This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Multi-resolution Common Fate Transform

Fatemeh Pishdadian and Bryan Pardo, *Member, IEEE,*

***Abstract*—The Multi-resolution Common Fate Transform (MCFT) is an audio signal representation useful for representing mixtures of multiple audio signals that overlap in both time and frequency. The MCFT combines the invertibility of a state-of-the-art representation, the Common Fate Transform (CFT), and the multi-resolution property of the cortical stage output of an auditory model. Since the MCFT is computed based on a fully invertible complex time-frequency representation, separation of audio sources with high time-frequency overlap may be performed directly in the MCFT domain, where there is less overlap between sources than in the time-frequency domain. The MCFT circumvents the resolution issue of the CFT by using a multi-resolution 2D filter bank instead of fixed-size 2D windows. This enables higher quality separation without the need to hand-tune the window size to the specific case. In this work, we describe the MCFT, discuss the properties of the MCFT with the aid of illustrative examples and provide definitions and objective measures for two desirable representation properties: *separability* of source signals and *clusterability* of components of each signal. The utility of the MCFT for source separation is illustrated by performing ideal masking on a comprehensive dataset of audio mixtures of musical tones played in unison, including audio samples from a wide pitch rage and a variety of instruments/playing techniques. Results show that the ideal masks made in the MCFT domain yield better separability than those made in commonly used time-frequency signal representations as well as the CFT. The use of the MCFT also results in more reliable clusterability than the CFT in most cases.**

***Index Terms*—Audio source separation, Multi-resolution Common Fate Transform, Separability, Clusterability.**

I. Introduction

AUDIO source separation refers to the process of estimating n source signals from m channel mixtures. It is an important enabling technology to a variety of applications, including: automatic speaker identification in a multi-speaker scenario [1], [2], speech recognition in noisy environments [3], musical instrument recognition in polyphonic audio [4], music remixing [5], music transcription [6], upmixing of stereo recordings to surround sound [7], [8], and lyric-music synchronization [9].

Underdetermined source separation is an important case, where the number of sources exceeds the number of recording channels. We focus on one of the most common underdetermined scenarios: performing separation on monophonic or stereo recordings of mixtures of two or more sound sources.

Many approaches to separation of underdetermined mixtures are applied to a time-frequency representation of the audio, such as the widely-used short-time Fourier transform (STFT). Dealing with high levels of energy overlap between sources is a major challenge faced by source separation algorithms that use time-frequency representations as their input. In general, regardless of the algorithm, if the input mixture is represented in the time-frequency domain, performance degrades as the time-frequency energy overlap between sources increases.

A number of source separation approaches map a time-frequency representation to another representation domain, so that the source separation problem can be solved through distance-based clustering. Clusters, which are assumed to correlate with sources, are then used to create time-frequency masks. Examples include approaches that perform the mapping with a mathematical formula, such as DUET [10] and Kernel Additive Modeling [11], as well as methods that learn a higher-dimensional embedding from data, such as Deep Clustering [12]. In all these cases, the final masking is performed in the time-frequency domain, which leaves the issue of time-frequency energy overlap unresolved.

The type of processing performed in human auditory system can inspire the development of richer representations that inherently increase the chance of better separation. For instance, a psychoacoustics principle called the *common fate* principle [13] states that spectral components with the same modulation properties (components moving up and down together in the time-frequency space) are more likely to be grouped into a single audio stream by human listeners.

The common fate principle has been employed by some methods such as Non-negative Tensor Factorization (NTF) [14] [15] at the algorithmic level, while leaving the underlying audio representation (magnitude spectrogram) unchanged. However, accounting for spectro-temporal modulation properties as explicit dimensions of the representation, would facilitate separation of sources with high time-frequency overlap without requiring the algorithm to grow too complex.

The method proposed by Abe et al. [16] is an early work that exploits modulation properties for source separation. In a recent attempt to address the difficulty in the separation of same pitch (unison), frequency-modulated sources, Stöter et al. [17] proposed a 4D representation, named the Common Fate Transform (CFT), which explicitly captures common fate. The use of the CFT for the separation of unison mixtures with different modulation properties has produced promising results [17].

The CFT is computed by dividing the complex STFT of an audio signal into overlapping 2D windows and then analyzing each windowed segment by the 2D Fourier transform. The main shortcoming of CFT is its use of the same fixed-size window over the entire STFT. This limits the transform-domain resolution, and hence affects the separation results for sources with close modulation patterns. To achieve maximal performance for a particular situation, a knowledgeable user must select the appropriate window size. It would, however, be preferable to attain good separation results without having the need for hand-tuning the window size.

The resolution issue of the CFT can be addressed by using a multi-resolution approach, i.e. analyzing the time-frequency representation over a range of window sizes, or equivalently

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

through a filter bank. The auditory model proposed by Chi et al. [18] transforms the audio signal into a 4D representation based on a multi-resolution analysis approach.

This auditory model emulates the important aspects of the cochlear and cortical processing stages in the auditory system of small mammals. The output representation captures spectro-temporal modulation patterns as two additional dimensions, named *scale* and *rate*. Since the main purpose of the model is to emulate the auditory system output, its cochlear processing stage includes non-linear operations and removal of phase information. As a result, the output representation is not fully invertible, which hinders its use in audio processing tasks, where the perfect reconstruction of time-domain audio signals is of crucial importance.

Krishnan et al. [19] proposed a source separation algorithm that uses the output of Chi's auditory model to build time-frequency-domain masks, but since it applies masking in the time-frequency domain, it remains susceptible to time-frequency overlap between sources. Mesgarani et al. [20] proposed a speech enhancement method based on filtering the noisy signal in the full 4D domain. The method is able to suppress noise with distinctive modulation patterns even in cases where there is time-frequency overlap between the speech and noise. However, to recover the acoustic signal they use the signal estimation algorithm accompanying the auditory model, which despite preserving the intelligibility of speech signals suffers from poor reconstruction quality.

In a pilot conference paper [21], we proposed a new representation, the Multi-resolution Common Fate Transform (MCFT), which combines the strengths of the CFT and of Chi's auditory model. The MCFT is computed based on a fully invertible complex time-frequency representation. It allows separation of sources with high time-frequency overlap in a 4D domain, where there is less overlap between sources. The MCFT circumvents the resolution issue of the CFT by using a multi-resolution 2D filter bank instead of fixed-size windows. This enables higher quality separation without the need to hand-tune the window size to the specific case.

In this paper, we substantially extend our pilot work. The mathematical formulation is described in greater detail. We give a detailed presentation of the principles based on which the common-fate-based representations, CFT and MCFT, are designed. We study the properties of the CFT and MCFT using illustrative examples. Furthermore, we discuss *separability* and *clusterability*, two desired properties for audio representations as well as objective metrics for evaluating these properties in different representation domains. We compare the efficacy for source separation of multiple audio representations on mixtures with significant time-frequency overlap.

The remainder of this paper is organized as follows: We introduce separability and clusterability, two important properties of audio representations affecting the performance of source separation algorithms in Sections II and III. The precursors to our work, the Common Fate Transform (CFT) and Chi's auditory model, are then studied in detail in Sections IV and V, respectively. In Section VI, we present the Multiresolution Common Fate Transform (MCFT) and discuss its important properties. Experimental results showing the separability and clusterability of a variety of representations are presented in Section VII. Section VIII concludes the paper and briefly discusses the significance of this work.

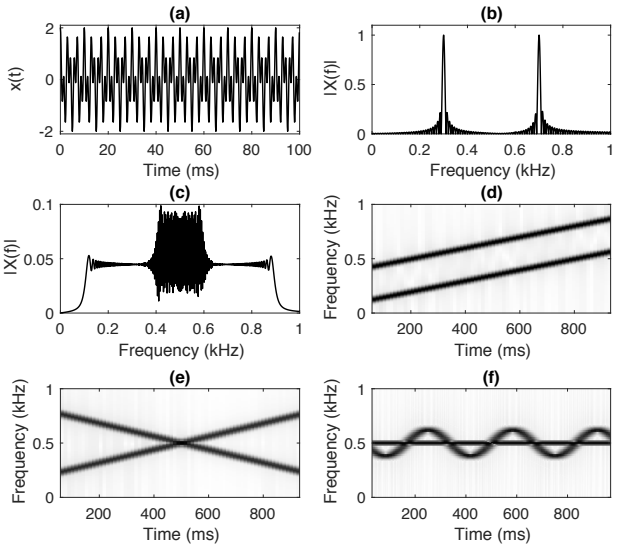

Fig. 1. Examples of simple audio mixtures with different levels of separability based on the representation used. Panels (a) and (b) display a mixture of two single frequency sinusoids, represented in the time domain and in the frequency domain respectively. In panels (c) and (d), a mixture of two linear chirps is demonstrated, in the frequency domain and in the time-frequency domain respectively. Panels (e) and (f) show mixtures with overlapping energy in the time-frequency domain (a mixture of two linear crossing chirps and a mixture of two sinusoids with different frequency modulation patterns).

