Contour representations of sound

Yoonseob Lim[1]*, Barbara Shinn-Cunningham*[2] *and Timothy Gardner*[3]

[1]Dept. Cognitive and Neural Systems, [2]Biomedical Engineering, [3]Biology, Boston University,
Boston, MA 02215, USA
{yslim, Shinn}@cns.bu.edu, timothyg@bu.edu

ABSTRACT

We consider how the continuity of form in natural sounds may be used to discover sparse time-frequency representations. To proceed, we describe a method to represent any sound as a collection of contours in the time-frequency plane. By analysing the signal in many time-scales, an over-complete set of shapes is generated for a given sound. From this redundant set of shapes the simplest, most parsimonious mathematical forms may be selected.

Index Terms— time-frequency analysis, sparse representation, kernel optimization, Gestalt principle

1. INTRODUCTION

Time-frequency analysis appears in fields ranging from speech recognition to radar analysis and the synthesis of electronic music. However, most time-frequency representations are overdetermined since every time scale of band-pass filtering results in a distinct representation [1]. The uncertainty principle dictates that the resolution in time Δt and resolution in frequency $\Delta\omega$ are reciprocally related: $\Delta t\Delta\omega > 1/2$. Therefore, analysis with high temporal resolution results in loss of spectral resolution, and vice-versa. For a given signal, some choices of Δt can lead to complex time-frequency representations, and others to simple representations. Various approaches exist for optimizing Δt . For example switched filter banks are used in MPEG AAC audio encoding, allowing for time-frequency representation that adapts to local signal content [2]. However, in most cases, computational optimization of Δt focuses on small patches of the time-frequency plane [1,3,4] which can provide little information about long range correlations in time and frequency. This limitation arises from that fact that standard time-frequency methods parcel the time-frequency plane into disconnected "atoms" of energy. Regions of the time-frequency plane that share-correlated phase derivatives are part of a single object – a tone, a click, or a frequency sweep, for example. However, associated portions of the time-frequency plane are not bound together into a single object by standard time-frequency methods . One method of sound representation known as the sinusoidal representation of speech avoids the "atomic" representation by explicitly tracking continuity of spectral peaks across time. The interpolated frequency tracks, or "contours" found in this approach can be used to resynthesize a speech signal by adding together time-varying sinusoids that track the time-frequency lines. While providing an important departure from standard methods, the sinusoidal representation of speech is a parameter dependent approximation that is applicable only to slowly changing sounds [5]. In particular, the method cannot efficiently represent sounds that may include fast transients or other broadband structure.

The focus of the present work is to outline a new method for a contour-based representation of sound that links together associated points of the time-frequency plane at the lowest levels of the representation. By analysing the signal in many time-scales, an over-complete set of contours is generated for a given sound. From this redundant set of contours the simplest, most parsimonious mathematical forms may be selected and form the basis of a practical method of signal analysis and synthesis. The method implies no constraints on the signal content and is applicable to any sound. The contour representation of sound provides new approaches to the problem of optimizing filter time-scales for a given signal. We propose the idea that representations built from the simplest shapes may prove to be useful in a variety of technological applications.

2. DEFINING A CONTOUR REPRESENTATION FOR SOUND

Our analysis starts with taking the Gabor transform of sound,

$$\chi(t,\omega) = \int e^{-(t-\tau)^2/\sigma_t^2} e^{i\omega(t-\tau)} x(\tau)\, d\tau = \left|\chi(t,\omega)\right| e^{i\phi(t,\omega)} \qquad (1)$$

where $x(t)$ is input sound. On the right side of this equation, the Gabor transform is written as an absolute magnitude (the ubiquitous spectrogram is just the square of this term) and a phase term. This phase term is often ignored in time-frequency analysis, but is important for the present method. For all calculations in this paper, we use a discrete version of the Gabor transform. For simplicity we take the number of time bins in the discrete Gabor transform to be equivalent to the number of samples in the original time-series. The variable σ_t defines the temporal resolution Δt of the analysis, discussed in the introduction.

