AUDIO SOURCE SEPARATION WITH TIME-FREQUENCY VELOCITIES

Guy Wolf, Stéphane Mallat

Department of Computer Science
École Normale Supérieure
45 rue d'Ulm, 75005 Paris, France

Shihab Shamma

Department of Cognitive Studies
École Normale Supérieure
45 rue d'Ulm, 75005 Paris, France

ABSTRACT

Separating complex audio sources from a single measurement channel, with no training data, is highly challenging. We introduce a new approach, which relies on the time dynamics of rigid audio models, based on harmonic templates. The velocity vectors of such models are defined and computed in a time-frequency scalogram calculated with a wavelet transform. Similarly to rigid object segmentation in videos, multiple audio sources are discriminated by approximating their velocity vectors with low-dimensional models. The different audio sources are segmented by optimizing a harmonic template selection, which provides piecewise constant velocity approximations. Numerical experiments give examples of blind source separation from single channel audio signals.

Index Terms— Audio source separation, harmonic templates, velocity, wavelets.

1. INTRODUCTION

Audio signals are usually given in mixtures that need to be separated to extract the information carried by each of them. Many works usually consider a setting where several measurement channels are available, which enables the utilization of various cues such as spatial information and cross channel correlations. Performing the separation from a single measurement channel is considered highly challenging, since such cues are unavailable, but the human brain has a remarkable ability to perform it. This task is referred to in literature as single-channel, cochannel or monaural source separation.

Several approaches have been applied to perform single channel source separation. Recent methods for this task often aim for specific settings or mixture types in order to utilize known properties of the mixed signals. Such settings include speech separation [1, 2], musical sound separation [3, 4], and singing voice separation [5]. More general methods have been proposed, for example in [6, 7], based on statistical learning techniques. However, these methods require a prior learning stage from a clean training set. Model-based methods with prior training were also suggested specifically for speech separation by applying a priori learned models, such as hidden Markov models [8] and Gaussian mixture models [9], to the sources in the mixture. These methods can obtain good separation results, but the requirement to have clean training data is often hard to meet in realistic settings. Furthermore, specific speech models, such as the ones used in [8, 9], may not be applicable in general settings that include non-speech sources.

In this paper, we aim for a broad framework that fits many mixture types without utilizing preexisting information about the signal. Therefore, we consider source separation as an unsupervised learning task, without relying on clean training data or prior knowledge about the type of source signals in the mixture. Several speech segregation methods (e.g., [10]) have been suggested for extracting a target speech signal from a mixture with various interferences, such as noise or interfering speech. While such methods are indeed suitable for several interference types, they are based on the existence of a dominant speaker in the mixture and only aim to extract it. The proposed method in this paper, on the other hand, treats the sources in the mixture as equally important and aims to extract all of them when possible.

Unsupervised single channel blind source separation can be interpreted as a segregation problem. We need to find a representation that separates components coming from different sources, and reconstruct separately each source by regrouping the components that belong to the same source. Human processing of auditory and visual have been shown to be related in psychophysical studies [11]. We apply similar reasoning to audio source separation, and we relate our approach for performing it to object segregation and image segmentation approaches in computer vision. Segregation is very difficult over static images, but it is much easier in videos. Indeed, the velocity vectors of image pixels of a rigid object belong to a low-dimensional affine space specified by the 3D motion tensor of the rigid object. Rigid object segmentation thus amounts to finding low dimensional models that approximate the velocity of large groups of image pixels [12].

Work supported by ERC InvariantClass 320959

This paper proposes a new approach for audio source separation based on the time dynamics of rigid audio models. It relies on the same principle as rigid object segmentation in video sequences. Like other audio processing methods, we use a time-frequency representation of the audio signal. The time evolution of the signal is interpreted as motion, and frequency bands take the same role as pixel positions in an image. Therefore, in this representation, an auditory scene is treated as a one-dimensional equivalent of a (two-dimensional) image video with changing luminosities, which correspond to audio amplitude modulations. Each source in a mixture signal is identified as a moving "rigid" template, where the introduced notion of rigidity is obtained by considering harmonic structures in the signal.

