2015

SYNCHRONY CAPTURE FILTERBANK (SCFB): AUDITORY-INSPIRED SIGNAL PROCESSING FOR FREQUENCY TRACKING

Vijay Kumar Peddinti
University of Rhode Island, vijay@ele.uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation

Peddinti, Vijay Kumar, "SYNCHRONY CAPTURE FILTERBANK (SCFB): AUDITORY-INSPIRED SIGNAL PROCESSING FOR FREQUENCY TRACKING" (2015). *Open Access Dissertations*. Paper 334.
https://digitalcommons.uri.edu/oa_diss/334

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

SYNCHRONY CAPTURE FILTERBANK (SCFB): AUDITORY-INSPIRED
SIGNAL PROCESSING FOR FREQUENCY TRACKING

BY

VIJAY KUMAR PEDDINTI

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL ENGINEERING

UNIVERSITY OF RHODE ISLAND

2015

ProQuest Number: 3689279

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality and completeness of the copy made available to ProQuest.

Distributed by ProQuest LLC (2015).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license or other rights statement, as indicated in the copyright statement or in the metadata associated with this work. Unless otherwise specified in the copyright statement or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA

DOCTOR OF PHILOSOPHY DISSERTATION
OF
VIJAY KUMAR PEDDINTI

APPROVED:

Dissertation Committee:

Major Professor	Ramdas Kumaresan
	G. Faye Boudreaux-Bartels
	David R. Heskett
	Peter A. Cariani
	Nasser H. Zawia
	DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2015

ABSTRACT

The mammalian auditory system is a more robust and versatile sound analyzer than any artificial system that has been developed to date. Nature found a simple yet elegant solution for the hearing mechanism. Incorporating some key aspects of the functional organization of the mammalian auditory system into artificial signal-processing systems may drastically simplify problems of auditory representation and scene analysis such that capabilities for acoustic signal separation, detection, classification, recognition and identification can be greatly improved.

The objective of the thesis is to mimic the functionality of the mammalian peripheral auditory system in a digital computer by developing a synchrony capture filterbank (SCFB) algorithm. This thesis is primarily inspired by two aspects of the peripheral auditory system: (1) synchrony capture, a phenomenon observed in the auditory nerve which involves the preferential synchronization of the discharges in a given frequency region of the cochlea to a single dominant frequency component in that region. In other words, a strong dominant frequency component suppresses any interfering weaker tones. (2) the spatial arrangement of the mammalian cochleae. The SCFB algorithm is used to track the frequency components of a speech signal, extract the pitch or fundamental frequency of quasi-periodic sounds.

To emulate synchrony capture, the proposed filterbank is designed as a two step process, which includes a coarse and a fine analysis. The first stage is a broad filter, followed by a bank of three adaptively tunable narrower bandpass filters, which resembles the basilar membrane and the three rows of outer hair cells in the inner ear. This filterbank attempts to emulate synchrony capture-like behavior using these adaptive filters, by creating a competition for different channels amongst frequency components that not only accurately reflect their relative magnitudes,

but is also invariant with respect to absolute signal amplitude. These 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, and it is shown that the VCOs can track the individual low frequency harmonics and the strongest harmonic present in each formant region.

Finally, these SCFB outputs are used to compute fundamental frequency or pitch, f_0 of quasi-periodic sounds present in the signal. Currently, auto-correlation based models are widely used for pitch extraction. Although there is overwhelming neurophysiological evidence for auto-correlation-like representations of sounds in the temporal firing patterns of neurons in the auditory nerve and brainstem, how the central auditory system makes use of these representations is still not well understood. Although neuronal populations that carry out a binaural cross-correlation operation have been long identified in the auditory brainstem, no obvious analogous neural time-delay architectures for monaural auto-correlation have yet been found. This motivates the search for an alternative signal processing strategy. An approach based on SCFB is proposed as a possible alternative to autocorrelation computation. The outputs of the SCFB are adaptively phase aligned with respect to a common time reference and added to compute a summary phase aligned function (SPAF), from which fundamental frequency or pitch, f_0 can then be extracted. Results show that component frequencies are f_0 are faithfully tracked.

ACKNOWLEDGMENTS

My research would not have been possible without support from several individuals. First, I would like to thank and express my deepest gratitude to my major professor, Dr. Ramdas Kumaresan, with whom I had the fortunate opportunity to work. His creating thinking and insights broadened my thought process; his intellect and patience are much appreciated. He is definitely one of the best mentors a student can ask. Without his support, this thesis would not have been possible. I am also fortunate to work with Dr. Peter Cariani, our research collaborator and have always enjoyed our meetings and discussions. Thanks to both Dr. Kumaresan and Dr. Cariani for introducing me to this research area, and also would to thank them for their help in editing my thesis.

Additionally, I would like to extend my deepest gratitude to all the professors in my thesis committee that includes Dr. Godi Fischer, Dr. David Freeman, Dr. Faye Boudreaux-Bartels, and Dr. David Heskett. I would like to thank them for their time, insightful comments and for their support throughout the thesis. I would like to thank all the professors in the department as well. Additionally, I would like to acknowledge Dr. Fischer and Dr. Freeman, who inspired me to pursue a Ph.D. This thesis would not have been possible without their support.

I have thoroughly enjoyed my time in both the Electrical engineering and Physics departments. I would like to thank the graduate students with whom I have interacted during my education at URI, including my lab mates over the years Juin Yee, Babak, Adrienne. In addition, I would like thank Meredith Leach-Sanders, our department secretary and all the student department assistants (Rebekah, Kristen, Melyssa, Andrea to name a few) over the years for all their assistance. I would like to thank the Electrical engineering department for providing teaching assistantship over the years, and also the Air Force Office of Scientific Re-

search (AFOSR) funding, which helped me in completing my Ph.D. and masters in Electrical engineering and Physics departments, respectively. Special thanks to Tim Toolan for his help, and Jim Vincent for his assistance with the labs. I have enjoyed conversations with Dan, our department custodian as well.

I am grateful to my parents, and brother for their support. Their love and support is very much appreciated. I also would like to acknowledge Larry McLaughlin, Scott Lachance and their families for their love and support. This paragraph is dedicated to all my friends who I could not mention due to space constraints.

DEDICATION

This thesis is dedicated to family and friends

PREFACE

In the past five years, with support from AFOSR# FA9550-09-1-0119 I worked with Dr. Kumaresan and Dr. Cariani in developing the auditory inspired algorithms for frequency tracking. The motivation for this work stems from the orderly structure of the auditory system and a phenomenon called “synchrony capture” which is observed in the auditory nerve. This phenomenon was originally reported by M. Sachs and E. Young (Sachs and Young, JASA-1979) and B. Delgutte and N.Y.S. Kiang (Delgutte and Kiang, JASA-1984). Synchrony capture means the auditory nerve fibers (ANFs) almost exclusively synchronize their firing rate to the dominant frequency component, in spite of the presence of other weaker nearby components in frequency. Our algorithm, called the synchrony capture filterbank (SCFB) consists of a bank of broadly tuned filters in cascade with narrower filters that adaptively lock onto locally-dominant frequency components to produce synchrony capture. The bank of broad filters is not unlike the basilar membrane (BM) and the narrow filters is not unlike outer hair cells action. The filterbank precisely tracks individual time-varying frequency components, such as low harmonics and formant frequencies in speech, in the midst of noise and auditory clutter. This precise tracking in turn can be used to enhance the separation of concurrent periodic sounds.

My research started with working on zero crossings of waveforms as a method for representation of signals, further continuing Dr. Kumaresan’s prior work. Our synchrony capture research idea started as a discussion on our way to an Acoustical Society of America (ASA) conference in Baltimore, 2010. We were intrigued by the striking similarities between synchrony capture and frequency capture. Since then we focused on emulating synchrony capture as an algorithm. Frequency capture is observed in signal processing mechanisms such as a bank of frequency modulation

(FM) receivers. Such receivers lock on to a strong signal even in the presence of nearby weaker signal components, similar to synchrony capture observed in the auditory nerve. Synchrony capture has several advantages. It assists in suppressing noise stimuli and enhances temporal representations of both individual harmonics and fundamental pitch, which is important for separating multiple speech signals. A goal of this thesis is to take advantage of these benefits to develop a workable algorithm which facilitates better pitch determination and improved acoustic signal separation, detection, identification, classification, and recognition.

In recent decades computational signal processing models based on the biology of the auditory system have been of significant interest. These models serve two main purposes: first, to leverage design principles of biological auditory systems for technological advances in audio signal processing, and second, to generate plausible functional hypotheses for reverse-engineering the auditory system. The hope is to uncover new design principles from nature that can equal or exceed human auditory performance. Even though in this thesis we try to emulate the function of the peripheral auditory system, the goal is not to model cochlear biophysics, but to develop a signal processing algorithm which emulates cochlear function.

Some of my work presented in the thesis already appeared in the following Acoustical Society of America (ASA) and the Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech and Signal Processing (ICASSP) publications and some of the later work is ready for submission. Additionally we are in the process of applying for a patent.

- Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2011) **“Multiple Pitch Identification Using Cochlear-like Frequency Capture and Harmonic Grouping,”** Proceedings of the ICASSP, Prague, CZ, pp-613-616, May 2011.

- Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2012) “**Synchrony Capture Filterbank (SCFB): An Auditory Periphery Inspired Method for Tracking Sinusoids**,” Proceedings of the ICASSP, Kyoto, Japan, pp-153-156, March 2012.
- Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2013) “**Synchrony Capture Filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech**,” Journal of ASA, Vol. 133, Issue 6, pp. 4290-4310, June 2013.
- Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2014) “**Auditory-inspired pitch extraction using a synchrony capture filterbank and phase alignment**,” Proceedings of the ICASSP, Florence, Italy, AASP-P8.11, May 2014.

In addition to these publications, I was fortunate to attend and present at the following ASA and ICASSP conferences.

- Poster Presentation: Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2010) “**Spatiotemporal coding of signals motivated by the auditory periphery**,” 159th ASA Meeting, Baltimore, Maryland, 4pPP9. Journal of ASA, Vol.127, Issue 3, pp. 1988-1988, April 2010.
- Conference Talk: Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2011) “**Synchrony capture filterbank (SCFB): A novel cochlear signal processing model**,” 161st ASA Meeting, Seattle, Washington, 1aPP3. Journal of ASA, Vol. 129, Issue 4, pp. 2383-2383, April 2011.
- Conference Talk: Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter

Cariani, (2012) “**Synchrony Capture Filterbank (SCFB): An Auditory Periphery Inspired Method for Tracking Sinusoids**,” Proceedings of the ICASSP, Kyoto, Japan, pp-153-156, March 2012.

- Conference Talk: Ramdas Kumaresan, Vijay Kumar Peddinti, and Peter Cariani, (2014) “**Auditory-inspired pitch extraction using a synchrony capture filterbank for speech signals**,” 167th ASA Meeting, Providence, 5aSP3. Journal of ASA, Vol. 135, Issue 4, pp. 2426-2426, May 2014.

An algorithm for tracking the frequency components in a speech signal has many applications. Some of the applications are: 1) designing superior hearing aids, to assist the hearing impaired or people with normal hearing loss with age, 2) more precise signal coding for cochlear implants, 3) in developing voice-to-text or speech recognition algorithms which could be used in “hands-free” communication such as GPS activation while driving, 4) speaker detection/identification, especially useful in the area of security, 5) and to eventually develop a more natural human-computer interface.

The thesis is presented in manuscript format, and organized in the following way:

- Manuscript 1 provides an overview of the auditory motivation. Peripheral auditory system, auditory phenomenon such as phase locking, synchrony capture and its possible advantages are reviewed in this chapter.
- Manuscript 2 presents our initial version of the synchrony capture filterbank (SCFB) algorithm. A brief review of synchrony capture is presented here. In addition a relationship to previous signal processing strategies and similarities to the cochlea and auditory nerve are discussed in detail.

- Manuscript 3 presents a synchrony capture filterbank (SCFB) based pitch extraction algorithm. The brief overview of the improved SCFB algorithm is also presented here.
- Manuscript 4 illustrates the motivation for an improved synchrony capture filterbank (SCFB) and an in-depth loop filter analysis which enables robust frequency tracking.

All the publications presented are in the original form apart from the style. Rest of the preface introduces the peripheral auditory system, and provides an overview of the auditory motivation. This serves purely as an overview, no new conclusions or original material is presented.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
DEDICATION	vi
PREFACE	vii
TABLE OF CONTENTS	xii
LIST OF FIGURES	xv
MANUSCRIPT	
1 Introduction	1
List of References	9
2 Synchrony capture filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech	11
2.1 Abstract	11
2.2 Introduction	12
2.2.1 Significance of synchrony capture	15
2.2.2 Design rationale for the SCFB algorithm	18
2.3 Tone followers and frequency capture	22
2.3.1 A simple tone follower (STF)[22]	22
2.3.2 Dominant tone follower (DTF)	29
2.3.3 A practical implementation of the frequency discriminator loop (FDL)	33
2.4 Synchrony capture filterbank (SCFB)	37

	Page
2.4.1 Bandpass filter triplet parameters	39
2.4.2 Frequency discriminator loop filter design $F(s)$	39
2.5 Simulation results	40
2.5.1 Dyads of synthetic harmonic signals	41
2.5.2 Speech signals	44
2.5.3 Isolated spoken letters	45
2.5.4 Continuous speech	46
2.6 Discussion	53
2.6.1 Relationship to previous signal processing strategies . . .	53
2.6.2 Similarities to response characteristics of the cochlea and auditory nerve	55
2.7 Conclusions	58
2.8 Acknowledgments	59
2.9 Appendix A: Alternate frequency error detectors	60
2.10 Appendix B: Expressions for the frequency discriminator con- stant k_s	61
List of References	62
3 Auditory-inspired pitch extraction using a synchrony cap- ture filterbank and phase alignment	67
3.1 Abstract	67
3.2 Introduction	68
3.3 Two Main ideas: Duals in time and frequency	73
3.4 Frequency tracking by the FDL/PLL block	76
3.5 Phase Alignment Loop (PAL)	77

	Page
3.6 Simulation results	77
3.7 Conclusion	79
List of References	81
4 Synchrony capture filterbank II: Auditory-inspired pitch extraction using synchrony capture filterbank and phase alignment	84
4.1 Abstract	84
4.2 Introduction	85
4.3 Pitch extraction	88
4.4 Design parameters for the frequency discriminator loop (FDL) .	90
4.5 Design parameters for the phase locked loop (PLL)	94
4.6 Simulations	96
4.7 Conclusions	105
List of References	105
 APPENDIX	
A Published ICASSP2011 paper	106
List of References	106
B Published ICASSP2012 paper	111
List of References	111
BIBLIOGRAPHY	116

LIST OF FIGURES

Figure		Page
1	Central auditory pathway [1].	2
2	Peripheral auditory system showing outer, middle and inner ear [1].	3
3	Cross section of cochlea showing basilar membrane (BM) and the organ of Corti, which consists of the haircells, supporting cells and tectorial membrane (TM) [2].	4
4	Electron microscope photographs showing inner and outer hair cells [1].	5
5	The three rows of the outer hair cells are modeled as three tunable bandpass filters shifted in frequency. Left image reference [4].	5
6	A typical channel, showing the broader filter which emulated the BM followed by the three tunable bandpass filters shown in Figure 5.	6
7	Phase locking illustration [10].	8
8	Two views of the representation of vowel-like sounds in the auditory nerve. a) Peristimulus time histograms for cat ANFs arranged by characteristic frequency in response to the five-formant vowel /a/ taken from the synthetic syllable "da." Reprinted from Secker-Walker and Searle (1990)[4]. (b) Distribution of synchronized rates in ANFs in response to a synthetic vowel /a/ with three formants F_1 , F_2 , and F_3 . $F_0 = 100\text{Hz}$. Reprinted from Sachs et al. (2002) [5].	13

- 9 Synchrony capture of adjacent partials for two frequency separations. The two neurograms show all-order interspike interval distributions for individual cat auditory nerve fibers as a function of CF in response to complex tone dyads presented 100 times at 60 dB SPL. Each tone of the pair consisted of equal amplitude harmonics 1-6. New analysis of dataset originally reported in Tramo et al. (2001) [6]. (a) Responses to a tone dyad a musical minor second apart (16:15, $\Delta F_0=6.6\%$). Vertical bars indicate CF regions where one predominant interspike interval pattern predominates. Fiber CFs: 153, 283, 309, 345, 350, 355, 369, 402, 402, 431, 451, 530, 588, 602, 631, 660, 724, and 732 Hz. Out of place interval patterns (single-asterisked histograms) are likely due to small CF measurement errors. (b) Response to a tone dyad a musical fourth apart (4:3, $\Delta F_0=33.3\%$). Three distinct interspike interval patterns associated with individual partials (440, 587, and 880 Hz) are produced in different CF bands, with abrupt transitions between response modes. One fiber shows locking to distortion product $2f_1 - f_2$ near its CF (double-asterisked histogram, $2f_1 - f_2 = 293$ Hz, CF = 283 Hz). Fiber CFs: 153, 283, 346, 350, 355, 369, 402, 402, 431, 451, 530, 588, 602, 631, 660, 662, 724, 732, and 732 Hz. 14
- 10 Synchrony capture filterbank (SCFB). (a) The filterbank architecture consists of K constant-Q gammatone filters whose logarithmically-spaced center frequencies span the desired audible frequency range. Each filterbank channel consists of a frequency discriminator loop (FDL) cascaded with each of the K gammatone filters. The output of each channel, $y_c(t)$, is obtained from its center filter. See sections 2.3 and 2.4 for details. Frequency responses of fixed and tunable filters in the SCFB. Bottom left panel (b) shows the frequency responses of fixed gammatone filters (the black dots indicate that not all filter responses are shown). Bottom right panel (c) shows the frequency responses of the tunable bandpass filter (BPF) triplets that adapt to the incoming signal. One BPF triplet is associated with each fixed filter, such that coarse filtering of the fixed gammatone filters is followed by additional, finer filtering by tunable filters. 19

Figure		Page
11	A generic frequency discriminator loop (FDL). The error signal $e(t)$ is a measure of the frequency difference between the input signal and the VCO. See Figures 12 and 15 for details of specific frequency error detectors.	22
12	Frequency error detector (FED) used in the simple tone follower (STF). Error signal $e(t)$ is computed using the formula $\frac{e_R(t)-e_L(t)}{e_R(t)+e_L(t)}$. The envelopes $e_L(t)$, $e_R(t)$, and $e_C(t)$, are obtained as $I^2 + Q^2$. The I and Q for center filter $H_C(\omega)$, are the outputs of the LPFs shown in (b). $H_L(\omega)$ and $H_R(\omega)$ have the same structure but with oscillator frequencies at $\omega_c - \Delta$ and $\omega_c + \Delta$ respectively. The discriminator transfer characteristics $S(\omega)$ (thick line) and magnitude responses of left and right filters (thin lines) are shown in (c).	24
13	Convergence of a BPF triplet on an input tone at ω_1 . (a) Frequency responses of BPF triplet filters in relation to an input tone. The input tone frequency is $\omega_1 = 2\pi \times 950$ Hz. Initially the L, C, and R filters are centered at $\omega_c - \Delta = 2\pi \times 859$ Hz, $\omega_c = 2\pi \times 901$ Hz and $\omega_c + \Delta = 2\pi \times 943$ Hz, respectively. Since initially $\omega_1 > \omega_c$, the initial envelope output $e_R(t)$ is greater than $e_L(t)$, so the normalized error $e(t)$ is positive. This positive value of $e(t)$ causes the VCO frequency ω_c to increase until ω_c equals ω_1 . (b) Time course of envelopes $e_L(t)$, $e_C(t)$ and $e_R(t)$. Note that the envelopes $e_R(t)$ and $e_L(t)$ become equal after some settling time and that $e_C(t)$ reaches a higher plateau, where $e_L(t)=e_R(t)=0.5e_C(t)$. (c) VCO frequency track for the C filter.	26
14	Linearized model of the frequency discriminator loop.	27
15	Frequency error detector (FED) for the dominant tone follower (DTF). The error signal $e(t)$ is computed using the formula $\log\left(\frac{e_R(t)}{e_L(t)}\right)$	31
16	Behavior of a DTF in response to two nearby tones of different amplitude. (a) Frequency response of BPF triplet filters and the input tones (vertical arrows, dominant tone at $\omega_1 = 2\pi \times 950$ Hz, plus a half-amplitude interfering tone at $\omega_1 = 2\pi \times 1050$ Hz. (b) Track of the VCO frequency for the center filter C. With minor fluctuations, the VCO tracks the stronger 950 Hz tone in-spite of the weaker 1050 Hz interferer.	33

Figure		Page
17	(a) Tunable cos-cos filter, (b) cos-sin filter, (c) Frequency responses $G_1(\omega)$ and $G_2(\omega)$ (without the scale factor j) are shown, (d) Frequency responses of the right and left filters, $H_R(\omega)$ and $H_L(\omega)$, obtained as sum and difference of $G_1(\omega)$ and $G_2(\omega)$ (Figure 18). 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 . The frequency response of $H_C(\omega)$, not shown, is centered around ω_c . All filters are linear phase filters.	35
18	Implementation of the frequency error detector and the frequency discriminator loop. The center filter $H_C(\omega)$ (not shown) is implemented using a cos-cos filter structure with $H(\omega)$ sandwiched between the multipliers as in Figure 12b.	37
19	A typical BPF Triplet centered at 1980 Hz. The broader frequency response corresponds to the gammatone filter centered around 1980Hz.	38
20	Filterbank responses to pairs of harmonic tones. Left. Responses to a note dyad separated by a minor second ($\Delta F_0=6.6\%$, $F_{0s} = 440$ & 469 Hz). Right. Responses to a note dyad separated by a perfect fourth ($\Delta F_0=33.3\%$, $F_{0s} = 440$ & 587 Hz). Top plots (a),(d). Frequency tracks of the VCOs (capturegram). Middle plots (b), (e). Half-wave rectified output waveforms of channel center filters (analogous to a post-stimulus time neurogram). Bottom plots (c), (f). Channel autocorrelations (compare with autocorrelation neurograms of Figure 9).	43
21	(a) Steps involved in the SCFB algorithm. The input speech signal $s(t)$ (after preemphasis) is processed by the 200 gammatone filters and the associated FDLs and the frequency tracks are plotted as capturegrams. The VCO frequency values and the associated envelopes are used to generate the frequency histograms from which dominant frequency tracks are derived. Results for ISOLET vowel /i/. (b) Spectrogram (c) Capturegram (d) Thresholded histogram plot.	49

Figure		Page
22	Results for TIMIT utterance, “Where were you while we were away?” (sx9) for male (left column) and female (right column) speakers. Top plots (a)(d). Spectrograms. Middle plots (b)(e). Capturegrams. Bottom plots (c)(f). Thresholded histogram plots. At low frequencies, all individual harmonics are tracked, whereas above 1000 Hz, only prominent formant harmonics are tracked.	50
23	Results for TIMIT utterance “The oasis was a mirage” (sx280) for male (left column) and female (right column) speakers. Plots as in the previous figure. High frequency frication above 4000 Hz in “oasis” not shown.	51
24	Results for two TIMIT utterances in 10dB SNR. “The oasis was a mirage” (sx280) for a female speaker (left column) and “Where were you while we were away?” (sx9) for a male speaker (right column). Plots as in the previous figure.	52
25	Distortion products. Spectrum of VCO output signal of a channel with center frequency of 1890 Hz in response to two pure tones at frequencies $f_1 = 1950$ Hz and $f_2 = 2050$ Hz with amplitudes $A_1 = 1$ and $A_2 = 0.5$ respectively. Note occurrences of distortion products at frequencies $f_1 \pm n(f_2 - f_1)$. These are generated in frequency discriminator loops when VCOs lock on to dominant tones at f_1 but are also frequency modulated by an error signals consisting of a weak tones at $\Delta f = f_2 - f_1$	56

