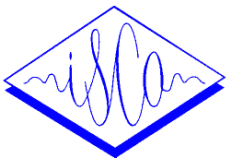

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

FORMANT AND PITCH-PULSE DETECTION USING MODELS OF AUDITORY SIGNAL PROCESSING

Thomas Holton[1,2], *Steven D. Love*[1] *and Stephen P. Gill*[1]

[1]Votan Corporation, Pleasanton, CA 94566 USA
and
[2]San Francisco State University, San Francisco, CA 94132 USA

ABSTRACT

Based on an analysis of a model of signal processing by the auditory system, we propose a *local time-domain phase-correlation* approach to the detection and categorization of sonorant speech features. In this approach, pitch pulses and formant frequencies are marked by characteristic patterns of phase-correlation in the output of groups of frequency-selective filters that correspond to the temporal sequence of firings of groups of nerve fibers in the cochlea. Algorithms for the detection of speech features based on the auditory model appear to improve upon conventional (i.e. spectrographic) techniques in several respects: they are resistant to additive noise, relatively independent of signal amplitude and spectral shaping of the input and speech-specific.

INTRODUCTION

Most current approaches for speech recognition are based on a spectrographic or *global frequency-domain energy* approach to feature extraction. In this approach, the energy in sequential frames of a speech signal is measured as a function of frequency, and parameters derived from the resulting spectrum are usually compared to a template or rule. Spectrographic techniques include fast Fourier transformation, power-spectral-density analysis, extraction of linear predictive coding and cepstral coefficients and processing by filter banks. Because spectrographic techniques are sensitive to anything that changes the relative magnitude of energies in different frequency bands, their performance is often severely degraded in situations of practical interest, for example conditions of reduced spectral bandwidth or the presence of high background or line noise.

In this work, 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 algorithms for robust computer detection of speech features. We have developed a comprehensive model of peripheral and early central auditory signal processing, which is described in detail elsewhere [1]. This model includes a detailed three-dimensional hydromechanical model of the cochlea, a biophysical description of mechano-electric transduction by cochlear hair cells and a description of the time-dependent synaptic chemistry of hair cells and auditory-nerve fibers. A comparison of predictions of the model with experimental physiological data obtained in response to simple (i.e. tonal) and complex (i.e. speech) stimuli suggests that the model accurately describes essential features of auditory signal processing.

Based on the analysis of this auditory model, we propose a *local time-domain phase-correlation* approach that provides noise-immune, efficient, speech-specific detection of pitch pulses and formant frequencies in sonorant speech. The approach is based on detecting spatially and temporally local patterns in the output of a number of parallel channels of a filter-bank derived from the auditory model.

RESULTS

The response of the auditory model

Fig. 1 shows the response of the auditory model to a voiced /a/ spoken by a male speaker.

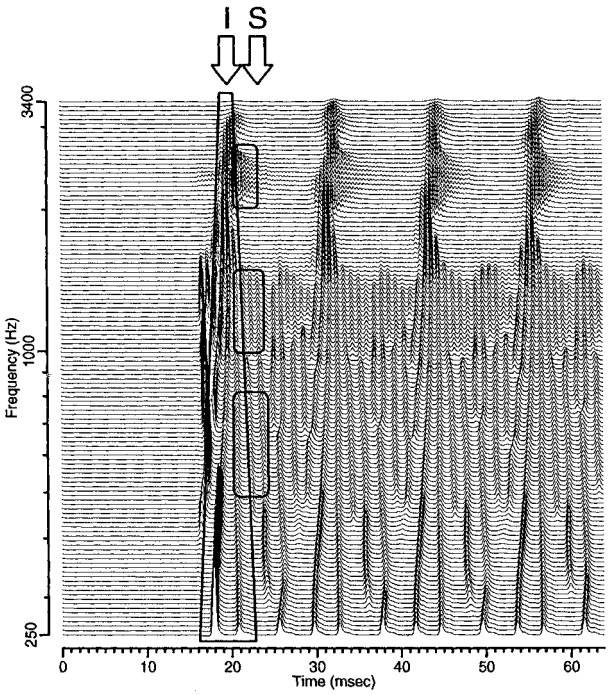

Fig. 1: The response of the auditory model to /a/. Waveforms represents the probability density function of discharge of 120 auditory-nerve fibers innervating positions spaced linearly along the basilar membrane. The center frequencies (CFs) of the model nerve fibers span the range 250 Hz to 3.4 kHz logarithmically, corresponding to a linear spacing of innervation along the basilar membrane. The responses of channels have been time aligned to remove neural response delay and mechanical group propagation delay of basilar-membrane motion. Arrows mark the impulsive, **I**, and synchronous, **S**, epochs.

