Recognition of Speech Features Using Models of Auditory Signal Processing

Tom Holton
School of Engineering
San Francisco State University
San Francisco, CA 94132
th@ernie.sfsu.edu

March 4, 1995

1 Introduction

Most current approaches for speech recognition are based on a spectrographic approach to feature extraction in which 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, conditions of reduced spectral bandwidth (over the phone) or 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 speech features. We have done this by

1. developing a comprehensive model of signal processing by the peripheral and early central auditory,
2. studying the response of this model to speech and other stimuli, and
3. distilling what we believe are the 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.

In the sections that follow, I'll like to briefly introduce our model, and describe how our analysis of the response of this model has lead us to the design of our algorithms.

2 A block diagram of the auditory model

Our model of auditory signal processing includes

- components describing the external and middle ears;
- a detailed three-dimensional hydromechanical 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 a model of the physiology of the hair-cell synapse.
- 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 to us that the model, in fact, describes quite accurately essential features of auditory signal processing.

3 The response of the auditory 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 consists of two distinct spatio-temporal patterns, occurring in alternation. I term these patterns the *impulsive epoch* and the *synchronous epoch*. In the impulsive epoch, in response to the pitch pulse, most fibers respond, and the response of the ensemble of fibers has a splayed pattern. In the synchronous epoch which follows, several groups of fibers appear to respond distinctly. Within each group, fibers respond at a frequency which corresponds to frequency of a proximal formant. We poetically term each group of fibers entrained to one formant an "island of synchrony". You can see in the figure that there appear to be at least three sharply delineated islands of synchrony. Fibers with CFs between approximately 400 and 900 Hz are synchronized to F_1; fibers with CFs between approximately 950 and 1400 Hz are synchronized to F_2; fibers with CFs above approximately 1800 Hz are synchronized to F_3.

