Sparse Contour Representations of Sound

Yoonseob Lim, Barbara Shinn-Cunningham, and Timothy J. Gardner

***Abstract*—Many signals are naturally described by continuous contours in the time-frequency plane, but standard time-frequency methods disassociate continuous structures into isolated "atoms" of energy. Here we propose a method that represents any discrete time-series as a set of time-frequency contours. The edges of the contours are defined by fixed points of a generalized reassignment algorithm. These edges are linked together by continuity such that each contour represents a single phase-coherent region of the time-frequency plane. By analyzing the signal across many time-scales, an over-complete set of contours is generated, and from this redundant set of shapes the simplest, most parsimonious forms may be selected. The result is an adaptive time-frequency analysis that can emphasize the continuity of long-range structure. The proposed method is demonstrated with a few examples.**

***Index Terms*—Adaptive filtering, kernel optimization, sparse representation, time-frequency analysis.**

I. Introduction

TIME-FREQUENCY analysis takes many forms but essentially involves the application of temporally localized band-pass filters to a time-series $x(t)$. The resulting time-frequency image is not unique since every function that localizes the filters 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: $\Delta t \Delta\omega > 1/2$, and the result of this trade-off is that fixed-filter methods cannot optimally represent signals with time-varying spectral content.

Various approaches have been developed for signal dependent adaptation of filters [2]–[6]. Applications can be found ranging from speech enhancement in noise [7] to radar target analysis [8]. A common approach to adaptive time-frequency analysis involves a search for minimum entropy decompositions that concentrate power either in a small number of positions in the time-frequency plane, or a small number of coefficients in a multi-scale wavelet decomposition [9]. The criterion of maximal local concentration provides little information about the coherence of long-range structure in the time-frequency plane since the "atoms" that represent long-range objects are not explicitly bound together. As a result, the simplicity of large-scale shapes cannot easily inform an adaptive filtering based on standard time-frequency methods.

The focus of the present work is to propose a new contour-based representation that links together associated points of the time-frequency plane at the lowest levels of representation. The method implies no constraints on the signal content and is applicable to any time-series. In this paper we outline the theory, and provide a few examples. The auditory contour analysis suggests a new framework for an old problem in signal processing: among a set of over-complete time-frequency representations, one can choose the parameters of analysis that will yield the simplest contours.

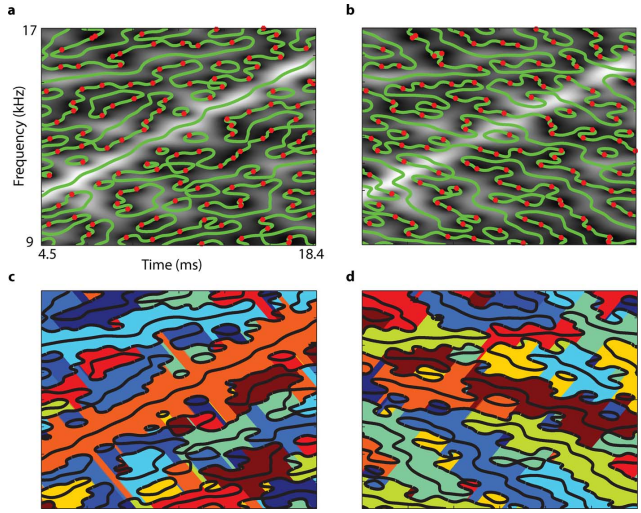

Fig. 1. Defining time-frequency contours; contours with angle selectivity $\theta = \pi/4$ and $\theta = -\pi/4$ in green in (a) and (b) respectively, superimposed on a gray-scale Gabor transform ($|\chi|$). The signal is a short sample of white noise combined with a linear frequency sweep. Red dots highlight the zeros of $|\chi|$. Territories corresponding to the contours in panels (a) and (b) are shown in (c) and (d) respectively.

