Non-negative Tensor Factorisation of Modulation Spectrograms for Monaural Sound Source Separation

Tom Barker, Tuomas Virtanen

Department of Signal Processing, Tampere University of Technology, Finland

Thomas.Barker@tut.fi, Tuomas.Virtanen@tut.fi

Abstract

This paper proposes an algorithm for separating monaural audio signals by non-negative tensor factorisation of modulation spectrograms. The modulation spectrogram is able to represent redundant patterns across frequency with similar features, and the tensor factorisation is able to isolate these patterns in an unsupervised way. The method overcomes the limitation of conventional non-negative matrix factorisation algorithms to utilise the redundancy of sounds in frequency. In the proposed method, separated sounds are synthesised by filtering the mixture signal with a Wiener-like filter generated from the estimated tensor factors. The proposed method was compared to conventional algorithms in unsupervised separation of mixtures of speech and music. Improved signal to distortion ratios were obtained compared to standard non-negative matrix factorisation and non-negative matrix deconvolution.

Index Terms: Non-negative matrix factorisation, non-negative tensor factorisation, modulation spectrogram, sound source separation

1. Introduction

Sound source separation is useful in many areas of speech enhancement, recognition and manipulation. There are currently numerous techniques for performing sound source separation, providing different performances for differing optimal conditions. Non-negative matrix factorisation (NMF) provides state-of-the-art single-channel blind source separation [1]. Basic NMF techniques decompose a mixture signal into a sum of components having a fixed spectrum and time-varying gain, which when multiplied approximate the mixture spectrogram.

The effectiveness of NMF stems from its ability to isolate redundant patterns in an unsupervised manner. One major shortcoming of basic NMF decomposition is that it does not fully utilise redundancy of sounds across frequencies, though. For example, random permutation of a spectrogram's frequency bins does not affect the outcome of regular NMF. Many natural sounds exhibit consistency in structure across their spectra, and it can be beneficial to take advantage of this fact. Current variations of NMF which take account of harmonic structure exist, but require a prior knowledge of source specific parameters [2]. Non-negative matrix deconvolution (NMD) [3] is an extension of NMF where a component spectrum is modelled as a convolution between a spectrum and filter. Source-filter NMF (SF-NMF) [4] models each component spectrum as multiplication of an excitation and filter spectrum. Both NMD and SF-NMF are capable of utilising the redundancy of audio spectra and modelling their structure. However, they require more parameters for representing spectra than conventional NMF and are therefore prone to over-fitting if no other restrictions are used.

Conventional NMF algorithms operate on the magnitude or power spectrogram, whilst in the human auditory system the sound is transduced to a representation based on the low frequency amplitude modulations within frequency bands, known as the modulation envelope. Modulation patterns are likely some of the cues utilised in higher level processing of stream information in the brain [5, 6] and it can be shown that these are present within harmonic sounds [7]. The spectrogram of a modulation envelope is termed the *modulation spectrogram* (MS), and it has been shown to be beneficial in representing speech signals [8].

In this paper, we propose the use of the MS to represent the signal so that structurally related content in different bands will exhibit similar features.

We propose to use non-negative tensor factorisation for decomposing the MS into component factors. The proposed non-negative tensor factorisation (NTF) technique utilises MS information redundancy between frequency bands. To our knowledge, this is the first use of NTF for a single-channel source separation problem; previous applications of NTF have focussed solely on multi-channel source separation [9].

2. Modulation spectrogram feature representation

The generation of the MS is based on a computational model of the cochlea, the structure in the ear where transduction of acoustic vibration to an electrical signal occurs. Components must be sufficiently distinct in frequency to be perceived as separate, partly due to the physical resonance properties of the inner ear structures. Components similar in frequency can therefore be considered to reside in the same auditory filter 'channel'. The cochlear output is modelled by a bank of overlapping filters whose output approximates the excitation of a particular physical location sensitive to a specific frequency. The firing rate of hair-cells attached to the basilar membrane roughly translates to the instantaneous excitation power of a filterbank channel output.

The method for obtaining the MS features used in factorisation is now described. The mixture signal is filtered using a gammatone filterbank, from the Patterson-Holdsworth cochlea model [10] and each band linearly spaced according to the equivalent rectangular bandwidth of the filter. This was implemented using Slaney's Auditory Toolbox [11]. Each band is half-wave rectified (hair cells within the ear can not have a negative firing rate) and low-pass-filtered to obtain the modulation envelope (ME) using a single pole recursive filter with -3dB bandwidth of approximately 26Hz.

