UNISON SOURCE SEPARATION

Fabian-Robert Stöter, Stefan Bayer, Bernd Edler

International Audio Laboratories Erlangen,*
Erlangen, Germany
{fabian-robert.stoeter}@audiolabs-erlangen.de

ABSTRACT

In this work we present a new scenario of analyzing and separating linear mixtures of musical instrument signals. When instruments are playing in unison, traditional source separation methods are not performing well. Although the sources share the same pitch, they often still differ in their modulation frequency caused by vibrato and/or tremolo effects. In this paper we propose source separation schemes that exploit AM/FM characteristics to improve the separation quality of such mixtures. We show a method to process mixtures based on differences in their amplitude modulation frequency of the sources by using non-negative tensor factorization. Further, we propose an informed warped time domain approach for separating mixtures based on variations in the instantaneous frequencies of the sources.

1. INTRODUCTION

Audio source separation is a very active research field with a large number of contributions. Applications are dependent on the context of the scenario, ranging from enhancements of speech signals to musically motivated analysis tasks.

The separation of sound sources from a single channel mixture is considered as an under-determined case which does not have a single solution. Knowing the way in which source signals are mixed together is crucial to the quality of separation systems. In the context of speech separation even unsupervised methods can lead to good results. This is due to the fact that mixtures of speech signals (like in a cocktail party environment) show a high degree of statistical independence. Mixtures of musical instruments, however, are highly correlated which is a desired aim of musical performances in general.

The Signal Separation Evaluation Campaign (SiSEC) is a solid indicator of the progress in research within the field of source separation [1]. The results from 2013 [2] show that for professionally produced music it is still difficult to achieve a high quality separation. One reason is due to the fact that the wide use of non-linear post-processing techniques (e.g. dynamic compression or effects like reverb) break assumptions that often are required to enable good performance of source separation algorithms. Another reason is that non-stationary effects like vibrato introduce additional problems [3].

In most scenarios for source separation of instrument signals it is common to assume that the spectral harmonics do only partially overlap. This enables algorithms like non-negative matrix factorization (NMF) to approximate the mixture from a lower-rank decomposition in an unsupervised way. Such systems are described in [4] and [5]. Additionally the popularity of the class of NMF algorithms can be explained by the intuitive way in which they work on time-frequency representations of the mixture signal.

In the context of musical instrument source separation, many researchers have focused on including prior information about the sources in their algorithms [6]. The availability and detail of such a-priori information varies. Often systems learn spectral as well as temporal cues from training data or parts of the mixture where only one instrument is active. One example of such informed source separation systems is described by Ewert and Müller [7]. It incorporates the pitch and onset information encoded in a MIDI file to improve the separation result.

Even the number of sources is a simple but very important information for source separation algorithms. One of the main drawbacks of many source separation systems is that they rely on this information. In some scenarios, like popular western music, the sources to separate are grouped into Melody + Bass + Drums and a residual signal. Constraining the system to such a scenario allows the results to be evaluated even if the set of sources being separated is incomplete. Constrained systems like these are also sufficient for real-world applications such as the eminent karaoke scenario. Limiting the number of desired sources helps not only to improve the performance of the algorithms but is also related to the fact that the number of sources humans can perceive is limited, too. Although a threshold has not been systematically addressed so far, a variety of experiments have been carried out. David Huron found [8] that the number of voices humans can correctly identify is up to three. When Stöter and Schoeffler et. al. [9, 10] asked participants to identify the number of instruments in a piece of music, the participants were only able to identify up to three, similar to Huron's voice experiments. There is very little chance that listeners are able to detect the presence of more than three sources. However in trials with fewer than three instruments, listeners tended to be very sensitive: One of the stimuli in the [9, 10] experiments with 1168 participants consisted of a mixture of Violin and Flute played in unison. The results showed that 76% of the participants correctly identified two instruments. Only 18% of the participants underestimated by one instrument, 6% overestimated by one instrument.

Since humans are able to reliably detect even instruments played in unison, this is a good motivation to expect the same from an algorithm. In this paper we want to address this scenario which has not been brought up so far. We believe creating and evaluating new algorithms for separating sources playing in unison will improve source separation systems in general.