Psychophysical evidences show that audio signals are perceived through projections of harmonic frequency templates, even though such harmonics may not be present in the signal [13]. Slow amplitude modulations and pitch frequency modulations also appear to be important perceptual cues for discriminating multiple sources. Section 2 introduces harmonic template audio models, which incorporate variable amplitude and frequency pitch modulations. Section 3 shows that the wavelet transform of these harmonic template models defines an image where the amplitude and pitch dynamics can be separated as two components of an audio velocity vector. A time-frequency velocity equation is derived in Section 4, which is similar to the optical flow velocity equation in images [14].

Similarly to computer vision algorithms (e.g., [14]), velocity vectors are computed by projecting the velocity equation over multiscale wavelets. The source separation is performed by finding low-dimensional models of the audio velocity vectors that result from the rigidity of the harmonic template models. Section 5 uses the harmonic template model to find the time-frequency supports of the different sources. Then, it uses non-overlapping regions of these supports to explicitly separate and demix overlapping time-frequency components, thus reconstructing the time-frequency representation of each separated source. We concentrate on the separation of two audio sources to explain the principles of the approach. This case captures important classes of applications, including signal versus background separation. Examples are shown in Section 6, and more results can be found in `www.di.ens.fr/data/scattering/BSS/`.

2. HARMONIC TEMPLATES

Amplitude and pitch frequency modulations are fundamental perceptual cues for audio source separation [15]. We introduce a harmonic template model, with amplitude and pitch frequency modulations, whose rate of change is defined by a velocity vector. These template models are used to locally approximate audio signals.

In order to formulate a harmonic template model, we first consider a harmonic excitation that is given by a Dirac comb with pitch frequency ξ. This excitation is modeled as

$$e(u) = \frac{\xi}{2\pi} \sum_n \delta\left(u - \frac{2\pi n}{\xi}\right) = \sum_k e^{ik\xi u},$$

where the last equality is due to the Fourier transform of a Dirac comb. Pitch frequency modulations are modeled by a time-warping $\theta(t)$, and the modulated excitation is thus modeled as

$$e_\theta(u) = e(\theta(u)) = \sum_k e^{ik\xi\theta(u)}\,.$$

This time warping modifies the pitch frequency of the excitation, and in a local neighborhood of time t it becomes $\xi(t) = \xi\theta'(t)$, thus the modulated excitation in such neighborhood can be written as

$$e_\theta(t+u) = \sum_k e^{ik\xi\theta(t)} \times e^{ik\xi\theta'(t)u}\,. \tag{2.1}$$

We consider pitch changes over log-frequency scales to get $\log\xi(t) = \log\xi + \log\theta'(t)$. Therefore, the resulting pitch velocity is given by

$$V_\xi(t) = \frac{d\log\xi(t)}{dt} = \frac{d\log\theta'(t)}{dt}\,. \tag{2.2}$$

A harmonic template model is obtained from the modulated excitation e_θ by first convolving it with a filter $h(t)$ that defines its spectral envelope. Then, the filtered signal is slowly modulated in amplitude by $a(t)$ to get the template model

$$x(t) = a(t)\,(e_\theta \star h)(t)\,. \tag{2.3}$$

Since audio amplitudes are typically perceived on a logarithmic scale, the resulting amplitude velocity is given by

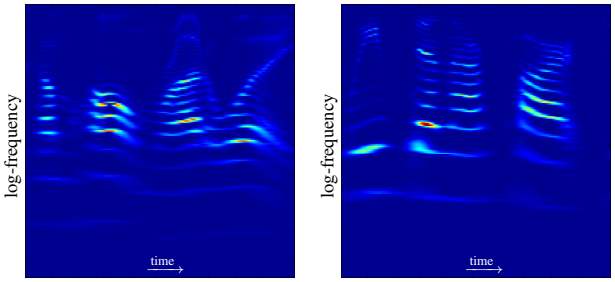
$$V_a(t) = \frac{d\log a(t)}{dt}\,.$$

The presented harmonic template model in (2.3) encompasses both amplitude and pitch modulations in the signal. The audio velocity of this model is defined by the vector $V(t) = (V_a(t), V_\xi(t))$.

Audio signals may not have a harmonic structure with a well defined pitch, as in unvoiced speech, or inharmonic frequency components. However, psychophysical studies show that audio signals are often perceived through approximations with multiple harmonic templates [13]. We shall similarly use projections on harmonic templates to separate multiple audio sources by their audio velocity.

