STABLE TIME-FREQUENCY CONTOURS FOR SPARSE SIGNAL REPRESENTATION

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

Dept. Cognitive and Neural Systems[1], Biomedical Eng.[2], Biology[3], Boston University
Boston, MA 02215 USA

ABSTRACT

Many signals cannot be resolved in time and frequency with a single time-scale of analysis and multi-band representations are needed that can adapt to the local signal content. Using a newly developed contour-based representation of signals, we show that efficient multi-band representations arise when long-range, structurally stable shapes are enhanced relative to background. For the examples provided here, resolution in time and frequency is distributed adaptively so that each component of a signal is represented in its most parsimonious form. The resulting representation is characterized by simple shapes in the time-frequency plane.

Index Terms**—** Time-frequency analysis, adaptive filtering, reassignment and sparse representation

1. INTRODUCTION

Sparse time-frequency methods typically search for a linear decomposition of signals through a minimal number of dictionary elements [1]. The dictionary elements are drawn from an over-complete set, which may be defined a-priori or adapted to a specific stimulus class [2]. Numerous iterative assembly processes are effective, but robust methods for single-pass adaptive time-frequency representations remain elusive, though numerous promising directions have been proposed [3]-[6]. The starting point of these and other time-frequency representations is the parcelation of the time-frequency plane into isolated "atoms" of energy with no intrinsic associations among them. From this basis, the structure of long-range shapes in time and frequency cannot easily guide adaptive algorithms, although many signals are naturally represented by coherent long-range forms, such as contours. Recently, a general time-frequency method was described whose elementary units are contours of varying shapes. These shapes fully represent any signal, but the details of the shapes depend on the signal content and on the time-scales of analysis. Each contour in the representation is a coherent object - a component of the signal whose boundaries are defined by a region of the Gabor transform that contains no analytic zeros [7]. The contours can be interpreted as the minimal coherent units of the signal from the perspective of the analytic Gabor transform. Their scales are typically much larger than the resolution limit of the analysis. Using this contour representation, a prior study demonstrated how measures of contour complexity could be used to optimize the time-scale of analysis, on average, for an entire signal [7]. The implicit assumption in that work was that parsimonious representations would involve contours of low curvature. The present work is motivated by the desire to define a more general principle for adaptive time-frequency analysis based on time-frequency contours. The principle is as follows: when a signal component is analyzed in its own natural time-scale, then the contours that represent the component are structurally stable - the details of the shapes do not change with small variations in the parameters of analysis. This hypothesis does not presume that contours should be simple in form, but only that they be structurally stable. A process that enhances structurally stable shapes provides a sparse multi-scale representation of complex signals. In the following, we outline the theory, and provide a few examples.

2. BACKGROUND

The contour description of sound [7] is based on a generalization of the reassignment process [8]. This involves the Gabor transform, (χ) and the associated transform (η) based on a window shape that is the derivative of a gaussian:

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

$$\eta(t,\omega)=\frac{1}{\sigma^2{}_t}\int(\tau-t)e^{-(t-\tau)^2/2\sigma_t^2}e^{i\omega(t-\tau)}x(\tau)d\tau \tag{2}$$

These transforms are applied to the acoustic signal of interest, $x(t)$, which is a function of time (t), to produce a representation that is a function of both time and frequency (ω). In this expression, σ_t defines the time-scale of the analysis window, therefore the resolution of the analysis. Contours edges are equivalent to the fixed points of the time-frequency reassignment process, subject to the constraint that reassignment moves along a fixed angle. By

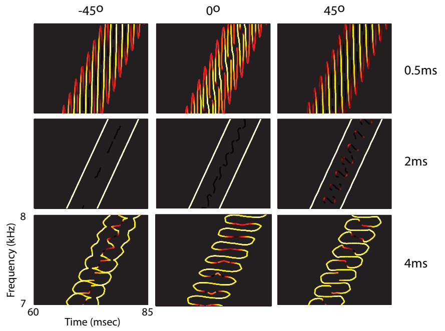

Fig. 1. The structural stability of contour shapes can guide an adaptive time-frequency analysis. In this example, the signal components (two frequency sweeps) are "separable" when 2ms filters are used in the Gabor transform. Contours are calculated for three time scales (rows) and three angles (columns.) At the optimum time-scale (middle row), contour shapes are robust to variations in the angle of analysis. Contour energy is drawn from and illustrated in hot color scale.

another definition, these points are stationary phase points for the resynthesis integral that produces the original signal from the Gabor transform [7].