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

4 The response of the auditory model to an impulse

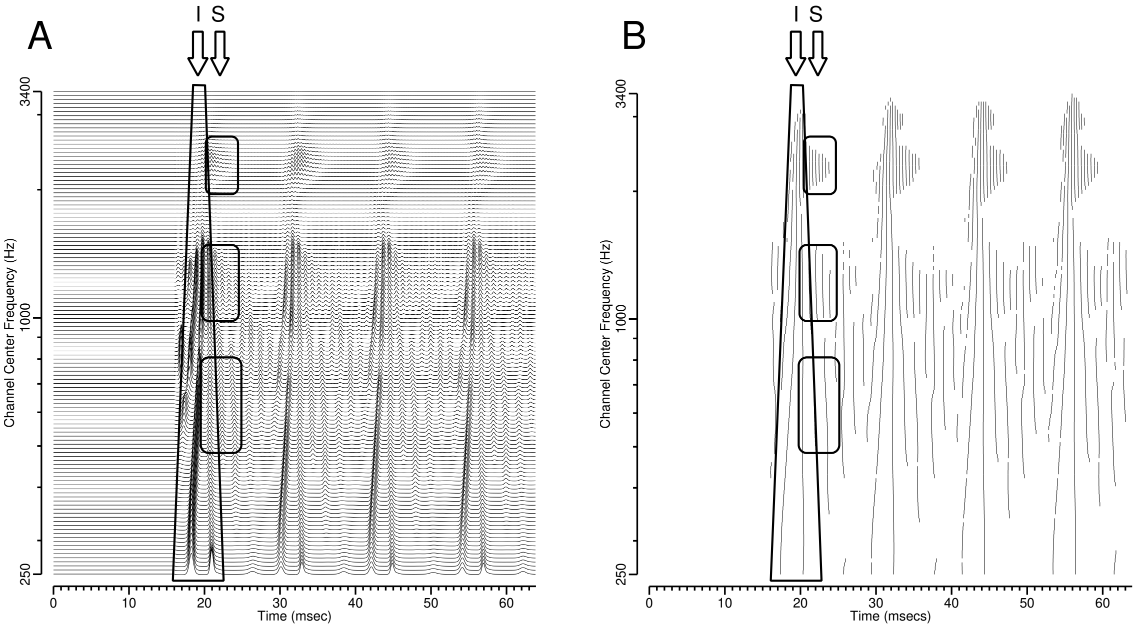

Figure 1: *A.* The response of 120 model fibers with CFs spanning most of the important speech range (at least, the range that corresponds to telephone bandwidth), from 250 Hz (below) to 3.4 kHz above. Each waveform represents the probability density function of neural discharge for an auditory nerve fiber innervating one location on the cochlea. Each location along the cochlea is maximally sensitive to particular characteristic frequency or CF. The responses of fibers has been time aligned to remove the delay that results from the transit time of sound along the basilar membrane plus the delay of neural response. *B.* The times at which the nerve fiber ensemble in A is most likely to fire. This picture results from processing the waveforms of A through a threshold-crossing algorithm, and placing a tick mark at the times at which each fiber is most likely to fire. This picture gives a stylized description of the pattern of timing information that this ensemble of fibers delivers to the brain in response to an /a/.

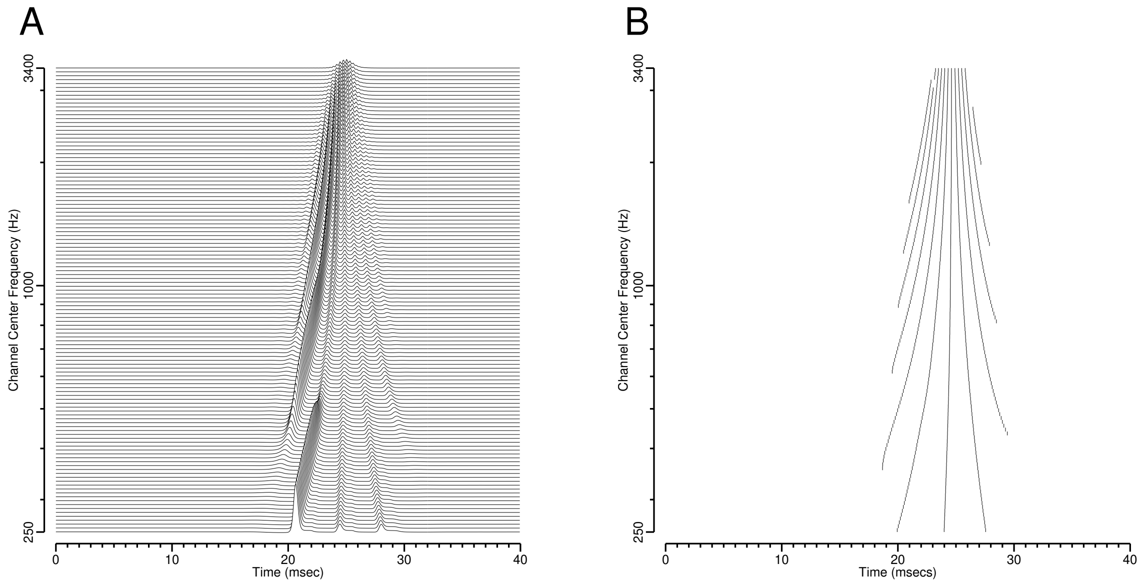

Figure 2: *A.* The response of the auditory model to an impulse. *B.* The result of processing the waveforms in *A* through a threshold-crossing algorithm that puts a tick mark at the times each fiber is most likely to fire.

Figure 2 shows the response of the auditory model to an impulse. The right panel shows the threshold-crossings of the response. Examined individually, every fiber tends to respond to an impulse at a rate equivalent to its own CF (higher CF fibers respond at a higher rate, lower CF fibers at a lower rate). To make a detector of this impulsive event, you might say that an impulse had been detected if, at any time, enough fibers were responding at a frequency equivalent to their own CF. To measure the frequency of response of each fiber an engineer might, for example, measure the time between firings, computing an interval histogram, take the inverse of this interval histogram to get a histogram of frequency, and extract the dominant frequency component by a transform method. Unfortunately, there is no evidence that the brain has any analogs to the processes of forming or inverting histograms, or performing transforms to extract frequency components.

What the brain *can* do is to detect patterns occurring in the response of a large number of parallel channels. What is interesting about this picture is *not* an individual fiber's response, but the pattern of response of the ensemble of fibers.