The model response to this utterance consists of two distinct spatio-temporal patterns occurring in alternation, which we term the *impulsive epoch* and the *synchronous epoch*. The impulsive epoch occurs in response to the glottal pulse. In this epoch, each fiber tends to respond at its own characteristic frequency (CF) for a brief period of time (several milliseconds), giving the response of the ensemble of fibers has a splayed appearance. In synchronous epoch which follows, several distinct groups of fibers appear to respond. Within each group, fibers respond synchronously at a rate which corresponds to the frequency of a proximal formant. We term each

group of fibers entrained to one formant an "island of synchrony". In Fig. 1, there appear to be sharply delineated islands of synchrony at frequencies corresponding to F_1 (600 Hz), F_2 (1100 Hz) and F_3 (2400 Hz).

The alternation of an impulsive pattern with a synchronous pattern is highly characteristic of the response to speech and suggests that these patterns of response could be used by the brain to detect and identify linguistically important features, such as the frequencies and times-of-occurrence of the formants as well as the times-of-occurrence of glottal pulses. Our approach is to build physiologically motivated "detectors" for impulsive and synchronous features and then combine these detectors to identify formants and glottal pulses.

We have derived detectors for impulsive and synchronous features from an analysis of patterns of the threshold-crossings of waveforms of neural response, such as those shown in Fig. 1 (see Holton et al. [1] for details). The splayed pattern that characterizes the impulsive epoch can be detected by an array of local impulse-detector "cells", each of which produces a response when a small group of adjacent primary fibers responds in sequence (i.e. a low-CF fiber fires before a high-CF fiber) for a period of time. Islands of synchrony can be detected by an array of local synchrony-detector "cells", each of which produces a response when a small group of adjacent primary fibers responds in synchrony. However, there are two technical drawbacks to this approach: temporal granularity and computational intractability. Temporal granularity means that threshold-crossings of the neural response of each fiber occur only at discrete times, which can be spaced milliseconds apart in model fibers with low CFs. Hence, the response of detectors of the impulsive and synchronous epochs, which require the input from several fibers, will be temporally coarse or granular. Computational intractability occurs because computing the complete response of the auditory model requires the solution of nonlinear equations for a large number of parallel channels. The computational load of performing these calculations at or near real-time is prohibitive.

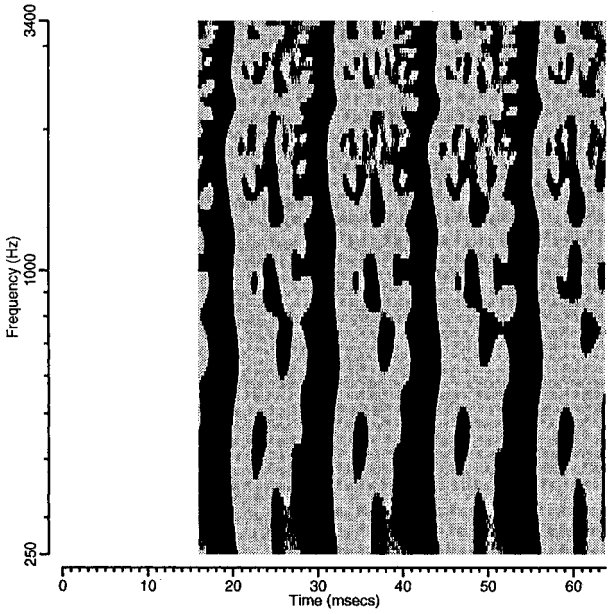

Fig. 2: The spatial phase-velocity, $d\phi/dx$, for 119 channels of the basilar-membrane model in response to /a/. Black indicates $d\phi/dx<0$; gray indicates $d\phi/dx>0$.

An analysis of the auditory model suggests an alternate approach to building detectors of speech features, which we term the *phase-coherence approach*. This approach, which is both temporally fine-grain and computationally tractable, is based on the finding that information on the relative timing of adjacent nerve fibers is equivalent to information derived from the spatial and temporal derivatives of the instantaneous phase of basilar-membrane velocity. Because the instantaneous phase is derived from the linear basilar-membrane component of the auditory model, computation is relatively efficient and can be performed with high time resolution. In the following sections, we propose the architecture of a *local impulse detector*, a *local synchrony detector* and a *local formant detector*, and show their response to speech.