Contour edges are defined by:

$$\Im\left((\eta/\chi)e^{i\theta}\right)=0 \quad (3)$$

where θ defines a contour preference angle in the time-frequency plane and $\Im(f)$ is the imaginary component of f. Intuitively, $(\eta/\chi)e^{i\theta}$ is an approximation to the derivative of the Gabor transform along a specific angle (θ) in the time-frequency plane - closely related to heuristic expressions for "spectral derivatives" based on multi-taper spectral analysis[9],[10]. The points that satisfy (3) form extended closed loops in the time-frequency plane that follow the ridges, valleys and saddle points of χ. To divide the contours into coherent units, contours are segmented whenever they cross zeros of the Gabor transform [7]. It is possible to analytically define a waveform for each contour such that the sum of all waveforms equals the original signal [7]. In all images shown here, the color scale for each contour is equal to the local value of $|\chi|$.

3. PARSING COMPLEX SOUNDS USING MULTIPLE TIME-SCALES

For every time-scale and angle of analysis, a distinct object-based decomposition exists. Every choice of time-scale σ_t and angle θ generates its own contour representation and associated territories - an over-complete family of valid contour representations, each of which fully captures the signal content (Fig. 1 in [7]). The complexity and structural stability of the contour shapes depend on how well the angle and time-scale parameters are matched to the signal content. Fig. 1 illustrates contour shapes derived for a simple signal, analyzed with multiple choices of time-scale and angle. The signal consists of two closely spaced, parallel frequency sweeps. In this figure, rows represent analysis in different time-scales and columns analysis in different angles.

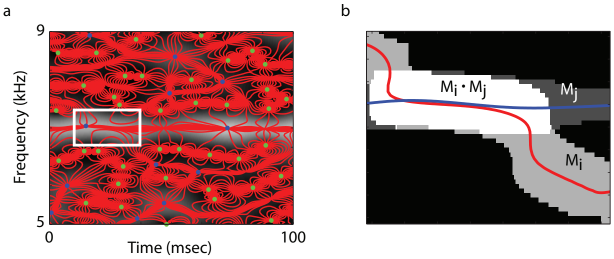

Fig. 2. Quantifying the structural stability of contours. (a) Contours calculated for a fragment of white noise with a superimposed 7kHz tone. Contours in red are calculated for a range of angles, at a single time scale. Blue dots are maxima of the signal. Green dots are minima. The contours that track the 7kHz tone are tightly bundled together - indicating local structural stability of contour shape across variations in the analysis parameters. In (b), two contours are extracted from the region marked by the white square in (a). The territories belonging to the two contours are shaded in gray-scale, and their overlap in white.

Although contour sets from each time-scale and angle produce a complete representation of the signal, 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 the optimal time-scale, each component is spaced by more than the resolution of the time-frequency uncertainty: in time and in frequency. Therefore, at this time-scale, the signal components are separable in the time-frequency plane.