The remainder of this paper is organized as follows: Section 2 describes the challenges of a unison source separation scenario. We propose techniques based on the modulation characteristics of the signal to address the separation scenario in Section 3 and Section 4. In Section 5 we introduce a data set for the unison scenario.

* The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen (IIS).

Further we present and discuss the results from our study and a comparison between the algorithms in Section 5.3. Conclusions are then presented in Section 6.

2. UNISON SOURCE SEPARATION SCENARIO

Up to date there are very few proposed source separation methods which perform good on a variety of input signals without making general assumptions or constraints. Most of the current state-of-the-art algorithms address specific scenarios like voice or melody extraction, or harmonic percussive separation. Additionally assumptions about the mixture itself are important, too. In this paper we consider the linear single channel case:

$$x(n) = \sum_{s=1}^{N} x_s(n). \tag{1}$$

Describing a source separation scenario includes the number of sources N and the number of desired sources D which is normally smaller than N when the desired sources contain groups of sources like instrument classes (strings, woodwinds, etc.).

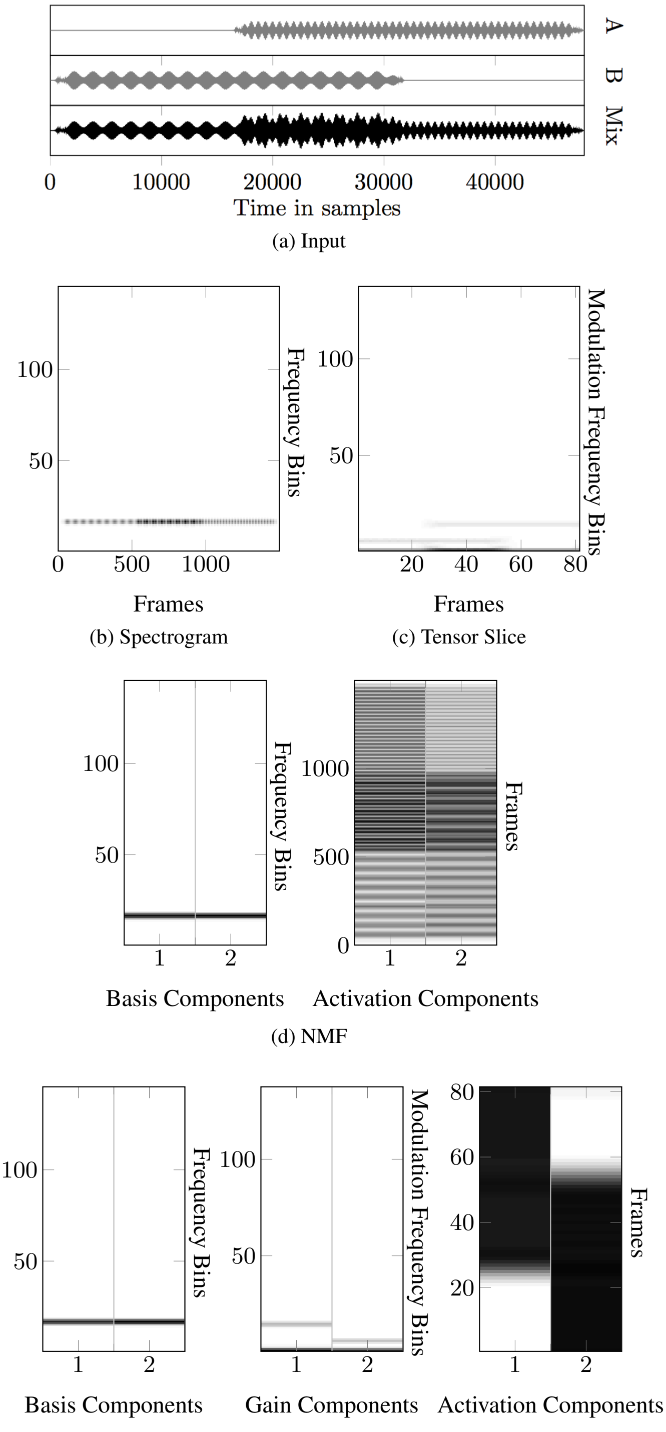
We propose a scenario where instruments play in unison. This means that they share the same fundamental frequency (regardless of the octave) so that the sources can overlap both in time and frequency. In fact unison[1] mixtures are meant to be as much overlapped as possible, hence they are very difficult to separate. However, due to masking effects, a relatively good subjective quality for the separated sources can be obtained, even if the other sources are not perfectly suppressed. As far as we know, there is no contribution to the source separation scene that focuses on mixtures of such unison sources.