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. All the brain needs to have to detect an impulse is an array of cells which scan the input from a number of adjacent channels, an produce an output when a particular pattern of firing is seen in the input. For example, each cell in this array would respond on detecting a pattern of sequential firing: first cell 1 fires, the cell 2 fires, then cell 3 fires. When an entire array of cells sees this pattern for a period of time, an impulse is said to be detected. 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 we could build a detector that would find impulsive features in the stimulus using the approach I've just outlined, there are two problems with this idea, which I will just mention. The first is *computational complexity*: in order to generate these waveform plots, we need to solve a system of nonlinear, time-varying differential equations, not a happy prospect. The second problem is what I call temporal granularity: because neural firings occur only at discrete times, our estimate of important events, such as the time of an impulse, will have a lot of temporal uncertainty or granularity.

To solve this problem we have used an important result derived from an study of the response of the model: stated simply, the preferred times of firings of the nerve correspond to zero-crossings in basilar-membrane motion. The nerve "tells" the brain what the state of the basilar membrane is. Furthermore, we can show that patterns of neural firings correlate with patterns of basilar membrane motion, specifically patterns in the spatial and temporal derivative of the instantaneous phase of the membrane's motion.

5 Construction of the impulse detector

Figure 3 shows a plot of the spatial derivative of the phase of basilar-membrane velocity, that is spatial phase velocity, $d\phi/dx$, as a function of time (abscissa) in response to an impulse. The pattern of phase velocity that corresponds to an impulse is very easy to detect. Figure $3B$ shows the response of an array of local impulse detectors. Each detector is designed to produce a response upon detecting the lagging-leading pattern of phase velocity in one channel or small group of channels.

6 The response of local impulse detector to /a/

Figure 3 shows the response of an array of local impulse detectors to /a/. The wavy lines correspond to the times of occurrence of the pitch pulses. I'd like to emphasize that each mark is obtained by examining local spatial and temporal patterns of phase velocity only over a small period of time and a small range of adjacent channels. Each detector only "looks" for an impulsive pattern temporally over about 1.5 msecs, and spatially over about 1 critical band.

7 The response of local formant detectors to /a/

It is also easily possible to make phase-based local synchrony detectors which find the "islands of synchrony" discussed earlier. Now, voiced speech comprises impulsive epochs and synchronous epochs occurring in alternation. We have built a candidate formant detector that detects this pattern.

Figure 5 shows the output of the local formant detector to the /a/. Our formant detector is an array of cells, each of which looks for an impulsive epoch followed by a synchronous epoch. This detector also does a bit of spatial lateral inhibition to sharpen up the spatial

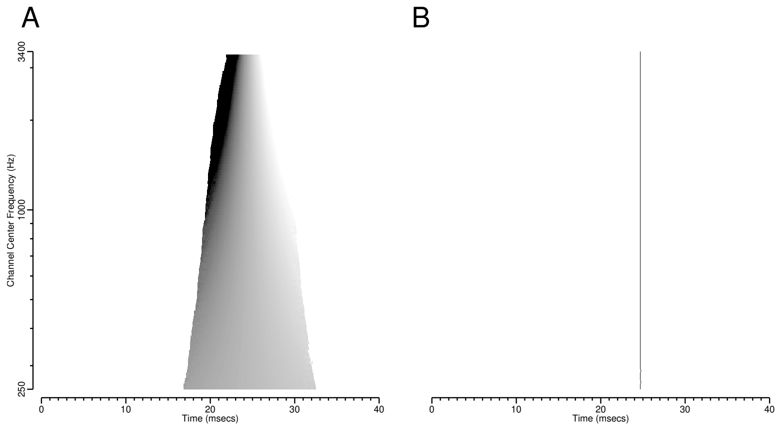

Figure 3: *A.* A plot of the spatial derivative of the spatial phase velocity of basilar-membrane velocity, $d\phi/dx$, as a function of time (abscissa) in response to an impulse. The ordinate corresponds to the 120 locations along the basilar membrane corresponding to CFs logarithmically distributed between 250 Hz (bottom) and 3.4 kHz (top). Negative phase-velocity is plotted dark and positive phase velocity is light. You can see that 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 of response of one channel. An impulse is said to be detected when $d\phi/dx$ 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, in which most channels in the ensemble initially respond before their higher-frequency neighbors.