How can one automatically select from the over-complete contour sets a representation of complex signals where each subcomponent of a signal is represented in its own natural time-scale and angle? An earlier publication suggested selecting contours with the simplest shapes [7]. Here we suggest a more general criterion - select structurally stable contours, regardless of their shapes. When a signal component is analyzed in its own natural time-scale and angle, then the long-range contours that represent the signal are structurally stable - the details of the shapes do not change with small variations in the parameters of analysis. Returning to Fig. 1, for example, one can observe that at the optimal time-scale, contour shapes for the chosen signal do not depend sensitively on parameter θ. For this simple signal, the contours in the middle row of Fig. 1 are the structurally stable contours. Fig. 2 and 3 illustrate how the structural stability of contours can highlight a tonal signal embedded in noise. The analysis reveals a quiver of similarly shaped contours that track the tonal component of the sound for an angle near $\theta \approx 0$. To quantify the structural stability of a contour, we (1) calculate a set of contours for a range of parameters, σ_t and θ. (2) Create a sparse time-frequency matrix, M_i representing "thickened" representations of each contour. The matrix for a contour is zero everywhere unless the pixel falls within a neighborhood of the contour defined by the resolution of the underlying Gabor transform used to generate the contour $(\Delta t = \sigma_t, \Delta f = 1/\sigma_t)$. (3) Define a consensus score for each contour

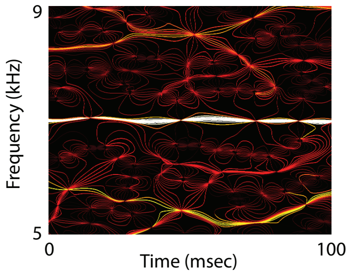

Fig. 3. Consensus scores enhance contours that follow signal rather than noise. The bundle of contours that track the embedded 7kHz sinusoid stand out from the noise in this analysis.

$$S_i = \max_j \left\{ M_i \cdot M_j \right\} \tag{4}$$

This maximization involves many sparse matrix multiplications (the computation is quadratic in the number of contours). In words- each contour is assigned a score defined by its maximal overlap with any other contour. In Fig. 2b, the light gray pixels represent the matrix for the red contour, the darker pixels the matrix for the blue contour, and their overlap $M_i \cdot M_j$ is the area of the white pixels. We call this the "consensus score" of the contour. This consensus score is not normalized by contour length, so the scoring system favors contours that are both long and highly overlapping with some neighboring contour.

In Fig. 3, the contours of Fig. 2a are recolored according to their "consensus scores," a process that highlights the signal region containing the sinusoid. Specifically, if the coordinates of i th contour are represented by a matrix, C_i in a discrete approximation to χ , then Fig 3 is a consensus image defined by

$$CI = \sum S_i C_i \tag{5}$$

Fig. 3 was calculated using a set of contours defined in a single time-scale and many angles; the more general approach used in the subsequent figures combines contours across variations in both time-scale and angle. A simple example illustrating this multi-band approach can be found in Fig. 4, which demonstrates the analysis of a click and a tone embedded in noise. The "consensus images" in this figure are produced by the pointwise histogram of all contours, weighted by their individual consensus scores. For this signal, the consensus image (Fig. 4a) accurately tracks both the click and the tone since each component is represented using information in its own natural time-scale. It must be emphasized that no a-priori information was applied to this figure. The consensus contour analysis also works for complex signals. Roughly speaking, as long as signal components are locally spaced by distances in time and frequency greater than the spread of the time-frequency uncertainty (for some time-scale), the method will highlight these components by emphasizing contours drawn from the appropriate time-scales.

Fig. 5 demonstrates how the consensus operation can reduce the representation of contours of low structural stability in a complex signal, revealing a parsimonious signal representation. Fig. 5a shows the standard spectrogram of a bird song. Fig. 5b shows the collection of all contours $\left(\sum C_i\right)$ of the same bird song, where contours are calculated over a narrow range of relevant time-scales. Even though the time-scales are already matched to zebra finch song, the summed image is visually dense with signal components multiply represented in different angles and time-scales. Fig. 5c shows the CI (Eq. 5) for that same song, which highlights the structurally stable features of the data. Fig. 5d further weights this image by the local spectrogram power, $\sum S_i C_i \cdot |\chi_i|$ where χ_i is the Gabor transform used to calculate i th contour. Fig. 6 applies the same consensus enhancement to the analysis of a human speech sample. In this analysis, consensus contours at low angles track some of the the formants, while consensus contours at steep angles track the glottal pulses.