The local impulse detector

The principle of the local impulse detector is to detect a pattern of spatial phase-velocity on the basilar-membrane motion that corresponds to the sequential pattern of neural activity evoked by an impulse. Fig. 2 shows the spatial phase-velocity, $d\phi/dx$, of the auditory model in response to /a/. At each glottal pulse, for every channel in the plot, the spatial phase-velocity goes from negative to positive. This pattern of spatial phase velocity is relatively easy to detect.

Fig. 3 shows the response of an array of local impulse detectors to the utterance /a/. Each impulse detector in the array produces a response (a tick-mark) when an interval of negative phase-velocity is followed by a transition to an interval of positive phase velocity. We term these detectors "local" to indicate that the output of each detector for a given channel k at time t is determined by examining the spatial phase velocity only over a small number of channels (generally one or two) around k, and only for a brief time interval (generally 1 to 2 msecs) around t.

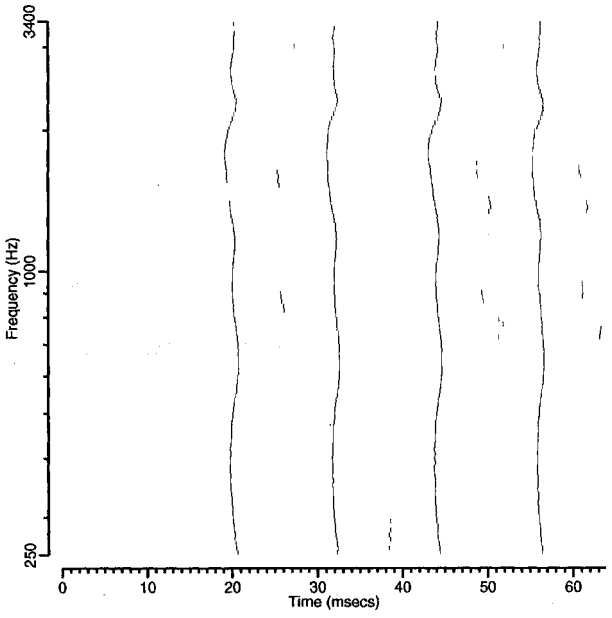

Fig. 3: The response of an array of 119 local impulse detectors to /a/.

The response of the local impulse detectors forms a series of wavy lines, each of which tends to occur near the time of a glottal pulse. We have designed robust voicing detectors based on the output of an array of local impulse detectors. Because impulse detection is temporally local, each glottal pulse is detected separately; hence, the stimulus need not be periodic to detect voicing. This contrasts with conventional voicing detectors based, for example, on autocorrelation techniques in which performance is degraded for non-periodic stimuli.

The local synchrony detector

The principle of the local synchrony detector is to detect a pattern of constant phase velocity of basilar-membrane motion that corresponds to the pattern of synchronous neural activity evoked by formants. Local synchrony is defined to occur in channel k at time t if the spatial phase velocity of a plurality of channels adjacent to k is nearly zero for a period of time around t. To detect this pattern of constant spatial phase-velocity we create an array of local synchrony detectors, each of which produces a response when the phase velocity of a small group of adjacent channels is roughly zero. The synchrony detectors are local in the sense that that they response only to a pattern of spatial phase velocity over a short interval of time and a small number of adjacent channels.

Fig. 4 shows the output of an array of local synchrony detectors to the utterance /a/. The islands of synchrony, defined in conjunction with Fig. 1, are visible as dark patches which persist several milliseconds after the glottal pulse.

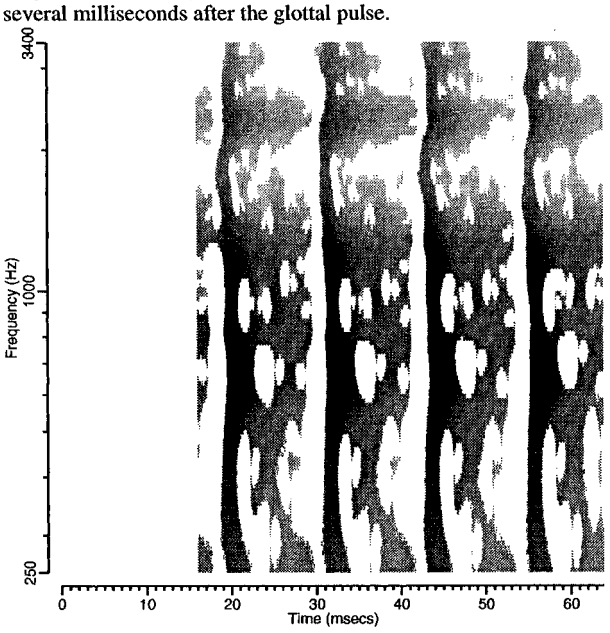

Fig. 4: The response of 119 local synchrony detectors to /a/. Response is shown for regions where $|d\phi/dx|$ is less than an equivalent phase velocity of 5 m/sec. The darkness of the plot is proportional to the instantaneous magnitude of the response.