26	a) Schematic of the Pitch Extraction Algorithm: The gammatone (GT) filters provide some spectral isolation in each channel. The frequency discriminator loop/phase locked loop (FDL/PLL) block achieves synchrony capture, i.e., the voltage controlled oscillator (VCO) in the block locks on to the strongest frequency component. The half wave rectified and low pass filtered (HWR/LPF) output of the channel is processed by the envelope (ENV) and the temporal fine structure (TFS) branches in parallel, which extract the envelope and phase align the dominant tonal signals, respectively. The outputs of all the channels, both the ENV and TFS branches, are summed to produce the Summary Phase Aligned Function (SPAF), the peak locations of which are used to determine f_0 . b) Phase alignment loop (PAL): The half wave rectified tone, $y(t)$, is delayed such that it overlaps symmetrically on the left ($w_L(t)$) and right ($w_R(t)$) windows (depiction shown in figure 28(b)). Analogous to the FDL, when $y(t)$ is centered around a time reference, t_0 , the error $e_t(t)$ goes zero and the loop reaches steady state. At this point, the half wave (HW) rectified tone is in cosine phase with respect to t_0 . See Section 3.5.	72
27	FDL and PLL: Together they track the frequency of the dominant tone in the passband of the associated GT filter.	74
28	a) Frequency measurement using stagger tuned filters: The left $H_L(f)$, right $H_R(f)$ (and center $H_C(f)$) filters are tuned by adjusting the frequency f_c of a voltage controlled oscillator (VCO), which is embedded in a feedback loop (See Figure 27). When $f_c = f_1$, the tone's amplitude at the output of the right and left filters is equal, hence the loop reaches steady state. b) Phase measurement using staggered time windows: The half wave (HW) rectified tone, $y(t)$ (shown in bottom panel), is multiplied separately by the left ($w_L(t)$) and right ($w_R(t)$) windows. $y(t)$ is delayed using a feedback loop until the areas under $y(t)w_L(t)$ and $y(t)w_R(t)$ are equal, and the loop reaches steady state. Then, the delayed HW rectified tone is in cosine phase with respect to a time reference, t_0	75
29	Single tone: (a) shows the frequency trajectories of the 4 VCOs near 800 Hz, (b) and (c) show the HW rectified channel outputs without and with phase alignment, respectively.	78

Figure		Page
30	Unresolved tones: Tracks of VCO frequencies for CFs around 2400 Hz with PLL (a) and without PLL (b, notice the bias.) . .	79
31	Frequency tracks of the Schroeder-phase signal	80
32	SPAF of Schroeder-phase signal	80
33	Synchrony capture filterbank (SCFB). (a) The filterbank architecture consists of K constant-Q gammatone filters whose logarithmically-spaced center frequencies span the desired audible frequency range. Each filterbank channel consists of a frequency discriminator loop (FDL) cascaded with each of the K gammatone filters. The output of each channel, $y_c(t)$, is obtained from its center filter. Frequency responses of fixed and tunable filters in the SCFB. Bottom left panel (b) shows the frequency responses of fixed gammatone filters (the black dots indicate that not all filter responses are shown). Bottom right panel (c) shows the frequency responses of the tunable bandpass filter (BPF) triplets that adapt to the incoming signal. One BPF triplet is associated with each fixed filter, such that coarse filtering of the fixed gammatone filters is followed by additional, finer filtering by tunable filters [2]. More details of this architecture are in [2] or manuscript 2 of this thesis.	86
34	Shows the bias in the voltage controlled oscillator (VCO) tracks for unresolved tones; here $f_1 = 2300$ Hz, $f_2 = 2500$ Hz, however the mean value of the estimates are off about approximately 15 Hz.	87
35	Improved frequency discriminator loop (FDL) block, which shows the addition of a phase locked loop (PLL). A weighted sum of the FDL and PLL errors is used to track the frequency of the dominant tone in the passband of the associated gammatone (GT) filter [1].	87
36	Frequency tracks of voltage controlled oscillator (VCO) using improved frequency discriminator loop (FDL) block. The mean value of the frequency estimates are same as the actual frequencies $f_1 = 2300$ Hz, $f_2 = 2500$ Hz. However it does add some variance to the estimates as compared to Figure 34.	88

Figure		Page
37	Single tone input at 800 Hz: (a) shows the frequency trajectories of the 4 VCOs near 800 Hz, (b) and (c) show the half-wave (HW) rectified channel outputs without and with phase alignment, respectively [1].	89
38	Pitch extraction algorithm using the improved synchrony capture filterbank (SCFB). The first stage of this architecture is the improved FDL (FDL/PLL). These outputs are phase aligned by phase alignment loop (PAL), which are added to produce the summary phase aligned function (SPAF), from which the f_0 can be determined.	90
39	Baseband equivalent circuit of the frequency discriminator loop (FDL).	90
40	Generic feedback circuit.	91
41	Analog equivalent circuit of phase locked loop (PLL).	94
42	Results for TIMIT utterance, “Where were you while we were away?” (sx9) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are capturegrams (frequency tracks superimposed on spectrograms). At low frequencies, all individual harmonics are tracked, whereas above 1000 Hz, only prominent formant harmonics are tracked.	97
43	Results for TIMIT utterance, “The oasis was a mirage” (sx280) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are Capturegrams(frequency tracks superimposed on spectrograms). High frequency frication above 4000 Hz in “oasis” not shown.	98
44	Results for TIMIT utterance in 10dB noise, “Where were you while we were away?” (sx9) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are capturegrams (frequency tracks superimposed on spectrograms).	99

Figure		Page
45	Results for TIMIT utterance in 10dB noise, “The oasis was a mirage” (sx280) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are Capturegrams(frequency tracks superimposed on spectrograms).	100
46	(a) Frequency tracks (b) SPAF for a two tone segment, first segment has a frequency 440 Hz from 0 to 0.425 seconds and the second segment has 587 Hz from 0.425 to 1 second respectively.	102
47	(a) Frequency tracks (b) SPAF for a two tone segment in 10 dB noise, first segment has a frequency 440 Hz from 0 to 0.425 seconds and the second segment has 587 Hz from 0.425 to 1 second respectively.	103
48	(a) Frequency tracks (b) SPAF for a speech utterance (male speaker) from TIMIT database “The oasis was a mirage” (sx280.wav).	104

Introduction

A brief overview of the peripheral auditory system and the motivating phenomenon for the synchrony capture filterbank algorithm is presented here.

Hearing is one of the five major senses possessed by most animals. It plays a vital role in day-to-day activities that are essential for their survival. The central auditory nervous system which is responsible for hearing functions is a complex multi-level sound processing network. Humans without hearing loss can detect frequencies from approximately 20 Hz to 20 kHz. A high level structure of the central auditory pathway from the cochlea through the different brain centers is shown in Figure 1.

The mammalian hearing organ in general, and the human auditory periphery in particular, is capable of analyzing time varying sounds and representing them in a spatio-temporal (in terms of space and time) array of neural discharges/spikes. Trains of these spikes are carried by auditory nerve fibers (ANFs) also referred to as cochlear nerve fibers and are further processed in many higher brain centers such as the cochlear nucleus, inferior colliculus, medial geniculate and the auditory cortex (Figure 1) to interpret the sound (extract characteristics (information) such as the pitch, loudness) and also assists in locating the sound source.

Let us look at the peripheral auditory system shown in Figure 2. It can be subdivided into the external ear, middle ear, and the inner ear.

The external ear, also known as the outer ear, is the most familiar part because it is the only visible part of the auditory system. The outer ear is connected to the middle ear, which is composed of the auditory canal (also called ear canal or external auditory canal), eardrum (also called the tympanic membrane), and a

Figure 1. Central auditory pathway [1].

neighboring cavity with the three smallest bones in the human body, the malleus, the incus, and the stapes. Together, these bones connect the middle ear to the inner ear [1, 2].

The primary function of the outer and the middle ear is to collect and transmit sound waves to the inner ear without losing energy and to protect the ear to a certain extent. The outer ear guards the ear from external environmental conditions and physical trauma. The ear canal contains tiny hairs and special glands that produce earwax to filter the dust particles. Other functions include maintaining proper temperature, humidity, impedance matching and pressure equalization. The inner ear is quite complex and houses the vestibulo-cochlear organs, which are responsible for hearing and balance. Sound is sensed by the cochlea and the

Figure 2. Peripheral auditory system showing outer, middle and inner ear [1].

body movements are detected by the vestibular system, which are converted into patterns of neural impulses. The brain later interprets these impulses for sound perception and to maintain balance. The cochlea is a bone-encased, fluid-filled, spiral-shaped, mechanically-tuned linear structure. The name “cochlea” derives from Greek, which means “snail-shape” or “spiral” [3].

One of the most important modules responsible for converting these sounds waves to impulses is the organ of corti. Figure 3 shows a cross section of the organ of corti. It consists of haircells, supporting cells and the tectorial membrane (TM) and is positioned on the basilar membrane (BM). There are approximately 3,500 inner hair cells (IHCs) in the human auditory systems and they form a single row. On the other hand, there are roughly 12,000 outer hair cells (OHCs) in humans, and they form three rows, as shown in Figure 4 [1].

The function of the basilar membrane is to vibrate in response to the incoming sound and decompose the signal into multiple frequency components. High fre-

Figure 3. Cross section of cochlea showing basilar membrane (BM) and the organ of Corti, which consists of the haircells, supporting cells and tectorial membrane (TM) [2].

quency sound vibrations mechanically resonate with the basilar membrane most strongly at its entrance (the base), whereas low frequencies resonate at its tip (the apex). These resonating frequencies at a given place it resonates are called the characteristic frequencies (CF). The physical spacing of the resonant frequencies along the length of basilar membrane is logarithmic (not linear) with frequency. The tectorial membrane follows the basilar membrane oscillations and consequently the cilia (hair) of hair cells bend back and forth, converting the vibrational sound signals into neural signals or spikes in the auditory nerve fibers after several inter-

Figure 4. Electron microscope photographs showing inner and outer hair cells [1].

mediate steps. These spike trains start in the auditory nerves and are transferred to the brain. Each human auditory nerve consists of 30,000 individual auditory nerve fibers that convey information about sounds from the cochlea to the central auditory system through these spike patterns.

Figure 5. The three rows of the outer hair cells are modeled as three tunable bandpass filters shifted in frequency. Left image reference [4].

Traditionally, following Ohm, Helmholtz, and von Békésy, the basilar membrane is viewed as an array of passive band-pass filters with logarithmically spaced center frequencies ranging from roughly 50-20,000 Hz. The OHCs are thought to support an “active process” which provides additional filtering and amplification of low intensity sounds [5].

It was the spatial arrangement of outer hair cells (OHCs) observed in mammalian cochleae [3] that originally inspired this particular design. Figure 5 shows the idea behind using three tunable bandpass filters slightly shifted in frequencies, derived from the three outer hair cell (OHC) rows. This orderly structure of cochlear hair cell arrays suggested to us that local lateral comparisons of signals produced by neighboring filters could be used for tracking individual stimulus frequency components. Design of this filterbank was both inspired by mono-pulse radar and cochlea anatomy. The basilar membrane can be viewed as a fixed broad filter, while the action of the three outer hair cells can be emulated as three tunable narrow BPF filters as illustrated in Figure 6.

Figure 6. A typical channel, showing the broader filter which emulated the BM followed by the three tunable bandpass filters shown in Figure 5.

Auditory filterbank models have a long history. One of the earlier filter archi-

ture traces back to Flanagan, who utilized a filter with a real pole and a pair of complex conjugate poles. Later, Lyon (with C. Mead) designed a cascade of filters as an electronic cochlea [6]. Other popular models include Meddis’ inner hair cell model, and Roy Patterson’s auditory image model (AIM). Additional auditory filter models include rounded exponential (roex) filters, gammatone filters and filter cascades [7, 8]. Our synchrony capture architecture falls under the gammtone filter bank category.

We proposed a novel signal analysis algorithm, which processes sounds the way the early stages of the auditory system does. In the past five years, we developed algorithms that emulate a phenomenon known as ”synchrony capture” observed in the auditory nerve.

The main basis for synchrony capture is related to phase locking and the volley principle. The neurons fire in response only to the positive phase (half-cycle) of the cochlear vibration patterns that are driving their associated inner hair cells. This can be seen in period histogram of Figure 7b, such that spikes are produced almost exclusively in phase. The spikes do not necessarily always respond in every positive phase of the waveform, as shown in Figure 7a. However when they respond the process can be described roughly as an Poisson process with dead time, whose instantaneous rate is proportional to the amplitude of the half-wave rectified signal.

When it became understood that auditory nerve fibers fire at much lower rates than the maximual frequencies observable in the cochlear potentials, Warren Wever proposed that the synchronized spikes of many auditory nerve fibers taken together could represent the whole half-wave rectified signal up to the frequency limits of phase-locking. This is known as the “volley principle” [9].

Synchrony capture enables the robust encoding of sound intensity, suppression of noise and tracking of individual signal components. Thus, it ought to serve as

Figure 7. Phase locking illustration [10].

a novel and critical underpinning of speech processing algorithms. Our proposed algorithm, emulates synchrony capture using frequency discriminator loop (FDL) which is an array of stagger-tuned, three bandpass filters that emulate the outer hair cell (OHC) array in the organ of Corti. The filterbank automatically identifies the locations of the dominant signal components and suppresses the nearby weaker components. This two-stage approach helps mask signal components that are weak and concentrate on the strong signal components. The filterbank potentially has potential applications in cochlear implants, hearing aids, front ends of speech recognizers, audio signal enhancement.

The filterbank can be extended for pitch computation. Even though it avoids the computation of auto-correlation, it still requires the calculation of the period between the peaks. Currently, it is determined by peak picking, which needs to be addressed. The ultimate goal would be to show the frequency tracks and pitch as a real-time signal.

List of References

- [1] Yost, *Fundamentals of hearing*, 4th ed. London, U.K.: Academic Press, 2000.
- [2] D. Emanuel and T. Letowski, *Hearing Science*. Wolters Kluwer Health/Lippincott Williams and Wilkins, 2009. [Online]. Available: <http://books.google.com/books?id=gj1bSAAACAAJ>
- [3] P. Dallos, “Overview: Cochlear Neurobiology,” in *The Cochlea*, ser. Springer Handbook of Auditory Research, P. Dallos, A. N. Popper, and R. R. Fay, Ed. New York, NY: Springer-Verlag, 1996, ch. 1, pp. 1–43.
- [4] S.A. Counter, E. Borg and L. Lofqvist, “Acoustic trauma in extracranial magnetic brain stimulation,” *Electroencephalography and Clinical Neurophysiology*, vol. 78, Issue 3, pp. 173–184, Mar. 1991.
- [5] G. A. Manley, “Evidence for an active process and a cochlear amplifier in nonmammals,” *J Neurophysiol*, vol. 86, no. 2, pp. 541–9, Aug 2001, manley, G A Review United States Journal of neurophysiology.
- [6] R.F.Lyon and C.Mead, “An electronic cochlea,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. assp-36, pp. 1119–1134, 1988.
- [7] L. C. Araújo, T. N. Magalhaes, D. P. Souza, H. C. Yehia, and M. A. Loureiro, “A brief history of auditory models,” *10 Simpósio Brasileiro de Computação Musical*, vol. 1, 2005.
- [8] R. F. Lyon, A. G. Katsiamis, and E. M. Drakakis, “History and future of auditory filter models,” in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on.* IEEE, 2010, pp. 3809–3812.
- [9] E. G. Wever, *Theory of Hearing*. New York: Wiley, 1949.
- [10] E. F. Evans, “Functional anatomy of auditory system,” in *The Senses*, ser. Volume 3 of Cambridge Studies in Philosophy, H. Barlow and J. Mollon, Eds. Cambridge University Press, 1982.

MANUSCRIPT 2

Published in The Journal of Acoustical Society of America, June 2013

**Synchrony capture filterbank: Auditory-inspired signal processing for
tracking individual frequency components in speech**

Ramdas Kumaresan¹, Vijay Kumar Peddinti¹, Peter Cariani²

¹ Department of Electrical, Computer and Biomedical Engineering,
University of Rhode Island, Kingston RI 02881

² Department of Otology & Laryngology, Harvard Medical School,
Boston, MA 02114

Corresponding Author: Ramdas Kumaresan

Department of Electrical, Computer-
and Biomedical Engineering
University of Rhode Island
Kingston, RI 02881

Phone: 401-874-4798

Email: kumar@ele.uri.edu

Synchrony capture filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech

2.1 Abstract

A processing scheme for speech signals is proposed that emulates synchrony capture in the auditory nerve. The role of stimulus-locked spike timing is important for representation of stimulus periodicity, low frequency spectrum, and spatial location. In synchrony capture, dominant single frequency components in each frequency region impress their time structures on temporal firing patterns of auditory nerve fibers (ANFs) with nearby characteristic frequencies (CFs). At low frequencies, for voiced sounds, synchrony capture divides the nerve into discrete CF territories associated with individual harmonics. An adaptive, synchrony capture filterbank (SCFB) consisting of a fixed array of traditional, passive linear (gammatone) filters cascaded with a bank of adaptively tunable, bandpass filter triplets is proposed. Differences in triplet output envelopes steer triplet center frequencies via voltage controlled oscillators (VCOs). The SCFB exhibits some cochlea-like responses, such as two-tone suppression and distortion products, and possesses many desirable properties for processing speech, music, and natural sounds. Strong signal components dominate relatively greater numbers of filter channels, thereby yielding robust encodings of relative component intensities. The VCOs precisely lock onto harmonics most important for formant tracking, pitch perception, and sound separation.

©2013 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4802653>]

PACS number(s): 43.72 Ar, 43.64 Bt, 43.64 Sj [MAH]

2.2 Introduction

For the past three decades there has been significant interest in developing computational signal processing models based on the physiology of the cochlea and auditory nerve (AN) [1]. The hope has been that artificial systems can be designed and built using signal processing strategies gleaned from nature that can equal or exceed human auditory performance. Our work in this area is motivated by neurophysiological observations of the synchrony capture phenomenon in the auditory nerve that was originally reported by Sachs et al. [2] and Delgutte et al. [3]. This paper proposes such a biologically-inspired signal processing strategy for processing speech and audio signals.

If one systematically examines the temporal representation of low harmonics of complex sounds in the auditory nerve, synchrony capture is a striking feature. Synchrony capture means that the dominant component in a given frequency band preferentially drives auditory nerve fibers innervating the entire corresponding frequency region of the cochlea [3]. Here, virtually all fibers innervating this cochlear place region, i.e. those with CFs in the vicinity of the frequency of the dominant component, synchronize exclusively to the dominant component, in spite of the presence of other nearby weaker components that may be closer to their CFs. 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. Because of the asymmetric nature of cochlear tuning, this dominant component mostly drives fibers whose CFs lie above it in frequency. Figures 8 and 9 provide examples of this phenomenon in slightly different forms. Figure 8a shows peristimulus time histograms (PSTHs) for a five-formant synthetic vowel sound. Sharp boundaries characteristic of synchrony capture are seen between the different CF regions driven by differ-

ent dominant, formant-region harmonics of the multi-formant vowel. Note that in Figure 8a other non-dominant harmonics in the vowel formant regions are not explicitly represented.

Figure 8. Two views of the representation of vowel-like sounds in the auditory nerve. a) Peristimulus time histograms for cat ANFs arranged by characteristic frequency in response to the five-formant vowel /a/ taken from the synthetic syllable "da." Reprinted from Secker-Walker and Searle (1990)[4]. (b) Distribution of synchronized rates in ANFs in response to a synthetic vowel /a/ with three formants F_1 , F_2 , and F_3 . $F_0 = 100\text{Hz}$. Reprinted from Sachs et al. (2002) [5].

Figure 8b summarizes temporal firing patterns observed in the cat auditory nerve in response to a three-formant synthetic vowel [5]. Relative synchronized rates of fibers to different component frequencies are shown as a function of fiber characteristic frequency (CF) or best frequency (BF). Sizes of squares indicate synchronized rates (larger squares = higher rates). The diagonal gray band shows regions where temporal firing periodicities match fiber BFs, and the dark horizontal swaths indicate capture of fibers over a range of fiber best frequencies by individual stimulus components. The most prominent swaths are the synchrony capture regions for the dominant harmonics associated with each of the three formants (enclosed boxes). In addition to capture by dominant harmonics in formant

regions, low-CF fibers show synchrony to less-intense, non-formant, low harmonics ($n=1-3$) when frequencies of those harmonics happen to be near their respective CFs (dark boxes within the gray diagonal band).

Figure 9. Synchrony capture of adjacent partials for two frequency separations. The two neurograms show all-order interspike interval distributions for individual cat auditory nerve fibers as a function of CF in response to complex tone dyads presented 100 times at 60 dB SPL. Each tone of the pair consisted of equal amplitude harmonics 1-6. New analysis of dataset originally reported in Tramo et al. (2001) [6]. (a) Responses to a tone dyad a musical minor second apart (16:15, $\Delta F_0=6.6\%$). Vertical bars indicate CF regions where one predominant interspike interval pattern predominates. Fiber CFs: 153, 283, 309, 345, 350, 355, 369, 402, 402, 431, 451, 530, 588, 602, 631, 660, 724, and 732 Hz. Out of place interval patterns (single-asterisked histograms) are likely due to small CF measurement errors. (b) Response to a tone dyad a musical fourth apart (4:3, $\Delta F_0=33.3\%$). Three distinct interspike interval patterns associated with individual partials (440, 587, and 880 Hz) are produced in different CF bands, with abrupt transitions between response modes. One fiber shows locking to distortion product $2f_1 - f_2$ near its CF (double-asterisked histogram, $2f_1 - f_2 = 293$ Hz, CF = 283 Hz). Fiber CFs: 153, 283, 346, 350, 355, 369, 402, 402, 431, 451, 530, 588, 602, 631, 660, 662, 724, 732, and 732 Hz.

Synchrony capture is most directly apparent when distributions of all-order in-

terspike intervals (spike autocorrelation histograms) produced by individual fibers are plotted as a function of fiber CF (cochlear place)[7]. Figure 9 shows fiber interspike interval patterns in response to two concurrent complex harmonic tones ($n=1-6$). For a stimulus in which pairs of harmonics are close together (Figure 9a, $\Delta F_0 = 6.6\%$ of F_0), all of the fibers in the region synchronize to the composite, modulated waveform. In this case, the temporal firing patterns in the whole CF region follow the beating of the adjacent partials, producing low-frequency fluctuations in firing rate that are associated with perceived roughness[6]. Here, when the adjacent partials are sufficiently close together there are no separate temporal, interspike interval representations of individual harmonics themselves. On the other hand, for a tone pair for which the lower harmonics are relatively well separated in frequency (Figure 9b, $\Delta F_0 = 33.3\%$ of F_0), different CF regions are captured by one or another partial. Thus each harmonic component drives a discrete region of the cochlea in which its temporal pattern dominates, with almost no zones of beating (right panel, there are different CF zones with different interval peak patterns). The result is that each individual partial drives its own swath of auditory nerve fibers that produce corresponding interspike interval patterns.

The foregoing examples indicate that auditory nerve fibers synchronize preferentially to dominant components in the signal. In signal processing terms, the peripheral auditory system appears to treat these dominant components as “carrier” frequencies. The effects of the weaker surrounding components (other harmonics) then manifest themselves as modulations on these carriers (as can be seen in Figure 8a).

2.2.1 Significance of synchrony capture

Synchrony capture may have implications for neural representations of periodicity and spectrum, as well as for F_0 -based sound separation and grouping.

Synchrony capture in the auditory nerve permits representation of relative intensity that is level-invariant, and thus is useful for representing the normalized power spectrum in a robust manner. The numbers of fibers locking onto particular frequency components give indications of the relative intensities of the corresponding components. This is a robust means of encoding their relative magnitudes using neural elements with limited dynamic ranges. The proposed SCFB algorithm [8] attempts to emulate this behavior using adaptive filters to create a competition for channels amongst frequency components that not only accurately reflects their relative magnitudes, but is also invariant with respect to absolute signal amplitude.

This signal processing strategy for encoding relative intensities has relevance for auditory nerve representations. Global temporal representations of lower-frequency sounds in the auditory nerve, called population-interval distributions or summary autocorrelations, implicitly utilize such principles to represent pitch and timbre (e.g. vowel formant structure) [7, 9, 10, 11]. The most direct signal processing analogues of these global temporal auditory nerve models are the ensemble interval histograms (EIHs)[12]. Essentially, dominant frequency components below 5 kHz that are present at any given instant partition the cochlear CF territory into swaths of auditory nerve fibers (ANFs) that have similar temporal discharge patterns (and hence similar interval distributions). In the context of global population-interval representations that sum together interspike intervals across the entire auditory nerve, relative intensities of partials are conveyed through relative numbers of all-order interspike intervals associated with their respective locally-dominant components rather than numbers of CF channels recruited. Whether through relative numbers of pooled intervals or of similarly-responding channels, this parcellation of the cochlea into competing synchronization zones efficiently utilizes the entire auditory nerve for signal representation.

Synchrony capture could also potentially be utilized by place-based brain-stem auditory representations that analyze excitation boundaries by using local across-CF comparisons of temporal firing patterns [13]. Here the abrupt temporal pattern discontinuities associated with synchrony capture increase contrast and the precision of boundary estimations in such coding schemes.

Further, synchrony capture may facilitate F_0 -pitch formation and sound separation by enhancing temporal representations of individual, resolved harmonics at the expense of those produced by interactions of multiple, unresolved harmonics. Synchrony capture has the effect of minimizing periodicities related to beatings of adjacent harmonics, as can be seen in the lack of composite interspike interval patterns when the harmonics are well separated (Figure 9b). The temporal auditory nerve representation of a harmonic complex with low, well-separated harmonics thus resembles a series of interspike interval patterns each of which resembles that of a pure tone of corresponding frequency.

