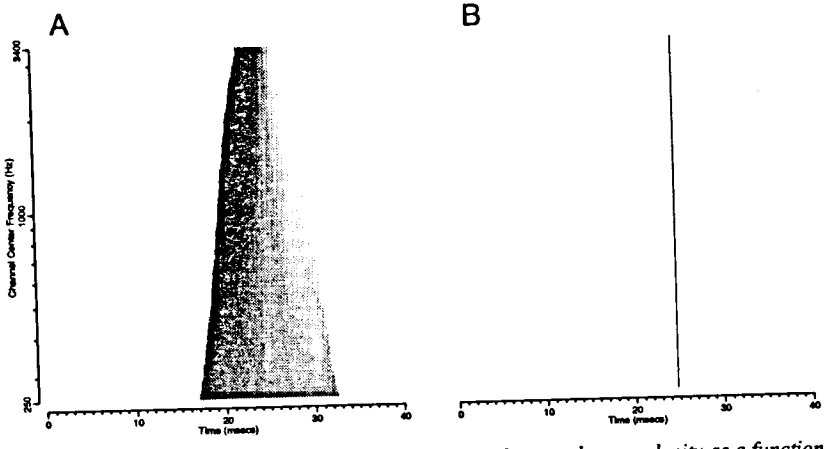

Figure 3. A. The spatial derivative of the phase of basilar-membrane velocity as a function of time in response to an impulse. The ordinate corresponds to the 120 locations along the basilar membrane with CFs logarithmically distributed between 250 Hz (bottom) and 3.4 kHz (top). Negative phase-velocity is plotted dark and positive phase velocity is light. Before the peak of the impulse the phase velocity is uniformly negative, and becomes uniformly positive after the peak of the impulse. B. The response of 119 local impulse detectors to an impulse. Each local impulse detector continuously examines the spatial phase velocity computed from the response of a pair of adjacent channels. An impulse is said to be detected when the spatial phase velocity becomes greater than zero after increasing monotonically for a period of at least one millisecond. This event corresponds to detecting the splayed pattern of nerve-fiber firings seen in Figure 2.

because model neural firings occur at discrete times, the estimate of the time of occurrence of the impulse has considerable temporal uncertainty or granularity.

To solve these problems we have used an important result derived from the study of the response of the cochlear model: patterns of neural firings correlate with patterns of basilar membrane motion; specifically, information about the sequential or simultaneous firings of groups of adjacent fibers reflects simple patterns in the spatial and temporal derivatives of the instantaneous phase of the basilar membrane's motion.

The local impulse detector

Figure 3A shows a plot of the spatial derivative of the phase of basilar-membrane velocity as a function of time in response to an impulse. At all points on the model cochlea, the spatial phase velocity is initially less than zero and increases monotonically over a period of time. This pattern of phase velocity is easy to detect. Figure 3B shows the response of an array of local impulse detectors. Each detector produces a response upon detecting the negative-to-positive pattern of spatial phase velocity.

Figure 4 shows the response of an array of local impulse detectors to */a/*. Each mark on the plot is derived by examining local spatial and temporal patterns of phase velocity over a small window of time (about 1.5 msec) and a small range of frequency (two adjacent channels, corresponding to about 0.3 critical bands). The wavy lines correspond to the times of occurrence of the pitch pulses.