The modulation spectrogram is obtained for each channel from the ME of each channel. Envelopes are segmented into a series of overlapping frames and (Hamming) windowed be-

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 290000 and Academy of Finland grant number 258708

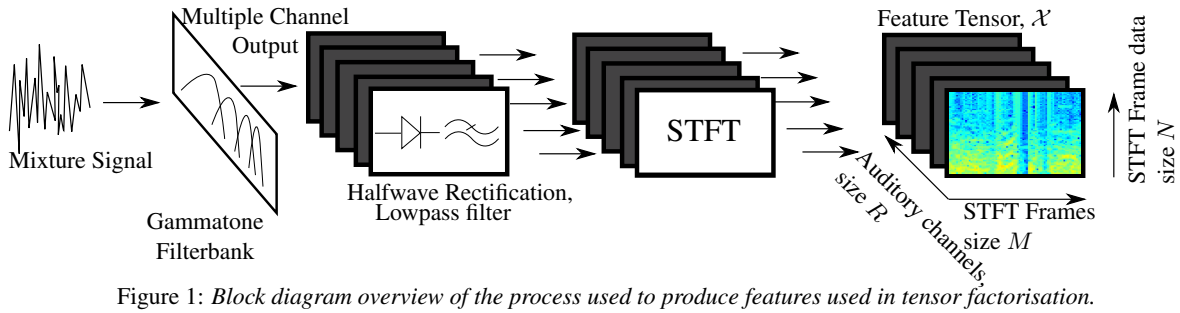
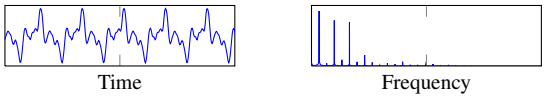
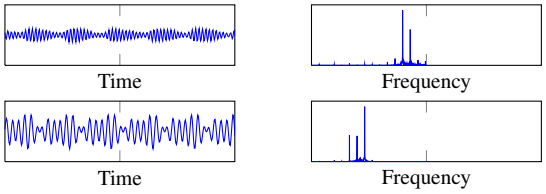


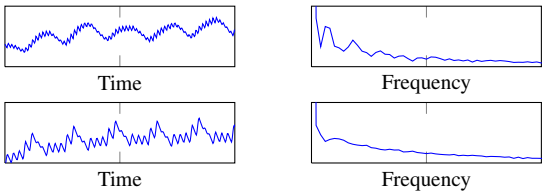
Figure 1: *Block diagram overview of the process used to produce features used in tensor factorisation.*



(a) *Time and frequency domain representations of a continuous note played on a clarinet.*



(b) *Time and frequency domain representations of two non-adjacent filterbank channel outputs.*



(c) *Time and frequency domain representations of two non-adjacent filterbank channel outputs after rectification and low-pass filtering. Signals have the same period. There is greater similarity between spectra as a result of the processing.*

Figure 2: *Demonstration of the use of modulation spectrograms on a continuous note played on a clarinet. The spectral magnitude content for non-adjacent filterbank channels with and without the modulation envelope as a feature is shown. (Axes scaled for clarity).*

fore the short-time discrete Fourier transform (STFT) is performed on each channel and the magnitude of the result retained. The output from the STFTs are truncated to 150 positive frequency bins, since low-pass filtering removes much of the high frequency content, which has no meaningful effect during the factorisation. The resulting data representation is therefore a 3-dimensional tensor, $\mathcal{X}$ (Figure 1) with dimensions of (number of filterbank channels x size of truncated STFT x number of observation frames). Harmonically related content has a similar modulation envelope and Figure 2 demonstrates the redundancy present between modulation envelope spectra for two non-adjacent channels of filterbank output compared with conventional magnitude spectra. Passing a harmonic signal (Figure 2a) through a filterbank produces a generally unrelated frequency magnitude spectrum output for each channel (Figure 2b) unless the ME is used as a feature (Figure 2c).