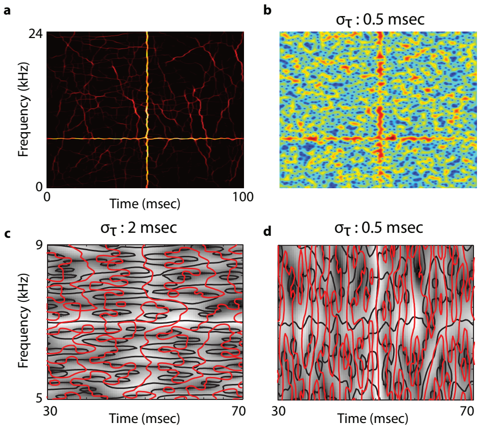

Fig. 4. Consensus highlights signal in noise. The analyzed signal is a fragment of white noise with an embedded sinusoid at 7kHz, and a click at t=50ms. (a) Contours weighted by consensus highlight the signal components with high temporal precision for the click and high frequency precision for the sinusoid. (b) Local amplitudes of the Gabor transform. In panels (c) and (d) individual contours are shown for 2ms and .5ms time-scale respectively. (Red, $\theta = \pi / 2$: Black, $\theta = 0$). The black contour in panel (c) tracks the sinusoid, while the red contour in panel (d) tracks the click. The two contours that track the signal components are structurally stable and stand out relative to noise in the consensus-weighted image in panel (a).

4. COMPUTATIONAL METHODS AND RESYNTHESIS

The analysis described here uses the Discrete Gabor transform (2048 frequency bins, Signal sampling rate 25 kHz or 48kHz) with a window overlap of 2038 samples. All matrixes used in contour calculations have the same resolution. For resynthesis, we do not synthesize exact waveforms for each contour as described previously [7], but use an overlapped inverse FFT for each column of the time-frequency consensus image. The purpose of the consensus representation is not to exactly represent the original signal but to capture the salient features of a signal as parsimoniously as possible - a sparse approximation to the original signal. The accuracy of the resynthesis can scale

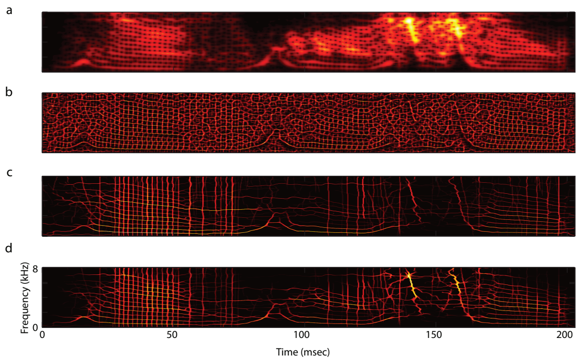

Fig. 5. Consensus process can reveal parsimonious representation of signal. Input signal is a short syllable of zebra finch song. (a) Standard spectrogram of signal analyzed for time-scales 0.3 ~ 2.2 ms, (b) Collection of all contours (c) Consensus image of the signal weighted by consensus score (d) Consensus image of signal weighted by consensus score and local power from spectrogram.

smoothly from a high quality perceptual match to compact, lower quality representations of sound, depending on the cutoff in contour consensus scores. The consensus image from Fig. 6b incorporating contours from all angles provides a fairly complete perceptual resynthesis of the speech sample. Resynthesis based on Fig. 6c remains intelligible, since many of the signal formants are captured by this population of contours.

5. CONSENSUS PROVIDES AN OBJECT-BASED SIGNAL ENHANCEMENT

In the consensus process, contours are never subdivided. If a contour contributes to the final representation, it does so in its entirety, even if some time-frequency points along the contour are not "in consensus" with some other contour. Fig. 7 reveals how the notion of contour consensus differs from a simple measure of contour density. For a double chirp signal, contours are calculated in time-scales in the range of 2-10ms.