II. Audio representation and separability

In this section, we introduce the concept of *separability* as a measurable property of audio mixtures. The separability of two signals depends on the properties of the signals and also the properties of the representation domain. What seems inseparable in one representation domain may be easy to separate in another. Figure 1 shows four simple mixtures represented in different domains. The first example, displayed in Panels (a) and (b), is a mixture of two single-frequency sinusoids in the time and frequency domains respectively. Clearly, the separation task is very difficult in the former case, while quite easy in the latter. Panels (c) and (d) present the second example, a mixture of two linear chirps, in the frequency and time-frequency domains respectively. Going from the frequency domain to the time-frequency domain decreases the energy overlap between the two sources, and thus helps the separation. The two examples in Panels (e) and (f) demonstrate that even the time-frequency domain is not immune to overlap between sources as mixtures become more complex. As a matter of fact, the example of Panel (f), where the mixture consists of sources with significant time-frequency overlap, presents one of the most challenging scenarios for the source separation task.

We now provide basic definitions for an underdetermined, linear mixture source separation problem. Let $x(t)$ denote a

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

mixture of N time-domain audio signals, that is

$$x(t) = \sum_{j=1}^{N} s_j(t), \tag{1}$$

where t is the time index, $s_j(t)$ is the amplitude of the j^{th} source in the mixture at time t, and $u_j(t)$ indicate the sum of all sources interfering with $s_j(t)$, i.e.

$$u_j(t) = \sum_{i=1, i\neq j}^{N} s_i(t). \tag{2}$$

Assume a linear transform, denoted by $\mathcal{T}$, is applied to the audio mixture and its constituent sources, taking them from the time domain to a k-dimensional representation domain $\mathcal{D}$. Let $\mathbf{d} = (d_1, d_2, ..., d_k) \in \mathcal{D}$ be an arbitrary point in $\mathcal{D}$, and let $S_j(\mathbf{d}) = \mathcal{T}\{s_j(t)\}$ and $U_j(\mathbf{d}) = \mathcal{T}\{u_j(t)\}$ indicate the transformed versions of $s_j(t)$ and $u_j(t)$, respectively. In general, we assume transformed signals to be complex valued.

An ideal binary mask separating the j^{th} source from the rest of the mixture in the transform domain can be defined as [22]

$$M_{j,\gamma}(\mathbf{d}) = \begin{cases} 1 & if \;\; 20\log_{10}\left(\frac{|S_j(\mathbf{d})|}{|U_j(\mathbf{d})|}\right) > \gamma \\ 0 & otherwise, \end{cases} \tag{3}$$

where γ, measured in deciBels (dB) is the masking threshold. Note that for the above formula to be valid, both $|S_j|$ and $|U_j|$ values are assumed to be nonzero. In practice, we use the expression $20\log_{10}\left[(|S_j(\mathbf{d})| + \epsilon)/(|U_j(\mathbf{d})| + \epsilon)\right]$ with $\epsilon << 1$ to avoid numerical errors.

Equation (3) simply states that the total mixture energy at each point in the representation is assigned to the j^{th} source if it dominates the total interference from other sources by γ dB. In other words, the j^{th} source "loses" its energy at a given point if the energy ratio between the source and the interference does not pass the masking threshold. It is possible to have points where none of the sources is dominant. The values of all masks at such points are set to zero, and thus the mixture energy is not assigned to any of the sources.

A measure of separability in a representation domain can be defined as the energy portion of the j^{th} source preserved through masking, normalized by the total energy of the original source, where both the original and masked signals are placed within that representation domain. A version of such a measure, named approximate W-disjoint orthogonality (WDO), introduced by Rickard et al. [22], is calculated by placing the mixture into a time-frequency representation. It should be noted that the use of this energy ratio measure is only appropriate when comparing different mixtures represented in the same domain. Due to the dimensionality mismatch between the representation domains discussed throughout this work and different types of analysis methods and parameters involved in their computation (e.g. fixed-size windowing versus multi-resolution filtering), the outputs of such a measure are not comparable across representations.

To measure how well different representations naturally separate sources in a mixture, we take an alternative approach, which makes comparison of different domains possible. Instead of measuring the preserved energy ratio in the transform domain, we infer the separability based on the quality of the time-domain reconstructed sources that were separated via ideal binary masking in different representation domains. Since the main assumption in computing ideal binary masks for a representation domain is the dominance of at most one source at each point, the quality of the separated sources using such masks would be highly correlated with the level of separability provided by the representation. For time-domain evaluation of the separation performance, we use the BSS-Eval [23] objective measures: Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), and Source to Artifact Ratio (SAR).

III. Audio representation and clusteribility

In this section, we present a measure of *clusterability*, an important property of representations that, to our knowledge, has been little studied in the context of source separation. We define *clusterability* as the tendency of a representation domain to map the energy of audio sources such that the distance between points belonging to one source (*intra-cluster* distance) is considerably smaller than the distance between points from two different sources (*inter-cluster* distance). This is known as *distance-based* clusteribility [24]. A representation that enhances the distance-based clusterability would make the source separation task more straightforward, as a simple distance-based clustering algorithm (e.g. Gaussian Mixture Models [25]) could be used to assign energy to sources. This insight has been exploited in multiple source separation approaches (e.g. Kernel Additive Modeling [11], DUET [10], Deep Clustering [12]). We are unaware, however, of any work using an objective measure to evaluate the clusterability tendency of the input audio representation.

In their approach to image segmentation as a graph partitioning problem, Shi et al. [26] proposed the *normalized cut*, a criterion that simultaneously measures the total similarity between nodes belonging to the same group and the total dissimilarity between nodes in different groups. Bach et al. [27] derived a loss function based on the normalized cut for *spectral clustering*, a graph partitioning technique, which relies on the eigenstructure of the similarity matrix in order to assign nodes with high similarity to the same cluster and those with low similarity to different clusters.

In this work, we use the normalized-cut-based loss function of Bach et al. [27] as a measure of the clusterability offered by a representation. The ideal binary masks in a representation are considered the outputs of an ideal clustering algorithm for that representation. We treat the mask points with a value of one as the nodes of an undirected weighted graph. The pairwise distance in the representation space defines edge weights. This lets us compute the value of the normalized cut for the partitioning of the high energy points produced by ideal binary masks corresponding to sources in a mixture. Low normalized cut values for a given representation imply high levels of distance-based clusterability.

In practice, treating every point as equally important can be problematic. Since the only criterion for passing a masking

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Transform	Input	Computation Steps	Output
CFT	$x(t)$	STFT $\rightarrow$ 2D windows centered at (Ω, T) $\rightarrow$ $\mathcal{FT}_{2D}$	$Y(s, r, \Omega, T)$
ICFT	$Y(s, r, \Omega, T)$	$\mathcal{IFT}_{2D}$ $\rightarrow$ 2D overlap and add $\rightarrow$ ISTFT	$x(t)$

TABLE I
AN OVERVIEW OF THE COMPUTATION STEPS IN CFT AND ICFT.

threshold is the dominance of the target source energy and not the absolute energy level, there can be a large number of low-energy points in each estimated source representation that can be counted in the source cluster without contributing much to the total signal energy. Thus, we apply magnitude thresholding to masks to remove low-energy points. A second motivation for the use of this thresholding stage is to lower the computational burden in the calculation of similarity matrices by removing points that contribute little. In our experiments, we set the threshold to 20 dB below the maximum magnitude value for each estimated source.

Let W denote the similarity matrix for a given set of high-energy points in a k-dimensional representation domain, $\mathcal{D}$. Following the framework in Bach et al. [27], we assume the similarity between two arbitrary points $\mathbf{d}_i$ and $\mathbf{d}_j$ to be defined as a diagonally scaled Gaussian function of the distance between the two points, i.e.

$$W_{ij} = \exp(-(\mathbf{d}_i - \mathbf{d}_j)^\top diag(\alpha)(\mathbf{d}_i - \mathbf{d}_j)), \tag{4}$$

where W_{ij} indicates the value on the i^{th} row and j^{th} column of the similarity matrix, $\alpha \in \mathbb{R}^k$ is a vector of positive weights, and $diag(\alpha)$ is a $k \times k$ diagonal matrix with diagonal α.

We use a formulation of the loss function from Bach et al. [27]. Let $v_n \in \mathbb{R}^m$ be the indicator vector for the n^{th} cluster, i.e. $v_n \in \{0, 1\}^m$ only has nonzero values for points belonging to the n^{th} cluster. With $V = (v_1, ..., v_N) \in \mathbb{R}^{m \times N}$ denoting the set of all indicator vectors associated with the N clusters, the loss function can be written as

$$\mathcal{L}(V, W) = \frac{1}{N-1} \sum_{n=1}^{N} \frac{v_n^\top (D - W) v_n}{v_n^\top D v_n}, \tag{5}$$

where $D = diag(W\mathbf{1})$, ($\mathbf{1} = (1, 1, ..., 1)^\top \in \mathbb{R}^m$), is a diagonal matrix, whose i^{th} diagonal element is the sum of all elements in the i^{th} row of W. The value of $\mathcal{L}(V, W)$ is always between zero and one, with lower values indicating higher clusterability. A more intuitive objective measure can be defined as $1 - \mathcal{L}(V, W)$, such that higher values are associated with better clusterability.