3. Tensor factorisation model

The mixture data modulation spectra tensor $\mathcal{X}$ has dimensions R x N x M which are the number of filterbank channels, truncated STFT length and number of observation frames, respectively. We model $\mathcal{X}$ as a sum of K components. Each component is modelled as a product of three factors $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$, each of which characterises one of the tensor dimensions. The model, $\hat{\mathcal{X}}$, for $\mathcal{X}$ is given as:

$$\mathcal{X}_{r,n,m} \approx \hat{\mathcal{X}}_{r,n,m} = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{A}_{n,k}\mathbf{S}_{m,k} \quad (1)$$

where $\mathbf{G}^{R\times K}$ contains the auditory channel dependent gain, $\mathbf{A}^{N\times K}$ the frequency basis function which models the spectral content of a modulation envelope feature, and $\mathbf{S}^{M\times K}$ is the time-varying activation of the component. The model is therefore able to describes a component's ME existing at different levels across channels, being activated at particular points in time. The model parameters are estimated by minimising the generalised Kullback-Leibler (KL) divergence D,

$$D(\mathcal{X}\|\hat{\mathcal{X}}) = \sum_{r,n,m} \mathcal{X}_{r,n,m} \log \frac{\mathcal{X}_{r,n,m}}{\hat{\mathcal{X}}_{r,n,m}} - \mathcal{X}_{r,n,m} + \hat{\mathcal{X}}_{r,n,m} \quad (2)$$

between X and $\hat{\mathcal{X}}$. KL divergence is widely used in non-negative tensor and matrix factorisation, [12] and provides effective results in comparison to other divergence measures for the baseline case to which this factorisation model is compared [1, 13].

Iterative update equations for minimizing the KL divergence can be derived as in [9, 14], and initialising $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$ to non-negative values ensures non-negativity throughout updates. The update equations use the definition of $\mathcal{C} = \mathcal{X}/\hat{\mathcal{X}}$, element-wise. The update rule for $\mathbf{G}$ is:

$$\mathbf{G}_{r,k} \leftarrow \mathbf{G}_{r,k} \frac{\sum_{n,m} \mathcal{C}_{r,n,m}\mathbf{A}_{n,k}\mathbf{S}_{m,k}}{\sum_{n,m} \mathbf{A}_{n,k}\mathbf{S}_{m,k}} \quad (3)$$

Similarly, the multiplicative update rule for $\mathbf{A}$ is:

$$\mathbf{A}_{n,k} \leftarrow \mathbf{A}_{n,k} \frac{\sum_{r,m} \mathcal{C}_{r,n,m}\mathbf{G}_{r,k}\mathbf{S}_{m,k}}{\sum_{r,m} \mathbf{G}_{r,k}\mathbf{S}_{m,k}} \quad (4)$$

$\mathbf{S}$ is updated using:

$$\mathbf{S}_{m,k} \leftarrow \mathbf{S}_{m,k} \frac{\sum_{r,n} \mathcal{C}_{r,n,m}\mathbf{G}_{r,k}\mathbf{A}_{n,k}}{\sum_{r,n} \mathbf{G}_{r,k}\mathbf{A}_{n,k}} \quad (5)$$

$\mathcal{C}$ is reevaluated between each update of $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$. In our tests, sufficient convergence to the KL solution had been reached after 200 iterations.

The total number of entries in the factor matrices $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$ is $K\times(M+R+N)$. N is a truncated discrete Fourier transform (DFT) result, so in practice is much smaller than the dimension of the DFT, (P) used in conventional NMF. A total of $K \times (M + P)$ entries is required for the NMF factor matrices and it

can be seen that in our implementations, where $R = 20$, $N = 150$ and $P = 513$ (redundancy in the 1024 bin DFT output allows removal of complex conjugates) , NTF requires fewer parameters than NMF, SF-NMF and NMD which can minimise over-fitting compared to the other approaches.