3. WAVELET SEPARATION

The time-frequency structures of audio signals are well revealed by Q-constant filter banks, which model the signal cochlea transformation [16]. This can also be written as a multiscale wavelet transform, whose modulus defines a time-frequency representation that is called a scalogram image, and



(a) Male (b) Female

Fig. 3.1. Scalograms $Ix(t,\lambda)$ of two audio segments produced by a male and a female.

it will be explained in details in this section. We show that a wavelet transform can partly separate the time-frequency supports of two different harmonic template models, and characterize the overlap of these time-frequency supports.

A complex wavelet transform is computed with a modulated wavelet filter $\psi(t) = g(t)\, e^{it}$, where $g(t)$ is a low frequency envelope, such as a Gaussian. Its Fourier transform $\hat{g}(\omega)$ has a bandwidth Q^{-1}. For audio applications, we choose $Q = 32$ [17]. It results that $\hat{\psi}(\omega) = \hat{g}(\omega - 1)$ has a support concentrated in $[1-Q^{-1}/2, 1+Q^{-1}/2]$. The wavelet transform of a signal $x(t)$ is defined by

$$Wx(t,\lambda) = x \star \psi_\lambda(t) \text{ with } \psi_\lambda(t) = 2^\lambda \psi(2^\lambda t)\,.$$

The dilated wavelet ψ_λ is a band-pass filter whose Fourier transform $\hat{\psi}_\lambda(\omega) = \hat{\psi}(2^{-\lambda}\omega)$ has support centered at $\omega = 2^\lambda$. For an appropriate envelope $g(t)$, one can prove that W has a stable inverse [17].

The scalogram image gives the energy of x in the neighborhood of the time t and of the frequency 2^λ:

$$Ix(t,\lambda) = |Wx(t,\lambda)| = |x \star \psi_\lambda(t)|\,.$$

Figure 3.1 shows two scalograms $Ix(t,\lambda)$ having harmonic structures, with amplitude and pitch frequencies that evolve in time. These time evolutions will be approximated by local harmonic template models.

Let us compute the wavelet transform of harmonic template model $x(t) = a(t)\,(e_\theta \star h)(t)$. The wavelet ψ_λ has a time support proportional to $Q2^{-\lambda}$. If 2^λ is sufficiently large, then $a(t)$ is approximately constant over this support, and one can derive that

$$Wx(t,\lambda) = x \star \psi_\lambda(t) \approx a(t)\,(e_\theta \star \psi_\lambda \star h)(t).$$

We saw in (2.1) that x has harmonics at frequencies $\xi(t) = k\,\xi\theta'(t)$, where $k \in \mathbb{N}$ is the harmonic index. If $Q^{-1}2^\lambda \leq k\,\xi\theta'(t)$ then each wavelet covers at most one harmonic. One can derive from (2.1) that

$$Ix(t,\lambda) \approx a(t)\,|\hat{h}(2^\lambda)|\,|\hat{\psi}(2^{-\lambda}k\xi(t))|\,. \tag{3.1}$$

For a fixed t, $Ix(t,\lambda)$ is a sum of harmonics at frequencies $\lambda \approx \log(k\xi(t))$, whose amplitudes are $a(t)\,|\hat{h}(2^\lambda)|$. Since $\hat{\psi}$ has a bandwidth of Q^{-1}, the harmonic frequency support of $Ix(t,\lambda)$ is thus:

$$H^t = \left\{\lambda \,:\, \exists k \in \mathbb{N}\,,\ |1 - 2^{-\lambda}\,k\,\xi(t)| \leq Q^{-1}/2\right\}\,.$$

This harmonic support is thus defined by the pitch frequency $\xi(t)$. In Figure 3.1, $Q = 32$ is sufficiently large to discriminate the first 15 harmonics of male and female speakers.