The decomposition of sources with overlapping partials are covered in several other publications like [3] and [11] which are based on non-negative matrix factorization. Lin et. al. [12] address the problem by defining invariant timbre based features. We propose to address the problem from a different perspective and focus on analyzing the non-stationarities of the source signals. For most musical instruments, the non-stationary features are intentionally created, for instance with vibrato or tremolo effects, which make them valuable to track. These non-stationarities can be modeled or learned from the signals themselves.

In this work we assume that we can separate overlapping partials of the sources based on differences in amplitude and/or frequency modulation, resulting in the following model for a signal with P commonly modulated partials

$$x(n) = \sum_{p=1}^{P} \Big[\big(1 + a(n)\big) \cdot \sin\Big(2\pi f_{p,0}\big(n + \frac{1}{f_{1,0}} \sum_{m=m_0}^{n} f(m)\big) + \phi_{p,0}\Big)\Big], \tag{2}$$

where effectively the amplitude modulation is $a(n)$ and the frequency modulation of the first partial is $f(n)$.



(a) Input

(b) Spectrogram

(c) Tensor Slice

(d) NMF

(e) Non-Negative Tensor Factorization

Figure 1: Example of separating a mixture of two amplitude modulated signals by NMF and Modulation-NTF.
(a) Mixture of two sinusoids at 440 Hz with AM of 4.7 Hz and 12.6 Hz (fs=8 kHz), **(b)** STFT (FFT length = 256), **(c)** Slice of Modulation Tensor (FFT length = 256), **(d)** $\mathbf{W} \times \mathbf{H}$ Result of Non-Negative Matrix Factorization ($\beta = 1$) after 100 iterations, **(e)** $\mathbf{G} \times \mathbf{A} \times \mathbf{S}$ Result of Non-Negative Tensor Factorization ($\beta = 1$) after 100 iterations,

[1] greek: with *one voice*

3. SEPARATING BY AMPLITUDE MODULATION

Amplitude modulation is normally not present, isolated in acoustical instruments. However electric pianos like Rhodes or Wurlitzers can generate a tremolo effect. Using the amplitude modulation to separate mixtures has already been done in [13] which makes use of the concept of *Common Amplitude Modulation*.

CAM is effectively the property of harmonics that share the same amplitude modulation across the bins. One way of analyzing it is a modulation spectrogram which is a frequency-frequency representation of a time domain input signal. There are also other ways to generate a modulation spectrum. A complete signal representation can be archived by a modulation tensor which holds the modulation spectrograms for each time frame. Barker and Virtanen [14] found a way to utilize the modulation tensor for single channel source separation. Standard NMF models the spectrogram by the sum of K components which are each factored into frequency/basis and time/activations components:

$$\mathbf{X}_{n,m} \approx \sum_{k=1}^{K} \mathbf{W_{n,k}} \times \mathbf{H_{k,m}}. \tag{3}$$

Non-negative Tensor factorization approximates a modulation tensor by a product of three matrices containing the frequency/basis, time/activation signals, and the modulation gain for each component. Compared to [14] we choose to generate the modulation tensor in way that is simpler and easier to invert. Barker and Virtanen use a Gammatone filter bank and rectification to model the characteristics of the human auditory system. We used a two-stage DFT filter bank where the modulation domain is based on magnitude spectrograms. Although this can give perceptually less optimal results, each step can be directly inverted by using the complex representation. Barker already showed that the NTF based approach gives better results on speech signals. We found that this approach can be used to separate two instrument mixtures by their amplitude modulation characteristics and is therefore ideal for the unison scenario.