4. Synthesis of components from factorised tensors

Following factorisation of the mixture signal into components, $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$, where separation is achieved in the modulation envelope domain, reconstruction is carried out in a Wiener-filtering-like reconstruction approach. The basis functions $\mathbf{A}$ are in the modulation envelope domain, but for effective Wiener reconstruction, parameters are required in the STFT domain of the original mixture signal. Full bandwidth spectral basis functions for reconstruction are estimated using the channel and temporal activations $\mathbf{G}$ and $\mathbf{S}$. A component synthesis tensor, $\mathcal{V}$ is generated by taking the STFT of the output of each auditory filterbank channel when filtering the original mixture signal. $\mathcal{V}$ is complex-valued, but only the magnitude of each value is used for factorisation. Conceptually, the generation of $\mathcal{V}$ is much like that illustrated in Figure 1, with the rectification and low-pass stage removed. Truncation of STFT frequency bins is not performed, so the resulting tensor dimensions are R x P x M. The matrix of signal reconstruction basis functions, $\mathbf{B}$ (dimensions: P x K) is estimated by minimising the Kullback-Leibler divergence between $|\mathcal{V}|$ and its approximation $|\hat{\mathcal{V}}|$. $|\hat{\mathcal{V}}|$ is calculated from components $\mathbf{G}$, $\mathbf{B}$ and $\mathbf{S}$:

$$|\hat{\mathcal{V}}|_{r,p,m} = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{S}_{m,k} \tag{6}$$

Defining $\mathcal{E} = |\mathcal{V}|/|\hat{\mathcal{V}}|$ allows repeated application of update rule:

$$\mathbf{B}_{p,k} \leftarrow \mathbf{B}_{p,k} \frac{\sum_{r,m} \mathcal{E}_{r,p,m}\mathbf{G}_{r,k}\mathbf{S}_{m,k}}{\sum_{r,m} \mathbf{G}_{r,k}\mathbf{S}_{m,k}} \tag{7}$$

after $\mathbf{B}$ has been randomly initialised with non-negative values. 200 iterations were used in our experiments. The Wiener filter formed from $\mathbf{G}$, $\mathbf{B}$ and $\mathbf{S}$ is applied to $\mathcal{V}$, to produce K separated components, $\hat{\mathcal{V}}^k$ in the STFT domain thus:

$$\hat{\mathcal{V}}^k_{r,p,m} = \mathcal{V}_{r,p,m} \frac{\mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{S}_{m,k}}{\sum_{k'} \mathbf{G}_{r,k'}\mathbf{B}_{p,k'}\mathbf{S}_{m,k'}} \tag{8}$$

Conversion of each of the K sets of STFTs back to the time domain frames is performed by the inverse DFT of the p dimension. Overlap-add reconstruction for successive frames generates sets of channel outputs which can be summed over r to produce the separated time-domain components from which separation performance can be measured.

5. Simulation experiments

The separation performance of the proposed algorithm was evaluated against existing techniques, with a variety component clustering approaches. The different approaches provide insight into the basic separation ability of the proposed algorithm alongside practical usage performance (blind clustering). Testing over a large set of randomly generated mixture signals was performed computationally with separation performance evaluated using the BSS Eval Matlab toolbox [15].

5.1. Acoustic material

Acoustic mixture material was generated by mixing speech data from the CMU Arctic database [16] with music data from the jazz and classical styles within the RWC database [17]. Material was re-sampled to 16kHz for these tests. A complete speech utterance was chosen at random from the CMU database, as was a piece of music from RWC. A portion of the music of the same length as the speech was randomly selected and extracted from the music. The RMS power of each signal was normalised, and the speech and music samples combined, retaining the originals for later evaluation of separation performance. A test set of 100 unique mixtures was used.

5.2. Evaluated algorithms

The proposed method was evaluated against a general monaural audio non-negative matrix factorisation of spectral magnitude, as well as non-negative matrix deconvolution [3] (NMD), performing convolution only in frequency. The basic NMF method is similar to that of Virtanen in [1] although excluding temporal continuity constraints. Enforced sparseness was implemented and trialled, but produced negligible difference compared to standard NMF. Signal reconstruction was performed using a Wiener-filtering-like approach, as in [18] and Equation 8. It should be noted that many modifications to basic NMF exist which can improve its performance for particular signal types. Extension of many NMF modifiers to the extra dimensionality of the tensor factorisation could also be performed in many cases.

A Hamming analysis window of length 1024 samples and 50% overlap was used in all methods. The length of the analysis window used affects the performance obtained; it was found through preliminary tests on development material that was not part of the test set that a length of 64ms (1024 samples at 16kHz) worked well and has also provided good results in previous similar separation tasks [18] by providing a fair compromise between temporal resolution and sufficient representation of low-frequency content. Reasonable performance could also be obtained with other window lengths, however. An auditory filterbank of 20 channels was used in the NTF approach, this number also being chosen after providing good performance on development material. All factorisations were run for 200 iterations as sufficient convergence was reached at this point. A convolution filter length of 10 frequency bins was used in the NMD implementation. KL-divergence was used as the minimisation criterion in all methods.