Let us now consider an audio mixture

$$x(t) = x_1(t) + x_2(t),$$

where x_1 and x_2 are locally approximated by harmonic template models. The wavelet transform is linear so $Wx = Wx_1 + Wx_2$. We shall recover x_1 and x_2 by estimating their wavelet transform Wx_1 and Wx_2 from Wx. For a fixed t, we denote by H_1^t and H_2^t the frequency supports of $Wx_1(t,\lambda)$ and $Wx_2(t,\lambda)$. The frequency overlap between both supports is

$$H_{1,2}^t = H_1^t \cap H_2^t\,.$$

Over the non-overlapping parts of their supports, we can simply compute Wx_1 and Wx_2 with

$$Wx(t,\lambda) = \begin{cases} Wx_1(t,\lambda) & \text{if } \lambda \in H_1^t - H_{1,2}^t \\ Wx_2(t,\lambda) & \text{if } \lambda \in H_2^t - H_{1,2}^t \end{cases} \tag{3.2}$$

If $\lambda \in H_{1,2}^t$ then the separation of Wx_1 and Wx_2 is more complex and requires to compute an approximation model of x_1 and x_2. We shall compute the harmonic supports H_1^t and H_2^t from the mixture Wx, by calculating the velocity vector of these supports. Section 5 explains the separation of Wx_1 and Wx_2 in the overlapping support.

4. AUDIO VELOCITY EQUATION

Similarly to rigid body segmentations in videos, the time-frequency supports of harmonic templates are segmented by calculating time-frequency velocity vectors over the scalogram image, with an audio velocity equation that we now introduce. Different harmonic supports are identified by computing piecewise constant approximations of these time-frequency velocities.

We first consider a single harmonic template model $x(t) = a(t)\,(e_\theta \star h)(t)$. The approximation (3.1) shows that the partial derivatives of $Ix(t,\lambda)$ satisfy the following velocity equation

$$\frac{\partial Ix(t,\lambda)}{\partial t} \approx Ix(t,\lambda)V_a(t) + \frac{\partial Ix(t,\lambda)}{\partial \lambda}V_\xi(t) \tag{4.1}$$

with $V_a = d\log a(t)/dt$ and $V_\xi = d\log\theta'(t)/dt$. This velocity equation is quite similar to the optical flow equation of image pixel velocities.

As in image optical flow equations, at each (t, ξ), a single audio velocity equation (4.1) relates the two coordinates of the velocity vector $V = (V_a, V_\xi)$. Following an algorithm developed in computer vision [14], one can transform this single equation into a system of two equations, by projecting it over a complex wavelets $\Psi(t, \lambda)$ defined over the scalogram image plane (t, λ). This wavelet $\Psi(t, \lambda)$ is computed as a separable product of a wavelet along t and a wavelet along λ. Its support size along t is chosen to be small enough so that $V_a(t)$ and $V_\xi(t)$ remain almost constant over this support. As a result, convolving both sides of (4.1) with Ψ gives:

$$\begin{aligned} Ix \star \frac{\partial \Psi}{\partial t}(t,\lambda) &\approx V_a\, Ix \star \Psi(t,\lambda) \\ &+ V_\xi\, Ix \star \frac{\partial \Psi}{\partial \lambda}(t,\lambda). \end{aligned} \tag{4.2}$$

Since Ψ is complex, the real and imaginary parts of this equation define a system of two equations with two unknowns, which yields a unique solution $V = (V_a, V_\xi)$ when it is not degenerated.

If x is a pure harmonic template then the solution $V(t, \lambda)$ only depends upon t. It is thus constant over the frequency support H^t of x. However, harmonic templates are only approximation models, and audio mixtures incorporate several harmonic templates, instead of a single one. It results that the computed audio velocity $V(t, \lambda)$ will depend both on t and λ.

5. SOURCE SEPARATION

For simplicity, we concentrate on audio mixtures $x = x_1 + x_2$ with only two sources. In principle, the algorithm can be extended to an arbitrary number of sources. However, when dealing with mixtures of multiple sources, relying solely on pitch-based templates may not be sufficient due to significant overlaps between their time-frequency supports. This issue can also occur when the mixture consists of sources with similar pitch modulations. In such cases, the definition of the identified templates can be extended by utilizing additional information, such as formant frequencies, spectral envelopes or other setting-specific distinguishing characteristics of the mixed sources. In this paper we focus on analyzing the simple case of pitch-based harmonic templates, which are sufficient for many two-sources mixtures, but the main principles of the presented method can also be extended to more complex templates that include such information.

We separate Wx_1 and Wx_2 from $Wx = Wx_1 + Wx_2$ by first estimating their time-frequency harmonic supports H_1^t and H_2^t. Each harmonic support H_j^t is defined by $H_j^t = \{\lambda : \exists k \in \mathbb{N}, |1 - 2^{-\lambda}\, k\, \xi_j(t)| \leq Q^{-1}/2\}$, where $\xi_j(t)$ is an unknown time-varying pitch. These harmonic supports are estimated by optimizing a piecewise constant approximation of the audio velocity vector $V(t, \lambda)$ from Section 4.