The enhancement of the representation of individual harmonics in turn has implications for F_0 -based sound separation. Most acoustic signals in everyday life are mixtures of sounds from multiple sources. In order to separate multiple concurrent sounds, human listeners mainly rely on differences in onset times and fundamental frequencies F_0 s. Results of psychophysical experiments suggest that separation of multiple auditory objects with different fundamentals, such as those produced by multiple voices or musical instruments, crucially depends on the presence of perceptually-resolved harmonics [14]. These resolved harmonics dominate in pitch perception and have high pitch salience [15].

In terms of interspike interval representations of individual partials (as seen in Figure 9), the effect of synchrony capture is to separate the interspike interval patterns of adjacent partials if they are separated by more than some threshold ratio,

or to fuse them together if they are not. It is therefore not unreasonable to hypothesize that the synchrony capture process might play a role in whether adjacent partials are fused together or separated perceptually. For frequencies for which there is significant phase-locking, synchrony capture behavior thus qualitatively parallels tonal separations and fusions that are associated with harmonic resolution and critical bands. These parallels notwithstanding, the size of psychophysically-measured critical bandwidths in cats, roughly twice those of humans, cast some doubt on a simple, direct correspondence[16].

The mechanism in the auditory pathway whereby the harmonically-related components of each of two concurrent harmonic complexes fuse together to produce two F_0 -pitches at their respective fundamentals is not yet understood. The two F_0 -pitches can be heard out, even if the harmonics of the two complexes are interleaved, provided that the unrelated, adjacent harmonics are sufficiently separated in frequency. In this context, synchrony capture minimizes temporal patterns associated with interactions between adjacent, harmonically-unrelated partials, thus eliminating interaction products that might otherwise degrade the representations of the individual harmonics and hinder their grouping and separation on the basis of shared interspike intervals.

For the above reasons, it seems reasonable to emulate synchrony capture in a signal processing algorithm.

2.2.2 Design rationale for the SCFB algorithm

A schematic of the proposed SCFB algorithm is shown in Figure 10a. It consists of a bank of K fixed, relatively broad filters in cascade with tunable, narrower filters that produce the synchrony capture behavior. This nesting of broad and narrow filters is not unlike coarse and fine gradations in a vernier scale. Tuning of the adaptive filters is carried out via frequency discriminator loops (FDLs) on time

scales of milliseconds to tens of milliseconds, making real-time frequency tracking possible.

Figure 10. Synchrony capture filterbank (SCFB). (a) The filterbank architecture consists of K constant- Q gammatone filters whose logarithmically-spaced center frequencies span the desired audible frequency range. Each filterbank channel consists of a frequency discriminator loop (FDL) cascaded with each of the K gammatone filters. The output of each channel, $y_c(t)$, is obtained from its center filter. See sections 2.3 and 2.4 for details. Frequency responses of fixed and tunable filters in the SCFB. Bottom left panel (b) shows the frequency responses of fixed gammatone filters (the black dots indicate that not all filter responses are shown). Bottom right panel (c) shows the frequency responses of the tunable bandpass filter (BPF) triplets that adapt to the incoming signal. One BPF triplet is associated with each fixed filter, such that coarse filtering of the fixed gammatone filters is followed by additional, finer filtering by tunable filters.

To a telecommunications engineer, the biological phenomenon of synchrony capture appears similar to the well known “frequency capture” behavior of traditional frequency modulation (FM) receivers such as FM discriminators and phase lock loops. Frequency capture [17] occurs when an FM receiver locks on to a strong

FM signal even in the presence of other interfering, relatively weaker FM signals. One such FM receiver circuit is a frequency discriminator [18](p.206) (with a limiter in front), which uses stagger-tuned bandpass filters whose output envelopes are differenced to obtain the demodulated baseband signal. Such circuits are known to exhibit frequency capture. The signal processing architecture proposed here was designed with both these circuits and possible cochlear analogues in mind.

Although the design of the SCFB was partially inspired by cochlear structure, its explicit goal is not to model cochlear biophysics but to emulate synchrony capture in the auditory nerve for purposes of artificial signal processing. However, some mention of broad parallels between the two is nevertheless useful to understanding the SCFB’s basic design.

In the SCFB architecture, the fixed gammatone filterbank with relatively coarse bandpass tunings ($Q = 4$) emulates the behavior of the passive basilar membrane whose stiffness decreases monotonically from base to apex. The bandwidths of the gammatone filters were chosen to approximate cochlear impulse responses and tuning characteristics observed for input signals at high sound pressure levels that are thought mainly to be consequences of passive mechanical filtering[19]. In the SCFB architecture, finer frequency tuning is achieved using a second layer of narrower bandpass filters (BPFs, $Q=8$) that emulate the filtering functions of outer hair cells (OHCs). In the cochlea, while inner hair cells (IHCs) are thought to be relatively passive mechano-electrical transducers, outer hair cells also have active electromechanical processes that permit them to change length under the influence of their transduction currents, thereby amplifying local mechanical vibrations [20].

The proposed adaptive bandpass filter (BPF) triplets that form the heart of the frequency discriminator loop (FDL) consist of three relatively narrowly tuned filters with slightly offset center frequencies that are in cascade with each fixed

filter of the passive gammatone filterbank. This arrangement contrasts with the situation in the cochlea, where OHCs with their active processes and narrower tunings are in bidirectional interaction with the more broadly tuned motions of the basilar membrane [19]. The BPF triplets are locally adaptive and are tuned based on differences in amplitudes of signals output by the filters in the triplet. Although broadly similar designs were available in the adaptive filtering literature[21, 22], independent of auditory modeling, it was the spatial arrangement of outer hair cells (OHCs) observed in mammalian cochleae [23] that inspired this particular triplet design. The lateral amplitude differencing process in each BPF triplet amounts to taking the spatial derivative of the local amplitude spectrum at that particular cochlear location. Such lateral differencing processes could conceivably be carried out via lateral interactions in intracochlear and olivocochlear neural networks [24](*p.15, Fig.1.13(A)*), [25], [26](*p.289, Fig.11*).

The tuned, oscillatory motility of outer hair cells inspired use of a voltage-controlled oscillator (VCO) to tune the filter triplets. Feedback control of triplet tuning could also be potentially implemented via other signal processing mechanisms. The action of hair cell stereocilia that open ion channels preferentially in one direction suggests half-wave rectification of the signal, an operation similar to envelope detection that is already commonly used in auditory modeling. The non-linear response characteristics of hair cells inspired the logarithmic compression of the envelope (see section 2.3.2) that is used by the frequency discriminator loop to capture dominant signals and suppress weaker ones. All of these design features stem from the general idea that many aspects of cochlear function and auditory nerve behavior can be emulated by frequency tracking circuits.

2.3 Tone followers and frequency capture

Frequency discriminator loops (FDLs) have been used for synchronizing transmitter and receiver oscillators in digital and analog communication systems for decades [27, 28]. Typically, in a communication receiver, an FDL brings the receiver oscillator frequency close to the transmitter frequency, i.e., within the lock-in range of a phase lock loop, such that it can lock the two oscillators[29]. The structure of the frequency tracking algorithms used here, called tone followers, are similar to the FDLs used in communication systems. The block diagram of a generic FDL is shown in Figure 11. It consists of a frequency error detector (FED), a loop filter and a voltage controlled oscillator (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 the frequency of the VCO output, 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 system function $F(s)$ of the the loop filter determines its dynamics and has the form $k_p + k_i/s$ where k_p and k_i are the proportional and integral gain factors[30], respectively (more details below in Section 2.3.1).

Figure 11. A generic frequency discriminator loop (FDL). The error signal $e(t)$ is a measure of the frequency difference between the input signal and the VCO. See Figures 12 and 15 for details of specific frequency error detectors.

2.3.1 A simple tone follower (STF)[22]

The frequency discriminator loop (FDL) (Figure 11) tracks the frequency of an input tone by using a frequency error detector (FED) that steers the center

frequencies of the VCOs of the triplet adaptive filters (Figure 12). Another type of FED is described in Appendix A. In principle, the FED consists of three identically shaped tunable band pass filters (BPFs), $H_R(\omega)$, $H_C(\omega)$ and $H_L(\omega)$, initially centered around frequencies $\omega_c + \Delta$, ω_c and $\omega_c - \Delta$, respectively. The subscripts R, C and L stand for the right, center and left filters, respectively. As ω_c , the frequency of the VCO (in Figure 11) is changed, the center frequencies of the BPFs' also change accordingly, such that these filters' response functions slide along the frequency axis. The spacing between triplet filters (Δ) is fixed. Only the left and right filters are used in calculating the error signal $e(t)$. The envelope detectors compute the (squared) envelope of the BPFs' outputs. When a tone, $A_1 \cos(\omega_1 t + \theta_1)$ is presented to the FED, the average values of the (squared) envelopes for right and the left filters are $e_R(t) = |A_1 H_R(\omega_1)|^2$ and $e_L(t) = |A_1 H_L(\omega_1)|^2$, respectively. (If the input tone's frequency changes with time then e_R and e_L are also functions of time t .) Then the error signal $e(t)$ is computed as the ratio of the difference of the envelopes ($e_R(t) - e_L(t)$) to their sum ($e_R(t) + e_L(t)$).

Note that the ratio eliminates the amplitude of the input signal A_1 from $e(t)$, and now $e(t)$ is just related to the frequency error $\omega_c - \omega_1$. Instead of computing the ratio, an AGC circuit at the input could have been used to normalize the amplitude. The principle is to move the frequency responses of the BPFs $H_R(\omega)$ and $H_L(\omega)$ (and $H_C(\omega)$) in tandem, under the control of the VCO frequency ω_c , such that when the error $e(t) = 0$, ω_c equals ω_1 . So, the VCO tracks the input frequency.

The frequency discriminator function $S(\omega) = \frac{|H_R(\omega)|^2 - |H_L(\omega)|^2}{|H_R(\omega)|^2 + |H_L(\omega)|^2}$ (also called the "S-curve" [29]), is shown in Figure 12c. When a tone $A_1 \cos(\omega_1 t + \theta_1)$ is applied as the input, then $e(t) = S(\omega_1)$. In the interval $\omega_c - \Delta < \omega < \omega_c + \Delta$ the error voltage $e(t)$ is approximately linear, so $e(t) \approx k_s(\omega_c - \omega_1)$. k_s is called the

Figure 12. Frequency error detector (FED) used in the simple tone follower (STF). Error signal $e(t)$ is computed using the formula $\frac{e_R(t) - e_L(t)}{e_R(t) + e_L(t)}$. The envelopes $e_L(t)$, $e_R(t)$, and $e_C(t)$, are obtained as $I^2 + Q^2$. The I and Q for center filter $H_C(\omega)$, are the outputs of the LPFs shown in (b). $H_L(\omega)$ and $H_R(\omega)$ have the same structure but with oscillator frequencies at $\omega_c - \Delta$ and $\omega_c + \Delta$ respectively. The discriminator transfer characteristics $S(\omega)$ (thick line) and magnitude responses of left and right filters (thin lines) are shown in (c).

frequency discriminator constant [29].

The tunable BPFs are built using the filter structure shown in Figure 12b (called “cos-cos” structure), which shows how $H_C(\omega)$ (centered at ω_c) is realized using two lowpass filters (LPFs). Identical LPFs with frequency response $H(\omega)$ are sandwiched between two multipliers in both the lower and upper branches of the circuit. Both the multipliers in the upper branch are supplied with $\cos \omega_c t$ (hence

the name cos-cos structure) and the lower branch are supplied with a $\sin \omega_c t$ from the same VCO with frequency ω_c . It can be easily shown that,

$$H_C(\omega) = H(\omega + \omega_c) + H(\omega - \omega_c). \quad (1)$$

Similarly, the BPF $H_L(\omega)$ (or $H_R(\omega)$) is implemented as a cos-cos structure with the same LPF filters but with the VCO frequency at $\omega_c - \Delta$ (or $\omega_c + \Delta$). Together the three filters shown inside the FED box in Figure 12a is called a *BPF triplet*. The frequency spacing between these filters, Δ , is kept fixed. Only the left and right filters are used in calculating the error signal $e(t)$.

The center filter envelope is used to declare a “track” condition, i.e. that the filter has converged on a tonal input. When this convergence occurs at the input tone frequency ω_1 , then the envelope of the center filter output $e_C(t)$ will satisfy the following condition,

$$e_L(t) = e_R(t) = \mu e_C(t) \quad (2)$$

for some constant μ . If the filter shapes are chosen such that $|H_R(\omega_c)| = |H_L(\omega_c)| = 0.707|H_C(\omega_c)|$ (i.e., 3-dB points of the right and left filter coincide with the center frequency of the center filter), then $\mu = 0.5$. If the above condition is satisfied, then the input is a tone whose frequency coincides with the VCO frequency ω_c , and a “track” condition is declared. Such channel outputs can be used to compute the pitch frequency of a complex tone. This FED structure requires three VCOs operating at $\omega_c - \Delta$, ω_c and $\omega_c + \Delta$ to realize the $H_L(\omega)$, $H_C(\omega)$, and $H_R(\omega)$ respectively.

An approximate linear equivalent circuit of the frequency discriminator loop can provide some insight into the behavior of the tone follower (Figure 14). Here the input tone and the oscillator output are replaced by their frequency values ω_1 and ω_c , respectively. Recall that the frequency error detector (FED) outputs a voltage level proportional to the frequency difference $\omega_1 - \omega_c$. Therefore, the FED

Figure 13. Convergence of a BPF triplet on an input tone at ω_1 . (a) Frequency responses of BPF triplet filters in relation to an input tone. The input tone frequency is $\omega_1 = 2\pi \times 950$ Hz. Initially the L, C, and R filters are centered at $\omega_c - \Delta = 2\pi \times 859$ Hz, $\omega_c = 2\pi \times 901$ Hz and $\omega_c + \Delta = 2\pi \times 943$ Hz, respectively. Since initially $\omega_1 > \omega_c$, the initial envelope output $e_R(t)$ is greater than $e_L(t)$, so the normalized error $e(t)$ is positive. This positive value of $e(t)$ causes the VCO frequency ω_c to increase until ω_c equals ω_1 . (b) Time course of envelopes $e_L(t)$, $e_C(t)$ and $e_R(t)$. Note that the envelopes $e_R(t)$ and $e_L(t)$ become equal after some settling time and that $e_C(t)$ reaches a higher plateau, where $e_L(t) = e_R(t) = 0.5e_C(t)$. (c) VCO frequency track for the C filter.

in Figure 12a is modeled by a proportionality constant k_s . Assuming that we operate the discriminator loop in the region $\omega_c - \Delta < \omega < \omega_c + \Delta$, this constant k_s is the gain factor representing the slope of the S-curve shown in Figure 12c. Assuming that the sandwiched LPF in Figure 12b has a system function $1/(s + \alpha)$, where α

represents its 3-dB bandwidth, it can be shown that the frequency error discriminator constant k_s is equal to $2\Delta/(\Delta^2 + \alpha^2)$ (see Appendix B). In addition, note that the calculation of the envelopes needed to estimate the frequency difference entails a group delay τ_g . This time delay is represented by its Laplace transform $e^{-s\tau_g}$ in Figure 14. At low frequencies the BPF filters are narrower, and hence τ_g is relatively large. At high frequencies $\tau_g \approx 0$. In Figure 14, $e^{-s\tau_g}$ is approximated (using Padé approximation [31]) by a ratio of first order s -polynomials,

$$e^{-s\tau_g} \approx \frac{1 - \gamma s}{1 + \gamma s} \quad (3)$$

where $\gamma = \tau_g/2$. The controller is a loop filter whose transfer function is $F(s) = k_p + k_i/s$ where k_p is the proportional constant and k_i is the integral constant ([30], page 254).

Figure 14. Linearized model of the frequency discriminator loop.

Then, the closed loop transfer function $H(s)$ of the linearized model is

$$H(s) = B(s)/A(s) \quad (4)$$

$$= \frac{\frac{1 - \gamma s}{1 + \gamma s} k_s \left(k_p + \frac{k_i}{s} \right)}{1 + \frac{1 - \gamma s}{1 + \gamma s} k_s \left(k_p + \frac{k_i}{s} \right)} \quad (5)$$

After some simplification we find that the denominator polynomial $A(s)$, which

determines the settling time τ_s of the loop, is given by the following expression,

$$A(s) = s^2 + \frac{(1 + k_s k_p - \gamma k_s k_i)}{(\gamma - \gamma k_s k_p)} s + \frac{k_i k_s}{(\gamma - \gamma k_s k_p)} \quad (6)$$

Using Routh's Stability Criterion, the conditions for stability are given by

$$\begin{aligned} (\gamma - \gamma k_s k_p) &> 0 \Rightarrow k_p < \frac{1}{k_s} \\ (1 + k_s k_p - \gamma k_s k_i) &> 0 \Rightarrow \gamma k_i - k_p < \frac{1}{k_s} \\ k_i k_s &> 0 \Rightarrow k_i > 0, (k_s \text{ is positive}) \end{aligned}$$

We need to find k_p and k_i such that the step response has a desirable settling time. This is done using the standard pole positioning method ([30], page 233) based on Bessel polynomials. For a second order system with a normalized settling time of 1 second, the Bessel roots of the closed loop system are at $-4.05 \pm j2.34$. And for a desired settling time of τ_s seconds, the roots are scaled by τ_s , i.e., $(-4.05 \pm j2.34)/\tau_s$. Hence the corresponding Bessel polynomial is $s^2 + (8.11/\tau_s)s + 21.90/\tau_s^2$. By comparing this polynomial with the $A(s)$ in Eq. 6, we can write the following two linear equations in terms of k_p and k_i :

$$\begin{aligned} a_1 k_i + b_1 k_p &= c_1 \\ a_2 k_i + b_2 k_p &= c_2 \end{aligned}$$

where

$$\begin{aligned} a_1 &= \tau_s \gamma k_s & b_1 &= -k_s (\tau_s + 8.11\gamma) & c_1 &= (\tau_s - 8.11\gamma) \\ a_2 &= \tau_s^2 k_s & b_2 &= 21.90\gamma k_s & c_2 &= 21.90\gamma \end{aligned}$$

Solving for k_p and k_i obtains

$$\begin{aligned} k_p &= \frac{1}{k_s} \frac{\beta - 1}{\beta + 1}, \\ k_i &= \frac{1}{k_s} \left(21.90 \frac{\gamma}{\tau_s^2} \right) \frac{2}{\beta + 1}, \end{aligned} \quad (7)$$

where $\beta = 8.11 \left(\frac{\gamma}{\tau_s} \right) + 21.90 \left(\frac{\gamma}{\tau_s} \right)^2$.

An example of the operation and convergence dynamics of a simple tone follower (STF) in response to a pure tone nearby in frequency is illustrated in Figure 13, and described in the caption. The step response of the linear equivalent circuit (step size is $950 - 901 = 49$ Hz) coincides almost exactly with that of the frequency track shown in Figure 13c.

2.3.2 Dominant tone follower (DTF)

The simple tone follower (STF) is suitable for tracking one tone, but in real world acoustic environments, pure tonal signals are only rarely encountered. Instead, the vast majority of signals are mixtures of complex sounds from multiple sources that can contain nearby partials or harmonics. Here a dominant tone follower (DTF) is needed that can track the frequency of a dominant partial in a signal even in the presence of other interfering ones, similar to the synchrony capture behavior observed in the auditory nerve. A simple modification of the STF described above that employs a nonlinearity in the feedback loop results in the dominant tone follower (DTF) described below.

Consider a signal $x(t)$ consisting of a tone at frequency $\omega_1 = 2\pi f_1$ and an interfering tone at $\omega_2 = 2\pi f_2$.

$$x(t) = A_1 \cos(\omega_1 t + \theta_1) + A_2 \cos(\omega_2 t + \theta_2) \quad (8)$$

Let us assume that $A_1 > A_2$, i.e., the tone at ω_1 is dominant. We rewrite $x(t)$ using complex notation as follows.

$$x(t) = \Re\{A_1 e^{j(\omega_1 t + \theta_1)} (1 + \frac{A_2}{A_1} e^{j\Delta\omega t + j\Delta\theta})\} \quad (9)$$

where $\Re$ stands for “Real part of”, $\Delta\omega = \omega_2 - \omega_1$ and $\Delta\theta = \theta_2 - \theta_1$, and $j = \sqrt{-1}$. Since $A_2/A_1 < 1$, (using the approximation that $e^y \approx 1 + y$ for $y < 1$, in the above

expression) we have,

$$x(t) \approx a(t) \cos(\phi(t)), \quad (10)$$

where the envelope is

$$a(t) \approx e^{\log A_1 + \frac{A_2}{A_1} \cos(\Delta\omega t + \Delta\theta)}, \quad (11)$$

and the phase function is

$$\phi(t) \approx \omega_1 t + \theta_1 + \frac{A_2}{A_1} \sin(\Delta\omega t + \Delta\theta). \quad (12)$$

The derivative of $\phi(t)$ (i.e., the instantaneous frequency (IF)[18], p. 180) and the log-envelope are as follows:

$$\frac{d\phi(t)}{dt} \approx \omega_1 + \frac{A_2}{A_1} \Delta\omega \cos(\Delta\omega t + \Delta\theta), \quad (13)$$

$$\log a(t) \approx \log A_1 + \frac{A_2}{A_1} \cos(\Delta\omega t + \Delta\theta). \quad (14)$$

The symbol $\log$ denotes natural logarithm. Note that the average value of IF is ω_1 , the dominant tone's frequency, and similarly, the average value of the log-envelope is the dominant tone's log amplitude. Either of these properties can be utilized for frequency discrimination purposes. An exact expression for the log-envelope of $x(t)$ can also be obtained as follows:

$$a^2(t) = |A_1 e^{j\omega_1 t + j\theta_1} + A_2 e^{j\omega_2 t + j\theta_2}|^2 = A_1^2 + A_2^2 + 2A_1 A_2 \cos(\Delta\omega t + \Delta\theta). \quad (15)$$

Taking logarithm and using the infinite series expansion for $\log(1+x)$ we have

$$\log a(t) = \log A_1 + \sum_{n=1}^{\infty} \frac{1}{n} \left(\frac{A_2}{A_1} \right)^n \cos(n\Delta\omega t + n\Delta\theta). \quad (16)$$

Note that Eq. 14 retains only the first term in the infinite sum above. Also note that the average value of $\log a(t)$ is $\log A_1$. On the other hand, the average value of the squared envelope $a^2(t)$ is $(A_1^2 + A_2^2)$.

A frequency discriminator can lock on to ω_1 by filtering the instantaneous frequency (IF, assuming that it is available) using a low-pass filter (LPF) with a cut off frequency $\Delta\omega$. Alternatively, the log-envelope can also be used to capture the dominant signal (Figure 15). In an FDL the logarithmically compressed envelope signal, $\log a(t)$, can be low pass filtered (with the same cut off frequency, $\Delta\omega$, as in the case of IF) to obtain $\log A_1$. This can then be used to lock on to the dominant tone in the input.

Figure 15. Frequency error detector (FED) for the dominant tone follower (DTF). The error signal $e(t)$ is computed using the formula $\log \left(\frac{e_R(t)}{e_L(t)} \right)$.

Compared to the simple tone follower, note that the envelopes in the dominant tone follower are now compressed using a logarithmic nonlinearity before they are low pass filtered (by the loop filter). If the input is just one tone ($x(t) = A_1 \cos(\omega_1 t + \theta_1)$) then the corresponding smoothed squared envelopes at the outputs of the right ($H_R(\omega)$) and left ($H_L(\omega)$) filters are $A_{1R}^2 = A_1^2 |H_R(\omega_1)|^2$ and $A_{1L}^2 = A_1^2 |H_L(\omega_1)|^2$ respectively. So, the error signal is $e(t) = 2 \log(A_{1R}/A_{1L})$. Note that

$e(t)$ is proportional to the frequency difference $\omega_1 - \omega_c$ and does not depend on the amplitude A_1 (as in STF).

Now, consider the case of an input $x(t)$ with two tones as in Eq. 8. Then, there are two cases. In the first case, assume that the same tone (either at ω_1 or ω_2) dominates both (right and left) filters' outputs. Then, clearly the (average) error is $2\log(A_{1R}/A_{1L})$ or $2\log(A_{2R}/A_{2L})$ depending on which tone dominates. Since the loop tends to drive this error to zero, the VCO frequency ω_c changes such that the left and right filter's log-amplitudes are equal. Thus ω_c tends to track the dominant tone. In contrast, if the nonlinearity is absent then the left and the right filters produce (squared, averaged) envelopes equal to $A_{1L}^2 + A_{2L}^2$ and $A_{1R}^2 + A_{2R}^2$, which result in ω_c settling in between ω_1 and ω_2 , i.e., no capture. Thus, the compressive non-linearity helps steer the VCO to the dominant signal's frequency.