The original signal can be exactly resynthesized with the following definition and integral:

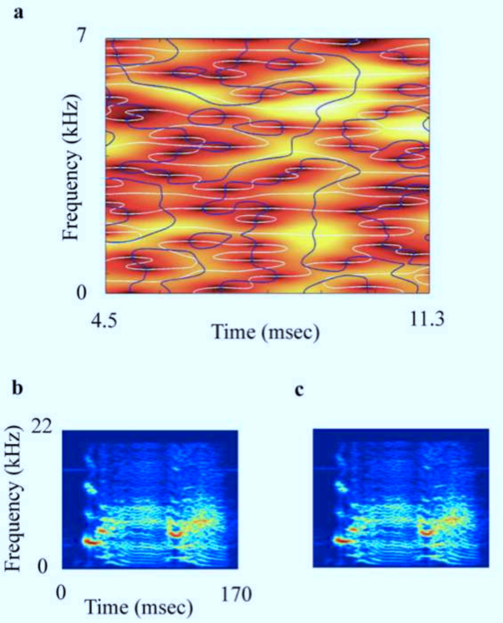

Figure 1: Contour representation. a. Blue and white lines are the zero-crossing points of $\Re(\eta/\chi)$ and $\Im(\eta/\chi)$ respectively. Background image is a sonogram of white noise represented in hot color scale; red and yellow show high values, and black shows zero. b. Sonogram of zebra finch analyzed at $\sigma_t = 2.0$ msec . c. Sonogram of zebra finch sound resynthesized with contours following the stationary phase approximation described in the text.

$$\Phi(\tau,t,\omega) = \phi(\tau,\omega) + \omega t - \omega\tau \tag{2}$$

$$x(t) = \iint |\chi(\tau,\omega)| e^{-(t-\tau)^2/\sigma_t^2} e^{i\Phi(\tau,t,\omega)} \, d\tau \, d\omega \tag{3}$$

A stationary phase approximation [5] indicates that the integral is dominated by points where $\partial\Phi/\partial\tau = 0$ and $\partial\Phi/\partial\omega = 0$. We define auditory contours to be the set of points that satisfy either $\partial\Phi/\partial\tau = 0$ or $\partial\Phi/\partial\omega = 0$. The phase derivatives $\partial\phi/\partial\tau$ and $\partial\phi/\partial\omega$ may be calculated in closed form by using a modified Gabor transform, defined as follows [3,4]:

$$\eta(t,\omega) = \frac{2}{\sigma_t} \int (\tau - t) e^{-(t-\tau)^2/\sigma_t^2} e^{i\omega(t-\tau)} x(\tau) \, d\tau \tag{4}$$

Taking the real and imaginary portions of the ratio, η/χ , it can be shown that [4]

$$\frac{\partial\phi}{\partial\tau} = \omega + \frac{1}{\sigma_t} \Im(\eta/\chi) \tag{5}$$

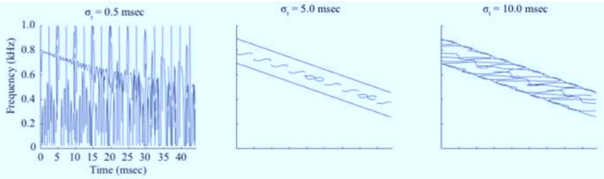

Figure 2: Simple and complex contour representations of Gaussian double chirp in three different time-scales. If analyzed in the optimal time scale (5 msec), the chirp signal is represented with simplest contours.

$$\frac{\partial\phi}{\partial\omega} = -\frac{1}{\sigma_t} \Re(\eta/\chi) \tag{6}$$

Therefore, using equations (2,5,6), we see that the contours ($\partial\Phi/\partial\tau = 0$ and $\partial\Phi/\partial\omega = 0$) lie along the real or imaginary zeros of the ratio, η/χ . The zeros of this ratio form extended closed loops in the time-frequency plane. These loops follow the ridges, valleys and saddle points of $\chi(\tau,\omega)$ (Fig. 1a). The constrained behavior of these contours result from the fact that a simple factor converts the Gabor transform into an analytic function [4,6].