In Figure 1 we show the factorization of a simple amplitude modulated input signal for comparison. The signal consists of two sinusoids which are linearly mixed. Both share the same carrier frequency but have different amplitude modulation frequencies. We choose a factorization into $K = 2$ components. From the activation components one can see that NMF is not able to separate the two signals sufficiently. NTF gives a smoother activation matrix and is able to generate the output with the separated amplitude modulations on each sinusoid. The modulation frequency gain matrix shows the two modulation frequency templates and the DC-component.

4. SEPARATING BY FREQUENCY MODULATION

Frequency modulation caused by vibrato is a very common playing style for string instruments but also for woodwind and brass instruments. Vibrato is an effect that is well studied especially in musicology. Performers tend to perform a vibrato in the same way when repeating a performance. This can be exploited in source separation scenarios. Typically, vibratos have modulation frequencies (rates) which vary between 4 and 8 Hz. Additionally vibrato rates vary across different instruments. In [15] the vibrato width (frequency deviation) was found to be significantly different between violinists and violists performers.

As with the amplitude modulated case NMF lacks the ability to model time varying frequencies since the $\mathbf{W}$ matrix is stationary. Several extensions for NMF have been proposed to improve the decomposition quality. [16] proposes frequency dependent activations matrices, [11] has developed a system which can be described as shift invariant NMF. Another approach is to model the spectral pattern changes by Markov chains [3]. All these approaches attempt to model the non-stationary effects within the decomposition model. In this paper we propose a method that increases the stationarity of the signal in preprocessing step and then use the standard NMF for the decomposition.

We make use of *time-warping* which refers to a mapping of the linear time scale t to a warped time scale τ via a mapping function $\tau = w(t)$. To ensure a unique mapping, the mapping function needs to be strictly increasing. For the discrete time case the mapping can be achieved by a time-varying re-sampling of the linear (i.e. regularly sampled) time signal under consideration. The instantaneous sampling frequency then corresponds to the first derivative of the mapping function. Although the mapping can be done from any time-span I on the linear time scale to any time span J on the warped time scale, in the discrete time case it is advantageous to have the same number of samples in the linear and warped time domain. This ensures that the average sampling frequency is the same in both domains. Such time-warping approaches have already been proposed for different purposes such as transform-based audio coding [17]. As in these applications, we derive the mapping function from the varying instantaneous fundamental frequency in such a manner that the variation of the frequency is reduced or removed. To be more precise the actual information needed is not the absolute instantaneous fundamental frequency but only its change over time. The discrete time warp map $w[n]$ is then simply the scaled sum of the relative frequencies $f[n]$:

$$w[n] = N \frac{\sum_{l=0}^{n} f[l]}{\sum_{k=0}^{N} f[k]} \qquad 0 \leq n < N, \tag{4}$$

where N being the number of samples of the signal under consideration. From the requirements for the mapping function it follows that the relative frequency $f[n]$ has to be positive at all instants and preferably should not exhibit large jumps. For the mapping from linear to warped time now the linear domain sample points $s[\nu]$ for the regularly spaced samples $x[\nu]$ in the warped domain are found by inverting $w[n]$. These sample points are then used to re-sample the linear time domain samples $x[n]$ to the warped time domain samples $x[\nu]$, in our case by employing an 128 times oversampled FIR low-pass filter. This processing leads to a sampling rate contour which is proportional to the pitch contour. Or in other words, a fixed number of samples are obtained in each period of the signal with the varying fundamental frequency. Mutatis mutandis the sample points $s[\nu]$ can be used for the re-sampling from warped time domain to linear time domain.

In this paper the time-warping was done globally over the full lengths of the signals under consideration. The globally time-warped sample sequence was then used in the further processing steps. In Figure 2 we show the results of the warping process in the time domain.