The formant detector

Voiced speech is characterized by impulsive and synchronous epochs occurring in alternation. Fig. 5 shows the output of an array of local formant detectors to the utterance /a/.

Each channel of the formant detector receives input from the corresponding channels of the local impulse detector (e.g. Fig. 3) and the local synchrony detector (e.g. Fig. 4). Each channel of local formant detector produces an output when a synchronous response occurs on that channel within a given time window after the occurrence of a local impulse. Additionally, the local formant detector employs some spatial filtering which emulates the neurophysiological process of lateral inhibition commonly used in cells throughout the nervous system to sharpen the spatial extent of responses. Fig. 5 shows distinct bands at frequencies corresponding to the formants at times near each glottal pulse. The response of the formant detector is local in that the output of each channel of the detector depends only upon the response of the impulse and synchrony detectors over a short period of time and a small range of frequencies. In the auditory-model approach, formants are detected on a pitch-pulse-by-pitch-pulse basis with simultaneously high time and frequency resolution.

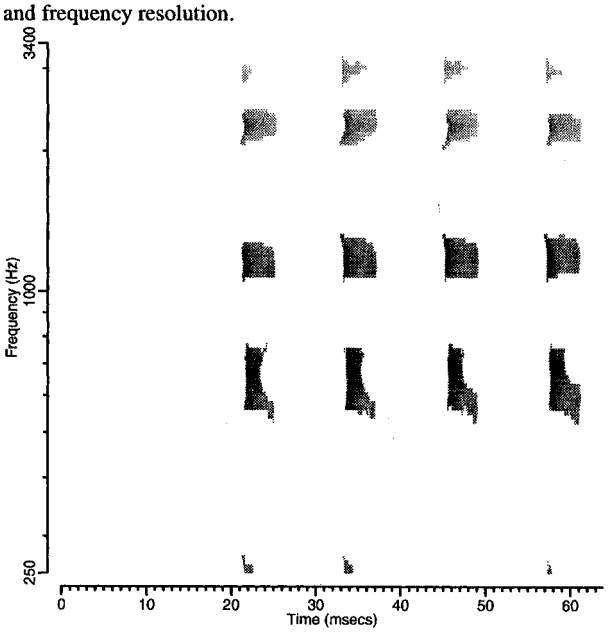

Fig. 5: The output of 119 channels of the local formant detector to /a/.

The feature detectors based on the phase-coherence approach demonstrate several important properties which we illustrate elsewhere [1]. The response of local impulse, synchrony and formant detectors is highly insensitive to noise. In the presence of moderate (e.g. S/N=6 dB), the representation of Fig. 5 shows substantial response at the formants, and little spurious response at other frequencies and times. This contrasts with the performance of spectrographic measures, in which the output of each channel is a function of the total energy (signal plus noise) within the channel's bandwidth. The output of the local detectors is not very dependent upon the amplitude or spectral shaping of the input stimulus, because the detectors operate by detecting patterns of response phase, rather than response magnitude. The response of the local formant detector is highly speech specific, because this detector requires sequences of impulsive and synchronous features that are characteristic of speech. This constrasts with spectrograms, whichsimply represent a frequency profile of the energy, in the stimulus, regardless of the source of that energy.

DISCUSSION

Implications for auditory signal processing

Several neurophysiological studies have demonstrated the existence of a stable representation of the formant frequencies of steady-state vowels in the temporal patterns of nerve-fiber discharge over a range of stimulus levels. The Average Localized Synchronized Rate (ALSR) measure of Sachs and Young [2] measures the spectral amplitude of period histograms of neural response for groups of fibers with CFs close to a given spectral component (or its harmonics). Profiles of ALSR show peaks at the formant frequencies which remain constant over a wide range of stimulus level. In a similar study, Delgutte [3] determined that the dominant spectral component of period histograms of the response to synthesized steady-state vowels tend to be found at either a formant frequency, the fundamental frequency (i.e. F_0) or the CF, depending on the proximity of the fiber's CF to the formant frequency. In our view, the computation of either the ALSR or the dominant component by the brain is not physiological plausible, since these schemes require the brain to compute a period histogram of the response to speech, perform a spectral analysis of the histogram and to select components at frequencies corresponding to

known CFs.