The local formant detector

It is possible to use the same auditory model concepts to make phase-based local

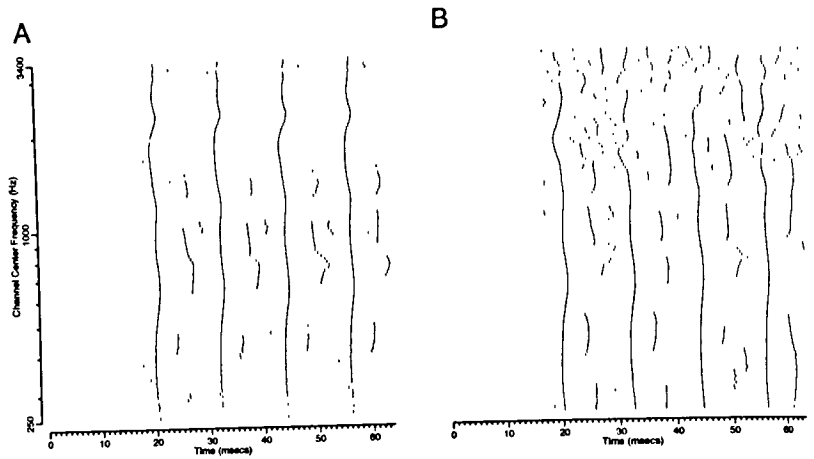

Figure 4. A. The response of the 119 local impulse detectors to /a/. B. The response of the impulse detectors to /a/ in the presence of white noise with S/N = 8 dB.

synchrony detectors which find the "islands of synchrony" discussed earlier. Specifically, it can be shown that the islands of synchrony correspond to spatio-temporal regions in which the spatial phase velocity of basilar-membrane motion is constant.

Voiced speech comprises impulsive epochs and synchronous epochs occurring in alternation. We have built a candidate formant detector that detects this pattern based on the detectors for the impulsive and synchronous epochs described above. The formant detector is an array of cells, each of which responds to an impulsive epoch in a given channel followed by a synchronous epoch.

Figure 5 shows the output of the local formant detector to /a/. Three formants (F_1, F_2, and F_3) plus a bit of F_4 are clearly represented. The representation of information in this plot is quite sparse; there is only information at frequencies corresponding to the formants and little elsewhere. None of the operations involved in generating this representation are either computationally complex or non-physiological, and none of the operations uses any of the conventional spectral techniques.

In natural speech, the frequencies of formants are not static, but change rapidly as a function of time depending on the consonantal context in which the vowel is embedded. Because all the stages of detection that generate this representation act on patterns which are temporally localized, the speech signal need not be periodic or quasi-periodic to determine the times of occurrence and frequencies of the formants. In this approach, formants are detected on a pitch-pulse-by-pitch-pulse basis with simultaneously high time and frequency resolution.

Human speech intelligibility, at least of vowels, is not very sensitive to additive background noise. Whereas spectrographic representations of speech are inherently sensitive to noise, the response of the local formant detector is relatively insensitive. Also, unlike spectrographic measures, it can be shown that the response of the formant detector is insensitive to pure tones and other non-speechlike stimuli.

A model of pitch

We have developed a theory for the detection and identification of pitch and

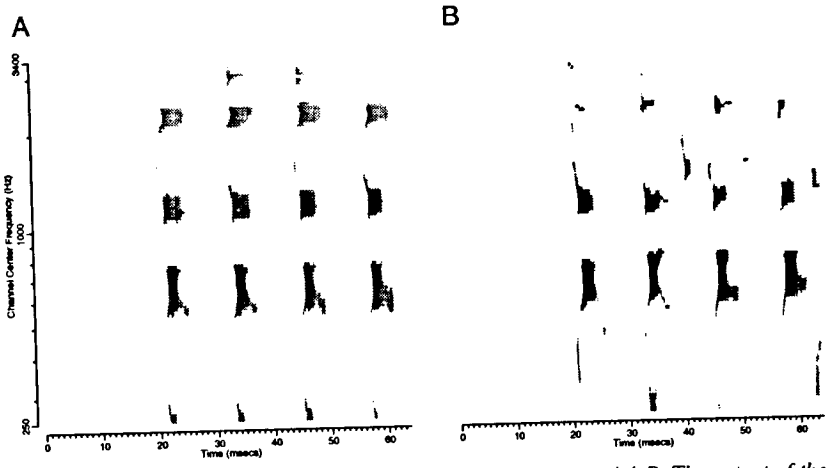

Figure 5. A. The output of an array of local formant detectors to /a/. B. The output of the formant detector to /a/ in the presence of additive noise, S/N=+8dB.

voicing based on the physiological model of auditory signal processing coupled with the idea of detecting spatially and temporally local patterns of response phase from a number of parallel channels[3].