The extended closed loops defined by these contours do not define coherent time-frequency structures. However, shorter fragments of the closed loop do form coherent objects in the following sense: Due to the analytic structure of the Gabor transform, the phase along a contour varies continuously until the contour passes through a zero of the Gabor transform, at which point the phase is effectively randomized. By terminating contours when they cross the zeros of the Gabor transform, we ensure smooth continuity of phase derivatives along the full length of every contour segment. (The zeros referred to here are the amplitude and phase singularities of the analytic Gabor function. In the spectrogram, these points appear to be black holes with no acoustic energy.) After segmentation, each contour represents a sub-component of the sound defined by an extended region of coherent phase.

As mentioned above, the stationary phase approximation [5] indicates that re-synthesis is dominated by points where $\partial\Phi/\partial\tau = 0$ and $\partial\Phi/\partial\omega = 0$. We find that a close approximation to the original sound is derived from integrating equation 3 only along the contours. (See figure 1b, 1c for resynthesis of a songbird zebra finch syllable that is spectrally complex. For human speech, this approximation is perceptually equivalent.)

In cases where exact resynthesis is required, it is possible to assign exact waveforms to each contour in the following manner: We observe that $\Re(\eta/\chi)$ and $\Im(\eta/\chi)$ define two different energy landscapes whose ridges separate distinct phase-locked regions of the time-frequency plane. If we simulate the movement of every point in the time-frequency plane through the landscapes defined by either $\Im(\eta/\chi)$ or $\Re(\eta/\chi)$, all points

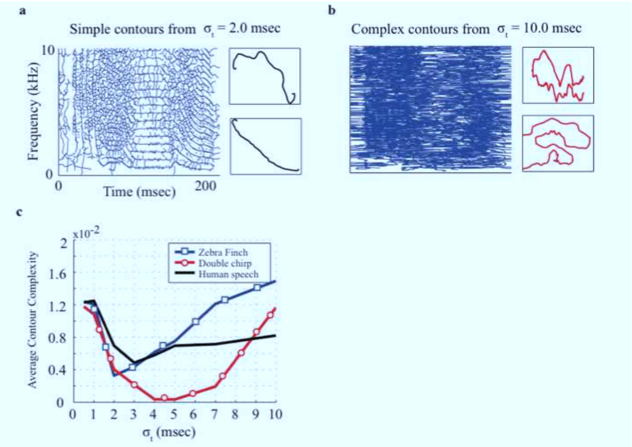

Figure 3: a,b. Examples of simple (black) and complex (red) contours of the zebra finch syllable. These contours are selected from the same sound, analyzed in optimal and non-optimal time-scales (2 and 10 msec respectively). c. The optimal time scale for contour simplicity depends on the details of the signal. Quantification of contour simplicity follows a contour shape measure described previously [7].

flow onto a contour – the energy minima. The "basin of attraction" for a contour is defined by the set of points that flow onto a given contour segment. For the contours defined by $\Re(\eta/\chi)=0$, the energy landscape is $\Re(\eta/\chi)$. For the contours defined by $\Im(\eta/\chi)=0$, the energy landscape is $\Im(\eta/\chi)$. To define a waveform for each contour, we simply apply the standard resynthesis integral (eq. 3) but limit the range of the integral to cover only the basin of attraction for that contour. The original signal can be recovered exactly by adding all contour waveforms together.

In summary, the contour analysis described here converts a time-frequency image (the Gabor transform) into a set of continuous contours. The original sound can be reproduced by exact or approximate means. Rather than a collection of time-frequency points, or "atoms," this representation is built from **shapes** that can have a short range or long range extent in the time-frequency plane. The simplicity of these shapes depends on how well the time scale of features in the signal match the time scale of the analysis. In what follows we examine how measures of contour simplicity can be used to discover sparse representations for a given signal.