II. Defining a Contour Representation for Sound

The proposed analysis begins with the complex Gabor transform,

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

where $x(t)$ is the input sound. The variable σ_t defines the temporal resolution of analysis, Δt. For all calculations in this paper, we use a discrete version of the Gabor transform with 1024 frequency bins, and signals sampled or synthesized at 44.1 or 48 kHz. For simplicity of exact resynthesis (but at high computational cost) 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. Matlab code to reproduce Fig. 1(a) and (b) is included in the supplement.

Manuscript received May 09, 2012; revised July 18, 2012; accepted July 19, 2012. Date of publication August 01, 2012; date of current version August 15, 2012. This work was supported in part by NIH Grant R01 DC009477 (Y. Lim and B. Shinn-Cunningham), the NSF Science of Learning Center CELEST (SBE-0354378), and by a Career Award at the Scientific Interface to T. J. Gardner from the Burroughs Wellcome Fund and a Smith family award to T. J. Gardner. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mads Graesboll Christensen.

This paper has supplementary downloadable materials available at http://ieeexplore.ieee.org, provided by the authors. These include 6 multimedia WAV format audio clips and MATLAB program for simulation. These materials are 353 kB and 98 kB in size respectively.

Y. Lim is with the Department of Cognitive and Neural Systems, Boston University, Boston, MA 02215 USA (e-mail: yslim@bu.edu).

B. Shinn-Cunningham is with the Department of Biomedical Engineering, Boston University, Boston, MA 02215 USA (e-mail: shinn@bu.edu).

T. J. Gardner is with the Department of Biology, Boston University, Boston, MA 02215 USA (e-mail: timothyg@ bu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2012.2211012

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

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

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

A stationary phase approximation to (3) indicates that the integral is dominated by points where $\partial\Phi/\partial\tau=0$ and $\partial\Phi/\partial\omega=0$ representing the local instantaneous frequency and points of zero group delay respectively. Points that satisfy both $\partial\Phi/\partial\tau=0$ and $\partial\Phi/\partial\omega=0$ are the fixed points of a method known as "reassignment" that concentrates power on the ridges of the Gabor transform [10]–[14]. Reassigned spectrograms demonstrate enhanced precision for many signals, and provide the basis for improved methods of separating signal components [15]. Like the standard spectrograms, reassigned spectrograms consist of separate pixels, and coherent signal objects are not intrinsically associated together. However, it is possible to define a representation that does link together associated time-frequency points as follows: first, we generalize the set of points defined by the stationary phase positions $\partial\Phi/\partial\omega=0$ and $\partial\Phi/\partial\tau=0$. These points are particular cases ($\theta=0$ and $\theta=\pi/2$) of an expression that generalizes to all angles:

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

$$\Im\left(\left(\frac{\eta}{\chi}\right)e^{i\theta}\right)=0 \tag{5}$$

where θ defines a contour preference angle in the time-frequency plane and $\Im(f)$ is the imaginary component of f. The points that satisfy (5) are equivalent to the fixed points of the standard reassignment method [16], subject to the constraint that reassignment operate only in the direction defined by the angle $\theta+\pi/2$. The points that satisfy (5) for some choice θ form extended contours in the time-frequency plane that follow the ridges, valleys and saddle points of $\chi(\tau,\omega)$ (Fig. 1). These contours do not branch, but terminate on the borders of the time-frequency axes, or form closed loops – a constraint of the analytic structure of $\chi(\tau,\omega)$ [16]. Also due to this analytic structure, the phase along a contour varies *smoothly* until it passes through a singularity on the zeros of $\chi(\tau,\omega)$. After defining contours by linking together the points that satisfy (5), the method then segments the contours whenever they cross the zeros of the Gabor transform. Splitting contours at the zeros ensures smooth continuity of phase derivatives along every contour segment. This process of linking edges into contours and segmenting at the zeros is the essential basis of the method. The relationship between contours and the analytic structure of the Gabor transform implies that a simple waveform can be assigned to each contour as described next.

