ADAPTIVE FILTERBANKS INSPIRED BY THE AUDITORY SYSTEM FOR SPEECH FEATURE EXTRACTION

Ramdas Kumaresan *Gopi Krishna Allu**

University of Rhode Island
Department of Electrical Engineering
Kelley Hall, Kingston RI 02881

Peter Cariani

Tufts Medical School
Department of Physiology
136 Harrison Ave., Boston, MA 02111

ABSTRACT

Using the human auditory system as a guide we propose a signal processing strategy for decomposing composite signals into bandpass signal components and extracting their features. We use two parallel filterbanks, one composed of a set of wideband overlapping filters and another consisting of a set of narrowband filters. The filterbanks cooperate to isolate regions of persistent and transient signal activity. The narrowband filters help identify spectral regions containing significant signal energy and groups of wideband filters in those regions are then optimally combined to isolate and track each of the bandpass signal components. Each group of wideband filters are combined to track one bandpass component while suppressing all other neighboring components. Narrowband filters in cascade with nonlinearities are used to characterize the transient components.

1. INTRODUCTION

The biggest barrier to widespread use of automatic speech recognition (ASR) systems in real-life situations is their unreliable performance in background noise and interference. In marked contrast to current artificial systems, human listeners are able to correctly identify speech utterances in many acoustically challenging contexts. We believe that critical examination of the auditory system and human auditory perception, with a focus on physiologically plausible signal processing mechanisms that is well grounded in the mathematics of signal processing, can lead to discovery of new functional principles of signal representation and processing that can improve ASR. Our overall goal is to develop a new robust signal processing front-end that will become part of many speech applications such as speech recognition in noisy environments.

Virtually every speech-recognition system that engineers have built uses framewise feature vectors derived from short-term spectral envelopes computed by spectral analysis or by using a bank of fixed bandpass filters. When speech is degraded by noise, interference, and channel effects (such as telephone speech, reverberation etc.,) perturbations at one frequency affect the entire feature vector. Also, frame-based processing reduces temporal acuity. This type of framewise spectral envelope extraction is entirely at odds with how the auditory system processes and recognizes speech. In the auditory system, sound components are spectrally and temporally separated to the extent possible, analyzed and subsequently fused into unified objects, streams and voices that exhibit perceptual attributes, such as pitch, timbre, loudness, and location.

2. SOME CLUES FROM THE AUDITORY SYSTEM

An understanding of the neural representation of complex sounds in early stages of auditory processing is critical for identifying the functional principles of its operation that are responsible for its high performance. After passing through the middle ear, the signals are bandpass filtered and compressed by the cochlea. For each place on the cochlear partition, both active or adaptive (outer hair cells) and passive processes (basilar membrane and inner hair cells) act to filter the signals. At low sound pressure levels (up to 40-50 dB SPL), cochlear filtering is dominated by an active, sharply tuned nonlinear process that seems to amplify frequency components near the resonant frequency. At higher levels, cochlear filtering is characterized by quasi-linear bandpass filters that are broadly tuned and asymmetric. Mechano-electrical transduction in roughly 3000 inner hair cells half-wave rectifies the signals. Each inner hair cell is innervated by 10-12 auditory nerve fibers (ANFs). As a consequence of these processes ANFs show responses that are narrowly tuned near their thresholds (20-50 dB SPL), but have very broad, overlapping and asymmetric frequency response areas at higher sound levels ($>$ 60 dB SPL). The 30,000 fibers that constitute the auditory nerve form the nexus through which virtually all information about the acoustic stimulus is transmitted to the central auditory system. However,the broad and shifting character of frequency tuning at higher sound intensities makes spectral representations based on

*Research funded by the National Science Foundation under the grants CCF-0105499 (to RK) and CCF-0130793 (to RK) and CCF-0130807 (to PC). Peter Cariani performed the work while at Massachusetts Ear and Eye Infi rmary, Boston, MA.

profiles of ANF spike rates highly unlikely. Instead, accumulated evidence suggests that most of the stimulus spectrum related information is conveyed through the auditory nerve via spike timing patterns. This "phase-locked" timing information, which extends to periodicities up to roughly 5 kHz, is critical both for fine auditory localization and frequency discrimination. It appears that this fine temporal information that is discarded (in the form of spectral phase) by artificial speech analyzers is utilized by the human auditory system for both auditory scene analysis and for the representation of periodicity (pitch) and spectrum (timbre, formant structure). We hypothesize that the temporal patterns of spikes convey information about the signal components' modulations and timing information related to the transient components. Hence we proceed to model and characterize these quantities.