IV. COMMON FATE TRANSFORM

In this section, we provide a brief introduction to the Common Fate Transform (CFT), proposed by Stöter et al. [17] and study the prominent characteristics of audio representation in this transform domain, which make its use beneficial for the task of audio source separation.

To formulate the transform, let us denote a single channel time-domain audio signal by $x(t)$ and its complex time-frequency-domain representation by $X(\omega, \tau) = |X(\omega, \tau)|e^{j\angle X(\omega,\tau)}$, where ω, τ, $|.|$, and $\angle(.)$ respectively denote frequency, time, the magnitude and phase operators. In the original version of CFT [17], $X(\omega, \tau)$ is defined as the STFT of $x(t)$. Due to the Hermitian symmetry of the Fourier transform of real signals, only the values of $X(\omega, \tau)$ for positive frequencies are stored for future processing.

In the following step, 2D windows, overlapped along both frequency and time axes are applied to $X(\omega, \tau)$. The 2D Fourier transforms of windowed segments are then computed and concatenated to form a 4D tensor. To keep the terminology and notation consistent throughout this paper, we refer to the 2D Fourier transform domain as the *scale-rate* domain. The scale and rate dimensions explicitly encode the spectro-temporal modulation information, where the former captures the spectral spread and the latter the modulation velocity over time (see Section VI). Let $Y(s, r, \Omega, T)$ denote the 4D representation generated by the CFT. Here, (s, r) indicates the scale-rate coordinate pair and (Ω, T) the 2D window centers along the frequency and time axes.

It should be noted that the CFT is perfectly invertible. The single-sided complex STFT, $X(\omega, \tau)$, can be reconstructed from $Y(s, r, \Omega, T)$ by taking the 2D inverse Fourier transform of all patches and then performing 2D overlap-and-add. Subsequently, the time-domain signal, $x(t)$, can be obtained by taking the 1D inverse Fourier transform of all time-frames and performing 1D overlap-and-add. The operations performed in the CFT and the inverse CFT (ICFT) computation are summarized in Table I.

As mentioned earlier, the CFT maps the signal energy from the time-frequency domain into a 4D space based on the common fate principle. The time-frequency components are, therefore, grouped based on their moving directions and mapped into different points in the target domain. Such a grouping property, which arises from the use of the 2D Fourier transform is in particular advantageous when dealing with mixtures of frequency-modulated harmonic signals. Since components of harmonic signals move up and down together in the time-frequency domain, they are likely to be mapped into the same locations in the scale-rate domain, causing harmonic elements of the same signal to group together in this representation. Such a mapping potentially increases the separability and/or clusterability of the data points, and hence makes it easier to isolate only those sound components belonging to the target source.

In the remainder of this section, we present illustrative examples of taking the 2D Fourier transform of a time-frequency representation. This will provide the reader a more intuitive understanding of this domain. In these examples, we consider the 2D representation domains in isolation and compare their properties. This approach is taken mainly due to the difficulty of higher-dimensional visualization. However, it is important to note that merely going from the time-frequency domain to the scale-rate domain does not necessarily result

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

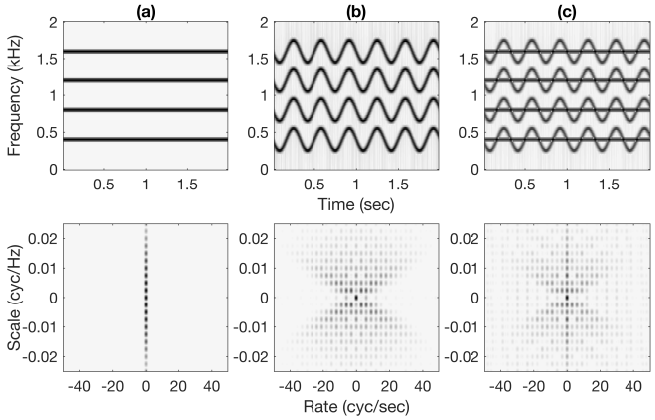

Fig. 2. Two harmonic signals, with and without frequency modulation and their mixture. The two signals have the same fundamental frequency (400 Hz). Columns (a) and (b) show the two separate signals and column (c) shows their mixture. Signals are represented in the time-frequency domain (top row) and in the scale-rate domain (bottom row).

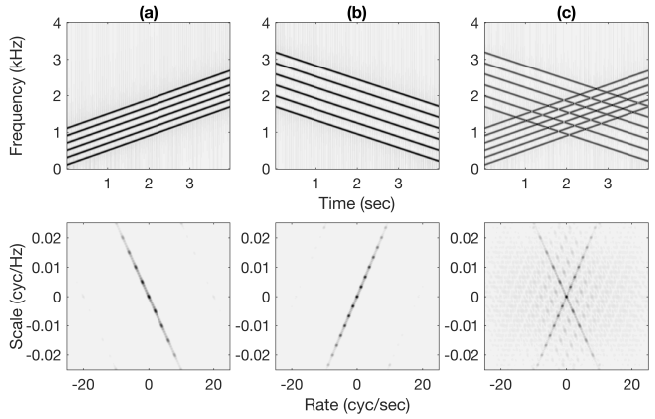

Fig. 3. Two linear group-chirps (linear chirps with the same slope) with upward and downward moving directions, each considered a separate source, and their mixture. The two sources have energy overlap at various time-frequency points. Columns (a) and (b) show the two separate signals and column (c) shows their mixture. Signals are represented in the time-frequency domain (top row) and in the scale-rate domain (bottom row). We note that the opposing directions of signal representations with respect to the vertical axis in the original and transform domains (i.e. upward in one and downward in the other) is adopted to be in consistence with the image processing literature, although we admit that it might seem counterintuitive to readers that are not familiar with multi-dimensional signal processing concepts.

in better separability or clusteribility. The power of the 4D representations studied in this work (CFT and MCFT) lies in combining the information from the scale-rate domain and the time-frequency domain. An analogy can be drawn to the example of Figure 1, Panels (c) and (d). Panel (c) shows that merely going from the time domain to the frequency domain does not completely solve the problem of overlapping energy. High separability is achieved when the time-domain information is processed over short windows and then combined with the frequency-domain information, resulting in a higher dimensional representation, displayed in Panel (d).

Figure 2 shows two harmonic signals, one with and one without frequency modulation, and their mixture. Both signals have a fundamental frequency of 400 Hz, and hence overlap significantly in the time-frequency domain. The magnitude

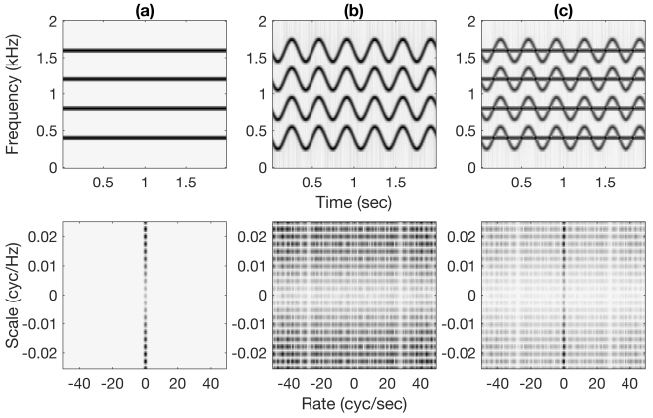

Fig. 4. The top row shows the two sources and their mixture in the time-frequency domain. The bottom row shows the 2D Fourier transform magnitude of the *complex* STFT of the signal. Compare this to Figure 2, which shows the 2D Fourier transform of the *magnitude* STFT of the same signal.

STFTs of the three signals are depicted in the top row and the 2D Fourier transforms of magnitude STFTs in the bottom row. As it can be seen, the energy of the non-modulated source, represented by horizontal lines in the time-frequency domain is mapped into the zero-rate line, whereas the energy of upward or downward moving ripples of the modulated source is mapped to points scattered over non-zero rate values.

Figure 3 illustrates two crossing linear group-chirps moving in opposite directions and their mixture. Each group of linear chirps with the same slope is considered as one source. The two sources overlap at various points in the time-frequency domain. The plots in columns (a) and (b) show the two sources in the time-frequency domain (top row) and scale-rate domain (bottom row) and the plots in column (c) show the mixture. Each line in the "X" shape pattern that emerges in the scale-rate-domain representation of the mixture corresponds to one moving direction. In this case, going from the time-frequency domain to the scale-rate domain increases separability to some extent by remapping the components based on their moving directions, and thus reducing the number of overlapping points down to one. One might argue that the clusterability is also increased since the energy from parallel lines in the time-frequency domain, regardless of their relative spacing, is mapped into a single line in the scale-rate domain.

In the above examples, we only considered the effect of applying the 2D Fourier transform to the magnitude STFT. Nevertheless, it should be noted that the CFT is computed from the complex STFT, where the inclusion of the phase would alternate previously observed patterns in the scale-rate domain. This is what renders the time-frequency-domain audio representation more challenging to analyze through the 2D Fourier transform than photographic images, which are typically 2D real signals.