In the second case, if the tone at ω_1 dominates the left filter output and the tone at ω_2 dominates the right filter output, then the error $e(t)$ is proportional to $\log(A_{2R}/A_{1L})$ and the VCO frequency is adjusted by the loop such that $A_{2R} = A_{1L}$. That is ω_c averages in between ω_1 and ω_2 . In summary, if one tone is sufficiently bigger than the other, then capture occurs, but if two tones are close in frequency and have equal or almost equal amplitudes, then the VCO locks on to a weighted average frequency. This behavior is similar to that seen in the auditory nerve (Figure 9b) for nearby partials.

The linear equivalent circuit for the DTF is essentially identical to that of the STF developed in section 2.3.1, except that the parameter k_s is slightly different $\left(k_s = \frac{4\Delta}{\Delta^2 + \alpha^2}\right)$ (see Appendix B). Figure 16 shows an example of a DTF homing in on a stronger tone in the presence of a nearby weaker tone (vertical arrows). Such dominant tone followers are used as the building blocks for the proposed filterbank algorithm described below in section 2.4.

Figure 16. Behavior of a DTF in response to two nearby tones of different amplitude. (a) Frequency response of BPF triplet filters and the input tones (vertical arrows, dominant tone at $\omega_1 = 2\pi \times 950$ Hz, plus a half-amplitude interfering tone at $\omega_2 = 2\pi \times 1050$ Hz. (b) Track of the VCO frequency for the center filter C. With minor fluctuations, the VCO tracks the stronger 950 Hz tone in spite of the weaker 1050 Hz interferer.

2.3.3 A practical implementation of the frequency discriminator loop (FDL)

This section presents the design of an FDL that incorporates a single VCO and matched BPF triplet filters. This implementation of the BPF triplet (and the FDL), which requires only one VCO, has several advantages over those described above. The filters that form the BPF triplet are implemented as linear phase filters. The BPF triplet is implemented with the help of odd/even prototype filters such that they result in perfectly matched, symmetrical, left ($H_L(\omega)$) and right ($H_R(\omega)$) filters. That is, their frequency response magnitudes are exactly equal at the VCO's frequency ω_c . Further, the computation of the envelopes $e_R(t)$ and $e_L(t)$ does not explicitly require in-phase (I) and quadrature phase (Q) signal components. Instead the envelope is simply obtained by taking the absolute value of the signal, i.e., the full-wave-rectified output, and low-pass filtering it. The

three bandpass filters that constitute the BPF triplet can all be synthesized from a single prototype noncausal, low-pass impulse response,

$$h(t) = e^{-\alpha|t|}, \quad (17)$$

$$H(\omega) = 2\alpha/(\omega^2 + \alpha^2). \quad (18)$$

Any other even impulse response function with unimodal low pass frequency response characteristics (such as, $h(t) = e^{-\beta t^2}$) can also be used as a prototype filter. Let $h_1(t)$ and $h_2(t)$ represent the impulse responses of frequency translated filters, given by

$$h_1(t) = e^{-\alpha|t|} \cos \Delta t, \text{ and } h_2(t) = e^{-\alpha|t|} \sin \Delta t, \quad (19)$$

where Δ is the translation frequency. So,

$$\begin{aligned} H_1(\omega) &= (H(\omega - \Delta) + H(\omega + \Delta))/2, \\ H_2(\omega) &= j(H(\omega - \Delta) - H(\omega + \Delta))/2, \end{aligned} \quad (20)$$

where $j = \sqrt{-1}$. 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 $G_1(\omega)$ and $G_2(\omega)$ shown in Figures 17a and 17b, respectively. $G_1(\omega)$ is called a cos-cos filter (same structure as Figure 12b) and $G_2(\omega)$ is named a cos-sin filter.

$$\begin{aligned} G_1(\omega) &= (H_1(\omega - \omega_c) + H_1(\omega + \omega_c))/2, \\ G_2(\omega) &= j(H_2(\omega - \omega_c) - H_2(\omega + \omega_c))/2. \end{aligned} \quad (21)$$

The frequency responses $G_1(\omega)$ and $G_2(\omega)$ are both real and even and are shown in Figure 17c. These frequency responses can be tuned by changing ω_c .

Assume for the moment, that the systems $H_1(\omega)$ and $H_2(\omega)$ sandwiched between the multipliers are identical. Then, note that the system functions of a

Figure 17. (a) Tunable cos-cos filter, (b) cos-sin filter, (c) Frequency responses $G_1(\omega)$ and $G_2(\omega)$ (without the scale factor j) are shown, (d) Frequency responses of the right and left filters, $H_R(\omega)$ and $H_L(\omega)$, obtained as sum and difference of $G_1(\omega)$ and $G_2(\omega)$ (Figure 18). 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 . The frequency response of $H_C(\omega)$, not shown, is centered around ω_c . All filters are linear phase filters.

generic cos-cos structure, $G_1(\omega)$, and cos-sin structure, $G_2(\omega)$, are related by the expression $G_2(\omega) = j \operatorname{sgn}(\omega) G_1(\omega)$ for sufficiently large ω_c . That is, cos-sin structure has an additional term which signifies a Hilbert transform when compared to cos-cos structure. This stems from the fact that the multipliers in the upper/lower branches of Figure 17b are cosine and sine unlike the cos-cos filter in Figure 17a.

This is a seemingly new way of realizing a band-pass Hilbert transformer. The outputs of the cos-cos and cos-sin filters are then added/subtracted (see Figure 18) to obtain the overall right/left filter responses $H_R(\omega)$ and $H_L(\omega)$ (Figure 17d), respectively. That is,

$$H_R(\omega) = G_1(\omega) - G_2(\omega), \text{ and } H_L(\omega) = G_1(\omega) + G_2(\omega). \quad (22)$$

Substituting for $G_1(\omega)$ and $G_2(\omega)$ in Eq. 22 from Eq. 21, we have,

$$\begin{aligned} H_R(\omega) &= (H_1(\omega - \omega_c) + H_1(\omega + \omega_c))/2 + j(H_2(\omega - \omega_c) - H_2(\omega + \omega_c))/2, \\ H_L(\omega) &= (H_1(\omega - \omega_c) + H_1(\omega + \omega_c))/2 - j(H_2(\omega - \omega_c) - H_2(\omega + \omega_c))/2. \end{aligned} \quad (23)$$

Further substituting for $H_1(\omega)$ and $H_2(\omega)$ in Eq. 23 from Eq. 20 and simplifying, we have

$$\begin{aligned} H_R(\omega) &= H(\omega - \omega_c - \Delta) + H(\omega + \omega_c + \Delta) \\ H_L(\omega) &= H(\omega - \omega_c + \Delta) + H(\omega + \omega_c - \Delta). \end{aligned} \quad (24)$$

Thus, the filters $H_R(\omega)$ and $H_L(\omega)$ (shown in Figure 17d) are the original prototype filter $H(\omega)$ shifted to center frequencies $\omega_c + \Delta$ and $\omega_c - \Delta$, respectively. They have purely real valued frequency responses (except for the linear phase introduced by requiring a causal impulse response) and are the ones used in frequency error detection. In practice, the filter impulse responses in Eq. 19 are symmetrically truncated and Hann windowed 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 , (shown in Figure 12b) 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

Figure 18. Implementation of the frequency error detector and the frequency discriminator loop. The center filter $H_C(\omega)$ (not shown) is implemented using a cos-cos filter structure with $H(\omega)$ sandwiched between the multipliers as in Figure 12b.

signal will drive the VCO to make ω_c equal to ω_1 and vice versa. The loop filter $F(s)$ determines the dynamics. The linear equivalent circuit described in section 2.3.1 is applicable to this implementation as well. The envelope detector shown in Figure 18 is a rectifier in cascade with a LPF. The logarithmic nonlinearity serves the same purpose as in DTF. This LPF increases the time delay τ_g around the loop and has to be included while calculating the loop filter constants k_p and k_i .

2.4 Synchrony capture filterbank (SCFB)

The proposed synchrony capture filterbank (SCFB) shown in Figure 10a consists of a bank of fixed filters each cascaded with a frequency discriminator loop (FDL). The filterbank consists of K logarithmically spaced gammatone filters that have been widely used in auditory system modeling[32]. Using physiologically-appropriate filter parameters (approximately constant, low Q filters), gammatone filterbanks effectively replicate the broadly tuned mechanical filtering characteris-

tics of the basilar membrane in the cochlea.

The gammatone filters used here were designed using the Auditory Toolbox developed by Malcolm Slaney [32], with further details of the cochlear model implementation discussed elsewhere [33]. In this implementation the number of gammatone channels K is 200. The constant-Q gammatone filters span center frequencies from 100-3940 Hz, with corresponding 3-db bandwidths ranging from 50 Hz to 905 Hz. Filter Q values (EarQ parameter) are all 4, and the order parameter is 1 [33]. The minBW used in computing the equivalent rectangular bandwidth (ERB) is 50 Hz. The sampling frequency is 16 kHz.

Figure 19. A typical BPF Triplet centered at 1980 Hz. The broader frequency response corresponds to the gammatone filter centered around 1980Hz.

An example of the frequency responses of one of the fixed filters and the associated three tunable filters of the SCFB are shown in Figure 19. Whereas the broadly tuned, fixed gammatone filters coarsely isolate the various frequency components in the incoming signal, the tunings of the more narrowly tuned bandpass

triplet filters in the frequency discriminator loops (FDLs) converge on the precise frequencies of the individual frequency components.

2.4.1 Bandpass filter triplet parameters

As mentioned earlier each triplet of tunable filters consists of left, center, and right filters, $H_L(\omega)$, $H_C(\omega)$ and $H_R(\omega)$, whose center frequencies are spaced by a constant ratio. All of them are derived from a single prototype filter $H(\omega)$ defined in Eq. 18, whose frequency response is

$$H(\omega) = \frac{2\alpha}{\alpha^2 + \omega^2}. \quad (25)$$

The parameter α is chosen to be equal to the spacing between the filters, i.e., $\alpha = \Delta$. Δ has been chosen to be one-fourth of the bandwidth (actually halfwidth) of the gammatone filter. Hence $\alpha = \Delta = B_{GT}/4$ determines the prototype filter, where B_{GT} stands for gammatone filter bandwidth. For example, Figure 19 shows a gammatone filter centered around 1980 Hz with bandwidth of 466 Hz. Individual left, center and right triplet filters have center frequencies 1864, 1980, and 2098 Hz, respectively. Their bandwidths and center frequency spacings are approximately 115 Hz. Bandwidths and spacings of fixed gammatone and adaptive triplet filters are approximately proportional to their center frequencies.

2.4.2 Frequency discriminator loop filter design $F(s)$

The typical loop filter used in our implementation is of the form $F(s) = k_p + k_i/s$. The proportional gain k_p is intended to improve the rise time of the step response. The VCOs that steer the tuning of the triplet filters are initially set to match the center frequency ω_c of their corresponding gammatone filter. Because the loop is initialized with the VCO frequency close to the input signal frequency, a consequence of the frequency selectivity of the associated gammatone filter, choosing $k_p = 0$ does not affect the loop's rise time performance significantly

and also simplifies its implementation. On the other hand, k_i is needed to keep track of the frequency changes in the input and drive the steady state error to zero. The value of k_i depends on the frequency discriminator constant, k_s , and also on the parameter τ_g that represents the group delay of the prototype filter (i.e., its causal approximation) plus any delay introduced (in smoothing the envelope) in the envelope detector in Figure 18. For each channel, the following values were used for the loop filter parameters, and they seem to work well in most circumstances (set $\beta = 1$ in Eq. 7):

$$k_p = 0$$

$$k_i = \frac{1}{k_s} \left(21.90 \frac{\gamma}{\tau_s^2} \right) = \frac{10.95 \tau_g}{k_s \tau_s^2}.$$

τ_s , the settling time, in seconds, is chosen to be approximately $\frac{50}{f_c}$, where f_c is the center frequency of a gammatone filter, in Hz. FDL operation is relatively insensitive to choice of particular parameter values.

2.5 Simulation results

The SCFB algorithm has been tested with appropriate parameter choices using several synthetic signals and speech signals drawn from the TIMIT database. Here simulation results are presented for one set of synthetic musical notes, an isolated utterance drawn from the ISOLET database, and a set of sentences of continuous speech from the TIMIT database with and without additive noise. For speech signals, the input signal is first subjected to spectral equalization by using a pre-emphasis filter and then processed through the filterbank and the self tuning FDL circuits. The frequencies of the VCOs in FDL modules indicate the frequency components that those modules are tracking at any given time. The outputs of the BPF triplets are available for further processing, and these can be used to classify whether the signal in local frequency bands are tonal or noise-like. For example,

if the envelope of the three filter outputs are larger than the background noise level and if the center filter has a significantly larger output when compared with the associated left and the right filters, then this implies that the corresponding channel has a tonal signal. Conversely, if the three envelopes are approximately equal in size then this implies that the channel output is non-tonal or locally white.

2.5.1 Dyads of synthetic harmonic signals

The filterbank response to synthetic harmonic signals is considered first. The stimulus consists of two notes of two harmonic complexes (equal amplitude harmonics, 1 to 6). In musical terms, these are two notes separated by a minor second (16:15) and a perfect fourth (4:3). They are the same signals that produced the auditory nerve interspike interval patterns depicted in Figure 9. The first note has two fundamentals (440 and 469 Hz) separated by 6.6%. The second has a frequency separation of 33.3% (with fundamental frequencies 440 and 587 Hz). Perceptually, for the minor second, human listeners hear only one pitch intermediate in frequency between the two notes, whereas for the perfect fourth, two note pitches can be heard.

Responses of the SCFB to these pairs of complex harmonic tones are shown in Figure 20. A "capturegram" plot of the resulting frequency tracks of the VCOs as a function of time shows the locking of groups of channels onto individual frequency components. The plots show only tracks of VCO frequencies of low frequency channels ($f_c < 1000$ Hz) to permit more direct comparison with the interspike interval histograms in Figure 9. Note that most of the VCO frequency tracks with CFs close to the dominant tone frequencies converge rapidly (within a few tens of milliseconds) to their steady state value.

The filterbank response for two closely spaced note dyads separated by 6.6% is shown in Figure 20a. This signal has 4 frequency components below 1000 Hz:

440, 469, 880, and 938 Hz. Here the filterbank does not resolve the pairs of nearby partials (440/469 and 880/938 Hz), but rather all the channels converge on the mean frequencies of the nearby partials (channels 53 to 88 fluctuate around 458 Hz, 89-112 fluctuate around 909 Hz). The pattern of frequency capture is similar to that in the interspike interval data in Figure 9a. Figure 20b shows rectified outputs of each channel's center filter and Figure 20c shows the autocorrelation of the rectified outputs (from time $t = 0.25$ to 0.5 seconds). In this case we can see the fluctuations in envelope are related to the beat frequency ($469-440=29$ Hz) (as seen in Figure 9a).

The filterbank response to the well-separated note dyad is shown in Figure 20d. This signal has 3 frequency components below 1000 Hz: 440, 587, and 880 Hz. Clearly each VCO is captured by the dominant partial in that channel's neighborhood. Channels with center frequencies between 300 and 525 Hz lock to 440 Hz, those with center frequencies between 525 Hz and 725 Hz lock to 587 Hz, and the rest are captured by the 880 Hz partial. Transitions of VCO frequency change from one dominant tone to the other is abrupt. For example, for center frequencies near 500 Hz, the channels are either captured by 440 Hz tone or the 587 Hz tone. Very similar behavior is also observed in the interspike interval histograms in Figure 9b where interspike intervals in the corresponding CF channels switch abruptly from interval patterns associated with 440 Hz to those associated with 587 Hz. Figure 20e shows rectified outputs of each channel's center filter and Figure 20f shows the autocorrelation of the rectified outputs after the frequency estimates, which are almost constant (in other words the channel's VCO are locked, in this case from time $t = 0.25$ to 0.5 seconds).

Figure 20. Filterbank responses to pairs of harmonic tones. Left. Responses to a note dyad separated by a minor second ($\Delta F_0=6.6\%$, $F_0s = 440 \text{ \& } 469 \text{ Hz}$). Right. Responses to a note dyad separated by a perfect fourth ($\Delta F_0=33.3\%$, $F_0s = 440 \text{ \& } 587 \text{ Hz}$). Top plots (a),(d). Frequency tracks of the VCOs (capturegram). Middle plots (b), (e). Half-wave rectified output waveforms of channel center filters (analogous to a post-stimulus time neurogram). Bottom plots (c), (f). Channel autocorrelations (compare with autocorrelation neurograms of Figure 9).

2.5.2 Speech signals

For synthetic signals, such as the musical notes in the previous subsection, the instantaneous frequency estimates obtained from the VCOs of nearby channels are essentially the same after the initial settling time. However, for natural signals like speech the frequency estimates of the partials tend to have some variability (as can be seen below). Clearly, some sort of clustering method is needed to obtain the average frequency tracks associated with each frequency component in the signal. Other well known auditory-inspired models such as the ZCPA (Zero-Crossing Peak Amplitude)[34] or EIH (Ensemble Interval Histogram)[12] use the upward-going zero or level crossing events in a signal (emanating from a filter channel) to estimate the frequency. The reciprocal of the time interval between adjacent zero/level crossing events is used as the instantaneous frequency estimate. Such frequency estimates obtained over a time window are collected to assemble a frequency histogram.

The frequency histograms across all filter channels are combined (in both ZCPA and EIH) to represent the output of the auditory model [34]. Further, in ZCPA the peak of the envelope that lies in between two consecutive zero-crossing events is used as a nonlinear weighting factor to a frequency bin to simulate the firing rate of the auditory nerve. Here a similar procedure is followed, except that the frequency estimates are not derived from the zero-crossing events but from the VCOs frequencies. The envelopes are obtained from the rectified and smoothed outputs of the center filter of each channel.

The frequency values corresponding to the 200 channels are binned into 40 logarithmically spaced frequency bins that lie between 100 and 4000 Hz. However, before binning the frequency values, a non-linear weighting factor ($\log(1+a)$, where a is the amplitude/envelope corresponding to that frequency value) was applied

as in ZCPA. Then histogram peaks with heights below a threshold (10% of peak amplitude) are eliminated. This will eliminate silent regions where amplitudes are very low. Only when the log-envelope value is above the threshold are the actual frequency estimates calculated for a bin, using $\frac{\sum_n \log(1 + a_n) f(n)}{\sum_n \log(1 + a_n)}$, where a_n and f_n represent the amplitude/envelope and frequency values that fall within a bin. The steps involved in the processing of speech signals are sketched in Figure 21a.

A histogram of the distribution of frequencies tracked by the VCOs is useful for assessing the degree to which channels have converged on particular frequencies. Here the number of channels converging on a particular frequency provides a robust, qualitative measure of its relative intensity. The running histogram of frequencies tracked (Figure 21a) provides a cleaner analysis of the time courses of dominant signal periodicities. Thresholding the running capture histogram keeps regions where multiples channels have converged on the same frequency and removes those where there is little agreement. Figures 21(b,c, and d), 22 and 23 demonstrate the character of this analysis.

2.5.3 Isolated spoken letters

The SCFB algorithm was applied to a vowel /i/ (as in “beet”)(file name: fskes0-E1-t.adc, male speaker) drawn from the ISOLET database. Figure 21(b,c,d) shows the simulation results. Figure 21b shows the spectrogram of the vowel utterance and 21c shows the capturegram, i.e. the raw frequency tracks of the 200 VCOs.

It can be seen that the FDLs track closely the frequencies of the individual partials up to at least 1000 Hz. Depending on the relative intensity of each partial, typically five to ten channels tend to converge on to the stronger partials’ frequency tracks. The first formant F_1 is located at around 300 Hz between the second and third harmonics. At higher frequencies (> 2000 Hz), where the filters (the

gammatone and BPFs tend to be wider) several channels tend to converge on the three higher formant frequencies which are located approximately at frequencies 2400, 2800 and 3800 Hz. Between the first and the second formant frequencies, where the signal energy is relatively low, there are no dominant tones, and hence, the VCO tracks tend to wander. Figure 21d shows the cleaned up tracks after the histogramming procedure outlined in Figure 21a is applied. This procedure tends to suppress meandering tracks and signal components with small envelope values.

2.5.4 Continuous speech

The SCFB algorithm was also applied to several continuous speech samples drawn from the TIMIT database. The speech signals were first pre-emphasized with a $H(z) = 1 - 0.95z^{-1}$ filter to equalize the spectrum to prevent strong low frequency components from swamping the weaker high frequency components. The sampling frequency is 16kHz. Capturegrams for two speech sentences, “Where were you while we were away?” (TIMIT sentence sx9, speakers mpcs0 and fgjd0) and “The oasis was a mirage” (TIMIT sentence sx280, speakers mdwk0 and fawf0) spoken by male and female speakers are shown in Figures 22 and 23, respectively.

Figures 22a and 22d show the spectrograms of the TIMIT sx9 utterances by male and female speakers. In Figure 22b and 22e the corresponding capturegram tracks for the 200 VCOs are superimposed on the spectrogram for the male and female utterances. Typically, for a strong low-frequency harmonic component, a handful of channels are captured by one harmonic. Note that at low frequencies and harmonic numbers ($f < 800$ Hz, $n < 8$) almost all the individual harmonics tend to be closely tracked by the FDLs. These frequency tracks together can provide a robust representation of the fundamental frequency (voice pitch). For higher frequencies and harmonic numbers, only dominant harmonics in formant regions are tracked. This behavior is due to the constant Q s of the filters, such

that FDL triplet filters with higher center frequencies have correspondingly larger bandwidths, and therefore cannot resolve individual harmonics. Instead these filters lock onto the nearest dominant harmonic component somewhere near the middle of a formant.

Similarly, Figures 23b and 23e show the capturegrams for the sentence TIMIT sx280 spoken by a male and a female, respectively. In both cases, the frequency transitions, especially at the higher frequency regions are precisely and robustly tracked. At lower frequencies, as one harmonic becomes weaker with respect to a nearby harmonic, the frequency tracks of channels in that neighborhood jump from the weaker harmonic to the stronger one due to the tendency of the FDL to track the stronger component (as in the time-frequency region $t = 1.0 - 1.45$ s, frequency < 1000 Hz) in Figure 23e. The last rows of the figures show the frequency tracks after the histogram thresholding procedure has been applied.

Previous analysis of cat auditory nerve responses had suggested that the synchrony capture effect is resistant to noise [35]. So, we tested the SCFB algorithm with noisy speech signals to determine its robustness to noise. Signal power P_s is calculated as the sum of squares of all the speech signal samples divided by the time duration of the speech signal. The variance σ^2 is obtained from the definition of signal to noise ratio (SNR) given below.

$$SNR = 10 \log_{10} \left(\frac{P_s}{\sigma^2} \right) dB. \quad (26)$$

The Gaussian distributed noise samples are generated with a variance σ^2 obtained from the above formula for an SNR of 10 dB. The generated noise samples are added to the speech signals, and are processed by the SCFB algorithm. Figure 24 shows the simulation results. Left column corresponds to “The oasis was a mirage” (sx280) for a female speaker, and the right column is for “Where were

you while we were away?” (sx9) by a male speaker. The spectrograms (a) and (d) are relatively darker than the spectrograms in Figures 22 and 23, because of the additive noise. Even in these noise corrupted cases, the formant and harmonics’ tracks (especially the formant transitions) are clearly visible. Capturegrams show that multiple channels still merge to the same frequencies and the histogram tracks are also relatively clean. Thus, qualitatively, the behavior of the SCFB in noise seems to parallel that seen in the cat auditory nerve.

Figure 21. (a) Steps involved in the SCFB algorithm. The input speech signal $s(t)$ (after preemphasis) is processed by the 200 gammatone filters and the associated FDLs and the frequency tracks are plotted as capturegrams. The VCO frequency values and the associated envelopes are used to generate the frequency histograms from which dominant frequency tracks are derived. Results for ISOLET vowel /i/. (b) Spectrogram (c) Capturegram (d) Thresholded histogram plot.

Figure 22. Results for TIMIT utterance, “Where were you while we were away?” (sx9) for male (left column) and female (right column) speakers. Top plots (a)(d). Spectrograms. Middle plots (b)(e). Capturegrams. Bottom plots (c)(f). Thresholded histogram plots. At low frequencies, all individual harmonics are tracked, whereas above 1000 Hz, only prominent formant harmonics are tracked.

Figure 23. Results for TIMIT utterance “The oasis was a mirage” (sx280) for male (left column) and female (right column) speakers. Plots as in the previous figure. High frequency frication above 4000 Hz in “oasis” not shown.

Figure 24. Results for two TIMIT utterances in 10dB SNR. “The oasis was a mirage” (sx280) for a female speaker (left column) and “Where were you while we were away?” (sx9) for a male speaker (right column). Plots as in the previous figure.