3. SIGNAL PROCESSING APPROACH

Our overall feature extraction method is as follows. In a real-world acoustic environment (e.g. an office or a crowded restaurant) acoustic sensors simultaneously receive sounds from multiple sources. To identify which sound comes from which source it is necessary to first decompose the composite signal into more manageable bandpass components and then group the components that belong together based on some of their common attributes (e.g. common harmonicity). With this goal in mind we model a sensor output $s(t)$ as a sum of N bandpass signal components. That is $s(t) = \sum_{k=1}^{N} x_k(t)$. Some of these bandpass components, say $x_1(t)$, $x_2(t)$, .. $x_P(t)$, $(P < N)$ constitute the desired speech signal ($x_1(t)$ to $x_P(t)$ are assumed to be its formants) and the other $N - P$ components are due to interfering signals. We propose an adaptive filterbank algorithm, called MVFB (Minimum Variance Filter-Bank), which decomposes such a complex signal (to the extent possible) into individual signal components. Once the signal components are separated, we analyze the details of each of these components. Some signal components are said to be *Spectrally-Compact*, that is, they have a prominent spectral peak and are relatively narrow band, such as some low-frequency speech formants. For such a component we extract and track its carrier frequency Ω_c, its amplitude A_c and further characterize its phase-envelope modulations [1]. Other signal components are deemed to be *Spectrally Diffuse*, that is, they are relatively compact in time and are better characterized by modeling them in the dual domain. After the signal components are separated and analyzed, we group the signal components that belong together based on their common harmonicity, onset/offset times and source direction dependent delays and isolate the (features of the) desired speech signal from those of the interfering signals.

3.1. Models for a Bandpass Signal

When a signal is persistent / continuous and its bandwidth is less than a specified value (say less than the critical bandwidth at a given frequency location e.g., about 200Hz bandwidth at a center frequency of 2kHz) then we call it Spectrally-Compact. Formants with narrow bandwidths fall into this category. On the otherhand wider bandwidth formants and transients are said to be Spectrally-Diffuse. We have developed models for such bandpass signals in reference [1]. We outline these models here.

It is convenient to work with the complex version of bandpass signals (or analytic signals). We denote an analytic signal component by $x_k(t)$, that is

$$x_k(t) = A_c\ a(t) e^{j(\Omega_c t + \phi(t))}. \tag{1}$$

Its envelope is $|x_k(t)|$ and its instantaneous frequency (IF) is $\frac{1}{2\pi}\frac{d}{dt}\angle x_k(t)$. Unfortunately, this signal model does not lead to any insight into the relationship between phase and envelope functions. To further understand the phase-envelope relationships, following Herbert Voelcker [2], we have invoked a certain type of duality between the envelope (and phase) of an analytic signal and the magnitude (and phase) response of a causal linear-time-invariant (LTI) system. That is, we model the the envelope and phase of analytic signals using pole/zero models (with poles and/or zeros located in the complex-time plane), in the same way that the LTI systems are modelled with poles and zeros located in the standard complex-frequency (the z or s) plane. Using this perfect dualism between complex-time representation of the signal and an LTI system's frequency response, we can then model an arbitrary bandpass analytic signal in Eq.(1) as a product of minimum phase (MinP) and maximum phase (MaxP) signals. That is

$$x_k(t) = A_c\ e^{j\ \Omega_c\ t}\ \underbrace{e^{\alpha(t)+j\hat{\alpha}(t)}}_{MinP}\ \underbrace{e^{\beta(t)-j\hat{\beta}(t)}}_{MaxP}, \tag{2}$$

where $\hat{\alpha}(t)$ is the Hilbert transform of $\alpha(t)$. Taking the natural logarithm on both sides of Eq.(2) and then its time-derivative we get

$$\frac{d}{dt}\log\left(x_k(t)\right) = j\underbrace{\Omega_c}_{AIF} + \underbrace{\dot{\alpha}(t) + j\ \dot{\hat{\alpha}}(t)}_{analytic} + \underbrace{\dot{\beta}(t) - j\ \dot{\hat{\beta}}(t)}_{antianalytic}, \tag{3}$$

where "dot" denotes the time derivative. Note that Ω_c (the Average Instantaneous Frequency (AIF)) and the analytic and antianalytic components all have essentially nonoverlapping spectra and hence can be separated by filtering. See Fig.4 in [3]. If the signal $x_k(t)$ has a dominant unimodal spectral peak then the AIF (measured either by using phase derivative of $x_k(t)$ or by counting zero-crossings) is a reliable feature of the bandpass signal. Further, low-pass filtering $\log|x_k(t)|$ yields $\log A_c$ which we call as the Average

Log-Envelope or ALE. Thus the ALE, AIF and the analytic and antianalytic components (which characterize the spectral tilt) of a bandpass signal are all obtained as features. Finally, note that the modulation analysis used here is the "time-domain" analog of cepstral analysis used to separate the causal and noncausal parts of a sequence except that it is applied here to each spectrally-compact component, not the entire signal.