In Fig. 7a and 7b the time-frequency points of highest pixelwise overlap fall between the two sweeps. Fig. 7c contains the result of the contour-based consensus (Eq. 5) . The take home message from this figures is that pointwise measures of contour density can fail to extract parsimonious representations. Any process of thresholding the images in Fig. 7a or 7b will fail to discover the parsimonious representation in Fig. 7c. To achieve the gains of the contour-based analysis, the method must amplify stable contours rather than just stable pixels. This contour or "object-based" time-frequency principle was absent from a prior definition of cross-bandwidth consensus [11]. The overlap of *reassigned* pixels in a multi-band analysis produces a figure similar to Fig. 7b. Reassignment alone does not provide the gains of an "object-based" time-frequency analysis, since reassignment does not link together associated points in the time-frequency plane.

6. LIMITATIONS

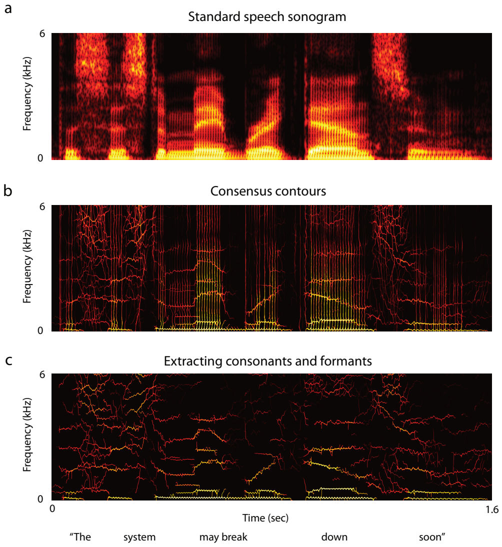

Fig. 6. Spectrogram and contour representation of human speech. (a) Spectrogram of a human speech sentence, "The system may break down soon." Gabor transform calculated with $\sigma_t = 3$ ms. (b) Top scoring consensus contours calculated for time-scales 1-4.5ms. (c) Highlighting the shallow- angle consensus contours drawn from 2-4.5ms time-scale. Many of these low-angle contours follow the formants, or vocal tract resonances essential to the perception of speech.

The consensus images (CI in Fig. 5c) combine qualitatively different forms of information, and for some applications, these should be kept separate. Specifically, the consensus score, S_i for a single contour includes information about both the structural stability of form and length of the contour. Furthermore, the consensus image (CI in Fig. 5c) is influenced by contour density at each point in time and frequency. The images that appear to be most useful add yet one more feature - the local weighting of consensus contours by the spectrogram power (Fig. 5d) For any quantitative analysis, a statistical understanding of the relative contributions of these features will be needed.

The consensus score depends on the granularity of the parameter space in time-scale (σ_t) and angle (θ) that is explored. The score also depends on the number of frequency channels, and the temporal overlap or step-size in the discrete approximation to the analytic Gabor transform (spectrogram). A principled approach to this analysis could compute contour score distributions in noise, and then select signal contours based on their likelihood in this background distribution. However, this noise distribution must be re-computed for the exact parameter settings used in each analysis.

Not only does the granularity of the parameter search affect the results, but a scoring process based on consensus scores requires awkward decisions such as how to rank a vertical contour that tracks a click relative to a horizontal contour that tracks a tone (The choice of discretization for the Gabor transform influences this relative weighting since it impacts contour length which is folded into the consensus scores).

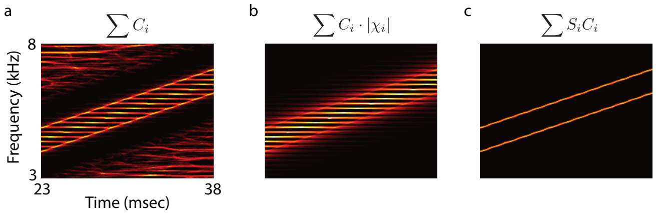

Fig. 7. Pixel-based consensus measures fail to extract a parsimonious representation. (a) Overlap of unweighted contours calculated in timescales 2-10ms. Pixels of highest overlap fall between the two sweeps. (b) Overlap of the same contours, weighted by the local power of the Gabor Transform. (c) Top scoring contours calculated by the consensus measure define here.