Figure 6 shows the architecture of a physiologically motivated pitch detector. For each point on the cochlea, we postulate the existence of an array of *local pitch detector* cells. Each cell in the array detects in the time domain a different fixed time periodicity in the output of the underlying local impulse detector cell. These pitch-detector cells could be implemented physiologically by a series of neural delay correlators, as originally proposed by Licklider[4]. For each point on the cochlea, cells in the pitch detector array respond when a pair of impulses is received in the same channel with a given fixed time delay. Cells in the current model are selective for time delays spanning the range of 1 to 15 msec with a resolution of .25 msec. The sum of the response of local pitch detectors serving the whole cochlea gives a global measure of the periodicity of the entire ensemble of channels, which we term the *global pitch* detector.

Using this pitch detector method, it is possible to track rapidly varying pitch of natural speech. Figure 7 shows the output of the global pitch detector, a representation of the instantaneous average pitch frequency as a function of time, for an utterance that has relatively constant formant structure but rapidly varying pitch. The response of this pitch detector can be shown to be robust in noise. While the pitch detector is particularly sensitive to impulsive stimuli, such as voiced speech, it is highly insensitive to pure tones and other non-speech-like input. The pitch detector also reproduces effects seen in the psychophysics of pitch perception, such as the recovery of the missing fundamental of resolved and unresolved harmonics.

The auditory-model pitch detector is computationally straightforward and physiologically plausible. Calculations correspond to the correlation of simple neural events. No continuous-time autocorrelation functions are explicitly computed, nor does the input stimulus need to be periodic for pitch to be detected and tracked.

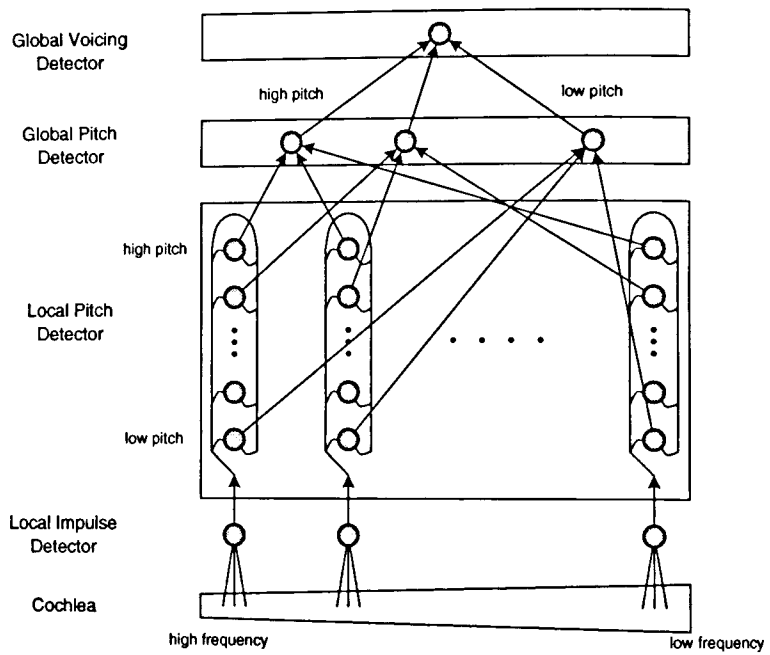

Figure 6. Architecture of the pitch detector

CONCLUSION

We have designed algorithms for the detection of important speech features based on an understanding of how the auditory system processes speech. Algorithms are computationally simple, physiologically reasonable and demonstrate performance that emulates that of humans.