III. Resynthesis of the Contour Representation

The definition of the contours can be motivated from a stationary phase approximation to integral (3), and this approximation is produced by simply integrating (3) along the contours, for any choice of angle θ. The accuracy of this approximation needs to be analyzed quantitatively, but in the human speech sample provided in supplement, readers will find near perceptual equivalence of resynthesized sound. In cases where exact resynthesis is needed, a distinct process can assign waveforms to all contours such that the sum of all waveforms is equivalent to the original signal. The exact method proceeds as follows: we first note that contours track the ridges and valleys of $\chi(\tau,\omega)$ (Fig. 1). For $\theta=0$ and $\theta=\pi/2$, $\Im\left((\eta/\chi)\,e^{i\theta}\right)$ provides the frequency and time "displacement vectors" used in the reassignment method to concentrate spectrogram power from the valley to the ridges of $\chi(\tau,\omega)$ [10]. By extension, for arbitrary θ, $\Im\left((\eta/\chi)\,e^{i\theta}\right)$ defines a local slope in direction $\theta+\pi/2$ perpendicular to a nearby contour at angle θ. The "watershed" structure of this function defines a segmentation of the time-frequency plane. That is, the *territory* for each contour is the set of all points that would "flow" onto the contour, following the slope defined by $\Im\left((\eta/\chi)\,e^{i\theta}\right)$. This is similar to applying the standard reassignment method iteratively, until all time-frequency points reach their fixed points, and then assigning a waveform to each fixed point by integrating (3) over the basin of attraction that flows to that fixed point. The only difference is that here we force every point to move along angle $\theta+\pi/2$ resulting in fixed points that form continuous lines – our contours. In practice, iterative reassignment is not needed to assign a territory to a contour segment- one need only observe the sign of $\Im\left((\eta/\chi)\,e^{i\theta}\right)$ at each point in the time-frequency plane and the nearest contour in the vector direction $\theta+\pi/2$ "owns" the point. Fig. 1(c),(d) demonstrates the arrangement of these contour territories in the time-frequency plane for the contours illustrated in Fig. 1(a),(b). Once the territory for a given contour is found, the corresponding waveform is derived by computing the resynthesis integral in (3), over the territory specific to that contour (i.e., integrating over one color zone in Fig. 1(c),or (d).) There are no gaps or overlaps in the territories, and this is true for all signals, so integral (3) remains exact, just computed piecewise for each contour territory. As a result, a complete set of contour waveforms that sum to the original signal is derived for any signal and for any single choice of σ_t and θ. By construction, no contour territory contains an analytic zero of χ, so the phase varies continuously within a contour territory. A quantitative assessment of contour waveforms is needed, but the absence of analytic zeros in each territory implies that these waveforms will all be simple.

IV. Gestalt Principles for the Discovery of Sparse Representations

The method defined here provides a family of contour representations of a signal, parameterized by angle and time-scale. The complexity of this representation will depend on how well the angle and time-scale parameters are matched to the signal content. Fig. 2 illustrates contour shapes derived for a simple signal, analyzed at angles $\theta=0$ and $\theta=\pi/2$ (This choice of angles is arbitrary and the following analysis would apply equally to other angles.) The signal consists of two closely spaced, parallel frequency sweeps. Although contour sets from each time-scale and angle produce a complete representation of the signal (Section III), a time scale of 2 ms, for this signal, yields the simplest contours and the most coherent long-range form. The underlying principle is simple: at a 2 ms time scale, each component is spaced by more than the resolution of the time-frequency uncertainty: Δt in time and $\Delta\omega$ in frequency. Therefore, at this time scale, the signal components are *separable* in the time-frequency plane [13].