A similar approach using frequency modulation to separate a harmonic source from a mixture was proposed in [18]. Here the individual lines are demodulated to the base band using a com-

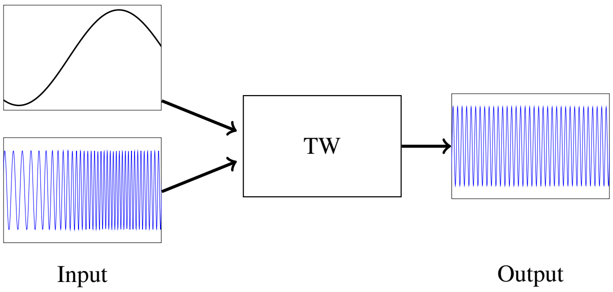


Figure 2: Example of applying warping to an input signal by using a frequency variation contour.

bined frequency tracking/demodulation approach. The difference to our approach is that first the absolute instantaneous frequency for every harmonic line has to be known instead of a relative frequency that is common to all harmonic lines of a single source. This relative frequency might be obtained easier than its absolute value for a mixed signal. Secondly every harmonic line has to be individually frequency demodulated while in our approach the full signal is frequency demodulated in one algorithmic step.

4.1. Pitch Variation Informed Source Separation

With the ability to remove the frequency modulation from a signal we can then include this system in a source separation system to address the non-stationarity issues of NMF based approaches. Figure 4 shows how this system works on a harmonic FM signal mixture. Plots (a) and (b) show the two input signals which are linearly mixed (c). For each source the warp contour needs to be calculated. The mixture is then warped with pitch variation estimates of source 1 (d) and source 2 (e). The actual separation/filtering of the sources is then done by using NMF which is not shown here. To separate the components from the warped mixture we used NMF on a spectrogram computed with a very long DFT (about 0.5 s). NMF can work unsupervised by detecting the more tonal **W** component by using a spectral flatness measure. The separated signals (f) and (g) then need to be warped back into the original time domain resulting in (h) and (i).

It is important to clarify that this approach would not be able to separate two modulating instruments playing in unison without having prior knowledge about the individual modulation functions. Although a pitch variation estimate might be difficult to achieve in a mixture our approach shows that such a system can make sense.

5. EXPERIMENTS

We wanted to evaluate the methods proposed in Sections 3 and 4 so that they show the fundamental differences in their separation quality. Like in [14] we choose not to address the problem of clustering the components after the matrix factorization operation. Instead of processing mixtures in a $A - B - AB$ or $A - AB - B$ paradigm we went for a supervised learning phase where we had access to the original source individually. In this *oracle* supervised approach for each of the sources we then learned the spectral, temporal (for NMF), and modulation gain components (for MOD-NTF) and concatenated them. The learned coefficients were then used to initialize the final factorization process. This way we can achieve the maximum possible quality.

Instrument	Vibrato	General MIDI #
Violin	yes	40
Viola	yes	41
Violon Cello	yes	42
Trumpet	no	56
Trombone	no	57
Horn	no	60
Bariton Sax	yes	67
Oboe	no	68
Clarinet	no	71
Flute	yes	73

Table 1: Instrument item test set

5.1. Test set

To build a test set we selected 10 instrumental items noted in Table 1. The items have each been generated by rendering C4 notes in a state of the art software sampler. All test have a duration of about three seconds. Items were equalized in loudness by using an iterative calculation of the loudness algorithm of the time varying Zwicker model. The implementation [19] was used. The 10 instrument items then generated 45 unique mixtures of two instruments each. The processing was done in 44.1 kHz / 16 bit.

5.2. Algorithms

The test set was processed by three algorithms: standard NMF, pitch variation informed NMF (PVI-NMF) (Section 4.1) and the non-negative tensor factorization based on modulation spectra (MOD-NTF) as described in Section 3). All factorizations for NMF and NTF were computed by minimizing the $\beta = 1$ divergence (Kullback-Leibler divergence). The Pitch Variation Informed-NMF (PVI-NMF) has been set up in the same way as the other algorithms. We choose to calculate results with $K = 2$ and $K = 4$. The pitch variation estimator is based on a method that was proposed by Bäckström in 2009 [20] with a subsequent post-processing to ensure the smoothness of the mapping.

Each of the algorithms did perform on the same filter bank and with the same sample rate. NMF approach did use a 2048 STFT with 512 samples hop size. For the MOD-NTF a second STFT based filter bank was used with 256 sample DFT size and 64 sample hop size. All methods use soft masking / wiener filtering for the actual synthesis.