2.6 Discussion

Our interest in synchrony-capture based filterbanks has been motivated by considerations of the functional anatomy and response characteristics of the cochlea, adaptive filtering signal processing strategies in radar and other artificial systems, and the possible role of synchrony capture in auditory nerve representation of complex sounds. The primary goal in this first stage of investigation has been to integrate these aspects into a workable algorithm for tracking the major frequency components present in an acoustic signal.

2.6.1 Relationship to previous signal processing strategies

As is often the case, the signal processing constituents of the SCFB algorithm proposed here have a long history. Frequency discriminator loops (FDLs) have been used in digital and analog communication systems for signal tracking for many decades [27]. The frequency error detector (FED) circuit (Figure 11) is a key component of the FDL that senses the difference between the frequency of the input signal and that of a local VCO in order to produce a proportional error voltage that can be used for steering purposes.

Basically there are two or three common types of frequency error detector circuits that are used in practice. The quadricorrelator [29, 28], briefly outlined in Appendix A, is often used in communication systems. The other type, which has been used here in the SCFB design, uses stagger-tuned filters and compares envelopes of filter outputs to derive running error voltages. Ferguson and Mantey [21] originally proposed the use of such adaptable stagger-tuned bandpass filters for frequency error detection. Alternately, frequency error detectors can also be implemented directly by using phase derivatives of a complex signal (see for example [36, 37]). Wang [38] has designed a harmonic locked loop to track the fundamental frequency of a periodic signal using this idea. However, these approaches require

a complex (Hilbert-transformed) signal for processing.

In their adaptive, stagger-tuned design, Ferguson and Mantey used the error voltage (envelope difference) to retune the bandpass filters directly by moving their pole locations. Such a design does not use VCOs to tune the filters. Based on this idea, one could imagine cochlear filters, where the frequency response of a filter is adjusted by changing a mechanical parameter such as stiffness depending on the envelope voltage difference between the left and the right filters. Costas [22] used a similar FED, but used the error voltage to change the frequency of a VCO that indirectly moved the left and the right bandpass filters in tandem. The approach proposed here is closer to Costas' method and its variants [22, 36, 38]. The main difference here is that a compressive (logarithmic) nonlinearity is used on the envelope of a signal to suppress nearby weaker signal components. Such compressive nonlinearities have the property of favoring a stronger component in the presence of other weaker ones. This is the primary reason that synchrony capture occurs.

The SCFB design is also related to adaptive formant tracking methods proposed earlier by Rao and Kumaresan [39, 40], and subsequently improved by Mustafa and Bruce [41]. However, in Rao-Kumaresan approach the adaptive formant filters were controlled by measuring the instantaneous frequency of a complex-valued signal. Further, as mentioned earlier, EIH and ZCPA algorithms also estimate the frequency of tonal signals based on the zero or level crossing intervals. However, these may be regarded as open loop methods for estimating instantaneous frequencies, unlike the closed loop methods like FDL.

2.6.2 Similarities to response characteristics of the cochlea and auditory nerve

Although the SCFB is not a biophysical model, its signal processing behavior bears many qualitative similarities to response patterns in the mammalian cochlea. First, the mammalian cochlea produces acoustic emissions, called spontaneous otoacoustic emissions (SPOAEs)[42]). The narrow spectral widths of these emissions suggest that they are generated by spontaneous oscillations in the cochlea, possibly in outer hair cells. This kind of behavior is also characteristic of voltage controlled oscillators that implement the FDL in the present architecture.

Second, it is also well known [42] (p.117) that the cochlea also produces acoustic emissions at additional frequencies when two tones of frequency f_1 and f_2 ($f_2 > f_1$) are presented. Listeners can often hear discordant faint tones not present in the original stimulus. The strongest of these cochlear distortion products, the cubic distortion product generated at $2f_1 - f_2$ Hz, is thought to be a direct byproduct of cochlear mechanics, in the form of a compressive nonlinearity in OHC response. The ensuing signal distortions are analogous to intermodulation products in communication systems. The FDL architecture produces similar combination tones as a byproduct of its operation. Consider the operation of the FDL as described in section 2.3.2 when two simultaneous tones with frequencies f_1 and f_2 and corresponding amplitudes A_1 and A_2 are applied as input. The spectrum of the VCO output for this stimulus is shown in Figure 25 for a channel with center frequency 1890 Hz. $f_1 = 1950$ Hz and $f_2 = 2050$ Hz, $A_1 = 1$ and $A_2 = 0.5$. Note that the VCO locks on to the stronger tone at f_1 Hz and that the left and the right filters of that channel adjust themselves such that their average envelopes are equal. Then the resulting error signal $e(t)$ is proportional to $C \cos(\Delta\omega t)$ where $\Delta\omega = 2\pi \times (f_2 - f_1)$ and C is a constant related to the ratio of amplitudes A_2/A_1 (see Eq. 14). This error signal then frequency modulates the VCO's carrier at the

dominant tone frequency f_1 . The resulting frequency modulated VCO output has sideband components at $f_1 \pm n(f_2 - f_1)$ [18] p.180-87. The output spectrum in Figure 25 shows some of the sidebands (for $n = 1$ and 2). Thus qualitative parallels exist between combination tones produced by live cochleae and the VCO-driven frequency capture circuits of the filterbank.

Figure 25. Distortion products. Spectrum of VCO output signal of a channel with center frequency of 1890 Hz in response to two pure tones at frequencies $f_1 = 1950$ Hz and $f_2 = 2050$ Hz with amplitudes $A_1 = 1$ and $A_2 = 0.5$ respectively. Note occurrences of distortion products at frequencies $f_1 \pm n(f_2 - f_1)$. These are generated in frequency discriminator loops when VCOs lock on to dominant tones at f_1 but are also frequency modulated by an error signals consisting of a weak tones at $\Delta f = f_2 - f_1$.

Two-tone suppression is a third nonlinear phenomenon. Like the cochlea, the proposed filterbank produces both rate- and synchrony-suppression. Two-tone rate suppression is generally regarded as a nonlinear property of the cochlea in which the average neural firing rate in the region most sensitive to a probe tone is reduced by the addition of a suppressor tone at a different nearby frequency. For

the filterbank, when dominant frequency components steer the tunings of local VCOs away from other frequencies, responses to less intense secondary tones at those frequencies are attenuated relative to those produced when the dominant tone is absent.

There is also the related phenomenon of synchrony suppression. The effects of two tonal inputs on temporal patterns of neural firing have been extensively studied. Auditory nerve fibers phase-lock in response to low frequency tones (< 5000 Hz), i.e. spikes are mainly produced at particular phase angles of the waveform [11]. The degree of synchronization of spikes to a given frequency can be quantified by computing the vector strength (“synchronization index”) of the spike distribution as a function of waveform phase. When the stimulus consists of two tones, Hind et al. [43] found that auditory nerve spikes may be phase locked to one tone, or to the other, or to both tones simultaneously. Which of these occurs is determined by the relative intensities of the two tones and their frequencies and spacings. Moore [11] summarizes these results as follows, “When phase locking occurs to only one tone of a pair, each of which is effective when acting alone, the temporal structure of the response may be indistinguishable from that which occurs when the tone is presented alone. Further, the discharge rate may be similar to the value produced by that tone alone. Thus the dominant tone appears to “capture” the response of the neuron. This (synchrony) capture effect underlies the masking of one sound by another”. The tone that is suppressed ceases to contribute to the pattern of phase-locking, and the neuron responds as if only the suppressing tone were present. The effect is that the synchronization index of a fiber to a given tone is reduced by the application of a second tone [44]. Similarly, in the filterbank, capture of a given channel VCO by a locally dominant component produces an output waveform having the frequency of the dominant tone, causing the vector

strength of the dominant component to increase at the expense of those of weaker secondary ones.

2.7 Conclusions

A striking feature of the phase-locked responses to complex sounds is the phenomenon of “synchrony capture” [3, 5], wherein an intense stimulus frequency component dominates the temporal firing patterns of auditory nerve fibers innervating the corresponding cochlear frequency region. The capture effect refers to the almost exclusive nature of the phase-locking to the dominant component, such that the output of whole subpopulations of auditory nerve fibers in a cochlear region respond in the same way.

An adaptive filterbank structure is proposed that emulates synchrony capture in the auditory nerve. This filterbank has two parts: a fixed array of traditional, passive linear (gammatone or equivalent) filters that are cascaded with a bank of adaptively tunable bandpass filter triplets. Envelope differences in the outputs of the filters that form the triplets are used in frequency discriminator loop (FDL) to steer their center frequencies with the help of a voltage controlled oscillator (VCO).

The resulting filterbank exhibits many desirable properties for processing speech and other natural sounds. First, the number of channels converging on a particular frequency yields a robust means of encoding the intensity of the driving frequency component. The VCOs track resolved harmonics, which are known to be essential in determining the pitch and for the separation of concurrent periodic sounds. For voiced speech, the VCOs track the strongest harmonic in each formant region, yielding precise features for formant tracking.

2.8 Acknowledgments

This work was supported by the Airforce Office of Scientific Research under the grant # AFSOR FA9550-09-1-0119. We thank Prof. R. Vaccaro for pointing out that the Laplace transform of a time delay operator $\delta(t - \tau_g)$ ($= e^{-s\tau_g}$) can be approximated (Padé approximation) by a ratio of s-polynomials. The authors thank the three reviewers for many suggestions that helped improve the manuscript.

2.9 Appendix A: Alternate frequency error detectors

The frequency error detector (FED) is a key component of the FDL (see Figure 11). In the tone followers described in section 2.3 we used the difference in (squared) envelopes (or log-envelopes) of the outputs of $H_R(\omega)$ and $H_L(\omega)$ as the error signal $e(t)$. $e(t)$ is proportional to the difference between the VCO frequency ω_c and the input (or dominant) tone frequency ω_1 . In section 2.3 the specific type of FED (that is, one that uses squared envelope differences) was chosen because of its apparent functional similarity to the functioning of cochlear hair cells. (The inner/outer hair cells act as halfwave rectifiers followed by low-pass filters). Disregarding such constraints, if computer implementation of a FDL is the primary goal, then many other FEDs are available. Of course, the frequency error signal could be positive or negative depending on whether ω_c is greater or smaller than ω_1 . Therefore, any method that is used to measure the frequency of a single tone can serve as a FED as long as it is also capable of detecting the sign of the frequency error. One such FED is called a Quadricorrelator [28]. The quadricorrelator (refer to Figure 3 in [28]) is input with a tone $A_1 \cos(\omega_1 t + \theta_1)$ and the VCO outputs $\cos(\omega_c t)$ and $\sin(\omega_c t)$. The low pass filters (LPF) (in Figure 3 in [28]) retain only the difference frequency outputs $\alpha_1 \cos(\Delta\omega t + \theta_1)$ and $\alpha_2 \sin(\Delta\omega t + \theta_1)$. The two differentiator outputs after cross multiplying (in Figure 3 in [28]) are added together to produce the error signal which retains the sign of the frequency error. Since in our simulations, in-phase and quadrature-phase signals (I and Q) are available, complex valued processing can also be used to estimate frequency error [37, 45, 38].

2.10 Appendix B: Expressions for the frequency discriminator constant k_s

k_s , defined in section 2.3.1, is the slope of the frequency discriminator function $S(\omega)$ at ω_c . $S(\omega)$ for the Simple Tone Follower (STF) is defined as

$$S(\omega) = \frac{|H_R(\omega)|^2 - |H_L(\omega)|^2}{|H_R(\omega)|^2 + |H_L(\omega)|^2} \quad (27)$$

where $|H_R(\omega)|^2 = |H(\omega - (\omega_c + \Delta))|^2$ and $|H_L(\omega)|^2 = |H(\omega - (\omega_c - \Delta))|^2$. Using $H(s) = \frac{1}{s + \alpha}$, $H(\omega) = \frac{1}{j\omega + \alpha}$, $|H_R(\omega)|^2$ and $|H_L(\omega)|^2$ are

$$|H_R(\omega)|^2 = \frac{1}{(\omega - (\omega_c + \Delta))^2 + \alpha^2} \quad (28)$$

$$|H_L(\omega)|^2 = \frac{1}{(\omega - (\omega_c - \Delta))^2 + \alpha^2} \quad (29)$$

Substituting Eqs. 28 and 29 in Eq. 27, we get

$$S(\omega) = \frac{2\Delta(\omega - \omega_c)}{\omega^2 + \omega_c^2 + \Delta^2 - 2\omega\omega_c + \alpha^2}. \quad (30)$$

k_s is obtained by taking the derivative of $S(\omega)$ with respect to ω and evaluating at $\omega = \omega_c$.

$$k_s = \left[\frac{dS(\omega)}{d\omega} \right]_{\omega=\omega_c} = \frac{2\Delta}{\Delta^2 + \alpha^2}. \quad (31)$$

Similarly, for the Dominant Tone Follower (DTF), k_s is obtained by taking the derivative of $S(\omega) = \log \frac{|H_R(\omega)|^2}{|H_L(\omega)|^2}$ and evaluating at $\omega = \omega_c$. It is easy to show that

$$k_s = \frac{4\Delta}{\Delta^2 + \alpha^2}. \quad (32)$$

List of References

- [1] R. M. Stern and N. Morgan, “Hearing is believing,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, 2012.
- [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, pp. 470–479, 1979.
- [3] B. Delgutte and N. Kiang, “Speech coding in the auditory nerve: I. Vowel-like sounds,” *J. Acoust. Soc. Am.*, vol. 75, pp. 866–878, 1984.
- [4] H. E. Secker-Walker and C. L. Searle, “Time-domain analysis of auditory-nerve-fiber firing rates,” *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1427–1436, 1990.
- [5] M. B. Sachs, I. C. Bruce, R. L. Miller and E. D. Young, “Biological basis of hearing-aid design,” *Annals of Biomedical Engineering*, vol. 30, pp. 157–168, 2002.
- [6] M. J. Tramo, P. A. Cariani, B. Delgutte, and L. D. Braida, “Neurobiological foundations for the theory of harmony in western tonal music,” *Ann N Y Acad Sci*, vol. 930, pp. 92–116, 2001.
- [7] P. Cariani and B. Delgutte, “Neural correlates of the pitch of complex tones.I. Pitch and pitch salience. II. Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch,” *J. Neurophysiol.*, vol. 76(3), pp. 1698–1734, 1996.
- [8] R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, “Synchrony capture filterbank (SCFB): An auditory periphery-inspired method for tracking sinusoids,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012, pp. 153–156.
- [9] P. Cariani, “Temporal coding of periodicity pitch in the auditory system: an overview,” *Neural Plasticity*, vol. 6, no. 4, pp. 147–172, 1999.
- [10] R. Meddis and L. O’Mard, “A unitary model of pitch perception,” *J Acoust Soc Am*, vol. 102, no. 3, pp. 1811–1820, 1997.
- [11] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. San Diego, CA: Academic Press, 1997, pp. 40, 90–103, 247.
- [12] O. Ghitza, “Auditory models and human performance in tasks related to speech coding and speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 115–132, 1994.

- [13] L. Cedolin and B. Delgutte, "Spatiotemporal representation of the pitch of harmonic complex tones in the auditory nerve," *J. Neurosci.*, vol. 30 (38), pp. 12 712–12 724, 2010.
- [14] C. J. Darwin, "Auditory grouping," *Trends in Cognitive Sciences*, vol. 1, no. 9, pp. 327–333, Dec. 1997.
- [15] C. J. Plack and A. J. Oxenham, "Psychophysics of Pitch," in *Pitch: Neural Coding and Perception*, ser. Springer Handbook of Auditory Research, C. J. Plack, A. J. Oxenham, R. R. Fay and A. N. Popper, Ed. New York, NY: Springer-Verlag, 2005, ch. 2, pp. 7–55.
- [16] J. O. Pickles, "Psychophysical frequency resolution in the cat as determined by simultaneous masking and its relation to auditory-nerve resolution," *J. Acoust. Soc. Am.*, vol. 66, pp. 1725–1732, Dec. 1979.
- [17] E. J. Baghdady, "Theory of stronger-signal capture in FM reception," in *Proceedings of Institute of Radio Engineers*, April 1958, pp. 728–738.
- [18] S. Haykin, *Communication Systems*, 2nd ed. New York, NY: John Wiley & Sons, 1983, pp. 180–187, 206.
- [19] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," *Journals of the American Physiological Society*, vol. 81, no. 3, pp. 1305–1352, July 2001.
- [20] W. E. Brownell, "The Piezoelectric Outer Hair Cell," in *Vertebrate Hair Cells*, ser. Springer Handbook of Auditory Research, R. A. Eatock, R. R. Fay and A. N. Popper, Ed. New York, NY: Springer-Verlag, 2010, ch. 7, pp. 313–347.
- [21] M. J. Ferguson and P. E. Mantey, "Automatic frequency control via digital filtering," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, no. 3, pp. 392–397, Sep 1968.
- [22] J. P. Costas, "Residual signal analysis -a search and destroy approach to spectral estimation," in *Proceedings of the First ASSP Workshop on Spectral Estimation*, Hamilton, Canada, Aug. 1981, pp. 6.5.1–6.5.8.
- [23] P. Dallos, "Overview: Cochlear Neurobiology," in *The Cochlea*, ser. Springer Handbook of Auditory Research, P. Dallos, A. N. Popper, and R. R. Fay, Ed. New York, NY: Springer-Verlag, 1996, ch. 1, pp. 1–43.
- [24] A. R. Møller, *Hearing: Its Physiology and Pathophysiology*, 1st ed. San Diego, CA: Academic Press, 2000.
- [25] Fabio A. Thiers, Joseph B. Nadol, and M. Charles Liberman, "Reciprocal synapses between outer hair cells and their afferent terminals: Evidence for a local neural network in the mammalian cochlea," *J. Assoc. Res. Otolaryngol.*, vol. 9, pp. 477–489, 2008.

- [26] H. Spoendlin, “Degeneration behaviour of the cochlear nerve,” *Archiv fr klinische und experimentelle Ohren-, Nasen- und Kehlkopfheilkunde*, vol. 200, pp. 275–291, 1971.
- [27] F. D. Natali, “AFC tracking algorithms,” *IEEE Trans. Commun.*, vol. Com-32, pp. 935–947, Aug. 1984.
- [28] F. M. Gardner, “Properties of frequency difference detectors,” *IEEE Transactions on Communications*, vol. Com-33, no. 2, pp. 131–138, 1985.
- [29] D. G. Messerschmitt, “Frequency detectors for PLL acquisition in timing and carrier recovery,” *IEEE Trans. Comm.*, vol. COM-27, no. 9, pp. 1288–1295, Sep 1979.
- [30] R. J. Vaccaro, *Digital Control: A State-Space Approach*, 1st ed. New York, NY: McGraw-Hill, 1995, ch. 6, pp. 233, 254.
- [31] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore and London: The Johns Hopkins University Press, 1996, ch. 11, pp. 572–574.
- [32] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, “Implementing a gammatone filter bank,” in *Annex C of the SVOS Final Report (Part A: The Auditory Filter Bank) MRC (Medical Research Council), APU (Applied Psychology Unit) Report 2341*, University of Cambridge, Cambridge, United Kingdom, Feb 1988.
- [33] M. Slaney, “An efficient implementation of the patterson-holdsworth auditory filter bank,” in *Apple Technical Report # 35, Perception Group - Advanced Technology Group*, Apple Computer Library, Cupertino, CA 95014, 1993, pp. 2–7.
- [34] D. Kim, S. Lee, and R. Kil, “Auditory processing of speech signals for robust speech recognition in real-world noisy environments,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, Jan. 1999.
- [35] M. B. Sachs, H. F. Voigt and E. D. Young, “Auditory nerve representation of vowels in background noise,” *J. Neurophysiol.*, vol. 50, no. 1, pp. 27–45, Jul. 1983.
- [36] R. Kumaresan, C. S. Ramalingam, and A. Rao, “RISC: an improved costas estimator-predictor filter-bank for decomposing multi-component signals,” in *Proceedings of the Seventh Statistical Signal and Array Processing Workshop*, Québec City, Canada, June 1994, pp. 207–210.
- [37] S. M. Kay, “A fast and accurate single frequency estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1987–1990, Dec. 1989.

- [38] A. L. Wang, “Instantaneous and frequency warped signal processing techniques and auditory source separation,” Ph.D. dissertation, Stanford University, Stanford, CA, August 1994.
- [39] 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.
- [40] A. Rao and R. Kumaresan, “On decomposing speech into modulated components,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.
- [41] K. Mustafa and I. C. Bruce, “Robust formant tracking for continuous speech with speaker variability,” *IEEE Trans. on Speech Audio Processing*, vol. 14, no. 2, pp. 435–444, 2006.
- [42] P. A. Fuchs, “Otoacoustic emissions and evoked potentials,” in *The oxford handbook of auditory science: The ear*, 1st ed., D. T. Kemp, Ed. Great Clarendon Street, Oxford OX26DP: Oxford University Press, 2010, ch. 4, pp. 93–137.
- [43] J. E. Hind, D. J. Anderson, J. F. Brugge and J. E. Rose, “Coding of information pertaining to paired low-frequency tones in single auditory nerve fibers of the squirrel monkey,” *Journal of Neurophysiology*, vol. 30, pp. 794–816, July 1967.
- [44] E. Javel, C. D. Geisler, and A. Ravindran, “Two-tone suppression in auditory nerve of the cat: Rate-intensity and temporal analyses,” *J. Acoust. Soc. Am.*, vol. 63, pp. 1093–1104, 1978.
- [45] R. Kumaresan and C. S. Ramalingam, “On separating voiced-speech into its components,” in *Proceedings of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1993, pp. 1041–1046.

MANUSCRIPT 3

*Published in the proceedings of the IEEE International Conference on Acoustics,
Speech and Signal Processing (ICASSP), May 2014*

**Auditory-inspired pitch extraction using a synchrony capture
filterbank and phase alignment**

Ramdas Kumaresan¹, Vijay Kumar Peddinti¹, Peter Cariani²

¹ Department of Electrical, Computer and Biomedical Engineering,
University of Rhode Island, Kingston RI 02881

² Hearing Research Center & Department of Biomedical Engineering,
Boston University, Boston, MA 02114

Corresponding Author: Ramdas Kumaresan

Department of Electrical, Computer-
and Biomedical Engineering

University of Rhode Island

Kingston, RI 02881

Phone: 401-874-4798

Email: kumar@ele.uri.edu

Auditory-inspired pitch extraction using a synchrony capture filterbank and phase alignment

3.1 Abstract

The question of how harmonic sounds produce strong, low pitches at their fundamental frequencies, f_0 s, has been of theoretical and practical interest to scientists and engineers for many decades. Currently the best auditory models for f_0 pitch, e.g. [1], are based on bandpass filtering (cochlear mechanics), half-wave rectification and low-pass filtering (haircell transduction and synaptic transmission), channel autocorrelations (all-order interspike interval statistics) aggregated into a summary autocorrelation, and an analysis that determines the most prevalent interspike intervals. As a possible alternative to autocorrelation computations, we propose an alternative model that uses an adaptive Synchrony Capture Filterbank (SCFB) in which groups of filters or channels in a filterbank neighborhood are driven exclusively (captured) by dominant frequency components that are closest to them. The channel outputs are then adaptively phase aligned with respect to a common time reference to compute a Summary Phase Aligned Function (SPAF), aggregated across all channels, from which f_0 can be easily extracted.

doi=10.1109/ICASSP.2014.6854747

3.2 Introduction

Pitch is an essential attribute of quasi-periodic acoustic signals in speech, music, and other listening contexts [2, 3, 4]. For a quasi-periodic sound, the dominant pitch is almost invariably heard at its fundamental frequency f_0 . Common periodicity (“harmonicity,” sharing of common subharmonics), along with common onset, play very strong roles in grouping frequency components into auditory objects, and separating out multiple objects, each evoking its own pitch. Neural pitch mechanisms thus appear to be intimately related to early auditory grouping mechanisms, which in turn render analyses of multiple objects and streams in an auditory scene much more tractable. Human listeners are presently far superior to artificial, machine listening systems when it comes to tracking and analyzing sounds in noisy, cluttered, real world acoustic environments. If the operating principles inherent in neural mechanisms for pitch and auditory grouping can be understood and emulated, better artificial speech and music recognition systems and auditory prostheses are likely to follow. With this in mind, a method for extracting pitches of harmonic sounds is proposed, that may have parallels with signal processing strategies employed by auditory systems.