Let $V(t, \lambda)$ be the velocity vector computed with equation (4.2). If x_1 and x_2 can be locally approximated by harmonic template models, then $V(t, \lambda)$ is respectively equal to velocity vectors $V_1(t)$ and $V_2(t)$, which do not depend upon λ, over the two non-overlapping supports $H_1^t - H_{1,2}^t$ and $H_2^t - H_{1,2}^t$. For $j = 1, 2$, we can thus approximate $V(t, \lambda)$ by its weighted average $\overline{V}_j(t)$ over $H_j^t - H_{1,2}^t$ defined by

$$\overline{V}_j(t) = \frac{\sum_{\lambda \in H_j^t - H_{1,2}^t} V(t,\lambda) |Ix \star \Psi(t,\lambda)|}{\sum_{\lambda \in H_j^t - H_{1,2}^t} |Ix \star \Psi(t,\lambda)|}.$$

This constant velocity vector provides a prediction model of the scalogram. Indeed, for $\overline{V}_j = (\overline{V}_{a_j}, \overline{V}_{\xi_j})$, equation (4.2) implies that $Ix \star \Psi(t, \lambda)$ can be predicted with a first order Taylor approximation for all $\lambda \in H_j^t - H_{1,2}^t$:

$$Ix \star \Psi(t + \Delta, \lambda) \approx 2^{\overline{V}_{a_j} \Delta}\, Ix \star \Psi(t, \lambda + \overline{V}_{\xi_j} \Delta)\,. \tag{5.1}$$

The two harmonic supports H_1^t and H_2^t, parameterized by a pitch pair (ξ_1, ξ_2), are jointly optimized in order to minimize the mean-square error of the scalogram prediction given by (5.1).

Once the harmonic supports H_1^t and H_2^t are optimized, according to (3.2), we directly derive the values of Wx_1 and Wx_2 on the non-overlapping supports $H_j^t - H_{1,2}^t$, $j = 1, 2$. It remains to estimate Wx_1 and Wx_2 from $Wx = Wx_1 + Wx_2$ in the overlapping support $H_{1,2}^t$. For each $(t, \lambda) \in H_{1,2}^t$, we estimate $Wx_1(t, \lambda)$ and $Wx_2(t, \lambda)$ from the parameters of the harmonic template models. We have already computed each pitch frequency $\xi_j(t)$, which defines H_j^t, and we need to estimate each amplitude modulation $a_j(t)$. Applying (3.1) shows that

$$a_j(t) = \frac{Ix_j(t,\lambda)}{|\hat{h}_j(2^\lambda)|\, |\hat{\psi}(2^{-\lambda} k \xi_j(t))|}\,.$$

If $\lambda_j \in H_j^t - H_{1,2}^t$ then one can verify that for any $\lambda \in H_j^t$

$$Wx_j(t,\lambda) = \tfrac{Wx(t,\lambda_j)}{\hat{\psi}(2^{\lambda_j} k' \xi_j)} \hat{\psi}(2^\lambda k \xi_j) e^{i(k'-k)\xi_j t} z_j\,, \tag{5.2}$$

where k and k' are the closest integers to $2^\lambda \xi_j^{-1}$ and $2^{\lambda_j}\, \xi_j^{-1}$ correspondingly, and $z_j = \hat{h}(2^\lambda)/\hat{h}(2^{\lambda_j})$.

To demix the overlapping coefficients, it remains to identify the unknown values z_1 and z_2. These variables are constant in time as long as each source is well approximated by the harmonic template model. To incorporate model errors, we replace these constants by time functions $z_1(t)$ and $z_2(t)$, and minimize the energy of their time variations

$$\int \left(\frac{dz_1(t)}{dt}\right)^2 + \left(\frac{dz_2(t)}{dt}\right)^2 dt$$