This example suggests that contour simplicity could vary systematically with the parameters of the analysis, for a given signal. To quantify contour simplicity, we first rank contours by power (power is defined by the integral of $\chi(\tau,\omega)$ along

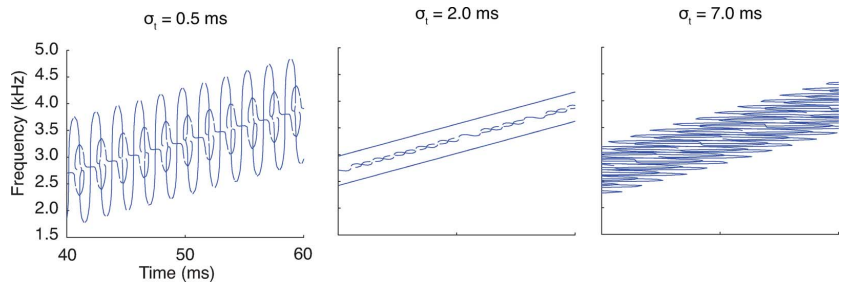

Fig. 2. Contour representations of a signal vary with the time-scale of analysis. Simple and complex contour representations of a double chirp, calculated for three different time-scales. If analyzed in the optimal time scale (2 ms), the chirp signal is represented with the simplest contours.

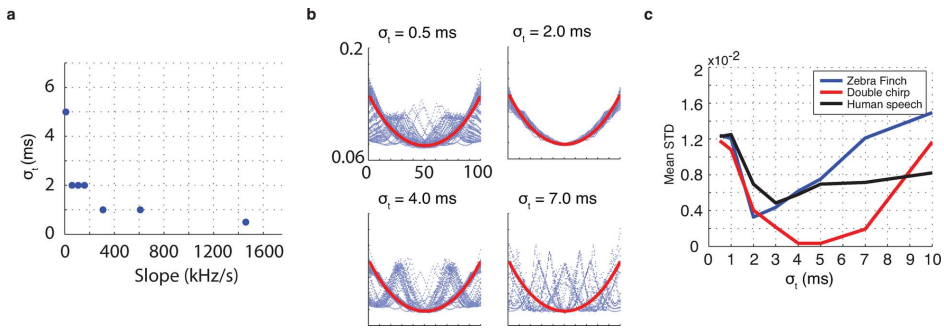

Fig. 3. Optimal time-scale selection is achieved through a quantitative measure of contour shapes. (a) The "optimal" time scale for analysis of the double chirp signal (shown in Fig. 2), as a function of chirp slope. (b) The spread of contour-shape eigenvectors for the double chirp signal, at different time-scales. Only the first eigenvector is shown. For reference, the red line shows the first eigenvector calculated for a straight line. (c) Contour simplicity, based on standard deviation of the first eigenvector, as a function of σ_t for three different signals.

the contour), then select as many of the top contours as needed to account for 95% of the signal power. A variety of objective shape-based measures could be applied to the collection of contours that remains. (Example measures include the quality of polynomial fits of a given order, contour persistence length, or average curvature.) In Fig. 3(a), we employ a measure designed for comparison of contour shapes in visual object recognition [17]. This measure is insensitive to scale and rotation. A matrix describing the shape of each contour is defined by the Euclidean distance between pairs of evenly spaced points along the contour; 100 points are sampled for each contour. Eigenvectors of this distance matrix are calculated for each contour. The spread of the eigenvector shapes provides a measure of the diversity of forms present in the population. Fig. 3(b) illustrates the spread of shapes for the frequency sweep analyzed in Fig. 2. In this example, we define the "optimal" time scale as the time-scale where the spread of contour shapes (based on the first eigenvector) is minimal.

The optimal time-scale depends on the signal. Changing the slope of the frequency sweep shifts the optimal time-scale of analysis as illustrated in Fig. 3(a). Steeper frequency sweeps require shorter filter time-scales. Fig. 3(c) demonstrates an example involving a sample of human speech, a zebra finch bird song, and a double chirp signal. The time-scale of analysis that produces the simplest contours, on average, differs for the three signal types.