5.3. Clustering of components

Three separate clustering approaches were used: Oracle clustering, blind clustering based on MFCCs, and blind clustering based on factor activations. Each approach allocated separated components to either of two sources, 'speech' or 'music'. The 'oracle' approach employed knowledge of the original speech and music signals in assignment of separated components which removed the clustering algorithm's influence from separation performance evaluation. In oracle clustering, each component was compared to both original sources of the mixture using the signal distortion ratio (SDR) of the BSS toolkit [15] and assigned to the source producing the higher SDR figure. Practically, an oracle approach can not be used in blind source separation though, and so blind clustering methods provide an indication of the potential application value of separation techniques. It should also be noted that using the oracle clustering approach with large numbers of bases will generally provide an unrealistically good separation performance since increasing numbers of bases reduces the minimal unit from which the sources can be reconstructed.

5.4. Blind component clustering

The use of blind clustering approaches for comparison of separation technique performance introduces a dependence on the selected clustering method into the evaluation. Where possible,

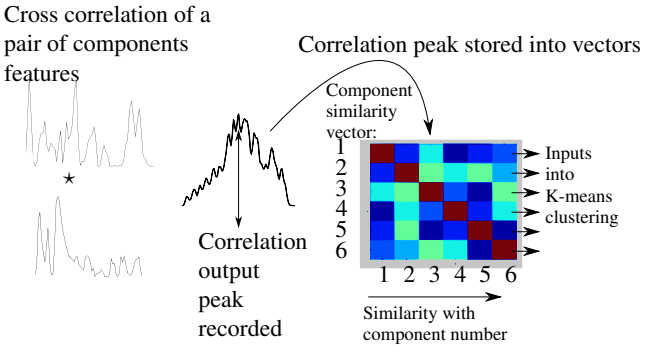


Figure 3: *Stylised representation of the generation of similarity vectors used in clustering of components*

features used in clustering should be equivalent across separation approaches, so that the clustering does not unduly affect the separation performance results obtained. Due to the different feature representations between factorisation approaches, equivalence is not always possible. The use of MFCCs calculated from time-domain components ensures equivalence of clustering features across several separation approaches, although may not be the most effective blind clustering method. Factors obtained from the NTF, NMF and NMD could also be used for clustering. Temporal activation factors (e.g columns of matrix $\mathbf{S}$) exist in both the NTF and NMF separations and so can be considered analogous for use in clustering, as can the mixing matrix obtained through NMD to a lesser extent. A feature vector of a component consists of its similarities to the other components, and the vectors of all the components are clustered using the k-means algorithm. Similarities between pairs of components are measured using the features mentioned previously: MFCCs or temporal activations.

The peak normalised cross correlation value between feature pairs are stored as shown in Figure 3. For the MFCC clustering, 13 channel MFCCs are calculated using Slaney's auditory toolbox [11]. Similarity vectors are calculated for each cepstral coefficient and concatenated to form the input into the k-means algorithm.

5.5. Results

The average of the separated speech and music SDRs was used to produce a single separation metric for each trial mixture. 100 mixtures were compared for each number of components for conventional NMF, NMD and the proposed modulation spectrogram NTF (MS-NTF) separation technique. Three separate clustering methods were used; the average SDR performance for each is shown in Figure 4. Oracle clustering highlights the baseline separation performance of each separation approach, where it can be seen that the separation ability of MS-NTF outperforms the other methods on trial below 10 components, after which NMF surpasses it. MS-NTF's superior basic separation performance for low component numbers translates into superior blind clustering performance since the correct assignment of all components becomes less likely with increasing component number. Of the blind clustering features used, temporal activations appear to be the more useful feature. It is interesting to note that once blind clustering is introduced, performance generally decreases with increasing numbers of components, except with temporal activations and NMF, where it tends to increase. Separating directly into 2 components with MS-NTF produces on average the best separation performance using a blind clustering approach.