Comparison with spectrographic approaches to feature extraction

Almost all current approaches to speech recognition are based on a spectrographic approach to feature extraction. These techniques include filter bank, fast Fourier transform (FFT), cepstral, power spectral density (PSD) and linear predictive coding (LPC) analysis. These spectrographic approaches are sensitive to anything that changes the magnitude of the input in a frequency band, for example by spectral shaping the input signal. Spectral approaches are sensitive to the frame size of analysis; a larger frame size may be used to average over pitch periods at the cost of coarser temporal and spectral resolution. Spectrographic approaches are also inherently noise sensitive, since they measure the energy in a frequency band, regardless of the source of that energy.

The auditory-model approach to detecting speech features differs in key respects from spectrographic methods. This approach, based on building detectors of spatially and temporally local patterns of response phase from a number of parallel channels, can be characterized as a *local time-domain phase-correlation* approach, in contrast with conventional spectrographic techniques, which can be characterized as examples of a *global frequency-domain energy* approach. Auditory-model algorithms for feature detection show noise insensitivity and amplitude independence, as well as

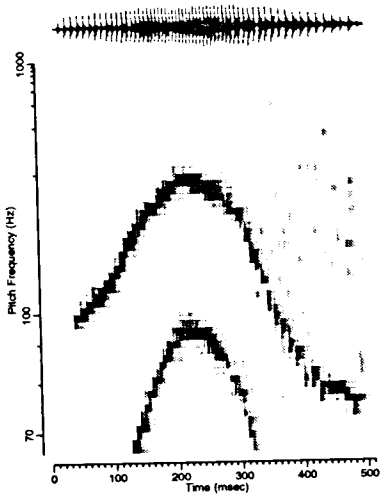

Figure 7. The response of the global pitch detector (lower plot) to an utterance /a/ spoken with rapidly increasing and then decreasing pitch (upper trace). The pitch plot shows bands at the fundamental pitch frequency, Fo, and at the first sub-harmonic, Fo/2.

selectivity for speech-like sounds. There are no inherent periodicity requirements for the stimulus, nor need the data be "framed" into arbitrary time segments as, for example, it must be prior to performing spectral analysis by Fourier transform or LPC coefficient extraction.

Comparison with other auditory model approaches

Several studies have used concepts of auditory physiology to motivate the design of algorithms for speech recognition. These approaches included the ensemble-interval histogram (EIH) method of Ghitza[5], the generalized synchrony detector (GSD) approach of Seneff[6] and the correlogram approach of Lyon[7]. All these methods are based on determination of the times of neural firings of a number of channels of a nonlinear auditory model. The response of each model fiber is then analyzed *individually*, for example by computing a period histogram of a fiber's response and then performing spectral or autocorrelation analysis of response. Global operations are then performed on the summed data from a number of individual channels to detect important features such formants. In addition to the drawbacks of temporal granularity and computational intractability discussed previously, the operations of accumulating histograms and performing spectral analysis are not likely to be physiological.

REFERENCES

[1] Deller, J.R. and Proakis, J.G. and Hansen, J.H.L. (1993). *Discrete-time processing of speech signals*, New York: Macmillan.

[2] Holton, T., Love, S.D. and Gill, S.P. (1991). "A fundamental approach to automatic speech recognition using models of auditory signal processing", *DARPA Technical Report DAAH01-91-C-R095.*

[3] Holton, T., Love, S.D. and Gill, S.P. (1994). "Robust pitch and voicing detection using a model of auditory signal processing", *Proc. ICSLP-94.*

[4] Licklider, J. (1951). "A duplex theory of pitch perception", *Experimentia*, vol. 7, pp. 128-131.

[5] Ghitza, O. (1986). "Auditory nerve representation as a front-end for speech recognition in a noisy environment", *Comput. Speech Lang.*, vol 1, pp. 109-130.

[6] Seneff, S. (1988). " A joint synchrony/mean-rate model of auditory speech processing", *J. Phonetics*, vol. 16, 55-76.

[7] Lyon, R.F. (1984). "Computational models of neural auditory processing", *Proc. IEEE ICASP*, pp. 1975-1978.