Currently three broad classes of f_0 pitch models exist: spectral pattern-matching models, residue models, and temporal autocorrelation models [5]. Spectral pattern-matching models first carry out a frequency analysis and then match patterns of resolved frequency components to harmonic spectral templates, so as to infer f_0 [6, 7]. Here, sinusoidal components that are exclusively represented in the output of a filter or channel are said to be “resolved,” whereas signal components that interact within the passband of a filter are referred to as “unresolved.” Residue models posit that f_0 pitch arises from (beating) interactions between (nearby) unresolved harmonics that are produced by broad cochlear filtering. A temporal

analysis of the resultant beating patterns produces an estimate of f_0 . Thus spectral pattern models predict only the f_0 pitches of resolved harmonics, whereas residue models predict only those of unresolved harmonics. Temporal autocorrelation models analyze patterns of all-order interspike intervals produced in the auditory nerve to identify patterns of interval peaks associated with different f_0 pitches [8, 9]. These models predict f_0 pitches produced by both resolved and unresolved harmonics. The underlying neuronal mechanisms proposed for temporal autocorrelation-based analysis and separation, utilize neural delay lines and coincidence detectors [10, 11, 12].

Meddis and co-workers [1, 13] have proposed a popular auto-correlation-based model that draws upon the original work of Licklider [10]. The model simulates cochlear action (cochlear bandpass filtering, transductive half-wave rectification, and synaptic low-pass filtering in each channel) to produce spike timing probability distributions (PSTHs, post-stimulus time histograms) for auditory nerve fibers (ANFs) of all characteristic frequencies (CFs). The autocorrelation of each fiber's PSTH is computed, and all of these auditory nerve frequency-channel autocorrelations are summed to produce the summary autocorrelation function (SACF) for the entire auditory nerve array. In effect, the SACF provides an autocorrelation-like representation of the acoustic signal. Major peaks in the SACF are identified, and the resultant f_0 pitch estimates successfully predict an extremely wide range of human pitch judgments. However, neuronal mechanisms by which the auditory system might analyze SACFs in the form of population-wide interspike interval statistics have yet to be found, motivating the search for alternative signal processing strategies that realize analysis operations similar to summary autocorrelation.

Recently, we have developed signal processing algorithms that emulate the synchrony capture phenomenon in the auditory nerve [14] that may afford alter-

native strategies for utilizing neural spike timing information. If one examines the representation of complex sounds in the auditory nerve, a striking feature is “synchrony capture,” wherein nerve fibers in an entire cochlear CF region are driven almost exclusively by one dominant local frequency component, (see, for example, [15]), such that the individual component imposes its (largely unmodulated) temporal fine structure (TFS) on the timing pattern of spikes in that region. At moderate and high sound pressure levels, a dominant harmonic can drive a large swath of auditory nerve fibers with CFs spanning an octave or more. For harmonics that are sufficiently separated, synchrony capture enhances their global temporal (interspike interval) representation by suppressing the temporal representation of beating interactions between harmonics. For harmonics closer together, within roughly a critical band or so, ANFs in surrounding CF regions are instead driven by the composite waveform pattern of the two interacting harmonics, such that the interspike interval representation of individual harmonics is severely degraded. Thus, neural synchrony capture appears to parallel perceptual frequency selectivity and harmonic resolution. Since resolved harmonics are known to permit separation of concurrent sounds by human listeners, synchrony capture in artificial systems may likewise be exploited for better sound separations.

The algorithm proposed here further develops the synchrony capture filterbank (SCFB) architecture presented in [14] and extends it to extract pitch frequencies of harmonic signals. The signal input may consist of resolved and/or unresolved components and additive noise. The key component of the algorithm, the SCFB architecture (see Figure 26), consists of a bank of broadly tuned filters (à la basilar membrane) in cascade with narrower filters (à la outer hair cells) that adaptively lock onto locally-dominant frequency components to produce synchrony capture behavior. The narrower filters constitute a frequency discriminator loop (FDL)

and are able to track individual time-varying frequency components, such as low harmonics and the dominant harmonics associated with formants in speech, in the midst of noise. In this article, we modified the algorithm to work even when there are unresolved tones in the input (see section 3.4).

Figure 26. a) Schematic of the Pitch Extraction Algorithm: The gammatone (GT) filters provide some spectral isolation in each channel. The frequency discriminator loop/phase locked loop (FDL/PLL) block achieves synchrony capture, i.e., the voltage controlled oscillator (VCO) in the block locks on to the strongest frequency component. The half wave rectified and low pass filtered (HWR/LPF) output of the channel is processed by the envelope (ENV) and the temporal fine structure (TFS) branches in parallel, which extract the envelope and phase align the dominant tonal signals, respectively. The outputs of all the channels, both the ENV and TFS branches, are summed to produce the Summary Phase Aligned Function (SPAF), the peak locations of which are used to determine f_0 . b) Phase alignment loop (PAL): The half wave rectified tone, $y(t)$, is delayed such that it overlaps symmetrically on the left ($w_L(t)$) and right ($w_R(t)$) windows (depiction shown in figure 28(b)). Analogous to the FDL, when $y(t)$ is centered around a time reference, t_0 , the error $e_t(t)$ goes zero and the loop reaches steady state. At this point, the half wave (HW) rectified tone is in cosine phase with respect to t_0 . See Section 3.5.

The schematic of the entire pitch extraction algorithm is shown in Figure 26 (a). The input signal, consisting of multiple, harmonically-related tones, is first filtered by a logarithmically spaced Gammatone (GT) filterbank. The output of each filter is then processed by a frequency discriminator loop/phase locked loop (FDL/PLL) block which determines the frequency of the dominant tone present in the passband of the associated GT filter. The details of this block are described in section 3.4 and shown in Figure 27. Each filter channel output is then halfwave-rectified and low pass filtered (by HWR/LPF block) and then delivered in parallel to the Envelope (ENV) branch and the Temporal Fine Structure (TFS) branch. In the ENV branch the HWR/LPF output is further low pass filtered to extract the envelope. In the TFS branch the halfwave rectified signals are aligned in phase with respect to a common time reference, t_0 , using a Phase Alignment Loop or PAL (see section 3.5). The FDL/PLL block, also provides a continuous estimate of the dominant frequency to the PAL. Once the channel outputs in the TFS branches are aligned in phase, the signals across the TFS *and* ENV branches of all channels are aggregated to obtain the Summary Phase Aligned Function (SPAF), which is then used to extract the f_0 information. Simulation results are presented in Section 3.6.

3.3 Two Main ideas: Duals in time and frequency

The proposed algorithm uses two basic signal processing strategies, one in the frequency domain and the other in time domain. Consider a sinusoidal signal $x(t) = A \cos(2\pi f_1 t + \theta_1)$. The first goal is to determine the frequency f_1 . It is determined using a frequency discriminator loop (FDL), the details are in Section 3.4. The basic principle is shown in Figure 28(a). The tone, shown as an impulse in Figure 28 (a), is fed as an input to stagger-tuned left ($H_L(f)$) and right ($H_R(f)$) bandpass filters (and a center filter $H_C(f)$). Assume that these

three frequency responses can be shifted in tandem along the frequency axis with the help of a VCO, whose frequency is adjustable using a feedback loop based on the difference in amplitude of the sinusoid at the output of $H_L(f)$ and $H_R(f)$. The loop ultimately settles to a steady state when the amplitude difference at the outputs of $H_L(f)$ and $H_R(f)$ is zero and the VCO frequency coincides with input frequency, i.e., $f_c = f_1$. Such FDLs have been used in automatic frequency control and communication systems for decades. We modified this basic FDL in Section 3.4 to work when the input consists of unresolved tones as well.

The second idea shown in Figure 28(b) is the time domain dual of the FDL. It is called a Phase Alignment Loop or PAL. The goal is to align the phase of the sinusoid such that it is in cosine phase with respect to an arbitrary time reference, t_0 . This adaptive phase alignment obviates the need to compute the autocorrelation function for each channel. In Figure 28(b) we assume that f_1 is known. The half wave rectified tone, named $y(t)$, is multiplied separately by the

Figure 27. FDL and PLL: Together they track the frequency of the dominant tone in the passband of the associated GT filter.

Figure 28. a) Frequency measurement using stagger tuned filters: The left $H_L(f)$, right $H_R(f)$ (and center $H_C(f)$) filters are tuned by adjusting the frequency f_c of a voltage controlled oscillator (VCO), which is embedded in a feedback loop (See Figure 27). When $f_c = f_1$, the tone's amplitude at the output of the right and left filters is equal, hence the loop reaches steady state. b) Phase measurement using staggered time windows: The half wave (HW) rectified tone, $y(t)$ (shown in bottom panel), is multiplied separately by the left ($w_L(t)$) and right ($w_R(t)$) windows. $y(t)$ is delayed using a feedback loop until the areas under $y(t)w_L(t)$ and $y(t)w_R(t)$ are equal, and the loop reaches steady state. Then, the delayed HW rectified tone is in cosine phase with respect to a time reference, t_0 .

left ($w_L(t)$) and right ($w_R(t)$) windows, which are centered around some arbitrary time reference, t_0 . The error signal, $e(t)$ is the difference in areas under the curves $y(t)w_L(t)$ and $y(t)w_R(t)$. Later $e(t)$ is smoothed and used to delay $y(t)$. Analogous to the FDL, when $y(t)$ is centered around t_0 , error approaches zero and the loop

reaches a steady state. See Section 3.5 for details. The delayed HW rectified tone is then in cosine phase with respect to t_0 .

3.4 Frequency tracking by the FDL/PLL block

The key element of the SCFB is the FDL/PLL block, the details of which are shown in Figure 27. It consists of three filters with frequency responses $H_L(f)$, $H_C(f)$, and $H_R(f)$ (which are spaced Δ Hz apart as shown in Figure 28 (a)). These filters are gang-tuned with the help of the VCO (See [14], our prior work for details). The difference between the log-amplitudes at the output of $H_L(f)$ and $H_R(f)$ (called FDL error) is used to adjust the VCO frequency which then moves these frequency responses in such a way to drive the log-amplitude difference to zero. If the input is a single tone then the FDL error approaches zero and the VCO frequency coincides with input frequency, i.e., $f_c = f_1$. These FDLs are used in cascade with a GT filterbank in the SCFB [14]. A key attribute of the FDL is that it exhibits the synchrony capture property similar to that seen in the auditory nerve. However, when the input signal consists of interfering sinusoids, i.e., more than one tone falls within the passband of a filter channel (unresolved case) the FDL tends to produce a biased estimate of the dominant frequency. Hence, we propose a combination of an FDL and a PLL which operates on the output of the center filter $H_C(f)$ (see Figure 27) to ameliorate this problem (details are described in [16]). The PLL shown within dashed lines in Figure 27 is a standard PLL, but the PLL error and the FDL errors are weighted to emphasize the importance of one or the other. For low frequency channels the GT filters are narrow and hence the interfering components are already sufficiently attenuated. Therefore, for these low frequency channels the FDL error alone is adequate. However, for high frequency channels the GT filters are wider and hence invariably have interacting tones within their passband. In this (high frequency channels) case, the FDL drives the VCO

within the lock-in range of the PLL and the PLL plays a vital role by homing in on the nearby dominant tone (See Figure 30 for simulation result). Hence, the PLL error is weighted more to reduce the bias in frequency estimate.

3.5 Phase Alignment Loop (PAL)

Figure 26 (b) shows the schematic of the PAL. The half wave rectified tone (output of the SCFB), $y(t)$, is multiplied by the left ($w_L(t)$) and right ($w_R(t)$) windows, which are centered around some arbitrary time reference, t_0 . These windows are supplied by the VCO at the rate of frequency f_1 , which is obtained from the FDL/PLL block. The area under the error waveform $e_t(t) = y(t)\{w_L(t) - w_R(t)\}$ is smoothed by the LPF and used to adjust the time delay δ . Analogous to the FDL, when the HW rectified tone $y(t)$ is centered around t_0 (actually, $t_0 + n/f_1$, where n is an integer), the error approaches zero and the loop reaches steady state. At that point the rectified tone is in cosine phase with respect to t_0 (simulation shown in Figure 29 (c)).

3.6 Simulation results

The pitch extraction algorithm has been tested with and without noise on several synthetic signals. Here, we present some results for a single tone, two unresolved tones and musical notes composed of several harmonic components, but without noise. The SCFB used here has 64 logarithmically spaced GT filters spanning 100 to 3827 Hz. The sampling frequency is 16 kHz. The filter Q values are all 6. Each GT filter is in cascade with the three filters in the FDL/PLL block, and each of the triplet filters have half the bandwidth of the GT filter. First, SCFB is input with a single tone at 800 Hz. Figure 29 (a) shows the converged frequency tracks of the VCOs of four channels in the neighborhood of 800 Hz (the center frequencies of the four GT filters and the VCOs' initial frequencies are 660, 696,

734, 773 Hz). As the tone is passed through the four GT filters it gets attenuated and phase delayed by each filter differently (figure 29 (b)). At the time reference, t_0 , the PALs start to phase align these four channel outputs. Figure 29 (b) & (c) shows the HW rectified tones before and after phase alignment. These phase aligned outputs (figure 29 (c)) can be coherently added (SPAF) to obtain the pitch of the tone.

Figure 29. Single tone: (a) shows the frequency trajectories of the 4 VCOs near 800 Hz, (b) and (c) show the HW rectified channel outputs without and with phase alignment, respectively.

In the second example, the input signal consists of two equal amplitude tones at frequencies 2300 and 2500 Hz. Around 2400 Hz, the 10 dB bandwidth of the GT filters is about 200 Hz. Since the frequencies fall in the same filter, this is an example of the unresolved case. Figures 30 shows the VCO frequency tracks for four channels with center frequencies 2234, 2338, 2447, and 2560 Hz, with and without the inclusion of PLL error, figure 30 (a) & (b) respectively. It can be seen that the bias in frequency estimates is zero when the PLL error is included, and there is a noticeable bias when PLL error is included {figure 30 (a) & (b) respectively}.

Figure 30. Unresolved tones: Tracks of VCO frequencies for CFs around 2400 Hz with PLL (a) and without PLL (b, notice the bias.)

The final example is a musical note with resolved harmonics, here the sum of the first six harmonics of the fundamental $f_0 = 440$ Hz. The tones are chosen to have a Schroeder phase, that is, the input is $x(t) = \sum_{k=1}^6 \cos(2\pi k f_0 t + \pi k^2/6)$. Such signals typically exhibit little envelope variability in a pitch period. Figure 31 shows the frequency tracks of the VCOs of different channels. The synchrony capture phenomenon, (i.e., the VCO frequencies in the neighborhood of a dominant tone lock on to that tone’s frequency,) is obvious. The sum of all the phase aligned signals, i.e., the SPAF is shown in Figure 32. The SPAF clearly shows peaked periodic pattern (unlike the input signal) and the pitch information can be extracted from the SPAF by finding the interval between the largest peaks.

3.7 Conclusion

We proposed a new Summary Phase Aligned Function (SPAF) as an alternative to Summary Autocorrelation Function (SACF) for computing the fundamental

Figure 31. Frequency tracks of the Schroeder-phase signal

Figure 32. SPAF of Schroeder-phase signal

frequency f_0 of a periodic signal. SCAF requires autocorrelation computations and SPAF does not. We also modified our previous SCFB algorithm [14] to improve tone resolution. The simulation results on synthetic signals are promising but need to be performed on real world signals.

List of References

- [1] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2866–2882, 1991.
- [2] C. J. Plack and A. J. Oxenham, "Psychophysics of Pitch," in *Pitch: Neural Coding and Perception*, ser. Springer Handbook of Auditory Research, C. J. Plack, A. J. Oxenham, R. R. Fay and A. N. Popper, Ed. New York, NY: Springer-Verlag, 2005, ch. 2, pp. 7–55.
- [3] A. de Cheveigne and H. Kawahara, "Yin: A fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, 2002.
- [4] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, pp. 684–697, May 1999.
- [5] A. de Cheveigne, "Pitch perception," in *The Oxford Handbook of Auditory Science*, C. J. Plack, Ed. Oxford University Press Inc, New York: Oxford, 2010, pp. 71–104.
- [6] J. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1496–1516, 1973.
- [7] P. Srulovicz and J. Goldstein, "A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum," *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1266–1276, 1983.
- [8] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. I. pitch and pitch salience. II. pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch," *J. Neurophysiology*, vol. 76, no. 3, pp. 1698–1734, 1996.
- [9] R. Meddis and M. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 233–245, 1992.
- [10] J. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. VII, no. 4, pp. 128–134, 1951.
- [11] P. Cariani, "Neural timing nets," *Neural Networks*, vol. 14, no. 6-7, pp. 737–753, 2001.
- [12] P. Cariani, "A temporal model for pitch multiplicity and tonal consonance," in *Proceedings of the 8th International Conference on Music Perception & Cognition (ICMPC8, Evanston, IL)*, S. Lipscomb, R. O. Ashley, R. O. Gjerdingen,

- and P. Webster, Eds. Adelaide, Australia: Causal Productions, 2004, pp. 310–314.
- [13] R. Meddis and L. O’Mard, “A unitary model of pitch perception,” *J Acoust Soc Am*, vol. 102, no. 3, pp. 1811–1820, 1997.
 - [14] R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, “Synchrony Capture Filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech,” *J. Acoust. Soc. Am.*, vol. 133, Issue 6, pp. 4290–4310, 2013.
 - [15] B. Delgutte and N. Kiang, “Speech coding in the auditory nerve: I. Vowel-like sounds,” *J. Acoust. Soc. Am.*, vol. 75, pp. 866–878, 1984.
 - [16] R. Kumaresan, V. K. Peddinti, and P. Cariani, “Synchrony Capture Filterbank II: Auditory-inspired pitch extraction using synchrony capture filterbank and phase alignment (To be submitted),” *J. Acoust. Soc. Am.*, 2013.

MANUSCRIPT 4

To be submitted to The Journal of the Acoustical Society of America

**Synchrony capture filterbank II: Auditory-inspired pitch extraction
using synchrony capture filterbank and phase alignment**

Ramdas Kumaresan¹, Vijay Kumar Peddinti¹, Peter Cariani²

¹ Department of Electrical, Computer and Biomedical Engineering,
University of Rhode Island, Kingston RI 02881

² Department of Otology & Laryngology, Harvard Medical School,
Boston, MA 02114

Corresponding Author: Ramdas Kumaresan

Department of Electrical, Computer-
and Biomedical Engineering
University of Rhode Island
Kingston, RI 02881

Phone: 401-874-4798

Email: kumar@ele.uri.edu

Synchrony capture filterbank II: Auditory-inspired pitch extraction using synchrony capture filterbank and phase alignment

4.1 Abstract

We developed a synchrony capture filterbank (SCFB) architecture that emulates temporal patterning of auditory nerve fiber spike trains in order to precisely and robustly track individual frequency components. The architecture implements an array of bandpass filters followed by adaptively-tuned filters that track component frequencies to achieve enhanced signal analysis and better separation of concurrent harmonic sounds, both from each other and from background noise. In this paper, we present an improved version of SCFB, and an in-depth analysis of the parameters. Primarily, the update includes an addition of a phase locked loop (PLL) to the existing frequency discriminator loop (FDL) block to eliminate any bias in the frequency estimates. The improved algorithm will determine the frequencies accurately even in the presence of unresolved tones in the input. The resulting frequency estimates were used to determine pitch or fundamental frequency f_0 . The results of the improved SCFB (with out the analysis) appeared in an International Conference on Acoustics, Speech and Signal Processing (ICASSP) publication [1].

4.2 Introduction

The basic structure of our synchrony capture filterbank (SCFB) algorithm is shown in Figure 33. The SCFB is a bank of broadly tuned filter (à la basilar membrane) in cascade with narrower filters (à la outer hair cells) that adaptively lock onto locally-dominant frequency components to produce synchrony capture like behavior. The three tunable bandpass filters (BPFs) together form a frequency discriminator loop (FDL). This resulting architecture can track individual frequency components at low frequencies and the dominant frequency components at high frequency. Gammatone (GT) filters are used in the first part (as broad filters) of the SCFB architecture, it is followed by three narrow filters as shown in Figure 33(b) and (c) [2]. More details of this architecture are in [2] or manuscript 2 of this thesis.

One setback of this preliminary SCFB structure is a slight bias in the frequency estimates, when unresolved signal components are present in the passband of the associated gammatone (GT) filter. In other words, more than one frequency is present in the corresponding passband of the GT filter. To show the bias, let us consider an input signal with two equal amplitude tones at frequencies 2300 and 2500 Hz. Around 2400 Hz, the 10 dB bandwidth of the GT filters is about 200 Hz. Since both tones are passed by this GT filter, this is an example of the unresolved case. Figure 34 shows frequency tracks of the voltage controlled oscillators (VCOs) of four channels with center frequencies 2234, 2338, 2447, and 2560 Hz. In Figure 34, the bias is clearly noticeable, the dotted line shows the actual frequency value and the solid line (red) shows the frequency estimate. There is approximately 15 Hz difference between the actual and the mean value of the frequency estimate.

To alleviate this problem, the existing frequency discriminator loop (FDL) in Figure 33(a) is replaced with a combination of a frequency discriminator loop

Figure 33. Synchrony capture filterbank (SCFB). (a) The filterbank architecture consists of K constant- Q gammatone filters whose logarithmically-spaced center frequencies span the desired audible frequency range. Each filterbank channel consists of a frequency discriminator loop (FDL) cascaded with each of the K gammatone filters. The output of each channel, $y_c(t)$, is obtained from its center filter. Frequency responses of fixed and tunable filters in the SCFB. Bottom left panel (b) shows the frequency responses of fixed gammatone filters (the black dots indicate that not all filter responses are shown). Bottom right panel (c) shows the frequency responses of the tunable bandpass filter (BPF) triplets that adapt to the incoming signal. One BPF triplet is associated with each fixed filter, such that coarse filtering of the fixed gammatone filters is followed by additional, finer filtering by tunable filters [2]. More details of this architecture are in [2] or manuscript 2 of this thesis.

(FDL) and phase locked loop (PLL) as shown in Figure 35. Tracking frequencies precisely is crucial for fundamental frequency/pitch (f_0) calculation [1]. This FDL/PLL block operates on the output of the center filter $H_C(f)$ as shown in Figure 35. The phase locked loop (PLL) shown within dashed lines in Figure 35 is a standard PLL. A weighted sum of the FDL and PLL errors is used to track the dominant frequencies in the signal. By using a weighted sum of the FDL and

Figure 34. Shows the bias in the voltage controlled oscillator (VCO) tracks for unresolved tones; here $f_1 = 2300$ Hz, $f_2 = 2500$ Hz, however the mean value of the estimates are off about approximately 15 Hz.

Figure 35. Improved frequency discriminator loop (FDL) block, which shows the addition of a phase locked loop (PLL). A weighted sum of the FDL and PLL errors is used to track the frequency of the dominant tone in the passband of the associated gammatone (GT) filter [1].

PLL errors, the importance of one or the other is emphasized accordingly. For low frequency channels the GT filters are narrow and hence the interfering components are already sufficiently attenuated. Hence, for these channels the FDL error alone is adequate for tracking these individual frequencies. However, for high frequency

channels the GT filters are wider and hence invariably have interacting tones within their passband. In this case, the FDL drives the VCO within the lock-in range of the PLL and the PLL homes in on the nearby dominant tone. In this case the PLL error is weighted more to reduce the bias in frequency estimate. Figure 36 shows the updated VCO frequency tracks for the same input signals presented in Figure 34. This figure clearly shows that the mean value of the frequency estimates is equal to the actual frequencies in the input signal with out any bias, however there more variance than the earlier case.

Figure 36. Frequency tracks of voltage controlled oscillator (VCO) using improved frequency discriminator loop (FDL) block. The mean value of the frequency estimates are same as the actual frequencies $f_1 = 2300$ Hz, $f_2 = 2500$ Hz. However it does add some variance to the estimates as compared to Figure 34.

4.3 Pitch extraction

Once the frequencies are tracked precisely, a phase alignment loop (PAL) is used to compute the fundamental frequency or pitch (f_0). More details about PAL are in [1] or manuscript 3 of this thesis. The idea is to adaptively phase align the filter outputs to a common time reference and sum them together. We referred to the resulting signal as “summary phase aligned function (SPAF)” [1] from which

the fundamental frequency or pitch (f_0) can be computed.

The main principle behind PAL is illustrated in Figure 37. Here, the input signal to SCFB is a single tone at 800 Hz. Figure 37 (a) shows the frequency tracks of four VCOs around 800 Hz (with 660, 696, 734, 773 Hz as their center frequencies). When the tone is subjected to the four gammatone (GT) filters, the resulting outputs get attenuated as well as phase delayed differently as shown in Figure 37 (b). At the time reference, t_0 , ($=0.2$ seconds, not shown in the figure) the PALs start to phase align these channel outputs. Figure 37 (b) & (c) shows the HW rectified tones before and after phase alignment. The phase aligned outputs presented in Figure 37 (c) are coherently added to obtain the summary phase aligned function (SPAF), from which the f_0 of the tone is obtained.

Figure 37. Single tone input at 800 Hz: (a) shows the frequency trajectories of the 4 VCOs near 800 Hz, (b) and (c) show the half-wave (HW) rectified channel outputs without and with phase alignment, respectively [1].

