SYNCHRONY CAPTURE FILTERBANK (SCFB): AN AUDITORY PERIPHERY INSPIRED METHOD FOR TRACKING SINUSOIDS

Ramdas Kumaresan, Vijay Kumar Peddinti

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

Peter Cariani

Department of Otology & Laryngology,
Harvard Medical School
Boston, MA 02114

ABSTRACT

We propose a novel algorithm for tracking multiple sinusoidal signals that is motivated by neural coding in the mammalian peripheral auditory system. A striking feature of auditory nerve activity is the phenomenon of "synchrony capture," whereby the most intense frequency components in the stimulus dominate the temporal firing patterns of whole subpopulations of auditory nerve fibers (ANFs). A novel adaptive filterbank structure that emulates key aspects of synchrony capture is presented. The proposed filterbank has two components: a fixed bank of traditional gammatone (or equivalent) filters that are cascaded with a bank of adaptively-tunable bandpass filter triplets. The bandpass filters are tuned by using a voltage controlled oscillator (VCO) whose frequency is steered by a frequency discriminator loop (FDL). The resulting filterbank is used to process synthetic signals and speech. It is shown that the VCOs can track the low frequency harmonics in speech that evoke voice pitch at their fundamental (F0). For vowels, the VCOs faithfully track the strongest harmonic present in each formant region.

Index Terms— auditory model, frequency capture, harmonics, cochlea, tunable filters

1. INTRODUCTION

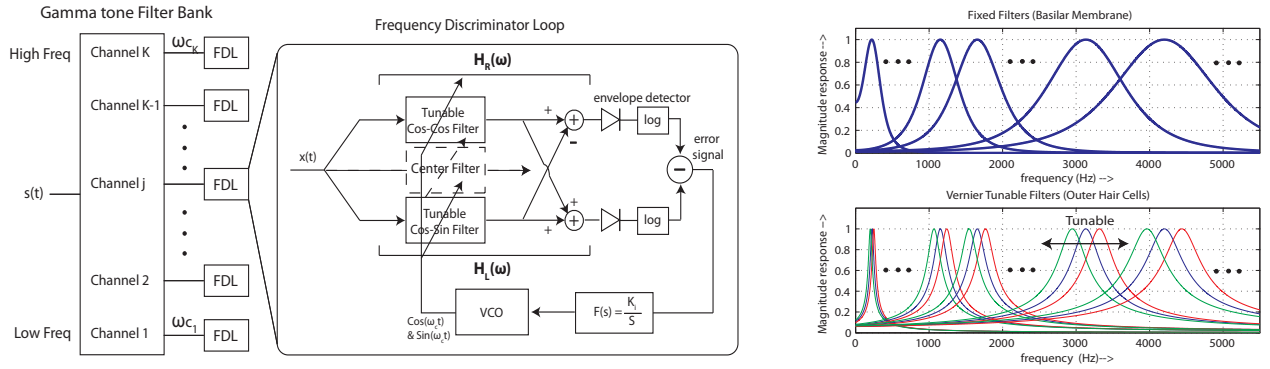
This paper proposes signal analysis algorithms for processing speech, music, and other audio signals that are inspired by the auditory system. For the past three decades there has been significant interest in developing computational signal processing models based on the neurophysiology of the auditory nerve [1]. Our work in this area is motivated by physiological observations of the synchrony capture phenomenon by Sachs and Young [2] and Delgutte and Kiang [3]. For vowel stimuli, the phase-locked , temporal firing patterns of fibers of an entire cochlear place region of nearby characteristic frequencies (CFs) are driven almost exclusively by one local, dominant frequency component, despite the presence of other, nearby weaker ones [3]. At moderate and high sound pressure levels, fibers spanning an entire octave or more of CF are typically driven at their maximal rates and exhibit firing patterns related to a single, dominant component in each formant region. From a signal processing perspective, capture by a dominant component while ignoring nearby weaker components resembles the well-known "frequency capture" behavior [4] of frequency modulation (FM) receivers. This mode of response permits FM devices to receive an FM signal with little distortion even when other, weaker FM signals nearby in frequency are also present. Traditional FM receiver circuits such as frequency discriminators, phase locked loops and ratio detectors exhibit this frequency capture property, suggesting possible signal processing analogies with the encoding of signals in the auditory nerve. In functional terms, one can conceive of hair cell stereocilia as soft rectifiers, outer hair cell active processes as voltage controlled oscillators, and hair cell membranes as lowpass filters. These functional analogies have motivated the signal processing architecture proposed here.