If a signal component $x_k(t)$ is transient in nature (assuming that it is composed of a single "bandpass pulse") then we model its Fourier transform as

$$X_k(\Omega) = A_c\ A(\Omega) e^{j(-\Omega\tau_k + \Phi(\Omega))}, \qquad (4)$$

where Ω is the frequency variable and τ_k is the location of the pulse with respect to a time reference. Note that this model is the dual of the model in Eq.(1). The delay, τ_k, is the counterpart of the AIF in Eq.(1) and the model can be decomposed into causal and noncausal parts analogous to that in Eq.(2).

3.2. MVFB Algorithm for Tracking Bandpass Components

The two branches of the proposed minimum-variance filterbank (MVFB) algorithm are shown in figure 1. We have space here to discuss in some detail the left branch only. The left branch has a set of wideband filters spanning the entire frequency range of interest. Similarly, the right branch has a set of narrowband filters. The input signal is simultaneously processed through both filterbanks. The envelopes of the filtered signals are continually monitored. When the envelopes of the narrowband filters' outputs exceed a certain threshold they indicate the "places" or spectral regions where signal power is significant. This information is then used to "enable" the groups of wideband filters that are located in those spectral regions. One such group consisting of L wideband filters is shown inside a dashed rectangular box (figure 1). The L filters are then linearly combined with weights w_1, ... ,w_L to form a **resultant** filter. The output of the resultant filter is $x_k(t) = \sum_{n=1}^{L} w_n s_n(t)$. The weights w_n are determined such that the frequency response of the resultant filter is constrained to be unity at a given frequency Ω_d (called the *steering frequency*) while the energy in $x_k(t)$ is minimized. This results in a simple minimization problem similar to the minimum variance distortionless receiver (MVDR) well known in adaptive beamforming [4]. Recall that in MVDR beamforming the array response is required to be unity in the direction of the steering vector while the power received from all other directions is minimized by adaptively placing nulls in the array response pattern. We adopt this same principle here by passing the desired bandpass component $x_k(t)$ without distortion while placing nulls at all other signal components' frequencies. The phase derivative of the bandpass signal $x_k(t)$ is averaged to obtain an estimate of the AIF which then serves as the steering frequency. Since speech formants slowly drift with time, the steering frequency (same as the AIF) helps move the resultant filter along the formant trajectory thereby tracking the formant frequencies. The MVFB algorithm is adaptive in two ways. First, it adaptively suppresses all other bandpass components while passing the desired bandpass component undistorted. This is crucial for estimating the AIF reliably since other components do not interfere with the AIF estimation. In this respect MVFB is a significant improvement compared to our previous method [5]. And secondly, the resultant filter with the help of the steering frequency positions itself right on top of the formant.

The spectrogram of the speech utterance "Three" and the formant tracks (AIFs) for three groups of filters obtained by the above procedure are shown in figure 2. Particularly noteworthy is the figure 3 which shows the tracks of the nulls of the time varying resultant filter whose passband stays centered on the third formant (located around 3 kHz). Note that the nulls of this filter are always centered over the first and the second formants, even though these formants themselves are slowly drifting. Similarly, the resultant filter working in the second formant region has nulls placed over the first and third formant locations (not shown here) and so on. The low frequency formant tracks obtained by the above procedure for a longer speech utterance "3o33951" are plotted over the spectrogram in figure 4.

Summarizing, since we have to filter bandpass components of unknown bandwidth and center frequency, it seems reasonable to try to synthesize the desired filters on line, based on the characteristics of the input signal itself. This is achieved in two steps. The narrowband filters first locate roughly the spectral regions with signal energy. Then wideband filters in those regions are combined such that the resultant filter is wide enough to pass a particular component, while nulling all other components by using a minimum variance criterion. The operation of the right branch is analogous, i.e., it allows a bandpass pulse located at time τ_k to get through undistorted, while suppressing other pulses in its neighborhood. The compressive nonlinearities tend to enhance the onsets and offsets of signal components.

4. REFERENCES

[1] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive-instantaneous frequency of signals and application to speech," *Journal of the Acoustical Society of America*, vol. 105 (3), pp. 1912–1924, (March) 1999.

[2] H. B. Voelcker, "Towards a unified theory of modulation part I: Phase-Envelope relationships," *Proceedings of the IEEE*, vol. 54, no. 3, pp. 340–354, 1966.