where the Wx_i in (5.2) satisfy

$$Wx(t,\lambda) = Wx_1(t,\lambda) + Wx_2(t,\lambda).$$

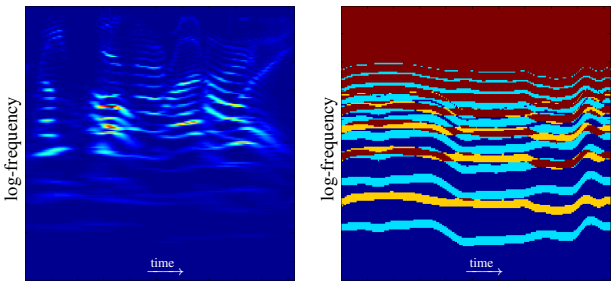
This minimization involves the resolution of a linear system whose solution specifies $z_1(t)$ and $z_2(t)$. Once we have estimated Wx_1 and Wx_2, an estimation of x_1 and x_2 is obtained by applying the inverse wavelet transform.

6. EXPERIMENTAL RESULTS

Figure 6.1(a) shows the scalogram $Ix(t, \lambda)$ of a mixture $x = x_1 + x_2$, where x_1 is the recording of a male speaker saying "Come home right away" and x_2 is a female speaker saying "We've done our part". Figure 6.2 gives the estimation of Ix_1 and Ix_2 obtained with our separation algorithm. It is computed by estimating the harmonic supports H_1^t and H_2^t shown in Figure 6.1(b). The separation quality can be evaluated by comparing the original source scalograms in Figure 3.1 with Figure 6.2. The corresponding audio reconstructions are available on the website `www.di.ens.fr/data/scattering/BSS/`, together with additional examples.

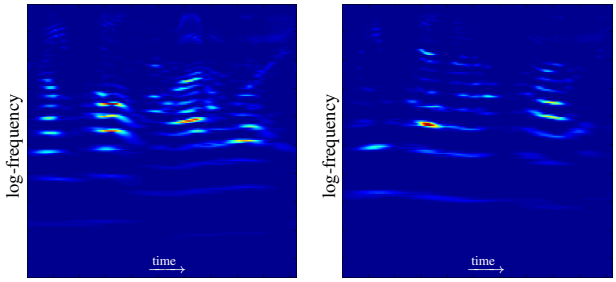
Additional evaluation of the quality of separated sources was performed using the "BSS eval" toolbox [18], which computes Source over Interference Ratios (SIR). Table 6 gives the results obtained on several mixture types, including the example of Figure 6.1(a), an additional mixture of two female speakers, a database of 100 male-female speech mixtures, and four additional mixtures of speech or singing voice with music or noise. The second column in the table compares the separated sources to the original signals. The comparison matches the signals by maximizing the mean SIR over the two sources. To provide a baseline for these performances we also compared the mixture signal to the two original sources. The results of these comparisons appear in the last column in the table. Despite the difficulty of the task, the large increase in SIR in all cases shows that an important improvement is obtained by the separation algorithm.

In order to compare the proposed algorithm to previous works we chose a representative unsupervised speech separation method that was recently suggested in [2] for separating two speakers. There, it was shown to obtain competitive results to prominent model-based and speaker-independent methods. This method follows a standard practice in speech separation of constructing binary masks on a time-frequency



(a) (b)

Fig. 6.1. (a) Scalogram of a mixture of two signals shown in Fig. 3.1. (b) This figure shows the harmonic supports H_1^t and H_2^t of each source: H_1^t is in cyan, H_2^t is in yellow, and the overlapping support $H_{1,2}^t$ is in dark red.



(a) Male (b) Female

Fig. 6.2. Separated scalograms computed from the mixture in Fig. 6.1(a)

Mixture	Prop. SIR	[2] SIR	Mix. SIR
Fig. 6.1(a)	12 & 14	8 & 8	2 & -1
Female & Female	12 & 17	6 & 7	2 & 0
Speech & Trumpet	19 & 9	9 & 4	8 & -5
Singing & Trumpet	17 & 13	8 & 7	-2 & 2
Speech DB (mean)	11 & 7	15 & 11	1 & -1
Singing & Cello	7	0	-3
Speech & Bubbles	11	6	7

Table 6.1. A comparison of the SIR (Source over Interference Ratio) values for the proposed method (Prop. SIR), the method from [2], and the mixture itself, in relation to the original signals. For the first four mixtures, both sources can be approximated by single harmonic templates and both SIR values are shown. For the database of 100 male-female speech mixtures, SIR values were computed for both sources and their means over the entire DB are shown. For the last two mixtures, only the speech or singing voice can be approximated and a single SIR is displayed. This is because a cello contains more than a single harmonic template, due to reverberations, and the bubble sound is completely inharmonic.