The proposed algorithm (an extension of our previous work [5]) can resolve closely spaced (low frequency) harmonics from interfering sounds in many cases, at least over short intervals. The nonlinearity in the feedback loop assists in this respect by locking onto the dominant component's frequency rather than finding a weighted average frequency of the two interacting signals. Frequency locking thus reduces distorting interference between nearby signals, which in turn can better support harmonic grouping operations that subserve separation of multiple concurrent voices. Signal processing strategies for automatic attenuation of weaker, interfering sounds thus seem attainable.

2. SYNCHRONY CAPTURE FILTERBANK

We propose a signal processing architecture (Figure 1) that uses an adaptive frequency locking mechanism to effect the capture of dominant frequency components in the stimulus. It consists of a bank of fixed, relatively broad bandpass filters (BPF) that emulate basilar membrane (BM) filtering, in

This work was supported by the Airforce Office of Scientific Research under the grant # AFSOR FA9550-09-1-0119



(a) (b)

Fig. 1. **Synchrony Capture Filterbank (SCFB):** (a) SCFB architecture. Rightmost is a bank of K logarithmically-spaced, constant (but low) Q gammatone filters whose center frequencies span the desired audible frequency range (emulating BM filtering). Next, a frequency discriminator loop (FDL) is cascaded with each of the K filters, with each such cascade henceforth being called a "channel." Each FDL is made up of three tunable bandpass filters. The right $H_R(\omega)$ and the left $H_L(\omega)$ filters' output envelopes are compared and their difference is used to drive the VCO after passing through an integrator. The VCO outputs are used to tune all three filters. The output of each channel is obtained from its center filter $H_C(\omega)$. (b)Frequency responses of fixed (top) and tunable (bottom) filters.

cascade with tunable narrower filters that produce the capture property. The proposed model is not unlike a vernier scale, in that the gross measurement of frequency is made by the fixed filterbank (á la BM), while more precise measurement is achieved by the second bank of tunable filters. Each secondary filter forms part of a frequency discriminator loop (FDL) whose hypothetical cochlear counterpart would be an outer hair cell/tectorial membrane/basilar membrane feedback loop. FDLs are basic tone trackers. Each FDL is made up of three tunable bandpass filters ("BPF triplet") whose arrangement was inspired by the triple-row geometry of outer hair cells on the basilar membrane. The tuning of all three BPFs is accomplished by a single VCO. The novel part of the SCFB is the design of the FDL, which is described in the next section. The frequency error detector (FED), the crucial part of the FDL uses matched right $H_R(\omega)$ and left $H_L(\omega)$ filters to compute frequency difference between its input tone and VCO frequency. Section 3 shows synthetic signals and speech processed using the SCFB. It is shown that for voiced part of speech signals the lowest frequency channels are captured by individual low harmonics, with higher frequency channels being captured by dominant harmonics in each formant region (not unlike what occurs in the auditory nerve).

2.1. Frequency Discriminator Loop (FDL)

Frequency Discriminator Loops (FDLs) have been used for decades to synchronize transmitter and receiver oscillators in digital and analog communication systems [6, 7, 8]. The structure of the proposed frequency tracking algorithm is similar to the FDLs used in communication systems. The block diagram of a generic FDL is shown in Figure 2. It consists of a frequency error detector (FED), a loop filter and a VCO. The FED outputs an error signal $e(t)$ that is proportional to the difference between the frequency of the input signal ω_1 and that of the VCO, ω_c. The loop filter provides the control voltage to the VCO and drives its frequency such that $\omega_c - \omega_1$ tends to zero. Typically the loop filter is an integrator, i.e., $F(s) = k_i/s$.

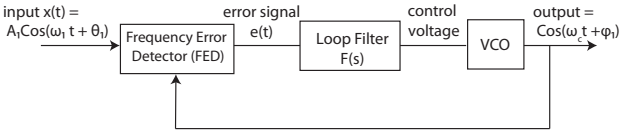


Fig. 2. **Generic FDL:** The error signal e(t) is a measure of the frequency difference between the input tone and the VCO output. The details of the frequency error detector are shown in figure 4.