3. GESTALT PRINCIPLES FOR THE DISCOVERY OF SPARSE REPRESENTATIONS

Figures 2,3a,3b Illustrate the contour shapes derived for a variety of signals, analyzed at different time scales. In Figure 2 the signal consists of two closely spaced, parallel frequency sweeps. For this simple signal, one particular time scale of analysis $(\sigma_t = 5 \text{ msec})$ yields the simplest contours and the most coherent visual form. Analyzed at other, less appropriate time scales, the representation consists of complex contours that have no coherent overall form. The underlying principle is simple: at a 5ms time scale, each component is spaced by more than Δt in time, and more than $\Delta\omega$ in frequency, and the signal components are separable in the time-frequency plane. We emphasize that all three representations are accurate time-frequency representations of the signal--- the original sound can be resynthesized from any of the three representations, not just the more parsimonious representation.

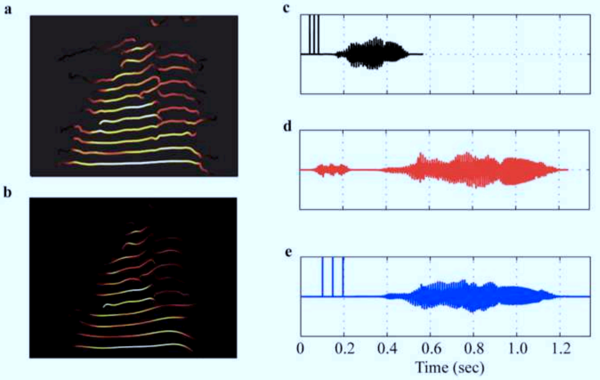

Figure 4: Application of contour representation. a. Contour representation of speech signal with the absolute value of contour amplitudes shown. b. Contours after polynomial fitting. Frequency and amplitude are fitted by low order polynomials (12th and 8th order respectively). c. Three clicks and human speech reconstructed by 30 contours from two time-scales (1msec and 10msec). Every click sound is represented by vertical contours at 1msec and speech is composed of horizontal contours at 10msec. d. Stretched (x2.37) sound by phase vocoder method. Fast transients are not preserved due to phase dispersion. e. Resynthesizing stretched contours preserves fast transients.

For a more complex sound like a songbird (zebra finch) syllable, simple contours are common when analyzed at relatively short time scales $(\sigma_t = 2 \text{ msec})$, while analysis using a 10ms time scale yields complex contours (Fig. 3a,b). The complex contours often violate rules of physical causality by forming loops in time (e.g., see the panels of Fig. 3a,b).

To quantify contour simplicity, a variety of objective measures can be applied, including the quality of polynomial fits of a given order, contour persistence length, or average curvature. In Fig. 3c, we employ a measure of contour similarity designed for comparison of contour shapes in visual object recognition [7] that is insensitive to scale and rotation.. Using this measure, we observe that the collection of contours have the simplest overall shapes at intermediate time-scales of filtering, but that the optimum time scale varies depending on the signal.

These observations lead to a simple principle for optimizing a time-frequency representation: analyze contour shapes across a range of time scales, and then pick the time scale that leads to the simplest shapes. This principle can be applied globally to optimize the time scale on average for the entire signal, or it can be separately applied locally in specific frequency bands or time epochs.

4. APPLICATIONS: COMPRESSION, AND TIME DILATION.

In what follows, we sketch two preliminary examples of possible applications

For sound compression, we observe that contours are smooth when the filter time-scale is locally well adapted to the signal. In a first test, we fit contours from a sample of voiced speech with 12th order polynomials and achieved a 20 fold decrease in the number of bits required to store the sound while preserving high perceptual quality. (The choice of 12th order was arbitrary, and we have not examined the sensitivity of the result to variations in this order. See figure. 4a,4b). Further compression could be achieved if we did not encode harmonics.independently.