The complete pitch extraction architecture using the improved frequency discriminator loop (FDL) and phase alignment loop (PAL) is shown in Figure 38.

In the next section the design parameters for the frequency discriminator loop

Figure 38. Pitch extraction algorithm using the improved synchrony capture filter-bank (SCFB). The first stage of this architecture is the improved FDL (FDL/PLL). These outputs are phase aligned by phase alignment loop (PAL), which are added to produce the summary phase aligned function (SPAF), from which the f_0 can be determined.

(FDL) and phase locked loop (PLL) are presented.

4.4 Design parameters for the frequency discriminator loop (FDL)

The baseband equivalent circuit of the FDL is shown in Figure 39. The low

Figure 39. Baseband equivalent circuit of the frequency discriminator loop (FDL).

pass filter (LPF), which is part of the envelope detector, is relatively broad and hence is not part of the equivalent circuit. The left and the right BPFs and the envelope detectors have been replaced by the linear discriminator (LD) with transfer function $k_s s$. Let us redraw this architecture in a simplified circuit shown

in Figure 40.

Figure 40. Generic feedback circuit.

By comparing Figures 39 and 40

$$G(s) = \left(\frac{1}{s + \alpha} \right) (k_s s) \left(\frac{1}{s} \right) \left(k_1 + \frac{k_2}{s} \right) = \left(\frac{k_s}{s + \alpha} \right) \left(k_1 + \frac{k_2}{s} \right) \quad (33)$$

Let us check the stability of the system, before deriving the parameters. To compute the steady state error, the error transfer function $E(s)$ is shown below.

$$\begin{aligned} E(s) = \frac{\omega_e(s)}{\omega_i(s)} &= X(s) \frac{1}{1 + G(s)} = \frac{X(s)}{1 + \frac{k_s}{s + \alpha} \left(k_1 + \frac{k_2}{s} \right)} = \frac{X(s)s(s + \alpha)}{s(s + \alpha) + k_s(k_1 s + k_2)} \\ &= X(s) \frac{s(s + \alpha)}{s^2 + s(\alpha + k_1 k_s) + k_2 k_s} \end{aligned} \quad (34)$$

Using the final value theorem, $\lim_{t \rightarrow \infty} e(t) = \lim_{s \rightarrow 0} sE(s)$, the steady state frequency error can be computed. For a step input: $w_i(t) = u(t) \leftrightarrow X(s) = \frac{1}{s}$, the output is the following.

$$\begin{aligned} e(\infty) &= \lim_{s \rightarrow 0} sX(s) \frac{s(s + \alpha)}{s^2 + s(\alpha + k_1 k_s) + k_2 k_s} \\ &= \lim_{s \rightarrow 0} s \left(\frac{1}{s} \right) \frac{s(s + \alpha)}{s^2 + s(\alpha + k_1 k_s) + k_2 k_s} \\ &= \lim_{s \rightarrow 0} \frac{s(s + \alpha)}{s^2 + s(\alpha + k_1 k_s) + k_2 k_s} = 0 \end{aligned} \quad (35)$$

Hence, the final value of $w_e(t)$, i.e., $\lim_{t \rightarrow \infty} w_e(t) = 0$ or the steady state frequency error is zero.

Similarly for a ramp input, $w_i(t) = tu(t) \leftrightarrow X(s) = \frac{1}{s^2}$,

$$\begin{aligned}
e(\infty) &= \lim_{s \rightarrow 0} sX(s) \frac{s(s + \alpha)}{s^2 + s(\alpha + k_1k_s) + k_2k_s} \\
&= \lim_{s \rightarrow 0} s \left(\frac{1}{s^2} \right) \frac{s(s + \alpha)}{s^2 + s(\alpha + k_1k_s) + k_2k_s} \\
&= \lim_{s \rightarrow 0} \frac{(s + \alpha)}{s^2 + s(\alpha + k_1k_s) + k_2k_s} = \frac{\alpha}{k_2k_s}
\end{aligned} \tag{36}$$

Hence, the steady state error is a constant ($= \frac{\alpha}{k_2k_s}$).

Now, to compute the loop filter parameters: proportional constant (k_1), integral constant (k_2), and the bandwidth, the closed loop transfer function $L(s)$ is computed. Its numerical value provides a clue to the loop characteristics.

$$\begin{aligned}
L(s) &= \frac{\omega_c(s)}{\omega_i(s)} = \frac{G(s)}{1 + G(s)} = \frac{\frac{k_s}{s + \alpha} \left(k_1 + \frac{k_2}{s} \right)}{1 + \frac{k_s}{s + \alpha} \left(k_1 + \frac{k_2}{s} \right)} = \frac{k_s(k_1s + k_2)}{s(s + \alpha) + k_s(k_1s + k_2)} \\
&= \frac{k_1k_ss + k_2k_s}{s^2 + s(\alpha + k_1k_s) + k_2k_s}
\end{aligned} \tag{37}$$

Let us rewrite $L(s)$ in a simplified form for convenience: $L(s) = \frac{c_1s + c_0}{d_2s^2 + d_1s + d_0}$, where $c_1 = k_1k_s$, $c_0 = k_2k_s$, $d_2 = 1$, $d_1 = (\alpha + k_1k_s)$, $d_0 = k_2k_s$. Often in loop design the equivalent noise bandwidth, B_n , of the loop is used as a design parameter and is defined as [3]

$$B_n = \frac{\int_0^\infty |L(j\omega)|^2 d\omega}{|L(0)|^2} \tag{38}$$

B_n has been analytically calculated for system functions up to order 4 in [4] (page 21), and is given by

$$B_n = \frac{c_1^2d_0 + c_0^2d_2}{4d_0d_1d_2}. \tag{39}$$

Hence the equivalent noise bandwidth, B_n , for the corresponding $L(s)$ becomes the following:

$$B_n = \frac{(k_1k_s)^2k_2k_s + (k_2k_s)^21}{4k_2k_s(\alpha + k_1k_s)1} = \frac{(k_1k_s)^2 + k_2k_s}{4(\alpha + k_1k_s)} \tag{40}$$

Further, by comparing the denominator of Eq.(37), given by $(s^2 + s(\alpha + k_1 k_s) + k_2 k_s)$ to the denominator of a standard second order transfer function, $(s^2 + 2\zeta\omega_n s + \omega_n^2)$, where ζ is the damping ratio and ω_n is the natural frequency, we can infer the following

$$\omega_n^2 = k_2 k_s \quad (41)$$

$$2\zeta\omega_n = \alpha + k_1 k_s \Rightarrow k_1 k_s = 2\zeta\omega_n - \alpha \quad (42)$$

$L(s)$ can be expressed in terms of ζ and ω_n , by substituting Eqs. 41 and 42 in Eq. 37 and is given by

$$L(s) = \frac{k_1 k_s s + k_2 k_s}{s^2 + s(\alpha + k_1 k_s) + k_2 k_s} = \frac{(2\zeta\omega_n - \alpha)s + \omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2}. \quad (43)$$

Similarly, B_n in terms of ζ and ω_n is given by

$$B_n = \frac{(k_1 k_s)^2 + k_2 k_s}{4(\alpha + k_1 k_s)} = \frac{(2\zeta\omega_n - \alpha)^2 + \omega_n^2}{8\zeta\omega_n}. \quad (44)$$

Assuming a value for B_n and ζ (is typically equal to 1), we can calculate the natural frequency, ω_n , and the proportional and integral constants k_1 and k_2 .

$$\begin{aligned} \frac{(2\zeta\omega_n - \alpha)^2 + \omega_n^2}{8\zeta\omega_n} &= B_n \\ (1 + 4\zeta^2)\omega_n^2 + \alpha^2 - 4\zeta\omega_n\alpha &= 8\zeta\omega_n B_n \\ (1 + 4\zeta^2)\omega_n^2 - 4\zeta(\alpha + 2B_n)\omega_n + \alpha^2 &= 0 \end{aligned} \quad (45)$$

ω_n is the positive root of the quadratic equation from Eq. 45. k_1 and k_2 are derived using the relation from Eqs. 41 and 42. The results are as follows.

$$\omega_n = \frac{4\zeta(\alpha + 2B_n) \pm \sqrt{(4\zeta(\alpha + 2B_n))^2 - 4(1 + 4\zeta^2)\alpha^2}}{2(1 + 4\zeta^2)} \quad (46)$$

$$k_1 = \frac{2\zeta\omega_n - \alpha}{k_s} \quad (47)$$

$$k_2 = \frac{\omega_n^2}{k_s} \quad (48)$$

Next let us consider the design parameters for the phase locked loop (PLL).

4.5 Design parameters for the phase locked loop (PLL)

The procedure for choosing the parameters of the phase locked loop (PLL) is similar to the procedure outlined in previous section 4.4. The equivalent PLL circuit is shown in Figure 41.

Figure 41. Analog equivalent circuit of phase locked loop (PLL).

The loop transfer function $H(s)$ is

$$\begin{aligned}
 H(s) = \frac{\hat{\theta}(s)}{\theta(s)} &= \frac{k'_0 k'_p \left(k'_1 + \frac{k'_2}{s} \right)}{s + k'_0 k'_p \left(k'_1 + \frac{k'_2}{s} \right)} \\
 &= \frac{k'_0 k'_1 k'_p s + k'_2 k'_0 k'_p}{s^2 + k'_0 k'_1 k'_p s + k'_2 k'_0 k'_p}, \tag{49}
 \end{aligned}$$

where k'_1 and k'_2 are the proportional and integral constants, k'_p and k'_0 are the phase detector and voltage controlled oscillator (VCO) constants. Here k' s (k primes) are used to distinguish them from the FDL parameters. $H(s)$ can be expressed in the standard form, by writing $k'_0 k'_2 k'_p = \omega_n'^2$ and $k'_0 k'_1 k'_p = 2\zeta' \omega_n'$.

$$H(s) = \frac{2\zeta' \omega_n' s + \omega_n'^2}{s^2 + 2\zeta' \omega_n' s + \omega_n'^2} \tag{50}$$

The equivalent noise bandwidth B'_n for phase locked loop (PLL) is derived in the

same fashion as in the frequency discriminator loop (FDL) derivations.

$$\begin{aligned} B'_n &= \frac{(2\zeta'\omega'_n)^2\omega_n'^2 + \omega_n'^2(1)}{4\omega_n'^2(2\zeta'\omega'_n)(1)} \\ &= \frac{\omega_n'}{2}(\zeta' + \frac{1}{4\zeta'}). \end{aligned} \quad (51)$$

Therefore, the natural frequency, ω'_n of PLL is

$$\omega'_n = \frac{2B'_n}{(\zeta' + \frac{1}{4\zeta'})}. \quad (52)$$

Since $k'_0 k'_2 k'_p = \omega_n'^2$ and $k'_0 k'_1 k'_p = 2\zeta'\omega'_n$, k'_1 and k'_2 in terms of natural frequency, ω'_n are the following:

$$\begin{aligned} k'_1 &= \frac{2\zeta'\omega'_n}{k'_0 k'_p} \\ k'_2 &= \frac{\omega_n'^2}{k'_0 k'_p}. \end{aligned} \quad (53)$$

The final expressions for proportional and integral constants, k'_1 and k'_2 in terms of bandwidth, B'_n and the damping factor, ζ' are obtained by substituting Eq. 52 in Eq. 53. The resulting expressions are as follows:

$$k'_1 = \frac{2\zeta'}{k'_0 k'_p} \left(\frac{2B'_n}{(\zeta' + \frac{1}{4\zeta'})} \right) \quad (54)$$

$$k'_2 = \frac{1}{k'_0 k'_p} \left(\frac{2B'_n}{(\zeta' + \frac{1}{4\zeta'})} \right)^2 \quad (55)$$

Assuming a value for $\zeta' = 1$, and let $k'_0 k'_p = c'$, a constant (k'_0 - phase detector and k'_p voltage controlled oscillator constant). We can simplify and express the proportional and integral constants, k'_1 and k'_2 as below:

$$k'_1 = \frac{3.2}{c'} B'_n \quad (56)$$

$$k'_2 = \frac{2.56}{c'} B_n'^2 \quad (57)$$

These parameters can also be obtained using an alternative approach outlined by Rice in [5] (page 732). The procedure is as follows: (1) The PLL analog

equivalent circuit, $H(s)$, and discrete equivalent circuit, $H(z)$, are determined. (2) The discrete equivalent function from $H(s)$ is calculated by substituting the bilinear transform, $s = \frac{2(1-z^{-1})}{T(1+z^{-1})}$ in the analog equivalent function. (3) Finally to derive an expression for the parameters, the denominators of these two discrete equivalent functions are equated. Only the denominator is considered because the bilinear transform preserves only the poles in the Z-domain, but not the zeros. We noticed that both methods resulted in the same expressions.

4.6 Simulations

The improved SCFB implemented for these simulations has $K = 64$ logarithmically spaced gammatone (GT) filters spanning 100 to 3827 Hz. The sampling frequency is 16 kHz. These filters are designed using the Auditory Toolbox developed by Malcolm Slaney [6]. The filter Q values are all 6. Each GT filter is in cascade with the three filters in the FDL/PLL block, and each of the triplet filters have half the bandwidth of the GT filter.

The improved SCFB is subjected to the same signals as our original SCFB algorithm [2] (equivalently manuscript 2 of this thesis). The speech signals were still first pre-emphasized with a $H(z) = 1 - 0.95z^{-1}$ filter to equalize the spectrum to prevent strong low frequency components from swamping the weaker high frequency components. The resulting frequency tracks overlapped on the spectrograms (capturegrams) for two speech sentences, “Where were you while we were away?” (TIMIT sentence sx9, speakers mpcs0 and fgjd0) and “The oasis was a mirage” (TIMIT sentence sx280, speakers mdwk0 and fawf0) spoken by male and female speakers are shown in Figures 42 and 43, respectively.

While Figure 43 and 44 shows the plots in a 10 dB noise. The Gaussian distributed noise samples are generated using a variance σ^2 , obtained from the signal to noise ratio (SNR) definition as below.

$$SNR = 10 \log_{10} \left(\frac{P_s}{\sigma^2} \right) dB.$$

Where P_s is signal power P_s , obtained as the sum of squares of all the speech signal samples divided by the time duration of the speech signal. The resulting plots show that the estimated frequency tracks are accurate even in the presence of noise.

Figure 42. Results for TIMIT utterance, “Where were you while we were away?” (sx9) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are capturegrams (frequency tracks superimposed on spectrograms). At low frequencies, all individual harmonics are tracked, whereas above 1000 Hz, only prominent formant harmonics are tracked.

Figure 43. Results for TIMIT utterance, “The oasis was a mirage” (sx280) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are Capturegrams(frequency tracks superimposed on spectrograms). High frequency frication above 4000 Hz in “oasis” not shown.

Figure 44. Results for TIMIT utterance in 10dB noise, “Where were you while we were away?” (sx9) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are capturegrams (frequency tracks superimposed on spectrograms).

Figure 45. Results for TIMIT utterance in 10dB noise, “The oasis was a mirage” (sx280) for male (left column) and female (right column) speakers. Top plots (a)(c) are spectrograms. Middle plots (b)(d) are Capturegrams(frequency tracks superimposed on spectrograms).

The pitch extraction algorithm has been tested with and without noise on several synthetic and speech signals. Some of the results were shown in [1] or 3 of this thesis. A few additional results are illustrated here.

Figure 46 (a) shows the frequency tracks for an input signal with two frequencies, 440 and 587 Hz (and the first six harmonics of each frequency), first segment has a frequency 440 Hz from 0 to 0.425 seconds and the second segment has 587 Hz from 0.425 to 1 second respectively.

Figure 46 (b) shows the corresponding summary phase aligned function (SPAF), and a zoomed is shown at the bottom of the panels. They exhibit a clear peaked periodic pattern. The f_0 from the SPAF spacing in these two segments is 440 and 588 Hz; while the actual frequencies in the input signal are: 440 and 587 Hz.

Figure 47 shows the result for the same signal in 10 dB noise. The SPAF still shows a peaked periodic pattern and the spacing still came out to 440 and 588 Hz.

Finally, Figure 48 shows the result for speech utterance: “The oasis was a mirage,” (sx280) taken from the TIMIT database. Even in this case, the SPAF still shows a periodic pattern, and the spacing varied around 100 Hz.

Figure 46. (a) Frequency tracks (b) SPAF for a two tone segment, first segment has a frequency 440 Hz from 0 to 0.425 seconds and the second segment has 587 Hz from 0.425 to 1 second respectively.

Figure 47. (a) Frequency tracks (b) SPAF for a two tone segment in 10 dB noise, first segment has a frequency 440 Hz from 0 to 0.425 seconds and the second segment has 587 Hz from 0.425 to 1 second respectively.

Figure 48. (a) Frequency tracks (b) SPAF for a speech utterance (male speaker) from TIMIT database “The oasis was a mirage” (sx280.wav).

4.7 Conclusions

An in-depth analysis of the parameters corresponding to our improved synchrony capture filterbank (SCFB) is presented in this paper. This improved SCFB tracks the frequencies precisely with out any bias even in the presence of unresolved tones in the input. This bias was observed in our original SCFB algorithm. In turn this improvement allows us to compute pitch accurately. More analysis is necessary for pitch determination for speech, especially in noise. Even though SPAF avoids the computation of auto-correlation, it still requires the calculation of the period between the peaks. Currently, it is determined by peak picking, which needs to be addressed.

List of References

- [1] R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, “Auditory-inspired pitch extraction using a synchrony capture filterbank and phase alignment,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, pp. 5959 – 5963.
- [2] R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, “Synchrony Capture Filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech,” *J. Acoust. Soc. Am.*, vol. 133, Issue 6, pp. 4290–4310, 2013.
- [3] F. M. Gardner, *Phaselock Techniques*. John Wiley & Sons, Inc., 1966.
- [4] J. Klapper and J. T. Frankle, *Phase-Locked and Frequency-Feedback Systems*. New York, NY: Academic Press, 1972.
- [5] M. Rice, *Digital Communications: A Discrete-Time Approach*, 1st ed. Prentice Hall, 2008.
- [6] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice , “Implementing a gammatone filter bank,” in *Annex C of the SVOS Final Report (Part A: The Auditory Filter Bank) MRC (Medical Research Council), APU (Applied Psychology Unit) Report 2341*, University of Cambridge, Cambridge, United Kingdom, Feb 1988.

APPENDIX A

Published ICASSP2011 paper

Our paper titled “Multiple pitch identification using cochlear-like frequency capture and harmonic grouping” appeared in the Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech and Signal Processing (ICASSP) publication [1]. It is presented here in the original format.

List of References

- [1] R. Kumaresan, Vijay Kumar 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.

MULTIPLE PITCH IDENTIFICATION USING COCHLEAR-LIKE FREQUENCY CAPTURE AND HARMONIC GROUPING

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

This work addresses the problem of identifying multiple fundamental frequencies in an acoustic signal. An auditory-inspired peripheral signal processing model is proposed that functions in a manner more like a bank of FM receivers rather than a traditional filterbank. Such receivers lock on to a strong signal (synchrony capture, frequency capture) even in the presence of nearby only slightly weaker signal components. Once the individual signal components are resolved, the model subjects them to an instantaneous nonlinearity and then performs harmonic grouping by cross correlating the isolated components. After the harmonically-related components are grouped, their pitches are computed using a standard summary autocorrelation approach.

Index Terms— Pitch, Harmonics, Cochlea

1. INTRODUCTION

Pitch is an essential attribute of periodic acoustic signals in speech, music, and other listening contexts. For a periodic sound, pitch is almost invariably heard at its fundamental frequency f_0 . Common onset and common periodicity ("harmonicity") are the two strongest factors in grouping frequency components into auditory objects, and separating out multiple objects, each with its own pitch. Neural pitch mechanisms thus appear to be intimately related to early auditory grouping mechanisms, which in turn render analyses of multiple objects and streams in the auditory scene much more tractable. Human listeners are presently far superior to artificial, machine listening systems when it comes to tracking and analyzing sounds in noisy, cluttered, real world acoustic environments. If the operating principles inherent in neural mechanisms for pitch and auditory grouping can be understood, better artificial speech and music recognition systems and auditory prostheses are likely to follow. With this in mind a method for identifying pitches of multiple sets of harmonic

sounds is proposed that may have parallels with signal processing strategies that are employed by auditory systems.

Three major classes of pitch models exist: spectral pattern-matching models, residue models, and temporal autocorrelation models [1]. Spectral pattern-matching mechanisms analyze patterns of resolved frequency components and group harmonically-related components into separate auditory objects (via template matching [2], neural nets, or subharmonic superpositions). Residue models rely on interactions of unresolved harmonics (beating) produced by broad cochlear filtering and carry out a temporal analysis of the resultant periodicities. Temporal autocorrelation and cancellation models analyze patterns of all-order interspike intervals produced in the auditory nerve to identify interval peaks [3, 4] or subpatterns associated with different pitches.

Spectral pattern theories depend entirely on resolved harmonics, and therefore cannot explain the somewhat weaker fundamental pitches produced by unresolved harmonics. Residue theories depend entirely on waveform interactions between unresolved harmonics, and therefore do not explain the stronger pitches produced by resolved harmonics. Temporal autocorrelation models operate on interspike intervals produced by both resolved and unresolved harmonics and pure tones, and therefore provide a unified account of pitches associated with periodicities below 4-5 kHz. [3].

Psychophysical research [5] indicates that separations of multiple auditory objects with different fundamentals crucially depend on the presence of resolved, non-interfering low harmonics. This implies some neural mechanism that depends on the separation of harmonics prior to the operation of harmonic grouping and formation of auditory objects.

Temporal discharge patterns in the auditory nerve provide a neurally- and psychophysically-plausible basis for central neural representations of frequency and periodicity, be they in either the frequency- [2] or time-domain [3]. Here the precise nature of temporal representations of resolved and unresolved harmonics can shed light on the neural mechanisms underlying separation of multiple concurrent harmonic sounds. If one examines the representation of low harmonics of complex

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

sounds in the auditory nerve, a striking feature is "synchrony capture", wherein nerve fibers in an entire cochlear region are driven almost exclusively by one local frequency component [6], such that the individual component imposes its temporal fine structure on the temporal patterning of spikes in that region. Synchrony capture occurs at moderate and high sound pressure levels, where auditory nerve fibers are typically driven at their maximal rates over a range of pure tone frequencies of an octave or more. For low or resolved harmonics, synchrony capture enhances the interspike interval representations of the individual harmonics in fibers whose characteristic frequencies (CF's) are nearby. When harmonics are close together in frequency, within a critical band or so, surrounding auditory nerve CF regions are instead driven by the composite waveform pattern of the two interacting harmonics, and the interspike interval representation of individual harmonics is degraded. Thus, neural synchrony capture appears to parallel perceptual harmonic resolution. Because resolved harmonics permit separation of concurrent sounds, synchrony capture in artificial systems may likewise be exploited for better sound separations.

2. SYNCHRONY/FREQUENCY CAPTURE MODEL

With this in view, a model is proposed for peripheral processing that uses a synchrony capture mechanism to effect the capture of individual frequency components. Rather than a simple filter bank, the mechanism behaves more like a bank of FM receivers, and hence tends to capture the strong frequency components in a signal and mask weaker ones. Such receivers are known to lock on (or frequency capture [7]) to a strong FM signal even in the presence of nearby, only slightly weaker signals. After individual components are separated via frequency capture, a harmonic grouping operation is performed by a Harmonic Relation Detector (HARD) (block in Figure 1). Here the channel outputs are subjected to a non-linear distortion, mutually cross-correlated, and then tested to determine if they are harmonically-related.

In the following, a simple adaptive signal processing approach is proposed that exhibits this type of frequency/synchrony capture behavior. The proposed implementation is motivated by the anatomy and biophysics of the cochlea, particularly the crystalline structure of the three rows of outer hair cells (OHCs) that ride on top of the basilar membrane. Motivated by these observations, an adaptive filter structure consisting of three bandpass filters (BPF) was envisioned, provisionally called a **BPF triplet**, that can home in on a dominant tone in an input signal. This idea was then used to synthesize a composite Synchrony Capture Filterbank (SCFB) that can be used as a front-end for further signal analysis. The BPF triplet structure is essentially the same as used by Costas [8].