5.3. Results

The results were evaluated by using commonly used evaluation measures provided by the PEASS Toolbox [21]. The evaluation measure are:

- Overall Perceptual Score (OPS)
- Target-related Perceptual Score (TPS)
- Interference-related Perceptual Score (IPS)
- Artifacts-related Perceptual Score (APS)
- Signal to Distortion Ratio (SDRi)
- Source to Interference Ratio (SIRi)

- Sources to Artifacts Ratio (SARi) [2]

The mean values of the PEASS evaluation are provided in Table 2. It can be seen that the SDR values give a different tendency than the OPS score, showing that the differences between both measures are substantial. Since unison mixtures are even very challening for humans to segregate we chose to focus on the psycho-acoustically weighted performance measures only. The results show a slightly better overall performance for the PVI-NMF. A more fine grained overview from the OPS results experiment is presented in Figure 3. It can be seen that results vary a lot between the mixtures. The modulation tensor factorization (MOD-NTF) performs good on mixtures like Clarinet-Viola (71-41) or Clarinet-Cello (71-42) where one source has vibrato and the other does not (see plots (e,f)). Although it performs well on average, MOD-NTF shows a high variance in the OPS results. The results have also been evaluated and confirmed subjectively by informal listening. Additionally we provide selected stimuli online on an acompanying webpage [3]. In general the PEASS scores give a good indication of quality. However the artifacts that are introduced by the standard NMF synthesis seem to be not well reflected. One possible reason is that PEASS toolbox has not been tested on artifacts from unison mixtures.

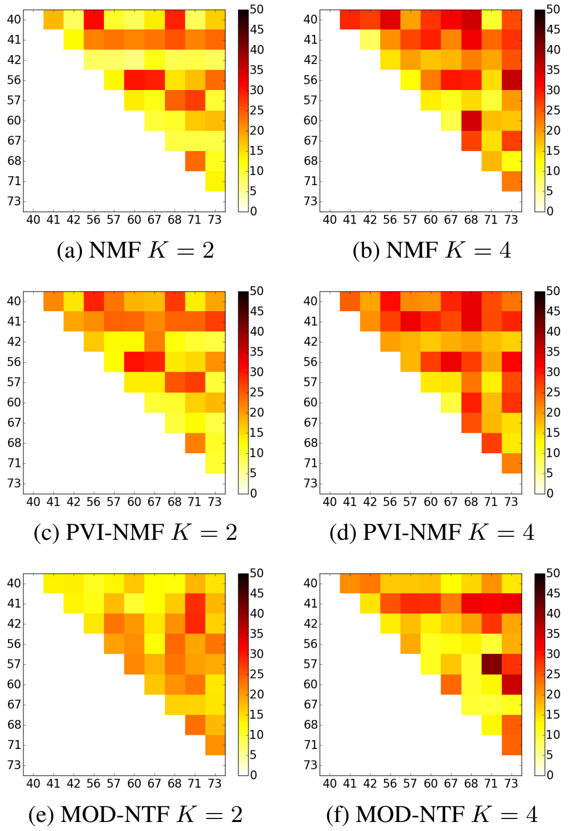
Future work could include a robust multi pitch variation estimator for musical instruments. Salamon and Gomez [22] describe the current state of the art of f0 estimation. Some approaches use source separation to estimate multiple f0 pitch tracks. Therefore our approach shows that a robust multi pitch f0 estimate can also help to improve source separation. In the future an iterative multi-step procedure could lead to better results in both problem domains.

6. CONCLUSIONS

This paper proposes a new source separation scenario for instruments played in unison. It highlights the time-varying aspects of the signal sources like amplitude or frequency modulations. By addressing these aspects, the separation quality for non-unison mixtures can generally be improved, too. Furthermore we present two methods to decompose those mixtures based on differences in the amplitude or frequency modulation of the sources. One is using a method already published based on a modulation tensor factorization. The other is a novel method that uses an estimate of the pitch variation of the two input sources to warp the mixture. Within the warped domain the frequency modulation of the desired source is removed so that the sources can be separated more easily from the mixture. The results of 45 mixtures have been evaluated by using the PEASS toolbox. The scores indicate an improvement of about 2 OPS points in favor of the pitch variation informed NMF compared to the standard NMF.