2.2. Frequency Error Detector (FED) based on Tunable Right, Left and Center Filters

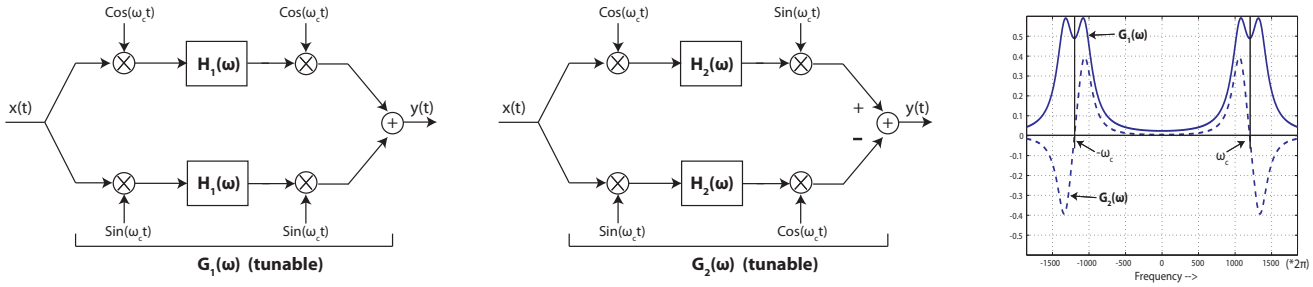
The three bandpass filters that constitute the FED (see Figure 4, $H_C(\omega)$ not shown) are all synthesized from a single prototype noncausal impulse response $h(t) = e^{-\alpha|t|}$. $H(\omega) = 2\alpha/(\omega^2 + \alpha^2)$. Only the right $H_R(\omega)$ and the left $H_L(\omega)$ filters are used in error detection. Let $h_1(t)$ and $h_2(t)$ be the impulse responses of frequency translated filters, given by

$$h_1(t) = h(t)\cos\Delta t, \text{ and } h_2(t) = h(t)\sin\Delta t, \qquad (1)$$

where Δ is the translation frequency. So,

$$\begin{aligned} H_1(\omega) &= (H(\omega-\Delta) + H(\omega+\Delta))/2, \\ H_2(\omega) &= j(H(\omega-\Delta) - H(\omega+\Delta))/2. \end{aligned} \qquad (2)$$

$j = \sqrt{-1}$. Δ is chosen equal to α, so that Δ is the 3-dB point of $H(\omega)$. The frequency responses $H_1(\omega)$ and $H_2(\omega)$ are purely real and imaginary, respectively. $H_1(\omega)$ and $H_2(\omega)$ are embedded as part of the tunable band pass filters $G_1(\omega)$



(a) (b) (c)

Fig. 3. : a) Tunable Cos-Cos filter and b) Cos-Sin filter. c) $G_1(\omega), G_2(\omega)$ (without the scale factor j) are shown in c.

and $G_2(\omega)$ shown in Figures 3a and 3b, respectively. $G_1(\omega)$ is called a Cos-Cos filter and $G_2(\omega)$ is named a Cos-Sin filter. The term Cos-Cos is used to denote that both the multipliers in the upper branch of $G_1(\omega)$ are supplied with $\cos\omega_c t$, whereas for the Cos-Sin filter the two multipliers in the upper branch are supplied with $\cos\omega_c t$ and $\sin\omega_c t$. It is easy to show that

$$\begin{aligned} G_1(\omega) &= (H_1(\omega-\omega_c)+H_1(\omega+\omega_c))/2, \\ G_2(\omega) &= j(H_2(\omega-\omega_c)-H_2(\omega+\omega_c))/2. \end{aligned} \quad (3)$$

The frequency responses $G_1(\omega)$ (real and even) and $G_2(\omega)$ (real and odd) are shown in Figure 3c. These frequency responses can be tuned by changing ω_c. Note that the system functions of a generic Cos-Cos structure and Cos-Sin structure (if we choose $H_1(\omega) = H_2(\omega)$) are related by the expression $G_2(\omega) = jsgn(\omega)G_1(\omega)$. That is, Cos-Cos structure has an additional term which signifies a Hilbert transform when compared to Cos-Sin structure. This stems from the fact that the multipliers in the upper/lower branches of Figure 3b are cosine and sine unlike the Cos-Cos filter in Figure 3a. The outputs of the Cos-Cos and Cos-Sin filters are added/subtracted (see Figure 4a) to obtain the overall right/left filter responses $H_R(\omega)$ and $H_L(\omega)$, respectively. That is,

$$\begin{aligned} H_R(\omega) &= G_1(\omega) - G_2(\omega), \\ H_L(\omega) &= G_1(\omega) + G_2(\omega). \end{aligned} \quad (4)$$