Figure 4 shows the same example as in Figure 2 along with the scale-rate-domain representation of the complex STFT. It can be observed that including the time-frequency-domain phase results in a shift in the scale-rate domain. The scale-rate-domain representation is still expected to offer more separability for the components overlapping in the time-frequency

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

domain, although it seems to have lost the nice clusterability property of the magnitude-only case. Our experimental findings discussed in Section VII confirm this expectation. That is, in going from the *complex* STFT domain to the CFT domain the results show an increase in separability, although there is a possibility for the loss of clusterability. We note that a general study of the time-frequency phase is beyond the scope of this work and would be the subject of our future research.

Similar to the frequency resolution of the STFT which is determined by the time-domain window size, the scale and rate resolutions are determined by the dimensions of the time-frequency-domain windows. Consequently, the choice of the 2D window size has a direct impact on the representation quality of the CFT in terms of the provided separability.

The effect of the window size on the transform-domain resolution is illustrated in Figure 5. Panel (a) presents the magnitude STFT of a frequency modulated harmonic signal with a fundamental frequency of 200 Hz. The 2D Fourier transforms of four windowed segments with different window dimensions are depicted in Panels (b)-(e). As it is clearly observed in the plots, an increase in the window size along the time or frequency axis results in an increased resolution along the rate or the scale axis respectively. In the case with the lowest resolution in both directions shown in Panel (b), the scale-rate-domain representation of the windowed segment is quite blurry and only a large peak at the center can be detected, whereas in higher resolution cases, e.g. Panel (e), a number of lower peaks associated with upward and downward moving components also appear in the plot. It can also be seen that each window, depending on its duration over time captures one or both moving directions. For instance, the upward direction is not emphasized by the short window of Panel (b) as strongly as it is by the longer windows of Panels (c) and (d).

No general guideline for choosing the window size is proposed by Stöter et al. [17], as the ideal window clearly depends on the signal content. In the two following sections, we show how our proposed multi-resolution approach in computing the time-frequency representation as well as the 4D representation largely eliminates the need to select the right window size.

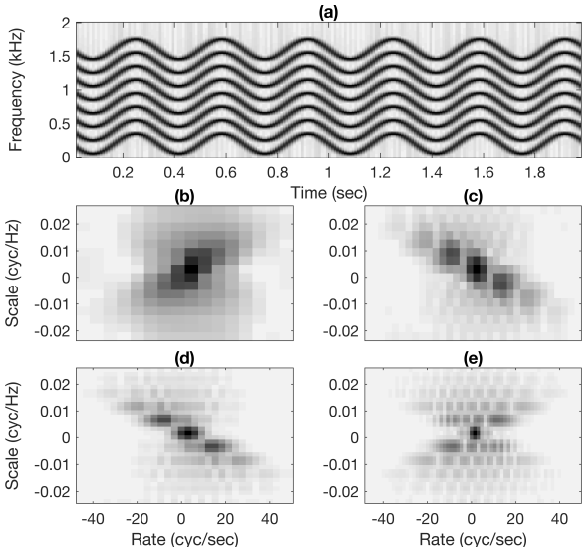

Fig. 5. The effect of 2D-window size on resolution of the scale-rate-domain representation of the magnitude STFT. Panel (a) shows the magnitude STFT of a harmonic, frequency-modulated signal with a fundamental frequency of 200 Hz. Panels (b-e) show the 2D Fourier transform magnitude of the signal over window sizes of 16×16, 16×32, 32×32, and 32×64 respectively.

V. The Auditory Model of Chi et al.

In the design of our representation, we were inspired by the multi-resolution aduitorty model of Chi et al. [18]. Recent studies on the primary auditory cortex of small mammals have shown the important role spectro-temporal modulation patterns play in audio perception and streaming [28] [29] [18]. In [18], Chi et al. present a computational model of early and central stages of the auditory system. The model outputs a 4D multi-resolution representation capturing spectro-temporal modulation patterns.

Their auditory model is composed of two stages: cochlear and cortical. The cochlear stage, as the name suggests, emulates the cochlear filter bank in performing spectral analysis on the input time-domain audio signal. The filter bank model is composed of 128 overlapping constant-Q bandpass filters, with logarithmically-spaced center frequencies. The collective passband of filters covers approximately 5.3 octaves. The goal of the cochlear stage in the model is to replicate, as accurately as possible, the time-frequency-domain representation of the audio signal generated by the cochlea and termed *auditory spectrogram*. To this end, additional operations such as high-pass filtering, nonlinear compression, half-wave rectification, and integration are performed on the output of the filter bank. These operations model the effect of processes taking place between the inner ear and midbrain.

The cortical stage replicates the type of analysis performed by the primary auditory cortex. The neuronal response of the primary auditory cortex to different spectro-temporal modulation patterns, termed Spectro-temporal Receptive Fields (STRFs), can be regarded as a bank of 2D filters. The role of the filter bank is to extract the spectro-temporal modulation patterns from the auditory spectrogram. Each filter within the filter bank is tuned to a particular modulation pattern. The time-frequency-domain impulse responses of the filters in the auditory model are modeled after STRFs ([28]).

An STRF is mainly characterized by: 1) its spectral spread (broad/narrow), referred to as *scale* 2) its frequency modulation velocity over time (slow/fast), referred to as *rate* 3) its moving direction in the time-frequency plane (upward/downward). Spectro-temporal modulation patterns are, therefore, described in terms of their scale and rate values, measured in cycles per octave and cycles per second, respectively. Scale and rate form the two additional dimensions (besides time and frequency) in the 4D output of the auditory model. The STRF models proposed in [18] play the central part in the multi-resolution analysis of modulation patterns. It is, therefore, important to go into some technical detail in this section about the computation of the model.

Let us denote an STRF that is tuned to an arbitrary scale-rate parameter pair (S, R) by $h(\omega, \tau; S, R)$ with ω and τ denoting

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the frequency and time respectively. Note that S and R are constant (scalar) values for a single filter and determine the filter characteristics (i.e. spectral spread, frequency modulation velocity, and moving direction). We denote the 2D Fourier transform of the STRF by $H(s, r; S, R)$, where the pair (s, r) indicates an arbitrary point in the transform (scale-rate) domain. The parameter pair (S, R), which is the same for h and H indicates the filter center in the scale-rate domain.

Mainly due to their diagonal movement in the time-frequency plain, STRFs cannot be modeled as separable functions of frequency and time, that is, $h(\omega, \tau)$ cannot be stated as $h(\omega, \tau) = f(\omega) \cdot g(\tau)$. In other words, more than one principal component would be required for describing the time-frequency-domain representation of an STRF. Nevertheless, the 2D Fourier transforms of STRFs are quadrant separable, meaning that they are separable functions of scale and rate in each quadrant of the scale-rate domain.

To derive the filter impulse response, first the spectral and temporal seed functions are to be defined. Chi et al. modeled the spectral seed function as a Gabor-like filter

$$f(\omega; S) = S \cdot (1 - 2(\pi S \omega)^2) e^{-(\pi S \omega)^2}, \tag{6}$$

and the temporal seed function as a gammatone filter,

$$g(\tau; R) = R \cdot (R\tau)^2 e^{-\beta R \tau} \sin(2\pi R \tau). \tag{7}$$

The dilation factors of the Gabor-like and gammatone filters in the above equations, S and R, are in fact the filter centers in the scale-rate domain. The dropping rate of the temporal envelop, or equivalently the filter bandwidth in the scale-rate domain, is controlled by the time constant of the exponential term, β. Since STRFs are not separable functions of frequency and time, the moving direction (up/down) of the time-frequency-domain components cannot be captured by a simple product of the seed functions. However, the quadrant separability of these functions allows computing their 2D Fourier transform as the product of the 1D Fourier transforms of the seed functions. This operation can be formulated as

$$F(s; S) = \mathcal{FT}_{1D}\{f(\omega; S)\}, \tag{8}$$

$$G(r; R) = \mathcal{FT}_{1D}\{g(\tau; R)\}, \tag{9}$$

$$H(s, r; S, R) = F(s; S) \cdot G(r; R), \tag{10}$$

where $\mathcal{FT}_{1D}$ denotes the 1D Fourier transform.