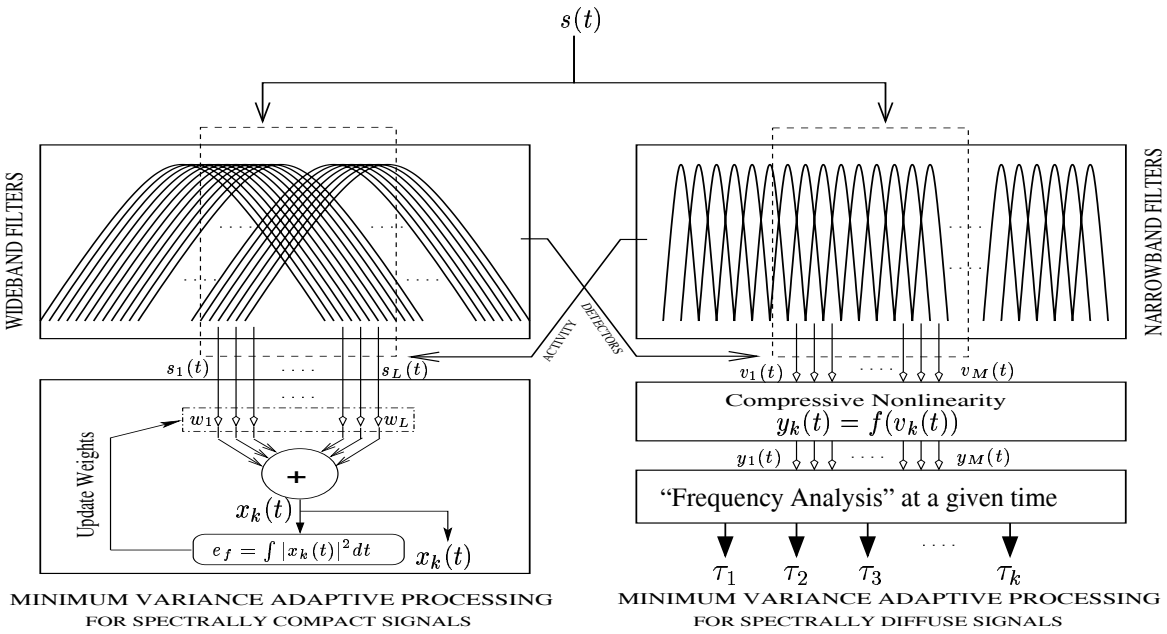

Fig. 1. Block diagram of the Minimum Variance Filterbank (MVFB) Algorithm

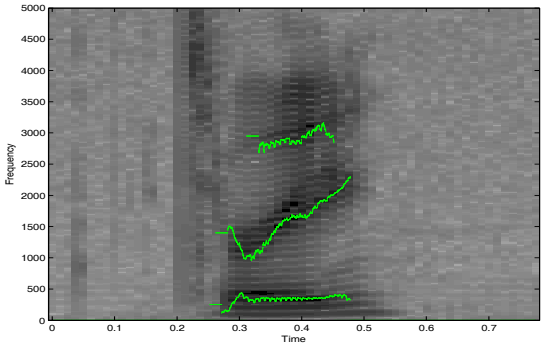

Fig. 2. AIFs plotted on top of spectrogram for the utterance "Three".

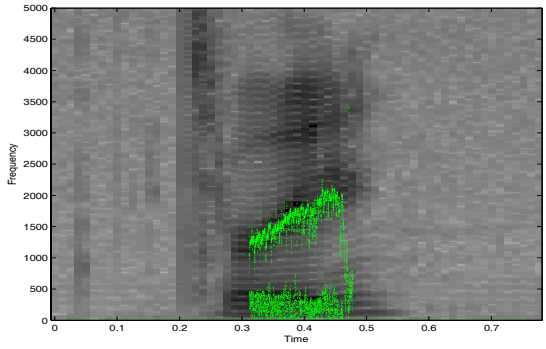

Fig. 3. Tracks of the nulls of the resultant filter while its passband is centered on the 3rd formant.

[3] Y. Wang, J. Hansen, Gopi K. Allu, and R. Kumaresan, "Average Instantaneous Frequency (AIF) and Average Log-Envelopes (ALE) for ASR with the Aurora 2 Database," in *Proceedings of Eurospeech-03*, Geneva, Switzerland, September 2003, pp. 25–28.

[4] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part IV ,Optimum Array Processing*, Wiley, New York, 2002.

[5] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.

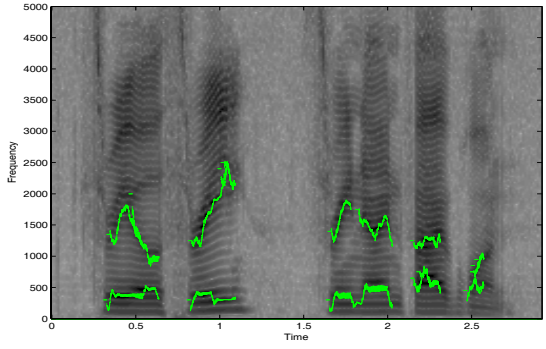

Fig. 4. AIFs plotted on top of spectrogram for the utterance "3o33951".