Algorithm	NMF	PVI-NMF	MOD-NTF
OPS	15.76	**17.64**	17.35
TPS	30.17	32.80	**34.03**
IPS	26.07	**27.03**	22.73
APS	46.14	**54.74**	46.06
SDRi	**2.96**	2.54	2.20
SIRi	2.31	1.80	**3.13**
SARi	22.87	23.35	**26.09**

Table 2: Results from Evaluation with PEASS 2.0 Toolbox [21]. Best performing algorithm is marked bold.



(a) NMF $K = 2$

(b) NMF $K = 4$

(c) PVI-NMF $K = 2$

(d) PVI-NMF $K = 4$

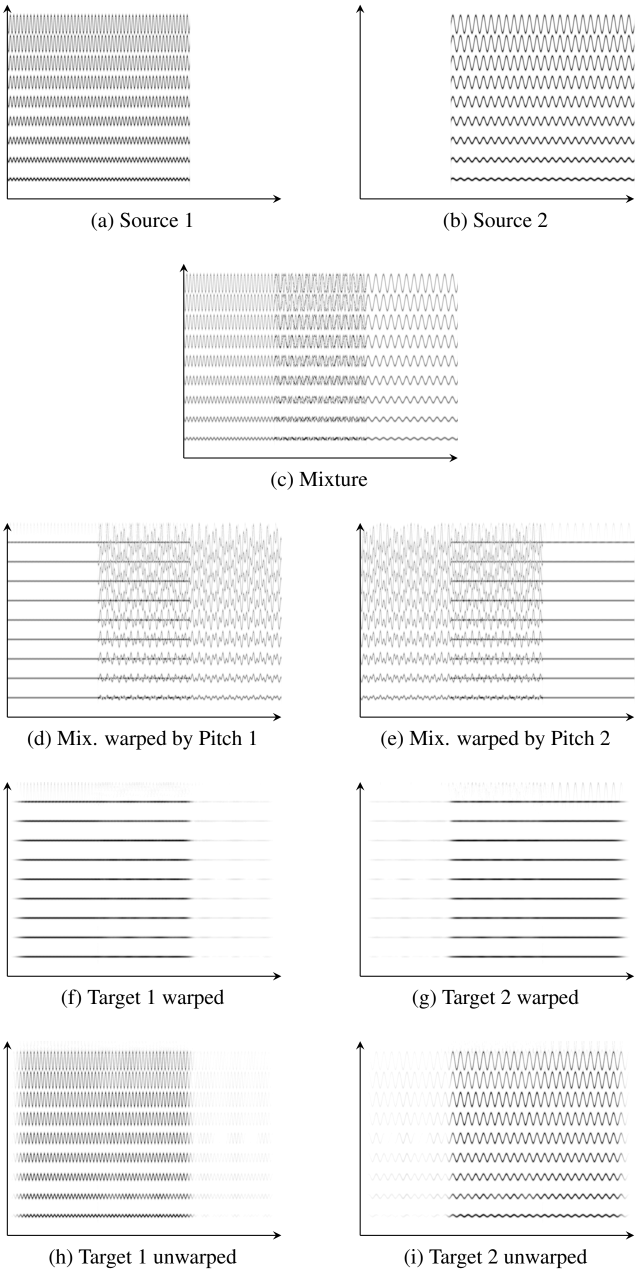
(e) MOD-NTF $K = 2$

(f) MOD-NTF $K = 4$

Figure 3: Results of Overall Perceptual Score. Each matrix represents the mean OPS values for each individual mixture of two sources. The x and y axis represent the instrument IDs in General MIDI notation (See Table 1).

[2] The i indicates that these scores have been calculated by decomposition with PEASS [21] instead of BSS EVAL.

[3] `http://www.audiolabs-erlangen.de/resources/2014-DAFx-Unison/`


(a) Source 1

(b) Source 2

(c) Mixture

(d) Mix. warped by Pitch 1

(e) Mix. warped by Pitch 2

(f) Target 1 warped

(g) Target 2 warped

(h) Target 1 unwarped

(i) Target 2 unwarped

Figure 4: Example of pitch variation informed NMF in the warped domain. *Time* is shown on horizontal axes. *Frequency* is shown on vertical axes.