The consensus representation is lossy (unless all contours are taken), and what is lost depends on the details of a complex scoring process.

Calculating contour sets requires little time beyond the calculation of the discrete spectrogram, but the present method repeats this calculation over a two dimensional parameter space and then adds to this a scoring processing involving sparse matrix multiplication that is quadratic in the number of contours. A significant advance would embody the principle of the contour stability analysis in a simpler process that was less computationally intensive.

7. DISCUSSION

The basis of the contour representation is the observation that any signal can be represented as a collection of contours in the time-frequency plane with associated simple waveforms. For spectrally dense signals such as white noise, local contour shapes change quickly with small changes in the time-scale or angle of analysis. However, signal components that are separable from the background can produce contour shapes that are stable to changes in the parameters of analysis. By emphasizing long, structurally stable contours, parsimonious signal representations can be found where separate components are represented in their own natural time-scales and angles of analysis.

The result is not just an adaptive time-frequency analysis, but provides an elementary form of stream segregation since contours that survive the winnowing process can be re-assembled element by element to capture chosen features of the original signal. The vertical and horizontal components of the signal in Fig. 4 can be separately resynthesized by taking only consensus contours from vertical or horizontal angles. Similarly, many of the formants of the speech sample are separated from glottal pulses by taking only the low-angle consensus contours (Fig. 6c).

The method described here is rooted in the analytic structure of the Gabor transform, but the principles will generalize to other transforms such as the chirplet transform [12]. Structurally stable forms are found when each component of a signal is represented in its own natural time-scale and angle. This principle can guide an adaptive time-frequency analysis, though quantitative benchmarks remain to be examined.

8. ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for comments on this manuscript. This work is supported in part by CELEST, a National Science Foundation Science of Learning Center (NSF OMA-0835976), 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.

9. REFERENCES

[1] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *Signal Processing, IEEE Transactions on*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] M. S. Lewicki, “Efficient coding of natural sounds,” *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[3] R. R. Coifman and M. V. Wickerhauser, “Entropy-based algorithms for best basis selection,” *Information Theory, IEEE Transactions on*, vol. 38, no. 2, pp. 713–718, 1992.

[4] D. Jones and R. G. Baraniuk, “A simple scheme for adapting time-frequency representations,” *Signal Processing, IEEE Transactions on*, vol. 42, no. 12, pp. 3530–3535, 1994.

[5] D. Jones and T. Parks, “A high resolution data-adaptive time-frequency representation,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 12, pp. 2127–2135, 1990.

[6] D. Rudoy, P. Basu, and P. J. Wolfe, “Superposition Frames for Adaptive Time-Frequency Analysis and Fast Reconstruction,” *Signal Processing, IEEE Transactions on*, vol. 58, no. 5, pp. 2581–2596, 2010.

[7] Y. Lim, B. Shinn-Cunningham, and T. J. Gardner, “Sparse contour representations of sound,” *IEEE Signal Processing Letters*, vol. 19, no. 10, pp. 684–687, Oct. 2012.

[8] F. Auger and P. Flandrin, “Improving the readability of time-frequency and time-scale representations by the reassignment method,” *Signal Processing, IEEE Transactions on*, vol. 43, no. 5, pp. 1068–1089, 1995.

[9] P. Mitra and H. Bokil, *Observed Brain Dynamics*. Oxford aUniversity Press, USA, 2007.

[10] D. J. Thomson, “Spectrum estimation and harmonic analysis,” presented at the Proceedings of the IEEE, 1982, vol. 70, no. 9, pp. 1055–1096.

[11] T. J. Gardner and M. O. Magnasco, “Sparse time-frequency representations,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 16, pp. 6094–6099, 2006.

[12] M. Aoi, Y. Lim, U. Eden and T. J. Gardner, “Object-based spectro-temporal analysis of auditory signals,” *Computational and Systems Neuroscience (COSYNE)*, Salt Lake City, 2013