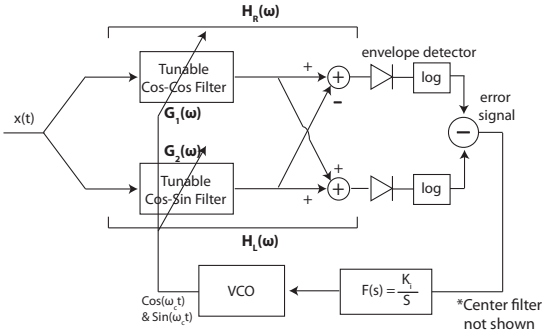
Substituting for $G_1(\omega)$ and $G_2(\omega)$ in Eq.4 from Eq.3, we have

$$\begin{aligned} H_R(\omega) = \; & (H_1(\omega-\omega_c)+H_1(\omega+\omega_c))/2 \\ & + j(H_2(\omega-\omega_c)-H_2(\omega-\omega_c))/2, \\ H_L(\omega) = \; & (H_1(\omega-\omega_c)+H_1(\omega+\omega_c))/2 \\ & - j(H_2(\omega-\omega_c)-H_2(\omega-\omega_c))/2. \end{aligned} \quad (5)$$

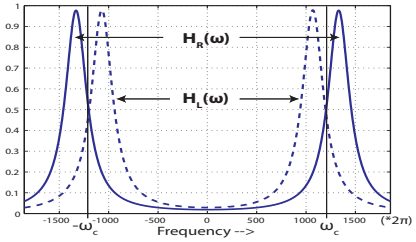
Further substituting for $H_1(\omega)$ and $H_2(\omega)$ in Eq.5 from Eq.2 and simplifying, we have

$$\begin{aligned} H_R(\omega) &= H(\omega-\omega_c-\Delta)+H(\omega+\omega_c+\Delta)) \\ H_L(\omega) &= H(\omega-\omega_c+\Delta)+H(\omega+\omega_c-\Delta)). \end{aligned} \quad (6)$$

Thus, the filters $H_R(\omega)$ and $H_L(\omega)$ (shown in Figure 4b) are the original prototype filter $H(\omega)$ shifted to center frequencies $\omega_c+\Delta$ and $\omega_c-\Delta$, respectively. They are purely



(a)



(b)

Fig. 4. **Frequency Error Detector and the FDL:** a)The filters $H_R(\omega)$ and $H_L(\omega)$ are obtained as sum and difference of $G_1(\omega)$ and $G_2(\omega)$. At low frequencies the envelopes are compressed using a logarithmic nonlinearity, where as at high frequencies the error is computed as a normalized envelope difference (envelope difference/envelope sum). The filters $H_R(\omega)$ and $H_L(\omega)$ are basically synthesized from a single prototype $H(\omega)$, and hence are perfectly matched and symmetric about ω_c. b) The frequency responses $H_R(\omega)$ and $H_L(\omega)$. $H_C(\omega)$, not shown, is centered around ω_c.

real valued. In practice, the filter impulse responses in Eq.1 are symmetrically truncated about the time origin and made causal by shifting them to the right resulting in linear phase filters. The center filter $H_c(\omega)$ (also tunable) centered around ω_c, not shown in Figure 4a or 4b, is synthesized using the Cos-Cos structure but with the prototype filter $H(\omega)$ sandwiched between the multipliers. Its output is not used in error signal calculation but is the channel output. If the input tone frequency ω_1 is less than the VCO frequency ω_c then the envelope at the output of $H_L(\omega)$ is larger than the envelope at the output of $H_R(\omega)$ and the error signal will drive the VCO to make ω_c equal to ω_1 and vice versa. The integrator gain k_i determines the dynamics.

3. SIMULATION

We have tested SCFB algorithms and adaptive parameters using several synthetic complex tones and speech signals from the TIMIT database. Here we show results of one synthetic and one speech simulation. The SCFBs used in these simulations have K=200 logarithmically spaced (roughly constant Q) gammatone filters spanning a frequency range of 0.1-5 kHz, which is standard fare in auditory system modeling [9], with sampling frequency of 16 kHz. Figure 1b (top) shows the magnitude response of the gammatone filter bank. Values of Δ for the tunable BPFs ranged from 19 Hz at low frequencies to a maximum of 226 Hz at the high frequency end. Details of the control loop design and effects of parametric variations will be presented elsewhere [10]. Figure 5 shows the frequency tracks of the center VCO when the input consists of tones at 440, 587 and 880 Hz with equal amplitudes. Clearly several channels are captured by the tones that dominate their frequency neighborhood. We call the running plots of VCO frequency tracks of the channels "capturegrams". It can be seen that 440 Hz dominates channels with center fixed frequencies from 380-500 Hz, 587 Hz dominates channels from 550-700 Hz, and 880 Hz dominates channels from 780-1000 Hz. Increasing the relative amplitude of a tone causes it to capture more channels in its neighborhood, which is akin to the synchrony capture phenomenon observed in the auditory nerve.