This discussion has highlighted how contour simplicity varies with the time scale of analysis. A complementary discussion could focus on how contour simplicity varies with angle of analysis, for a fixed time-scale. As illustrated in Fig. 1(c), the orange contour at angle $\theta = \pi/4$ is matched to the angle of the rising frequency sweep, and tracks the signal content well. In Fig. 1(d), the angle of analysis is mismatched to the signal content and no contour follows the frequency sweep. A systematic analysis is needed, but we suggest that the search for parameters that yield the simplest representations should consider the contours as cross-sections of high dimensional manifolds in the four dimensional space of time, frequency, time-scale and angle. Smooth three-dimensional manifolds would indicate parameter regions where the shape of contours is simple and not sensitively dependent on the choice of time-scale or angle of analysis. A practical method to optimize the contour representation that takes into account both time-scale and angle remains to be defined.

In Fig. 3, we examined how contour shape depends on time-scale, on average, for an entire signal. The more powerful approach would combine contours from multiple time scales and angles in a single image. An example of a multi-band analysis is shown in Fig. 4. The signal consists of a mixture of a zebra finch bird song and a double-frequency sweep. In the figure, we display only those contours whose power exceeds a bandwidth – dependent threshold, and whose shapes are simpler than a carefully chosen cutoff, as measured by eigenvector spread. The top contours from the 5 ms time-scale, shown in red, exclusively pertain to the frequency sweep. The top contours in the 2 ms time-scale exclusively pertain to the zebra finch sound. Separating contours from the two time-scales and resynthesizing produces a perceptually clean separation of the mixture (Sounds are provided in supplement.) This separation is simple given the length and power of the double chirp signal. It remains to be seen whether this approach to signal separation is applicable to more complex mixtures.

V. Conclusion

In standard time-frequency representations, the division of a signal into "atoms" of localized power makes it difficult to infer associations between distant regions of the time-frequency plane. The key element of this proposal is the process of building a time-frequency representation based on extended shapes that are linked together by continuity. In the present method, contour edges are equivalent to the fixed points of a reassignment

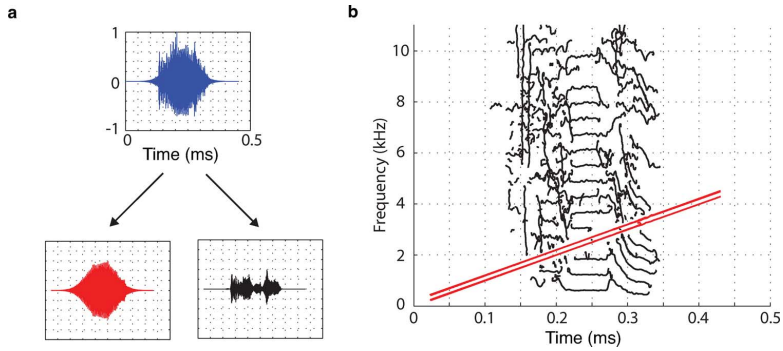

Fig. 4. Separating sound mixtures based on contour statistics. (a) A signal consisting of a zebra finch bird song syllable superimposed on a double chirp is split into separate components. Short time-scale contours pertain to the bird song syllable, (panel b, 2 ms, black) and long time scale contours correspond to the frequency sweep (panel b, 5 ms, red). By setting thresholds for power and contour complexity appropriately, the mixed signal can be cleanly separated.

process, constrained to move along a specific angle. Contour segmentations are derived from the analytic structure of the Gabor transform resulting in a number of useful properties; every contour represents a zone of coherent phase derivatives in the time-frequency plane, and the waveform associated with each contour is simple because the territory that defines the waveform contains no analytic zeros. Exact resynthesis is achieved by taking the direct sum of all contour waveforms.

By computing contours for multiple filter-banks and multiple angles θ, an overcomplete set of shapes is found. From this set, the simplicity of contours in each time scale and at each angle θ can be explicitly examined to inform an adaptive time-frequency analysis. The potential benefit of this approach is the representation of sub-components of a sound in their own simplest forms.