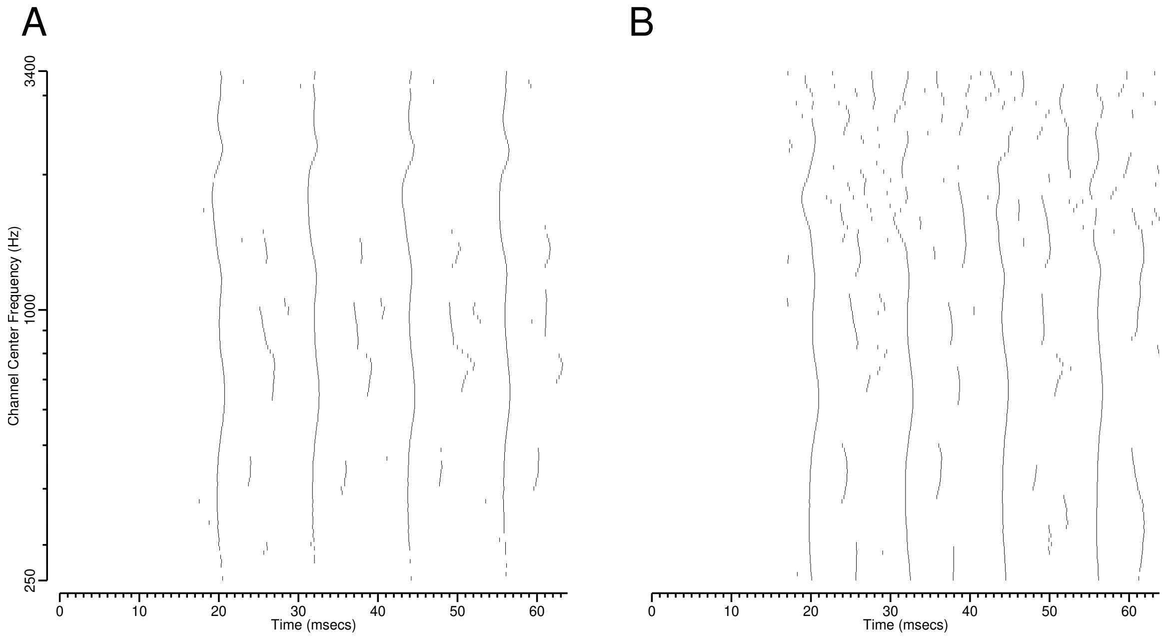

Figure 4: *A.* The response of the 119 local impulse detectors to /a/. The local impulse detector are described in the caption to Figure 3. *B.* The response of the impulse detectors to /a/ in the presence of white noise with S/N = 8 dB.

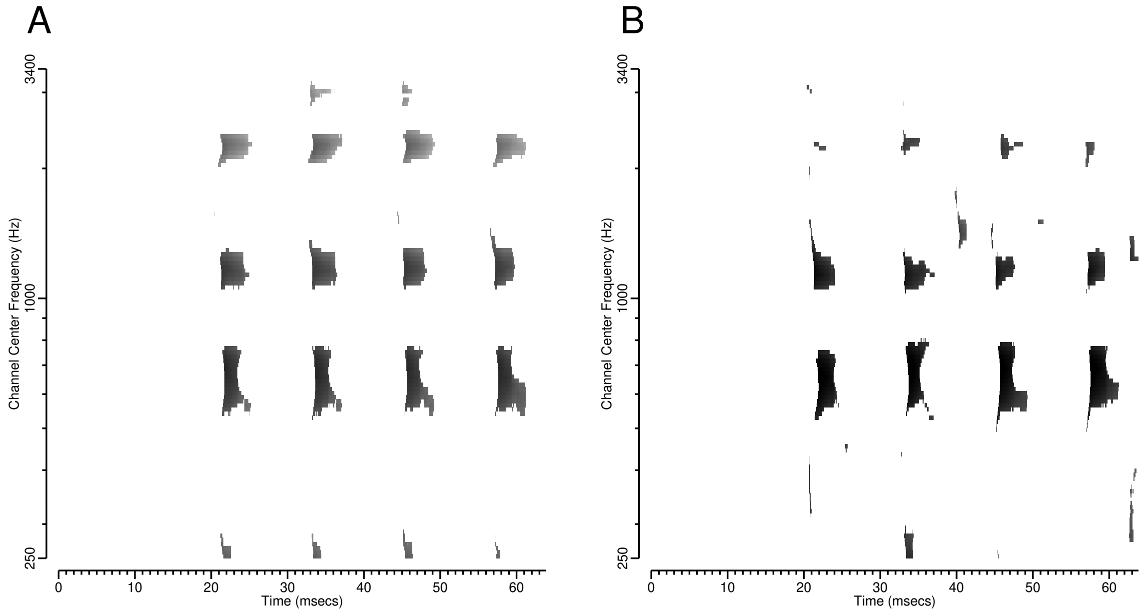

Figure 5: *A.* The output of an array of local formant detectors to utterance /a/. *B.* The output of the formant detector to /a/ in the presence of additive noise, S/N=+8 dB.

extent of the response. But 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.