To generate the time-frequency-domain representation of an up-/down-ward moving filter, the value of H over a pair of opposing quadrants must be set to zero. The scale-rate domain response of the upward-moving filter, indicated by $H^{\Uparrow}(s, r; S, R)$ is defined as

$$H^{\Uparrow}(s, r; S, R) = \begin{cases} H(s, r; S, R) & (s \geq 0, r \leq 0) \\ H(s, r; S, R) & (s < 0, r > 0) \\ 0 & otherwise. \end{cases} \tag{11}$$

Similarly, the response of the downward filter, $H^{\Downarrow}(s, r; S, R)$ can be defined as

$$H^{\Downarrow}(s, r; S, R) = \begin{cases} H(s, r; S, R) & (s \geq 0, r \geq 0) \\ H(s, r; S, R) & (s < 0, r < 0) \\ 0 & otherwise. \end{cases} \tag{12}$$

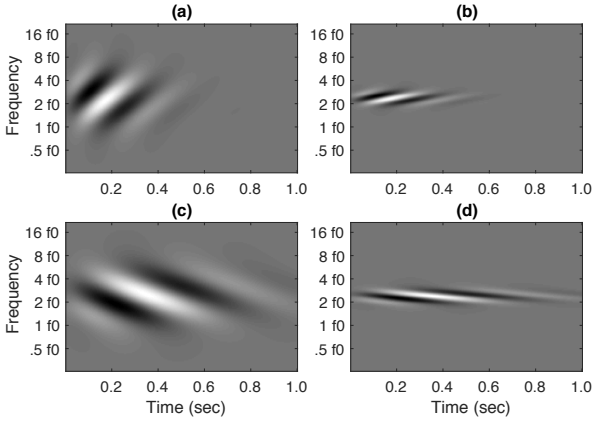

Fig. 6. Impulse responses, known as Spectro-temporal Receptive Fields (STRFs), of four filters from the 2D filter bank: (a) Upward-moving STRF $h^{\Uparrow}(\omega, \tau; S = 0.5, R = 4)$ (low scale, high rate). (b) Upward-moving STRF $h^{\Uparrow}(\omega, \tau; S = 2, R = 4)$ (high scale, high rate). (c) Downward-moving STRF $h^{\Downarrow}(\omega, \tau; S = 0.5, R = 2)$ (low scale, low rate). (d) Downward-moving STRF $h^{\Downarrow}(\omega, \tau; S = 2, R = 2)$ (high scale, low rate). The frequency is displayed on a logarithmic scale based on a reference frequency f_0.

In the next step, the impulse responses are computed as

$$h^{\Uparrow}(\omega, \tau; S, R) = \Re\{\mathcal{IFT}_{2D}\{H^{\Uparrow}(s, r; S, R)\}\} \tag{13}$$

$$h^{\Downarrow}(\omega, \tau; S, R) = \Re\{\mathcal{IFT}_{2D}\{H^{\Downarrow}(s, r; S, R)\}\} \tag{14}$$

where $\Re\{.\}$ denotes the real part of a complex value, and $\mathcal{IFT}_{2D}\{.\}$ the 2D inverse Fourier transform.

Examples of 2D filter impulse responses (STRFs) for different values of S and R are presented in Figure 6. Panels (a) and (b) correspond to upward moving filters, both with a rate of 4 cycles per second (a full cycle of the sinusoidal pattern covers 0.25 seconds). Panels (c) and (d) show downward moving filters with a rate of 2 cycles per second. In all panels, the frequency is shown on a logarithmic scale based on a reference frequency f_0, which maps frequencies that are separated by multiple octaves (an octave is a power of 2 relationship between frequencies) to a linear scale. The scale value for Panels (a) and (c) is 0.5 cycles per octave (a full cycle of 2 octaves), while Panels (b) and (d) demonstrate filters with a scale of 2 cycles per octave.

To compute the output representation, a bank of 2D filters, computed as described for various (S, R) values and different moving directions is applied to the auditory spectrogram. We denote the 4D output of the cortical stage by $Z(S, R, \omega, \tau)$, where (S, R) give the filter center in the scale-rate domain. Note that since the fast Fourier transform has a lower computational complexity than convolution, filtering can be performed more efficiently in the scale-rate domain.

The main disadvantage of the output representation of the auditory model, which hinders its use in signal processing tasks such as source separation, is the lack of invertibility. The non-linear operations in the cochlear stage and the removal of phase-related information makes perfect reconstruction impossible. An algorithm for estimating the time-domain signal from the 4D output representation is proposed in [18]. Unfortunately, the quality of the estimated audio signal is not acceptable for audio processing applications.

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Transform	Input	Computation Steps	Output
MCFT	$x(t)$	CQT $\rightarrow$ $\mathcal{FT}_{2D}$ $\rightarrow$ 2D filters centerd at (S,R) $\rightarrow$ $\mathcal{IFT}_{2D}$	$\tilde{Z}(S,R,\omega,\tau)$
IMCFT	$\tilde{Z}(S,R,\omega,\tau)$	$\mathcal{FT}_{2D}$ $\rightarrow$ 2D inverse filters centered at (S,R) $\rightarrow$ $\mathcal{IFT}_{2D}$ $\rightarrow$ ICQT	$x(t)$

TABLE II
AN OVERVIEW OF THE COMPUTATION STEPS IN MCFT AND IMCFT.

VI. MULTI-RESOLUTION COMMON FATE TRANSFORM

In this section, we propose a new representation, which circumvents the shortcomings of the CFT and the auditory model output and combines their strengths. To address the invertibility issue, we replace the auditory spectrogram by a fully invertible complex time-frequency representation with log-scale frequency. The Constant-Q Transform (CQT) is a multi-resolution time-frequency representation, where the resolution is progressively more coarse-grained as frequency increases. The log-scale frequency spacing of the CQT is similar to the frequency spacing of the auditory spectrogram in Chi et al. auditory model (see Section V). Unlike the auditory spectrogram, however, the CQT captures the phase. In our implementation, we use the CQT as proposed by Schörkhuber et al. [30], which is fully invertible back to the time domain.

To compute the new 4D representation, the cortical filter bank of the auditory model is applied to the complex CQT of the audio signal. This new representation is termed the Multi-resolution Common Fate Transform (MCFT), and denoted by $\tilde{Z}(S,R,\omega,\tau)$. The MCFT addresses the resolution issues of the CFT in the time-frequency domain as well as the scale-rate domain. The linear-scale frequency of the STFT offers a fixed resolution for the whole range of musical notes. Given that the fundamental frequency of musical notes are distributed on a logarithmic scale, the STFT would not be able to resolve low-frequency notes as effectively as high-frequency notes.

The use of a multi-resolution 2D filter bank instead of fixed size 2D windows in the spectro-temporal modulation analysis stage, results in an improvement in the scale-rate domain resolution of the MCFT compared to the CFT. The difference between the modulation analysis stages in the MCFT and CFT is analogous to the difference between the frequency analysis stages in the CQT and STFT, in that one of the transforms performs the short-term analysis through fixed-size windowing in the original domain, while the other by multi-resolution filtering in the transform domain.

The time-domain signal can be reconstructed from $\tilde{Z}(S,R,\omega,\tau)$ in two steps. First, the time-frequency representation is reconstructed from $\tilde{Z}(S,R,\omega,\tau)$ by inverse filtering:

$$\hat{X}(\omega,\tau) = \mathcal{IFT}_{2D}\left\{\frac{\sum_{S,R}^{\Uparrow\Downarrow}\tilde{z}(s,r;S,R)H^*(s,r;S,R)}{\sum_{S,R}^{\Uparrow\Downarrow}|H(s,r;S,R)|^2}\right\}, \tag{15}$$

where $*$ is complex conjugate, $\tilde{z}(s,r;S,R)$ denotes the 2D Fourier transform of $\tilde{Z}(\omega,\tau;S,R)$ for a particular (S,R), and $\sum_{S,R}^{\Uparrow,\Downarrow}$ indicates summation over the whole range of (S,R) values and all up-/down-ward filters. The time domain signal is then reconstructed from $\hat{X}(\omega,\tau)$ using the inverse CQT method proposed in [30]. Table II gives a summary of operations performed in the MCFT and IMCFT computation.

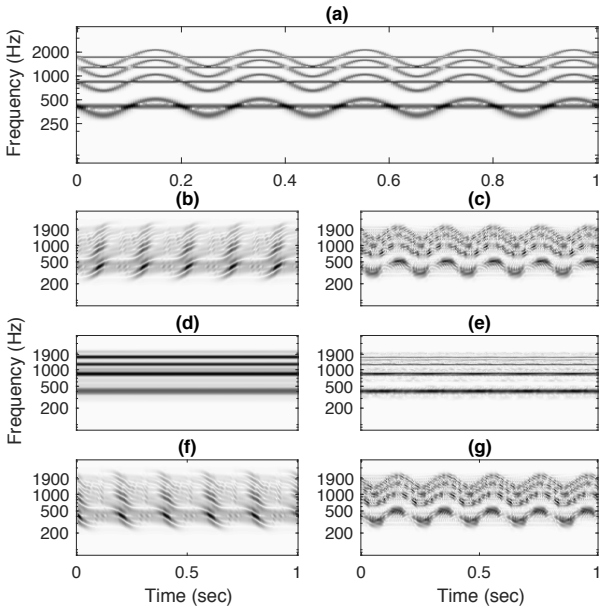

Fig. 7. (a) Magnitude spectrogram of a mixture of two harmonic sources one with and one without frequency modulation. (b,d,f) Magnitude spectrograms of the filtered mixture. Filters are applied to the magnitude spectrogram. (c,e,g) Magnitude spectrogram of the filtered mixture. Filters are first modulated with the mixture phase and then applied to the complex spectrogram.