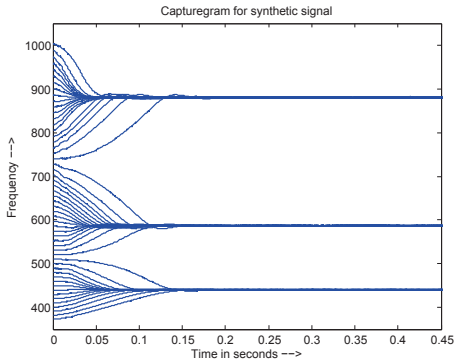


Fig. 5. **Capturegram** for a synthetic signal with tones of equal amplitudes at 440, 587 and 880 Hz

Figure 6 shows output of the SCFB in response to the TIMIT speech waveform (sx9), "Where were you while we were away?" which is spoken by a male speaker. The traditional spectrogram is plotted in color, and the capturegram showing all 200 VCO frequency trajectories is overlaid in black. No information about amplitudes of channel outputs was used in obtaining the capturegram. Simply, if a harmonic in voiced speech signal (after passing through a gammatone filter) is large compared to its neighbors, then the VCOs of channels in that neighborhood tend to lock on to the frequency of that component. As can be seen in Figure 6, at low harmonic numbers all individual harmonics are tracked, whereas at higher harmonic numbers, only one prominent harmonic in each formant region is tracked.

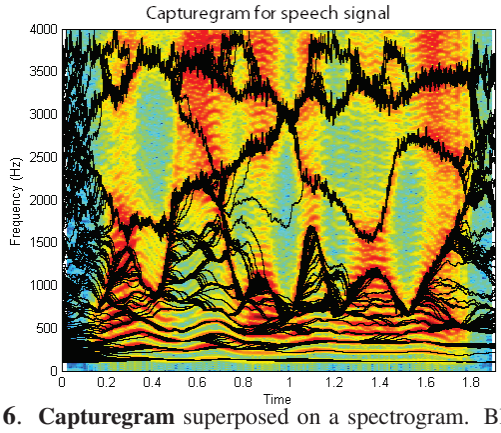


Fig. 6. **Capturegram** superposed on a spectrogram. Black lines show frequency tracks of the 200 filterbank VCOs.

4. REFERENCES

[1] C Kim, Yu-H Chiu, and R M. Stern, "Physiologically-Motivated Synchrony-Based Processing for Robust Automatic Speech Recognition," *INTERSPEECH 2006 - ICSLP*, Sep. 2006.

[2] M. Sachs and E. Young, "Encoding of Steady-State Vowels in the auditory nerve: representation in terms of discharge rate," *J. Acoust. Soc. Am.*, vol. 66, no. 1, pp. 470–479, 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] E. J. Baghdady, "Theory of Stronger-Signal Capture in FM Reception," in *Proceedings of Institute of Radio Engineers*, Aalborg, Denmark, April 1958, pp. 728–738.

[5] R. Kumaresan, V. K. Peddinti, and P. Cariani, "Multiple pitch identification using cochlear-like frequency capture and harmonic grouping," in *Proceedings of the ICASSP*, Prague, May 2011, pp. 613–616.

[6] Francis D. Natali, "AFC Tracking Algorithms," *IEEE Trans. Commun.*, vol. Com-32, pp. 935–947, Aug. 1984.

[7] J. P. Costas, "Residual Signal Analysis," *Proceedings of the IEEE*, vol. 68, pp. 1351–1352, Oct. 1980.

[8] M.J.Ferguson and P.E.Mantey, "Automatic frequency control via digital filtering," *IEEE Transactions on Audio and Electroacoustics*, vol. AU-16, no. 3, pp. 392–397, Sept. 1968.

[9] J. Holdsworth, and I. Nimmo-Smith, and R. Patterson, and P. Rice , "Implementing a GammaTone filter bank," *MRC Applied Psychology Unit Tech. Rep*, Feb 1988.

[10] R. Kumaresan, V. K. Peddinti, and P. Cariani, "Synchrony Capture Filterbank (SCFB): An Auditory System Inspired Method for Tracking Sinusoidal Signals (in preparation)," *J. Acoust. Soc. Am.*, 2012.