First, a simple model is formulated for the log-envelope and phase derivative of a signal consisting of two tones, which is useful in explaining the synchrony capture phenomenon. Using this model, how a tunable BPF triplet can be designed

Fig. 1. Multiple pitch estimator: The SCFB (Synchrony Capture Filterbank) block consists of K peripheral filters, each of which is followed by a Bandpass filter Triplet shown in Figure 2. Each such cascade is called a channel. SCFB helps separate and capture the dominant frequency components in $x(t)$. HARD thus determines which harmonic components belong in a group, i.e. in a separate auditory object. The ACF unit computes the autocorrelation function for each channel output and obtains a pitch estimate for each group separately by computing the Summary ACFs (SACFs)[4]. K is the number of channels. N is the number of harmonic groups. $K \gg N$.

to follow a dominant tone in the input is outlined. Consider a signal $s(t)$ consisting of a tone at frequency $\omega_1 (=2\pi f_1)$ and an interfering tone at $\omega_2 (=2\pi f_2)$. $s(t)$ is a model for the output of one of the peripheral filters in the SCFB module.

$$s(t) = A_1 \cos(\omega_1 t + \theta_1) + A_2 \cos(\omega_2 t + \theta_2) \quad (1)$$

Let us assume that $A_1 > A_2$. Then, it is easy to show that $s(t)$

$$s(t) \approx a(t) \cos(\phi(t)) \quad (2)$$

where the envelope is

$$a(t) = e^{\ln A_1 + \frac{A_2}{A_1} \cos(\Delta\omega t + \Delta\theta)}, \quad (3)$$

and the phase function is

$$\phi(t) = \omega_1 t + \theta_1 + \frac{A_2}{A_1} \sin(\Delta\omega t + \Delta\theta), \quad (4)$$

and $\Delta\omega = \omega_2 - \omega_1$ and $\Delta\theta = \theta_2 - \theta_1$. The instantaneous frequency (IF) and the log-envelope are as follows.

$$\frac{d\phi(t)}{dt} = \omega_1 + \frac{A_2}{A_1} \Delta\omega \cos(\Delta\omega t + \Delta\theta), \quad (5)$$

$$\ln a(t) = \ln A_1 + \frac{A_2}{A_1} \cos(\Delta\omega t + \Delta\theta). \quad (6)$$

Note that the average value of IF is ω_1 , the dominant tone's frequency. A system can lock on to it by filtering the IF using a low-pass filter (LPF) with a cut off frequency $\Delta\omega$. This is the common frequency capture phenomenon [7] that occurs in traditional FM receivers such as FM discriminators and Phase

Lock Loops (PLLs). Alternatively, envelope or amplitude information can also be used to capture the dominant signal. That is, the compressed envelope signal, $\ln a(t)$, can be low pass filtered to obtain $\ln A_1$. This can then be used to home in on the dominant tone in the input as outlined in Figure 2.

Fig. 2. Synchrony/Frequency Capturing BPF Triplet: $s(t)$ is the output of a peripheral filter in SCFB. Each peripheral filter inputs to a BPF triplet shown above. The left (L), Center (C) and Right (R) bandpass filters are centered at $f_c - \Delta$, f_c and $f_c + \Delta$ Hz respectively. The center frequencies of the BPF triplets can be moved together by changing the frequencies of the VCOs which are used to tune them. The envelopes of the BPF filter outputs are filtered to get their slowly changing amplitudes. These are denoted by L, C and R. The quantity $\frac{R-L}{R+L}$ is used by the servo loop to change the frequencies of the VCOs. Thus, the filters are moved such that centre frequency f_c of the center filter tracks the frequency of the dominant tone in $s(t)$. Track is achieved when $R = L$. The C filter output, $y_c(t)$, is the channel output.

3. HARMONIC RELATION DETECTOR (HARD)

Once the tones are isolated how do we determine if two tones $\cos(2\pi f_1 t + \theta_1)$ and $\cos(2\pi f_2 t + \theta_2)$ have a harmonic relationship, that is, they have a common fundamental f_0 ? Clearly, the range of f_0 must be restricted to avoid arbitrarily low f_0 values. Here f_0 has been restricted to the range of periodicity pitch, between 20 to 2000 Hz. Therefore, we need to determine if f_1 and f_2 have a harmonic relationship over a maximum interval of $T_i \approx 50\text{ms}$ ($=1/20$ Hz.). Let us assume that f_1 and f_2 are indeed harmonically related, that is $f_1 = n f_0$ and $f_2 = p f_0$ where n and p are integers. We then subject the tones to an instantaneous nonlinearity such that they produce tones of all the higher integer multiples of f_1 and f_2 . Chebyshev polynomials serve as a convenient non-linear function [9]. That is, $g_m(t)$ is defined as the output of the nonlinearity (denoted by the NL box in Figure 3) when the input is a tone $\cos(2\pi f_m t + \phi_m)$.

$$g_m(t) = \sum_{k=1}^{N_m} T_k(\cos(2\pi f_m t + \phi_m)), \quad m = 1, 2 \quad (7)$$

where T_k are the Chebyshev polynomials. Notice that the nonlinear operation is such that it produces overtones with frequencies that are both odd and even multiples of f_1 (or f_2). Note also that the dc component is intentionally eliminated. Nonlinearities other than Chebyshev polynomials can

also be used. The nonlinearly distorted tones are then mul-

Fig. 3. Non-linear Interchannel Crosscorrelator: This unit is the building block for the harmonic relationship detector (HARD) in Figure 1. It produces the overtones of two sinusoids and computes their normalized crosscorrelation γ .

tiplied together and integrated over an interval of T_i secs to compute the normalized cross correlation γ .

$$\gamma = \frac{\int g_1(t)g_2(t)dt}{\sqrt{\int g_1^2(t)dt \int g_2^2(t)dt}}. \quad (8)$$

As an example, let f_1 and f_2 be 220 and 550 Hz respectively. $f_1 = n f_0$ and $f_2 = p f_0$ where $n = 2$, $p = 5$ and $f_0 = 110\text{Hz}$. Then the two distorted tones $g_1(t)$ and $g_2(t)$ have common frequencies 1100, 2200, 3300 etc which are integer multiples of $\text{LCM}(n,p) \times f_0$, where LCM is the least common multiple. Thus γ will tend to be large. If the frequencies are unrelated then γ will be small.

4. SIMULATION EXAMPLE

In this paper, for lack of space, we restrict ourselves to applying the proposed multipitch identification algorithm to a synthetic signal $x(t)$ which consists of only two sets ($N=2$) of harmonics. The two fundamental frequencies are $f_{01} = 109$ Hz and $f_{02} = 123$ Hz.

$$x(t) = \sum_{k=1}^6 A_k \cos(2\pi f_k t + \theta_k), \quad (9)$$

where the frequencies f_1 to f_3 are 218, 436 and 981 Hz (integer multiples of f_{01}), while the frequencies f_4 to f_6 are 492, 738 and 861 Hz (integer multiples of f_{02}). θ_k are chosen randomly. All A_k are unity. $x(t)$ is input to a gamma tone filter bank, the details of which are described elsewhere [10]. The filter bank consists of 50 filters spanning a frequency range of 100 to 1500 Hz. The filters are relatively broad with $Q = 4$.

The output of each filter is fed to a BPF triplet shown in Figure 2. For each sinusoid in $x(t)$ with frequency f_k , the channels in that neighborhood stay locked on to it. Typically the channel with the center frequency closest to f_k will have the maximum amplitude. For example, if we consider the case of $f_1 = 218$ Hz, the channel #s 5 to 10 (with center frequencies ranging from 164 Hz and 253 Hz respectively) are locked on to 218 Hz, and channel #7 (whose center frequency is 223Hz) has the highest amplitude as it is closest to 218 Hz. The outputs of those channels with maximum amplitude in a neighborhood are fed to the ACF calculator and the harmonic relation detector(HARD) in parallel. See Figure 1.

In the HARD block in Fig. 1 the channel outputs are nonlinearly distorted to produce some of the higher harmonics

(see Eq. (7)). The nonlinearly distorted signals are then cross-correlated to obtain the γ values using formula in Eq.8. As an example, let us consider the spectra of distorted signals $g_m(t)$ in two different cases: (1) 218 and 981 Hz, (2) 218 and 738 Hz. 218 and 981 Hz are harmonically related, whereas 218 and 738 Hz are not (within a time window of 50ms). Figure 4 shows the spectra of the corresponding distorted signals. Notice that the spectral lines corresponding to the higher harmonics of 218 and 981 Hz coincide at many locations (top panel) whereas such coincidences or near coincidences are less common in the case of 218 and 738 Hz (bottom panel). As a result the γ for 218 and 981 Hz tones are higher than for 218 and 738 Hz. Hence 218 and 981 Hz belong in group 1.

Fig. 4. Spectra of NL distorted signals corresponding to frequencies 218 and 981 Hz (top) and 218, 738 Hz (bottom). If the overtone frequencies are the same then the ‘ ϕ ’ and ‘ $\ast$ ’ coincide. No exact coincidences are found in the bottom panel.

The 3D plot in Figure 5 displays the γ values for all the 15 possible pairs of frequencies. Generally, if two frequencies are from the same group (either f_{01} group or the f_{02} group) then the corresponding γ values tend to be larger. Note that harmonically related tones like 436 Hz and 981 Hz have large γ values whereas 218 Hz and 738 Hz or 861 and 981 Hz have relatively small γ s. Thus we can use these γ values to group the tones that belong to a harmonic set. Once the tones are grouped, then the autocorrelation functions for the channel outputs for these groups can be summed. Figure 6 shows the Summary Autocorrelation for groups 1 and 2. The pitch periods corresponding to the two groups can then be determined by picking the first global peak away from lag zero.

5. REFERENCES

- [1] A. de Cheveigne, “Pitch perception,” in *The Oxford Handbook of Auditory Science*, Christopher J. Plack, Ed. Oxford, Oxford University Press Inc, New York.
- [2] P. Sruлович and J. Goldstein, “A central spectrum model: A synthesis of auditory nerve timing and place cues in monaural communication offrequency spectrum,” *J Acoust Soc Am*, vol. 73, pp. 1266–1276, 1983.
- [3] P. A. Cariani and B. Delgutte, “Neural correlates of the pitch of complex tones. I. pitch and pitch salience. II.

Fig. 5. γ as a function of frequencies of pairs of tones. Note that harmonically related tones like 436 Hz and 981 Hz have large γ s whereas 218 Hz and 738 Hz or 861 and 981 Hz have smaller γ s.

Fig. 6. Summary Autocorrelation for groups 1 and 2

pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch,” *J. Neurophysiology*, vol. 76, no. 3, pp. 1698–1734, 1996.

- [4] R. Meddis and M. Hewitt, “Modeling the identification of concurrent vowels with different fundamental frequencies,” *J Acoust Soc Am*, vol. 91, no. 1, pp. 233–245, 1992.
- [5] J. F. Culling and C. J. Darwin, “Perceptual separation of simultaneous vowels: within and across-formant grouping by f_0 ,” *JASA*, vol. 93, no. 6, pp. 3454–67, 1993.
- [6] 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.
- [7] 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.
- [8] J. P. Costas, “Residual signal analysis,” *Proceedings of the IEEE*, vol. 68, pp. 1351–1352, Oct. 1980.
- [9] John P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover Publications, Inc., 31 East 2nd Street, Mineola, New York 11501, 2000.
- [10] R. Patterson J. Holdsworth, I. Nimmo-Smith and P. Rice, “Implementing a gammatone filter bank,” *MRC Applied Psychology Unit Tech. Rep*, Feb 1988.

APPENDIX B

Published ICASSP2012 paper

Our paper titled “Synchrony Capture Filterbank (SCFB): An Auditory Periphery Inspired Method for Tracking Sinusoids” appeared in the Institute of Electrical and Electronics Engineers (IEEE) International Conference on Acoustics, Speech and Signal Processing (ICASSP) publication [1]. It is presented here in the original format.

List of References

- [1] R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, “Synchrony capture filterbank (SCFB): An auditory periphery-inspired method for tracking sinusoids,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012, pp. 153–156.

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 preserve harmonic grouping operations that subserve separation of multiple concurrent voices. Strategies for automatic attenuation of weaker, interfering sounds thus seems plausible.

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

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 (à 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

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

where Δ is the translation frequency. So,

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

$j = \sqrt{-1}$. 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 $G_1(\omega)$

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 $G_2(\omega) = j \operatorname{sgn}(\omega) G_1(\omega)$. 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

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

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

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, whereas 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.

BIBLIOGRAPHY

- Araújo, L. C., Magalhaes, T. N., Souza, D. P., Yehia, H. C., and Loureiro, M. A., “A brief history of auditory models,” *10 Simpósio Brasileiro de Computação Musical*, vol. 1, 2005.
- Baghdady, E. J., “Theory of stronger-signal capture in FM reception,” in *Proceedings of Institute of Radio Engineers*, April 1958, pp. 728–738.
- C. J. Plack and A. J. Oxenham, “Psychophysics of Pitch,” in *Pitch: Neural Coding and Perception*, ser. Springer Handbook of Auditory Research, C. J. Plack, A. J. Oxenham, R. R. Fay and A. N. Popper, Ed. New York, NY: Springer-Verlag, 2005, ch. 2, pp. 7–55.
- Cariani, P., “Temporal coding of periodicity pitch in the auditory system: an overview,” *Neural Plasticity*, vol. 6, no. 4, pp. 147–172, 1999.
- Cariani, P., “Neural timing nets,” *Neural Networks*, vol. 14, no. 6-7, pp. 737–753, 2001.
- Cariani, P., “A temporal model for pitch multiplicity and tonal consonance,” in *Proceedings of the 8th International Conference on Music Perception & Cognition (ICMPC8, Evanston, IL)*, Lipscomb, S., Ashley, R. O., Gjerdingen, R. O., and Webster, P., Eds. Adelaide, Australia: Causal Productions, 2004, pp. 310–314.
- Cariani, P. and Delgutte, B., “Neural correlates of the pitch of complex tones.I. Pitch and pitch salience. II. Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch,” *J. Neurophysiol.*, vol. 76(3), pp. 1698–1734, 1996.
- Cariani, P. A. and Delgutte, B., “Neural correlates of the pitch of complex tones. I. pitch and pitch salience. II. pitch shift, pitch ambiguity, phase-invariance, pitch circularity, and the dominance region for pitch.” *J. Neurophysiology*, vol. 76, no. 3, pp. 1698–1734, 1996.
- Costas, J. P., “Residual signal analysis -a search and destroy approach to spectral estimation,” in *Proceedings of the First ASSP Workshop on Spectral Estimation*, Hamilton, Canada, Aug. 1981, pp. 6.5.1–6.5.8.
- D. G. Messerschmitt, “Frequency detectors for PLL acquisition in timing and carrier recovery,” *IEEE Trans. Comm.*, vol. COM-27, no. 9, pp. 1288–1295, Sep 1979.

- Dallos, P., "Overview: Cochlear Neurobiology," in *The Cochlea*, ser. Springer Handbook of Auditory Research, P. Dallos, A. N. Popper, and R. R. Fay, Ed. New York, NY: Springer-Verlag, 1996, ch. 1, pp. 1–43.
- Darwin, C. J., "Auditory grouping," *Trends in Cognitive Sciences*, vol. 1, no. 9, pp. 327–333, Dec. 1997.
- de Cheveigne, A., "Pitch perception," in *The Oxford Handbook of Auditory Science*, Plack, C. J., Ed. Oxford University Press Inc, New York: Oxford, 2010, pp. 71–104.
- de Cheveigne, A. and Kawahara, H., "Yin: A fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, 2002.
- Delgutte, B. and Kiang, N., "Speech coding in the auditory nerve: I. Vowel-like sounds," *J. Acoust. Soc. Am.*, vol. 75, pp. 866–878, 1984.
- E. Javel, C. D. Geisler, and A. Ravindran, "Two-tone suppression in auditory nerve of the cat: Rate-intensity and temporal analyses," *J. Acoust. Soc. Am.*, vol. 63, pp. 1093–1104, 1978.
- Emanuel, D. and Letowski, T., *Hearing Science*. Wolters Kluwer Health/Lippincott Williams and Wilkins, 2009. [Online]. Available: <http://books.google.com/books?id=gj1bSAAACAAJ>
- Evans, E. F., "Functional anatomy of auditory system," in *The Senses*, ser. Volume 3 of Cambridge Studies in Philosophy, Barlow, H. and Mollon, J., Eds. Cambridge University Press, 1982.
- F. D. Natali, "AFC tracking algorithms," *IEEE Trans. Commun.*, vol. Com-32, pp. 935–947, Aug. 1984.
- F. M. Gardner, "Properties of frequency difference detectors," *IEEE Transactions on Communications*, vol. Com-33, no. 2, pp. 131–138, 1985.
- Fabio A. Thiers, Joseph B. Nadol, and M. Charles Liberman, "Reciprocal synapses between outer hair cells and their afferent terminals: Evidence for a local neural network in the mammalian cochlea," *J. Assoc. Res. Otolaryngol.*, vol. 9, pp. 477–489, 2008.
- Ferguson, M. J. and Mantey, P. E., "Automatic frequency control via digital filtering," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, no. 3, pp. 392–397, Sep 1968.
- Gardner, F. M., *Phaselock Techniques*. John Wiley & Sons, Inc., 1966.
- Ghitza, O., "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 115–132, 1994.

- Goldstein, J., “An optimum processor theory for the central formation of the pitch of complex tones,” *J. Acoust. Soc. Am.*, vol. 54, no. 6, pp. 1496–1516, 1973.
- Golub, G. H. and Loan, C. F. V., *Matrix Computations*, 3rd ed. Baltimore and London: The Johns Hopkins University Press, 1996, ch. 11, pp. 572–574.
- Haykin, S., *Communication Systems*, 2nd ed. New York, NY: John Wiley & Sons, 1983, pp. 180–187, 206.
- J. E. Hind, D. J. Anderson, J. F. Brugge and J. E. Rose, “Coding of information pertaining to paired low-frequency tones in single auditory nerve fibers of the squirrel monkey,” *Journal of Neurophysiology*, vol. 30, pp. 794–816, July 1967.
- J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice , “Implementing a gammatone filter bank,” in *Annex C of the SVOS Final Report (Part A: The Auditory Filter Bank) MRC (Medical Research Council), APU (Applied Psychology Unit) Report 2341*, University of Cambridge, Cambridge, United Kingdom, Feb 1988.
- Kay, S. M., “A fast and accurate single frequency estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1987–1990, Dec. 1989.
- Kim, D., Lee, S., and Kil, R., “Auditory processing of speech signals for robust speech recognition in real-world noisy environments,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55–69, Jan. 1999.
- Klapper, J. and Frankle, J. T., *Phase-Locked and Frequency-Feedback Systems*. New York, NY: Academic Press, 1972.
- Kumaresan, R., Peddinti, V. K., and Cariani, P., “Synchrony Capture Filterbank II: Auditory-inspired pitch extraction using synchrony capture filterbank and phase alignment (To be submitted),” *J. Acoust. Soc. Am.*, 2013.
- Kumaresan, R. and Ramalingam, C. S., “On separating voiced-speech into its components,” in *Proceedings of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1993, pp. 1041–1046.
- Kumaresan, R., Ramalingam, C. S., and Rao, A., “RISC: an improved costas estimator-predictor filter-bank for decomposing multi-component signals,” in *Proceedings of the Seventh Statistical Signal and Array Processing Workshop*, Québec City, Canada, June 1994, pp. 207–210.
- Kumaresan, R. and Rao, A., “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.

- L. Cedolin and B. Delgutte, "Spatiotemporal representation of the pitch of harmonic complex tones in the auditory nerve," *J. Neurosci.*, vol. 30 (38), pp. 12 712–12 724, 2010.
- Licklider, J., "A duplex theory of pitch perception," *Experientia*, vol. VII, no. 4, pp. 128–134, 1951.
- Lyon, R. F., Katsiamis, A. G., and Drakakis, E. M., "History and future of auditory filter models," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on.* IEEE, 2010, pp. 3809–3812.
- M. B. Sachs, H. F. Voigt and E. D. Young, "Auditory nerve representation of vowels in background noise," *J. Neurophysiol.*, vol. 50, no. 1, pp. 27–45, Jul. 1983.
- M. B. Sachs, I. C. Bruce, R. L. Miller and E. D. Young, "Biological basis of hearing-aid design," *Annals of Biomedical Engineering*, vol. 30, pp. 157–168, 2002.
- Manley, G. A., "Evidence for an active process and a cochlear amplifier in non-mammals," *J Neurophysiol*, vol. 86, no. 2, pp. 541–9, Aug 2001, manley, G A Review United States Journal of neurophysiology.
- Meddis, R. and Hewitt, M., "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 233–245, 1992.
- Meddis, R. and Hewitt, M. J., "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2866–2882, 1991.
- Meddis, R. and O'Mard, L., "A unitary model of pitch perception," *J Acoust Soc Am*, vol. 102, no. 3, pp. 1811–1820, 1997.
- Møller, A. R., *Hearing: Its Physiology and Pathophysiology*, 1st ed. San Diego, CA: Academic Press, 2000.
- Moore, B. C. J., *An Introduction to the Psychology of Hearing*, 4th ed. San Diego, CA: Academic Press, 1997, pp. 40, 90–103, 247.
- Mustafa, K. and Bruce, I. C., "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. on Speech Audio Processing*, vol. 14, no. 2, pp. 435–444, 2006.
- P. A. Fuchs, "Otoacoustic emissions and evoked potentials," in *The oxford handbook of auditory science: The ear*, 1st ed., Kemp, D. T., Ed. Great Clarendon Street, Oxford OX26DP: Oxford University Press, 2010, ch. 4, pp. 93–137.

- Pickles, J. O., "Psychophysical frequency resolution in the cat as determined by simultaneous masking and its relation to auditory-nerve resolution," *J. Acoust. Soc. Am.*, vol. 66, pp. 1725–1732, Dec. 1979.
- R. Kumaresan, Vijay Kumar 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.
- R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, "Synchrony capture filterbank (SCFB): An auditory periphery-inspired method for tracking sinusoids," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012, pp. 153–156.
- R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, "Synchrony Capture Filterbank: Auditory-inspired signal processing for tracking individual frequency components in speech," *J. Acoust. Soc. Am.*, vol. 133, Issue 6, pp. 4290–4310, 2013.
- R. Kumaresan, Vijay Kumar Peddinti and P. Cariani, "Auditory-inspired pitch extraction using a synchrony capture filterbank and phase alignment," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, pp. 5959 – 5963.
- Rao, A. and Kumaresan, R., "On decomposing speech into modulated components," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.
- R.F.Lyon and C.Mead, "An electronic cochlea," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. assp-36, pp. 1119–1134, 1988.
- Rice, M., *Digital Communications: A Discrete-Time Approach*, 1st ed. Prentice Hall, 2008.
- Robles, L. and Ruggero, M. A., "Mechanics of the mammalian cochlea," *Journals of the American Physiological Society*, vol. 81, no. 3, pp. 1305–1352, July 2001.
- S.A. Counter, E. Borg and L. Lofqvist, "Acoustic trauma in extracranial magnetic brain stimulation," *Electroencephalography and Clinical Neurophysiology*, vol. 78, Issue 3, pp. 173–184, Mar. 1991.
- Sachs, M. and Young, E., "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate," *J Acoust Soc Am*, vol. 66, pp. 470–479, 1979.
- Secker-Walker, H. E. and Searle, C. L., "Time-domain analysis of auditory-nerve-fiber firing rates," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1427–1436, 1990.

- Slaney, M., “An efficient implementation of the patterson-holdsworth auditory filter bank,” in *Apple Technical Report # 35, Perception Group - Advanced Technology Group*, Apple Computer Library, Cupertino, CA 95014, 1993, pp. 2–7.
- Spoendlin, H., “Degeneration behaviour of the cochlear nerve,” *Archiv fr klinische und experimentelle Ohren-, Nasen- und Kehlkopfheilkunde*, vol. 200, pp. 275–291, 1971.
- Srulovicz, P. and Goldstein, J., “A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum,” *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1266–1276, 1983.
- Stern, R. M. and Morgan, N., “Hearing is believing,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 34–43, 2012.
- Tramo, M. J., Cariani, P. A., Delgutte, B., and Braida, L. D., “Neurobiological foundations for the theory of harmony in western tonal music,” *Ann N Y Acad Sci*, vol. 930, pp. 92–116, 2001.
- Vaccaro, R. J., *Digital Control: A State-Space Approach*, 1st ed. New York, NY: McGraw-Hill, 1995, ch. 6, pp. 233, 254.
- W. E. Brownell, “The Piezoelectric Outer Hair Cell,” in *Vertebrate Hair Cells*, ser. Springer Handbook of Auditory Research, R. A. Eatock, R. R. Fay and A. N. Popper, Ed. New York, NY: Springer-Verlag, 2010, ch. 7, pp. 313–347.
- Wang, A. L., “Instantaneous and frequency warped signal processing techniques and auditory source separation,” Ph.D. dissertation, Stanford University, Stanford, CA, August 1994.
- Wang, D. L. and Brown, G. J., “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Transactions on Neural Networks*, vol. 10, pp. 684–697, May 1999.
- Wever, E. G., *Theory of Hearing*. New York: Wiley, 1949.
- Yost, *Fundamentals of hearing*, 4th ed. London, U.K.: Academic Press, 2000.