A second sketch of an application is time dilation of a complex sound. The original phase-vocoder method for time-dilation of sound [8,9] does not meaningfully parse connected structures in the time-frequency plane, and relative phase between frequency bands are not preserved. As a result, the processed signals exhibit dispersion of phase coherence between bands which converts fast transients in the time-stretched signal into bursts of noise (Fig. 4e) A number of current methods address this phase dispersion and have improved on the original phase-vocoder algorithms [10]. In the present method, transients are represented by lines with steep slopes. Modeling the contours with sinusoids and dilating the time axis preserves the phase relationships between points that are linked along the contour. After time-dilated re-synthesis of contours, fast transients are preserved while still accomplishing a perceptual stretching of harmonic sounds.

5. CONCLUSION

Auditory contour analysis provides a new framework for an old problem in signal processing: how to optimize the time scale of analysis to produce the simplest, most efficient time-frequency representation. By generating contours from parallel filterbanks, each working at a distinct time scale, an over-complete representation is derived. The coherence or simplicity of long-range structures in each time scale can be explicitly examined and quantified. Using this approach, sounds can be represented in a manner that emphasizes each component in its own simplest form, retaining high precision in time and frequency estimates. As such, the method may be applicable to a range of time-frequency analysis issues; denoising, compression, time dilation, and other signal manipulations for audio effects.

We note in closing that neural auditory processing could conceivably involve a similar contour representation for sound. Each stage in the analysis is a plausible operation for neurons: parallel and redundant primary processing streams in multiple bandwidths, grouping neurons locally in time by phase coherence, and linking groups together over extended times by continuity [11,12].

6. ACKNOWLEDGMENT

This work is supported in part by NIH grant R01 DC009477 (YL and BGSC) and the NSF Science of Learning Center CELEST (SBE-0354378) and by a Career Award at the Scientific Interface to T.G. from the Burroughs Wellcome Fund and a Smith family award to T.G.

7. REFERENCES

[1] L. Cohen, “Time-frequency distributions - a review,” Pro. IEEE, 1998, vol. 77, pp. 941–981.

[2] D. Pan, “A tutorial on MPEG/audio compression,” IEEE Multimedia, 1995, vol. 2, pp. 60-74.

[3] F. Auger and P. Flandrin, “Improving the readability of time-frequency and time scale representations by the reassignment method,” IEEE Trans. Signal Processing, 1995, vol. 43, pp. 1068–1089.

[4] T. J. Gardner and M. O. Magnasco, “Sparse time-frequency representations,” Proc. Natl. Acad. Sci., 2006, vol. 103, pp. 6094–6099.

[5] P. Flandrin, *Time-Frequency/ Time scale Analysis*, Academic Press, 1999.

[6] E. Chassande-Mottin, I. Daubechies, F. Auger, and P. Flandrin, “Differential reassignment,” IEEE Signal Processing Letters, 2002, vol. 4, pp. 293-294.

[7] R. C. Veltkamp, "Shape Matching: Similarity Measures and Algorithms," in Shape Modeling and Applications, International Conference on, p. 0188, International Conference on Shape Modeling & Applications, 2001.

[8] J. L. Flanagan, and R. M. Golden, “Phase Vocoder,” Bell Syst Tech. J., 1966, vol. 45, pp. 1493-1509.

[9] L. R. Rabiner, and R. W. Schafer, *Digital Processing of Speech Signals,* Prentice Hall, 1978.

[10] U. Zölzer, *DAFX: Digital Audio Effects*. Chichester, UK: John Wiley & Sons, Ltd, 2002.

[11] E. D. Young, "Parallel processing in the nervous system: Evidence from sensory maps," Proceedings of the National Academy of Sciences, 1998, vol. 94, pp. 933-934.

[12] A. S. Bregman, *Auditory scene analysis*. Cambridge, USA: MIT Press, 1990.