In previous sections, we mostly studied the scale-rate-domain behavior of the magnitude of the time-frequency representation. Including the phase in the time-frequency domain results in a shift in the location of scale-rate-domain components (see Figure 4). Including the phase, however, allows for invertibility of the CFT and MCFT back to the time domain, which in turn allows separation to be performed in the 4D domain, where separability is improved, compared to the time-frequency domain. The cost is that including phase may introduce some scattering to the patterns in the representation that could potentially reduce clusterability. In practical use, this potential reduction in clusterability is outweighed by the improvement for separability, as illustrated in the experiments in Section VII. An in depth study of the phase behavior for all types of audio signals is beyond the scope of this work.

The method we use to deal with the effect of phase is shifting the filter components in the scale-rate domain in accordance with the shift in the location of mixture components. This can be achieved through modulating the filters with the phase of the mixture CQT, i.e. using filters with impulse responses equal to $h(\omega,\tau;S,R)e^{j\angle X(\omega,\tau)}$. Panel (a) in Figure 7 shows the magnitude CQT of a mixture of harmonic and non-harmonic signals. Panels (b), (d), and (f) present the output of three filters applied to the magnitude CQT. The upward and downward moving components of the modulated

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Instrument	ModulationTechnique	Note	Instrument	Modulation Technique	Note
piano	-	C2, C3, C4, C5, C6, C7	english horn	vibrato, major trill, minor trill	C4, C5
contrabassoon	vibrato	C2	clarinet	major trill, minor trill	C4, C5, C6
contrabass	vibrato	C2, C3, C4	oboe	vibrato, major trill, minor trill	C4, C5, C6
bassoon	vibrato, major trill, minor trill	C2, C3, C4, C5	trumpet	vibrato	C4, C5, C6
cello	vibrato	C2, C3, C4, C5, C6	saxophone	major trill, minor trill	C5
viola	major trill, minor trill	C3, C4, C5, C6	trombone	tremolo	C5
tuba	minor trill	C3, C4	piccolo trumpet	major trill, minor trill	C5, C6
tuba	major trill	C4	piccolo flute	vibrato, major trill, minor trill	C6, C7
saxophone	tremolo	C4	violin	vibrato, major trill, minor trill	C7
flute	vibrato	C4, C5			

TABLE III
SINGLE SOUND SOURCES USED IN GENERATING THE MIXTURE DATASETS. INSTRUMENTS ARE ORDED BY THE PITCH OF THE LOWEST NOTE USED.

source are clearly separated from the components of the non-modulated source. Panels (c), (e), and (g) demonstrate the outputs of three modulated filters applied to the complex CQT. Although the emerged modulated patterns looks slightly different from the output of the original filters in the left column, they are still successfully separated from the non-modulated components. Furthermore, it is worth noting that due to phase preservation, the CQTs in the right column are fully invertible to the time domain, while this is not the case for the CQTs in the left column and not true for the auditory model of Chi et al.

VII. EXPERIMENTS

In this section, we examine the separability and clusterability of unison mixtures of instrumental sounds played with different techniques when they are encoded in four different representations. Two are commonly used for source separation: the STFT and the CQT. The other two are common-fate-based representations: the CFT and the proposed MCFT. We do not compare to the auditory model of Chi et al. because audio encoded in this model cannot be perfectly reconstructed.

A. Dataset

In our experiments, we primarily focus on evaluating the efficacy of the MCFT in capturing spectro-temporal modulation patterns as higher dimensions and in using them as source separation cues in cases with high energy overlap in the time-frequency domain. Mixtures of instrumental sound sources played in unison (same pitch) but with different frequency modulation techniques (e.g. vibrato versus trill) are a good example of such cases. Such mixtures also happen to be one of the most challenging cases for state-of-the-art audio source separation algorithms. Our next goal is to study the effect of the multi-resolution property of the MCFT, in the frequency domain as well as the scale-rate domain, on the separation quality and to compare its performance to CFT, which has fixed resolution at both stages. To this end, we include a wide range of musical octaves and a variety of modulation techniques in our dataset.

The testing dataset in our prior work [21] included a single pitch from a middle octave (D4 with a fundamental frequency of 293.66 Hz). In this work, the pitch range is extended to two lower and three higher octaves (6 octaves in total). The set of single sources we used to generate our mixture dataset is composed of 68 orchestral instrument samples generated by the EastWest Symphonic Orchestra sampler [1], 7 samples selected from the Philharmonia Orchestra [2], and 6 piano samples recorded on a Steinway grand (81 samples in total). All samples are 2 seconds long and are sampled at 44.1 kHz.

We chose the note C as a representative pitch class over octaves 2 to 7 (65.41 Hz to 2093 Hz). Table III presents the list of all instruments included in our dataset along with their playing techniques and octave coverage. The playing techniques include vibrato: continuous frequency modulation, trill: frequency modulation alternating between two adjacent pitches in the chromatic scale, and tremolo: amplitude (and sometimes frequency) modulation. We note that a *single sound source* in our experiments refers to a single note played by an instrument-technique pair, e.g. a C4-viola-major trill is considered a different source than a C4-viola-minor trill. It can be clearly observed that the number of samples per octave follows a bell-shaped distribution (there are respectively 7, 9, 21, 22, 15, 7 samples in octaves 2 to 7). This is because orchestral instruments have a limited pitch range, as a result of which the number of samples for all pitch classes is much larger in middle octaves than in high/low octaves.

Due to the imbalance in the number of sources per octave, there is a large difference between the number of mixtures per octave. For instance, the total number of two-source mixtures ranges from 21 for the second and seventh octaves to 231 for the fifth octave. We keep the number of mixtures the same for all octaves by randomly selecting 21 mixtures (minimum number) in octaves 3 to 6. This gives rise to a testing dataset of size 126 two-source mixtures.

To study the behavior of representations as the number of sources increases, we create three-, four-, and five-source-mixture datasets, each of size 126, following the same procedure described for two-source mixtures.

It should be taken into account that the MCFT is designed to explicitly capture frequency modulation. Our dataset is thus almost entirely composed of frequency-modulated samples (vibrato, major trill, minor trill), such that we can attribute the dominant effect on the behavior of separability and clusterability results to frequency-modulation. Since tremolo is

[1] http://www.soundsonline.com/symphonic-orchestra
[2] www.philharmonia.co.uk

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

sometimes defined as amplitude and sometimes as frequency modulation, we expect the MCFT to provide improvement only if there is frequency modulation that is greater than the minimum detectible frequency change, which is controlled by the resolution of the underlying transform.

B. Audio Representations

In our experiments, the window length and overlap ratio of the STFT are set to 93 ms (4096 samples) and 75% respectively. At its time-frequency representation stage, the CFT uses the same parameter values.

To study the effect of the 2D window on separability and clusterability of the CFT, we experiment with a grid of values including all combinations of $L_\omega \in \{2, 4, 8, 16, 32\}$ (21.6 Hz - 344.5 Hz) and $L_\tau \in \{4, 8, 16, 32, 64, 90\}$ (93 ms - 2 sec) and present the results for the best and worst window sizes. There is 50% overlap between windows in both dimensions.

For computation of CQTs, we use the MATLAB toolbox in [30]. The minimum frequency, maximum frequency, and frequency resolution of the CQT are respectively set to 61.74 Hz (note B1), 4435 Hz (note C#8) , and 96 bins per octave.

The same parameter values are used in the time-frequency representation stage of the MCFT. In the modulation analysis stage, the MCFT uses a spectral filter bank $F(s; S)$ including a lowpass filter centered at 2^{-4} (cyc/oct), 6 bandpass filters at $2^0, 2^1, ..., 2^5$ (cyc/oct), and a highpass filter at $2^{5.5}$ (cyc/oct). The temporal filter bank $G(r; R)$ is composed of a lowpass filter centered at 2^{-2} (cyc/sec), 5 bandpass filters at $2^0, 2^1, ..., 2^4$ (cyc/sec), and a highpass filter at $2^{4.5}$ (cyc/sec). The time constant parameter, β, is set to 1. The product of F and G gives rise to a 2D filter response, which is then split into two analytic filters (see Section V). Since one advantage of the MCFT is that it is inherently multi-resolution, we used only the single setting described above, rather than experimenting with 30 settings, as was done with the CFT. We have provided an implementation of the MCFT and audio examples from our experimental results in the accompanying website [3].

C. Separability Results

Figure 8 presents an example of source separation via ideal binary masking in different representation domains for one of the mixtures in our dataset. For easier visual comparison, all signals are presented in the STFT domain. The top row shows the mixture and the original signal, and the next two rows show the results of separation in the 4D and 2D domains. It can be observed that the MCFT preserves more of the signal energy and harmonic structure, and introduces fewer masking artifacts than other representations.

The separability of the representations is evaluated over the testing dataset through ideal binary masking (see Section II). We use a range of threshold values (0 dB to 30 dB with a step of 5 dB) in the computation of ideal binary masks in each representation domain. We perform separation through ideal binary masking in each representation domain and then compare the preserved energy level and separation quality of all reconstructed signals in the time domain. For time-domain evaluation of the separation performance, we use the BSS-Eval [23] objective measures: SDR, SIR, and SAR. The mean SDR values over the whole dataset are used as a measure of separability. In the separability results the "CFT-best-sep" and "CFT-worst-sep" correspond to the window sizes (drawn from the set of 30 window sizes we used) that resulted in the best and worst mean SDR values.