The practical utility of the approach remains hypothetical, and a quantitative analysis is needed, particularly for the quality of the stationary phase resynthesis based on integrating (3) along the contours. In certain obvious examples, the contour representation is lacking: for instance, overtones of a harmonic stack can be represented by separate contours and the method will not bind these separate components into a single object, based on their harmonic relationships.

In closing, we note the connection between this representation of sound and the theories of the "Gestalt" psychologists who maintained that every sensory percept is represented in its own most parsimonious form. In early stages of vision, this principle can be seen in the perceptual enhancement of forms demonstrating good continuity and low curvature along the boundaries of objects. Human auditory processing also introduces a bias toward the perception of continuity in sound streams [18]. The method described here provides one means of enhancing the continuity of time-frequency representations. Each stage in the contour analysis is a plausible operation for neural systems: computation of parallel and redundant early auditory streams, binding together of phase-coherent channels, and linking groups together through time by continuity.

Acknowledgment

The authors would like to thank P. Mehta and L. Sherbakov for suggesting many improvements in the manuscript.

References

[1] L. Cohen, "Time-frequency distributions-A review," *Proc. IEEE*, vol. 77, pp. 941–981, Jul. 1989.
[2] D. L. Jones and T. W. Parks, "A high resolution data-adaptive time-frequency representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, pp. 2127–2135, Dec. 1990.
[3] D. L. Jones and R. G. Baraniuk, "A simple scheme for adapting time-frequency representations," *IEEE Trans. Signal Process.*, vol. 42, pp. 3530–3535, Dec. 1994.
[4] R. N. Czerwinski and D. L. Jones, "Adaptive short-time Fourier analysis," *IEEE Signal Process. Lett.*, vol. 4, pp. 42–45, Feb. 1997.
[5] H. K. Kwok and D. L. Jones, "Improved instantaneous frequency estimation using an adaptive short-time Fourier transform," *IEEE Trans. Signal Process.*, vol. 48, pp. 2964–2972, Oct. 2000.
[6] D. Rudoy, P. Basu, and P. J. Wolfe, "Superposition frames for adaptive time-frequency analysis and fast reconstruction," *IEEE Trans. Signal Process.*, vol. 58, pp. 2581–2596, May 2010.
[7] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 2064–2074, Nov. 2006.
[8] I.-S. Choi and I.-K. Rhee, "Performance comparison of time-frequency analysis methods for radar signature analysis," in *Proc. 2008 Second Int. Conf. Future Generation Communication and Networking (FGCN)*, 2008, pp. 24–27.
[9] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inf. Theory*, vol. 38, pp. 713–718, Mar. 1992.
[10] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, pp. 1068–1089, May 1995.
[11] A. K. Fitz and L. Haken, "On the use of time–frequency reassignment in additive sound modeling," *J. Audio Eng. Soc.*, vol. 50, pp. 879–893, Nov. 2002.
[12] S. A. Fulop and K. Fitz, "A spectrogram for the twenty-first century," *Acoust. Today*, pp. 26–33, Jul. 2006.
[13] D. J. Nelson, "Cross-spectral methods for processing speech," *J. Acoust. Soc. Amer.*, vol. 110, pp. 2575–2592, Nov. 2001.
[14] T. J. Gardner and M. O. Magnasco, "Sparse time-frequency representations," *Proc. Nat. Acad. Sci. USA*, vol. 103, pp. 6094–6099, Apr. 2006.
[15] D. J. Nelson and D. C. Smith, "A linear model for TF distribution of signals," *IEEE Trans. Signal Process.*, vol. 54, pp. 3435–3447, Sep. 2006.
[16] E. Chassande-Mottin, I. Daubechies, F. Auger, and P. Flandrin, "Differential reassignment," *IEEE Signal Process. Lett.*, vol. 4, pp. 293–294, Oct. 1997.
[17] R. C. Veltkamp, "Shape matching: Similarity measures and algorithms," in *Proc. SMI 2001 Int. Conf. Shape Modeling and Applications*, 2001, pp. 188–197.
[18] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.