In contrast, the detectors of the auditory-model approach are based on variations on a single physiologically plausible operation: finding the correlation of response patterns of groups of auditory-nerve fibers. Local impulse detectors detect sequential patterns of the firings of groups of adjacent nerve fibers. Local synchrony detectors detect simultaneous firings of groups of fibers. These detectors do not require the computation of period histograms of response, or spectral analysis of the results, nor does any channel need to "know" its CF. We suggest that detectors of speech features using this simple correlation approach may provide a reasonable model for signal processing by cells in the auditory central nervous system that process speech.

Comparison with other auditory-model approaches

Several workers have used concepts of auditory physiology to motivate the design of algorithms for speech recognition. In the approach of Ghitza [4], speech is frequency-analyzed by a bank of filters derived from a model of basilar-membrane motion. The output of each filter band is passed through a series of threshold detectors and an interval histogram is formed from the times between successive threshold crossings. Interval histograms for a plurality of filter bands are then combined to produce an ensemble histogram, from which a profile of the dominant average frequency components of the input signal is generated by means of conventional spectral-analysis techniques.

The Generalized Synchrony Detector (GSD) approach of Seneff [5] measures the periodicity of time-domain responses of a plurality of channels of a non-linear auditory model, but requires a series of detectors, each of which is effectively tuned to an individual channel's CF.

Lyon[6] uses a two-dimensional array of autocorrelators to extract synchronous information from the output of a non-linear cochlear model. This approach differs from that of Seneff in that none of the detector's channels explicitly represents or "knows" the channel's underlying CF. However, the problems of temporal granularity and computational complexity apply to this approach as well.

All the preceding methods are based on determination of the times of neural firings of non-linear auditory models. These methods thus have the drawbacks of temporal granularity and computational intractability which we have discussed previously. Furthermore, as we have argued, the operations of accumulating histograms and performing spectral analysis used by some methods are not likely to be physiological. The essential feature of all the preceding signal-processing approaches is to analyze the response of each model fiber individually, that is, without reference to the response of other fibers, (for example, by computing a period histogram of fiber response and then performing spectral analysis, or by performing an autocorrelation analysis of response). Then, further global operations may be performed on data from a number of individual fibers to detect important features (i.e. peaks in the response of histograms near their CFs). In contrast, the essential signal-processing strategy of our auditory-model approach is to measure simple correlations among the responses of groups of individual fibers. We believe we have taken this approach further than provious approaches, to the point of having designed speech-specific detectors for formants based on the detection of combinations of low-level features such as impulses and synchronous epochs.

Comparison with spectral approaches

The performance of the auditory-model algorithms appears to improve upon conventional spectrographic techniques in several respects:

• Accurate estimation of formant frequencies and voice pitch. The auditory-model algorithms represent the frequencies and times-of occurrence of formants accurately, 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 the measurement of response phase. They are highly independent of signal amplitude whereas 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 speech from non-speech.

• Computational simplicity. The operations of the local detectors involve simple correlations of instantaneous response phase among a small number of adjacent channels over a small number of samples. These operations are amenable to highly parallel signal processing architectures. Finally, unlike spectral-analysis techniques, processing is continuous and asynchronous. There are no inherent framing requirements for the data.

We conclude that algorithms for automatic speech recognition based on an auditory model are an effective alternative to spectrographic processing schemes for formant detection and may yield insight into the mechanism of speech processing by the auditory system.

This research supported by DOD contracts MDA904-90-C-3331 (Holton) and DAAH01-91-C-R095 (Votan).

REFERENCES

[1] T. Holton, S.D. Love and S.P. Gill. "A Fundamental Approach to Automatic Speech Recognition Using Models of Auditory Signal Processing." *DARPA Report.* Contract #DAAH01-91-C-R095, 1991.

[2] E.D. Young and M B. Sachs. "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers." *J. Acoust. Soc Am.*, vol. 66, pp. 1381-1403, 1979.

[3] B. Delgutte and N. Y.-S. Kiang. "Speech coding in the auditory nerve: I. Vowel-like sounds." *J. Acoust. Soc. Am.*, vol 75, pp. 866-878, 1984.

[4] O. Ghitza. "Auditory nerve representation criteria for speech analysis/synthesis" *IEEE Trans ASSP*, vol 35(6), pp. 736-740, 1987.

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

[6] R. Lyon. "Computational models of neural auditory processing." *IEEE-ICASSP*, 1984.