[3]https://interactiveaudiolab.github.io/MCFT

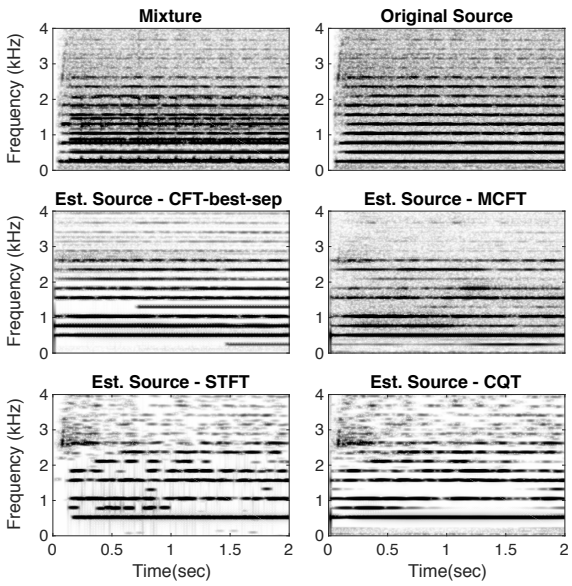

Fig. 8. An example of separation via ideal binary masking with a threshold of $\gamma = 25$ dB for a mixture of C4-clarinet-major trill and C4-flute-vibrato. (a,b) Magnitude spectrograms of the mixture and C4-flute-vibrato. (c-f) Magnitude spectrograms of the estimated source by applying the mask respectively in the CFT-best-sep, MCFT, STFT, and CQT domains.

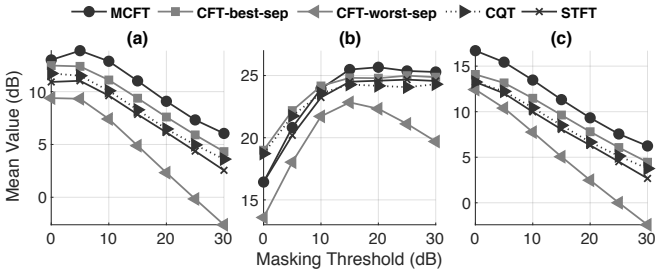

Fig. 9. Measuring Separability for two-source mixtures as a function of masking threshold. Higher values are better. Mean (a) SDR, (b) SIR, and (c) SAR for 2D and 4D representations versus masking threshold, γ. The results for 2 out of 30 2D window sizes tried in CFT computation are presented: CFT-best-sep (4×64) and CFT-worst-sep (32×4).

Figure 9 presents the mean values of separability metrics as a function of masking threshold used for the ideal binary mask, γ, for two-source mixtures. It can be clearly seen that the MCFT outperforms all other representations in terms of SDR and SAR at all threshold values and in terms of SIR for threshold values above 10 dB (i.e. the source energy must be 10 dB louder than the interference to be included in the ideal binary mask).

The reason why all representations have a better SIR performance for middle thresholds (15-20 dB) can be explained

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

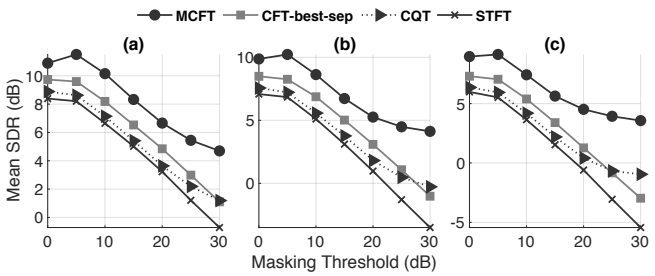

Fig. 10. Mean SDR versus masking threshold for 2D and 4D representations over (a) three-, (c) four-, and (d) five-source mixture datasets. Higher values are better. Only the results for the best two-dimensional window size used in CFT computation (4×64) are presented.

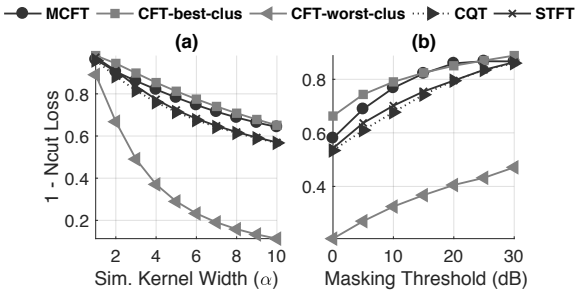

Fig. 11. Mean clusterability for 2D and 4D representations versus similarity kernel width, α (a) and masking threshold, γ (b). Higher values are better. The results for 2 out of 30 2D window sizes tried in CFT computation are presented: CFT-best-clus (2×90) and CFT-worst-clus (32×8).

by considering the fact that low threshold values would let in a large amount of noise and interference along with the energy from the target source and high threshold values would remove a significant potion of the target signal energy, both resulting in a decrease in SIR. The performance of the CFT depends heavily on the 2D window size and ranges from much worse than the STFT to better than the CQT. Such dependency makes the use of the CFT less reliable in blind source separation scenarios, since it is highly sensitive to data-dependent settings to achieve maximal performance.

The mean SDR values over masking threshold for datasets with more than two sources per mixture are shown in Figure 10. While the performance of all representations degrades in general with an increase in the number of sources, the MCFT stays strictly dominant in all cases. Moreover, the MCFT shows the slowest dropping rate over increasing threshold values for mixtures with more than two sources.

D. Clusterability Results

We use the clusterability measure defined in Section III to measure how well each representation groups together elements of a single source. Higher values are better. The Gaussian kernel in Equation (5) along with the Euclidean distance measure are used in the computation of similarity values in our experimental results.

An increase in the similarity kernel width assigns higher weights to points farther from the center of each cluster and thus increases the likelihood of mislabeling points from neighboring clusters. On the other hand, increasing the masking threshold means removing lower-energy points, which are presumably located towards the boundaries of neighboring classes and therefore producing wider inter-cluster margins. We study the effect of the similarity kernel width, α as well as the masking threshold, γ. Figure 11 demonstrates the mean clusterability values for two-source mixtures versus these two parameters. As it can be observed in Figure 11, an increase in the similarity width results in a drop in clusterability values for all representations, whereas an increase in masking threshold causes an increase in clusterability values. The MCFT seems to outperform the 2D representations over α values larger than 2 and γ values below 30 dB. The performance of the CFT is again dependent on the window size and can vary dramatically as shown by the results for the best- and worst-performing window sizes, where it goes from outperforming to underperforming all the other representations.

An interesting difference between the MCFT curve and others is that it almost levels out after 20 dB while the others keep increasing. This behavior is not unexpected since the MCFT tends to project the signal energy to a larger number of points in the higher-dimensional space and thus preserve a much larger portion of signal energy for higher thresholds compared to other representations (see Figure 9). This behavior is more noticeable in Figure 12 for mixtures composed of more than two sources. The slower decrease of the MCFT separability and slower increase of the MCFT clusterability over thresholds higher than 20 dB compared to other sources demonstrate the existence of a tradeoff between these two properties (preserving more signal energy versus creating wider inter-cluster margins).

Note, however, that higher separability (i.e. sources are not overlapped) is a neccessary precursor to high-quality source separation while higher clusterability (sources are in separate regions of the representation space) is strongly desireable, but not technically neccessary, depending on the sophistication of the separation algorithm.

The mean SDR and mean clusterability over the whole two-source dataset and all parameter values are respectively shown on the y-axis and x-axis in Figure 13. The plot depicts the mean performance for the STFT, CQT, MCFT, and CFT (all 30 2D window sizes). The bold dashed lines delimit the range of values that are inferior to the MCFT performance across both dimensions. The MCFT clearly outperforms all other representations in terms of separability and only underperforms the CFT in terms of clusterability for 2 out of 30 different window sizes. Even in these two cases, the clusterability is similar between CFT and MCFT, while MCFT strongly dominates the CFT on separability.

The bar plots of mean SDR and mean clusterability measure for all representations and all mixture types are demonstrated in Figure 14. The error bars indicate the range of values between the first and third quartiles.

Note that, as the number of sources increases, MCFT's separability (as measured by SDR) degrades much less than other representations. All other representations drop by more than half in their SDR values when moving from 2 to 5 sources, making the MCFT the obvious choice of representation when

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

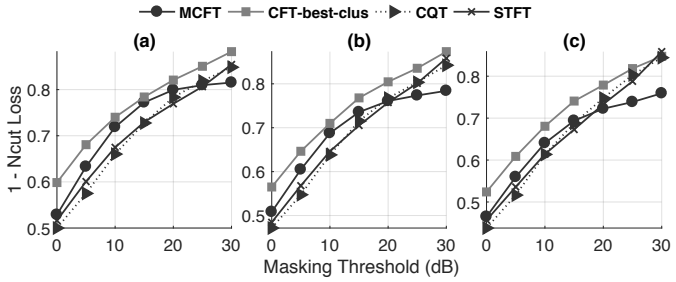

Fig. 12. Mean clusterability versus masking threshold for 2D and 4D representations over (a) three-, (b) four-, and (c) five-source mixture datasets. Higher values are better. Only the results for the best 2D window sizes used in CFT computation (2×90) are presented.