7. REFERENCES

[1] Emmanuel Vincent, Shoko Araki, Fabian Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, Vikrham Gowreesunker, Dominik Lutter, and Ngoc Q. K. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[2] Nobutaka Ono, Zbynek Koldovsky, Shigeki Miyabe, and Nobutaka Ito, "The 2013 signal separation evaluation campaign," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[3] Masahiro Nakano, Jonathan Le Roux, Hirokazu Kameoka, Yu Kitano, Nobutaka Ono, and Shigeki Sagayama, "Non-negative matrix factorization with markov-chained bases for modeling time-varying patterns in music spectrograms," in *Latent Variable Analysis and Signal Separation*, pp. 149–156. Springer, 2010.

[4] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[5] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[6] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[7] Sebastian Ewert and Meinard Müller, "Using score-informed constraints for NMF-based source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012, pp. 129–132.

[8] D. Huron, "Voice denumerability in polyphonic music of homogeneous timbres," *Music Perception*, pp. 361–382, 1989.

[9] Fabian-Robert Stöter, Michael Schoeffler, Bernd Edler, and Jürgen Herre, "Human ability of counting the number of instruments in polyphonic music," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2013, vol. 19.

[10] Michael Schoeffler, Fabian-Robert Stöter, Harald Bayerlein, Bernd Edler, and Jürgen Herre, "An experiment about estimating the number of instruments in polyphonic music: a comparison between internet and laboratory results," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

[11] Paris Smaragdis, Bhiksha Raj, and Madhusudana VS Shashanka, "Sparse and shift-invariant feature extraction from non-negative data.," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2008, pp. 2069–2072.

[12] Yiju Lin, Wei-Chen Chang, Tien-Ming Wang, Alvin WY Su, and Wei-Hsiang Liao, "Timbre-constrained recursive time-varying analysis for musical note separation," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx)*, 2013, pp. 2–6.

[13] Yipeng Li, John Woodruff, and DeLiang Wang, "Monaural musical sound separation based on pitch and common amplitude modulation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1361–1371, 2009.

[14] Tom Barker and Tuomas Virtanen, “Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation,” in *Proceedings of INTERSPEECH*, 2013.

[15] Rebecca Bowman MacLeod, “Influences of dynamic level and pitch height on the vibrato rates and widths of violin and viola players,” 2006.

[16] Romain Hennequin, Roland Badeau, and Bertrand David, “NMF with time–frequency activations to model nonstationary audio events,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[17] Bernd Edler, Sascha Disch, Stefan Bayer, Fuchs Guillaume, and Ralf Geiger, “A Time-Warped MDCT Approach to Speech Transform Coding,” in *126th AES Convention*, Munich, Germany, May 2009, Preprint 7710.

[18] Avery Wang, “Instantaneous and frequency-warped techniques for source separation and signal parametrization,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ASSP)*, 1995, pp. 47–50.

[19] “GENESIS S.A.: Loudness toolbox (version 1.2),” 2012.

[20] Tom Bäckström, Stefan Bayer, and Sascha Disch, “Pitch variation estimation,” in *Proceedings of INTERSPEECH*, 2009, pp. 2595–2598.

[21] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[22] Justin Salamon and Emilia Gómez, “Melody extraction from polyphonic music signals using pitch contour characteristics,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759–1770, 2012.

[23] Romain Hennequin, Roland Badeau, and Bertrand David, “Time-dependent parametric and harmonic templates in non-negative matrix factorization,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2010, pp. 246–253.

[24] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] Estefanía Cano, Christian Dittmar, and Gerald Schuller, “Rethinking sound separation: Prior information and additivity constraint in separation algorithms,” in *Proceedings of the 16th Int. Conference on Digital Audio Effects (DAFx)*, 2013.

[26] Alexey Ozerov, Ngoc Q. K. Duong, and Louis Chevallier, “Weighted nonnegative tensor factorization: on monotonicity of multiplicative update rules and application to user-guided audio source separation,” Tech. Rep., 2013.

[27] Kazuyoshi Yoshii, Ryota Tomioka, Daichi Mochihashi, and Masataka Goto, “Beyond NMF: Time-domain audio source separation without phase reconstruction,” in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2013.