You can see that three formants (*F1*, *F2* and *F3*) are clearly represented, plus a bit of *F4*. The representation is quite sparse; there is only information at frequencies that correspond to the formant frequencies.

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. The response of the local formant detector is also quite robust in noise.

8 Other features of the auditory model

We also showed in class that the response of the formant detector is highly speech specific, namely that it does not respond to tones or combinations of tones, even when these tones are located at the canonical frequencies for a formant. We also demonstrated a pitch detector and vocing detector, and demonstrated that these detectors could track the rapidly changing pitch of natural speech.

Finally, we deomstrated the response of the pitch and voicing detectors to a simulated telephone call that comprised several sections containing high-energy periodic stimuli, such as DTMF, modem and ringback stimuli, as well as speech. The voicing detector only responded forcefully to the voiced parts of natural speech.

9 Conclusion

Our approach to detecting speech features is based on building detectors of spatially and temporally local patterns of response phase from a number of parallel channels; that is, you can characterize this as a *local time-domain phase-correlation* approach. This contrasts markedly with the conventional spectrographic techniques, which can be characterized as examples of a *global frequency-domain energy* approach.

We have tried to design algorithms for the detection of important speech features based on an understanding of how the auditory system processes speech. As far as possible, we have tried to make our candidate algorithms to be computationally simple, physiologically reasonable and demonstrate performance that emulates the performance of humans. Specifically, the auditory-model algorithms show the following properties:

- *Accurate estimation of formant frequencies.* 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. The auditory-model algorithms track formants accurately as a function of time, on a pitch-pulse-by-pitch-pulse basis.

- *Noise insensitivity.* The auditory-model algorithms appear robust in noise. Spectrographic algorithms based on spectral energy measures are inherently sensitive to noise.

  Amplitude independence. The auditory-model algorithms are based on measurement of response phase. They are highly independent of signal level. Algorithms based on the measurement of the energy in spectral bands are inherently sensitive to signal amplitude.

- *Selectivity for speech.* The auditory-model algorithms are designed to be selective for speech-like sounds. Spectrographic algorithms are not inherently capable of distinguishing between speech and non-speech.

- *Accurate estimation of voice pitch.* Voice pitch is not constant; it varies markedly during a voiced utterance. The auditory-model algorithms detect each pitch pulse "on the fly", and thus can track pitch dynamically and accurately.

- *Computational simplicity.* The auditory-model approach offers several important computational advantages over conventional approaches.

  1. Processing is local and highly parallel. The approach of local detection appears to help make the auditory-model algorithms highly immune to noise and independent of stimulus amplitude.
  2. Processing is continuous and asynchronous. There are no inherent periodicity requirements for the stimulus, nor are the data "framed" into arbitrary time segments as, for example, is data prior to performing spectral analysis by Fourier transform or LPC coefficient extraction. In conventional techniques, the presence of pitch pulses can be viewed as an impediment to the task of obtaining an accurate representation of the formant frequencies. Techniques like cepstral analysis are used in an attempt to "deconvolve" the excitation function (i.e. the pitch-pulse train) from the transfer function of the vocal tract to get an estimate of the underlying formant frequencies. This task is complicated by the non-periodic nature of the pitch-pulse train in speech. In the auditory-model approach, pitch pulses are *necessary* for formant detection: the pitch-pulse detector gate-enables the synchrony detector. Thus, in the auditory-model approach, data are "naturally" framed by the occurrence of the pitch pulses.

We conclude that algorithms for automatic speech recognition based on an auditory model offer an effective alternative to spectrographic processing schemes for formant detection. Our future task is to refine these algorithms, extend them to allow the detection of other speech features, and incorporate them into speech recognition systems.