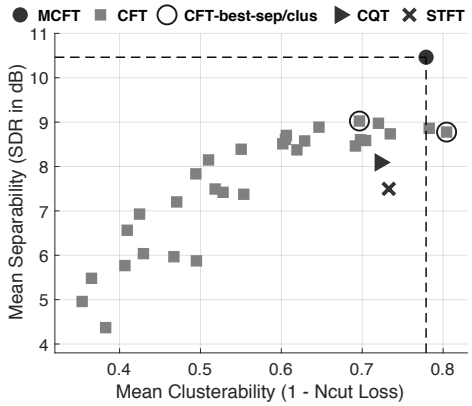

Fig. 13. Mean SDR versus mean clusterability over all samples and masking thresholds for two-source mixtures. The results for all 30 2D window sizes used in CFT computation are presented, along with the results for the MCFT, CQT and STFT. Higher values are better in both dimensions.

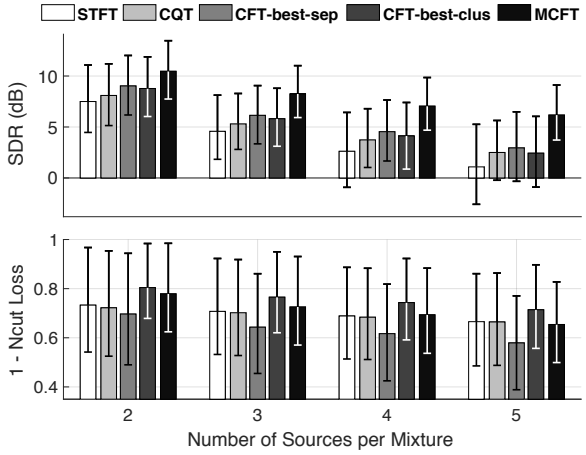

Fig. 14. Bar plots of mean separability measured by SDR (top panel) and mean clusterability measured by normalized cut loss (bottom panel) over all masking thresholds and similarity kernel parameter values for two-, three-, four-, and five-source mixtures. Higher values are better. Number of samples is $n = 882$ in the top panel and $n = 8820$ in the bottom panel. Error bars indicate the range of values between the first and third quartiles. The results for CFT window size with the best separability and the CFT window size with the best clustering are presented.

there are many sources. This improved separability can come at a price, however. Although many more points associated with a single source are not overlapped in the MCFT, these points can become interspersed with points from interfering sources, which reduces clusterability.

To compare the result distributions, we used the Wilcoxon rank sum test. The SDR values for the MCFT show significant improvement over all other representations and for all mixture types with $p \leq 0.0001$ in all cases. The MCFT performs significantly better on clusterability than CFT-best-sep for all mixture types with $p \leq 0.0001$, significantly better than CQT for two-, three-, and four-source mixtures with $p \leq 0.05$ in the worst case, and significantly better than the STFT for two-, and three-source mixtures with $p \leq 0.0001$. Although MCFT outperforms most other representations on clusterability, the CFT-best-clus improves on the MCFT in all cases with $p \leq 0.0001$.

Note that the superior clustering of CFT-best-clus is due to a careful selection of the window size in the presence of ground truth, which is typically not possible in real-world use. Moreover, separability for CFT-best-clus remains worse than the separability of the MCFT. Therefore, even if it is easier to cluster energy from a single source, the resulting separation will have a ceiling of performance that is lower for other representations than for the MCFT.

VIII. Conclusion

The efficacy of source separation algorithms can be limited by the representation used for the input audio. A representation that reduces overlap and interspersal of sources can simplify the separation process and improve results. We presented the Multi-resolution Common Fate Transform (MCFT), a representation that is fully invertible and increases the separability of audio signals with significant time-frequency-domain overlap, through explicitly representing spectro-temporal modulation patterns. We placed it in the context of two existing common-fate-based models: the Common Fate Transform (CFT) and the auditory model of Chi et al. The MCFT, by being multi-resolution and fully invertible combines the strengths of both approaches.

We also introduced and provided metrics for two desirable properties of audio representations for source separation: separability and clusterability. Experiments on a dataset of unison mixtures of musical instrumental sounds showed that the MCFT strictly dominates the other representations on separability. It also outperforms other representations on clusterability in the majority of cases, without requiring data-dependant parameter setting to achieve these results. Given these results, the MCFT is a promising representation to be used as the input to source separation algorithms. Moving forward, we plan to use the MCFT as the input representation to state-of-the-art separation methods such as Common Fate Modeling (CFM) [17] and deep clustering [12].

IX. Acknowledgements

This work was supported by United States National Science Foundation award number 1420971.

References

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech & Language*, vol. 24, no. 1, pp. 1–15, 2010.

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2018.2878616, IEEE/ACM Transactions on Audio, Speech, and Language Processing

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural computation*, vol. 17, no. 9, pp. 1875–1902, 2005.

[3] F. Ehlers and H. G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Transactions on Signal processing*, vol. 45, no. 10, pp. 2608–2612, 1997.

[4] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." in *International Society for Music Information Retrieval conference (ISMIR)*, 2009, pp. 327–332.

[5] J. F. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation." in *International Society for Music Information Retrieval conference (ISMIR)*, 2006, pp. 314–319.

[6] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic music transcription and audio source separation," *Cybernetics & Systems*, vol. 33, no. 6, pp. 603–627, 2002.

[7] S.-W. Jeon, Y.-C. Park, S.-P. Lee, and D.-H. Youn, "Robust representation of spatial sound in stereo-to-multichannel upmix," in *Audio Engineering Society Convention 128*. AES, 2010.

[8] D. Fitzgerald, "Upmixing from mono-a source separation approach," in *2011 17th International Conference on Digital Signal Processing (DSP)*. IEEE, 2011, pp. 1–7.

[9] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals," in *Multimedia, 2006. ISM'06. Eighth IEEE International Symposium on*. IEEE, 2006, pp. 257–264.

[10] S. Rickard, "The duet blind source separation algorithm," *Blind Speech Separation*, pp. 217–237, 2007.

[11] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, 2014.

[12] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[13] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[14] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," 2005.

[15] E. Creager, N. D. Stein, R. Badeau, and P. Depalle, "Nonnegative tensor factorization with frequency modulation cues for blind audio source separation," *arXiv preprint arXiv:1606.00037*, 2016.

[16] M. Abe and S. Ando, "Auditory scene analysis based on time-frequency integration of shared fm and am," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 1998, pp. 2421–2424.

[17] F.-R. Stöter, A. Liutkus, R. Badeau, B. Edler, and P. Magron, "Common fate model for unison source separation," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016.

[18] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[19] L. Krishnan, M. Elhilali, and S. Shamma, "Segregating complex sound sources through temporal coherence," *PLoS Comput Biol*, vol. 10, no. 12, p. e1003985, 2014.

[20] N. Mesgarani and S. Shamma, "Denoising in the domain of spectrotemporal modulations," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 3, p. 3, 2007.

[21] F. Pishdadian, B. Pardo, and A. Liutkus, "A multi-resolution approach to common fate-based audio separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 566–570.

[22] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–529.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[24] S. Epter, M. Krishnamoorthy, and M. Zaki, "Clusterability detection and cluster initialization," in *Proceedings of the Workshop on Clustering High Dimensional Data and its Applications at the 2nd SIAM International Conference on Data Mining*, 2002, pp. 47–58.

[25] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 2. IEEE, 2004, pp. 28–31.

[26] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[27] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 1963–2001, 2006.

[28] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[29] M. Elhilali and S. A. Shamma, "A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3751–3771, 2008.

[30] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *53rd International Conference on Semantic Audio*. AES, 2014.

Fatemeh Pishdadian is a Ph.D. candidate in Electrical Engineering and Computer Science at Northwestern University. She received her B.Sc. in Electrical Engineering from Ferdowsi University of Mashhad, Iran, and her M.Sc. in Electrical and Computer Engineering from George Mason University. Her research interest lies in the application of signal processing and machine learning methods to the analysis of audio/music. More specifically, the focus of her doctoral research has been on audio/music source separation.

Bryan Pardo is an associate professor in the Northwestern University Department of Electrical Engineering and Computer Science. Prof. Pardo received a M. Mus. in Jazz Studies in 2001 and a Ph.D. in Computer Science in 2005, both from the University of Michigan. He has authored over 100 peer-reviewed publications. He has developed speech analysis software for the Speech and Hearing department of the Ohio State University, statistical software for SPSS and worked as a machine learning researcher for General Dynamics. While finishing his doctorate, he taught in the Music Department of Madonna University.

2329-9290 (c) 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.