representation of the mixture, thus essentially assigning each frequency band at each time frame to one of the sources. Due to the use of time-frequency binary masks, this method and other similar ones do not take into account possible overlapping frequency bands. This makes such methods ill-suited when considering general mixtures, which may include musical instruments. Indeed, one of the main challenges in musical source separation is resolving such local frequency overlaps [19]. There are several musical source separation methods that deal with such overlaps explicitly (e.g., [19]) or implicitly (e.g., [3]). However, such methods often work under the assumption that the musical instrument signals consist of constant-pitch notes. Since human speech and singing voice do not fit this assumption, such methods would be ill-suited for mixtures that contain such sources.

The SIR values obtained by the method from [2], using the implementation provided on its authors website, are pre-

sented in the third column of Table 6, where they can be compared with the results of the proposed method. In most cases, both methods show an increase in SIR compared to the original mixture. In the case of the speech database, the method from [10] outperforms the proposed method, however this is to be expected since it is designed specifically for the case of speech separation, unlike the approach presented in this paper. In all other cases, the proposed method obtains a better SIR gain than the method from [2].

7. CONCLUSION

This paper introduces the notion of an audio velocity vector that gives the rates of change of amplitude and pitch frequency modulations. This velocity vector is computed with an audio velocity equation defined in a wavelet time-frequency plane. Complex audio sources are separated by approximating their audio velocities with rigid harmonic template models, which amounts to computing piecewise constant approximations. This approach extends the principles of rigid object segmentations in videos. It can be generalized to other types of signals by defining signal models that have low dimensional time dynamics. Psychophysical models indicate that such harmonic template approaches are also relevant for audio perception [13].

8. REFERENCES

[1] Q. Huang and D. Wang, "Single-channel speech separation based on long–short frame associated harmonic model," *Digital Signal Processing*, vol. 21, no. 4, pp. 497–507, 2011.

[2] K. Hu and D.L. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 122–131, 2013.

[3] S. Kirbiz and B. Gnsel, "Perceptually enhanced blind single-channel music source separation by non-negative matrix factorization," *Digital Signal Processing*, vol. 23, no. 2, pp. 646 – 658, 2013.

[4] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[5] P.-S. Huang, S.D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. of ICASSP '12*, 2012, pp. 57–60.

[6] G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *J. Mach. Learn. Res.*, vol. 4, pp. 1365–1392, 2003.

[7] S. Hochreiter and M.C. Mozer, "Monaural separation and classification of mixed signals: A support-vector regression perspective," in *ICA2001: 3rd International Conference on ICA and BSS*, 2001.

[8] J. Barker, A. Coy, N. Ma, and M. Cooke, "Recent advances in speech fragment decoding techniques," in *Proc. of INTERSPEECH '06*, 2006, pp. 85 – 88.

[9] Y. Shao and D.L. Wang, "Sequential organization of speech in computational auditory scene analysis," *Speech Commun.*, vol. 51, no. 8, pp. 657 – 667, 2009.

[10] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, 2004.

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

[12] T. Li, V. Kallen, D. Singaraju, and R. Vidal, "Projective factorization of multiple rigid-body motions," in *Proc. of CVPR '07*, June 2007, pp. 1–6.

[13] S. Shamma and D. Klein, "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system," *Jour. of the Acoust. Soc. of America*, vol. 107, no. 5, pp. 2631–2644, 2000.

[14] C.P. Bernard, "Discrete wavelet analysis for fast optic flow computation," *Applied and Computational Harmonic Analysis*, vol. 11, no. 1, pp. 32–63, 2001.

[15] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "Speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621 – 633, 2013.

[16] T. Chi, P. Ru, and S.A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Jour. of the Acous. Soc. of Amer.*, vol. 118, no. 2, pp. 887–906, 2005.

[17] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, to be published, DOI: 10.1109/TSP.2014.2326991.

[18] C. Févotte, R. Gribonval, and E. Vincent, *BSS_EVAL toolbox 2.0*, 2005.

[19] Y. Li, J. Woodruff, and D.L. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 7, pp. 1361–1371, 2009.