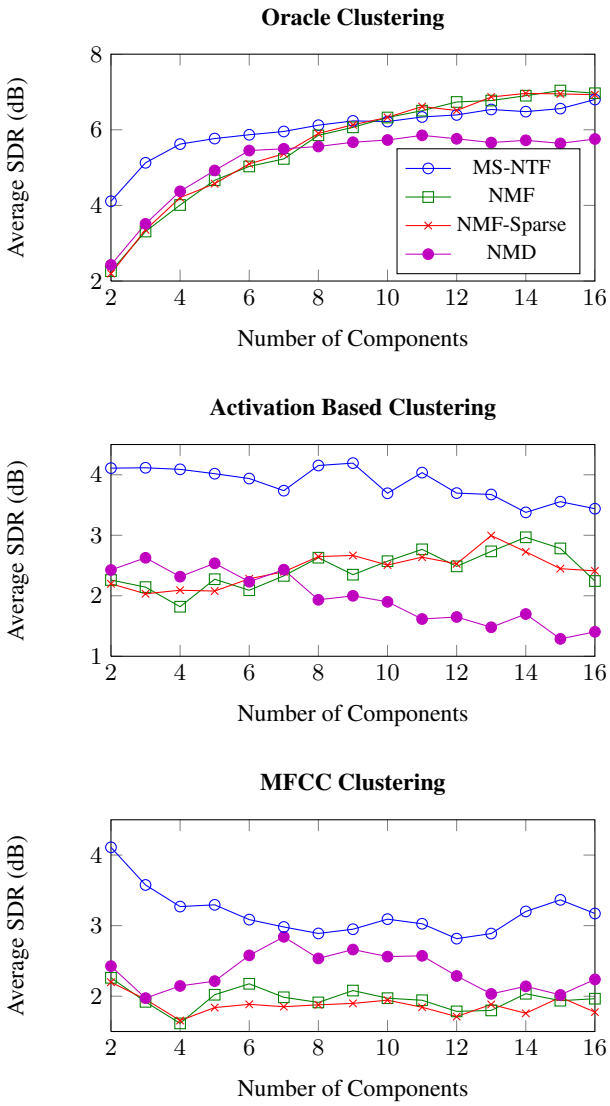


Figure 4: *Separation performance across different separation and clustering approaches. Legend in the top panel applies to all plots.*

6. Conclusions

This paper has proposed a novel algorithm for monaural sound source separation which used the modulation spectrogram as a feature in non-negative tensor factorisation. The model makes use of redundancy in spectral similarities across frequencies during the factorisation of a mixture signal into its constituent components. The proposed approach was evaluated against other source separation techniques on the decomposition of mixtures of 2 sound sources into varying numbers of components. For low numbers of components, MS-NTF produced better source separation performance than conventional NMF and NMD for the speech-music mixture signals under test. With oracle clustering, separation performance increased with increasing component numbers, and NMF approaches outperformed the proposed MS-NTF algorithm above 10 components. The utility of the proposed method is demonstrated by results obtained with lower number of components, where in the practical use case of blind clustering, superior performance is obtained compared to the established techniques of NMF and NMD.

7. References

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

[2] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcription," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 538 –549, March 2010.

[3] M. Schmidt and M. Mørup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation," in *Independent Component Analysis and Signal Separation, International Conference on*, ser. Lecture Notes in Computer Science (LNCS), vol. 3889. Springer, April 2006, pp. 700–707.

[4] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in Models for Acoustic Processing: Neural Information Processing Systems Workshop*, 2006.

[5] D. E. Broadbent and P. Ladefoged, "On the fusion of sounds reaching different sense organs," *Journal of the Acoustical Society of America*, vol. 29, no. 6, pp. 708–710, 1957.

[6] C. Plack, *The Sense of Hearing*. Lawrence Erlbaum Associates, 2005, ch. 10, pp. 199–201.

[7] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 2, pp. 255 –266, February 2008.

[8] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on, ICASSP-97.*, vol. 3, April 1997, pp. 1647 –1650 vol.3.

[9] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. of Irish Signals and Systems Conf.*, 2005.

[10] R. D. Patterson, K. Robinson, J. Holdsworth, D. Mckeown, C. Zhang, and M. Allerhand, "Complex Sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Honer, Eds., Oxford, 1992, pp. 429–443.

[11] M. Slaney, "Auditory toolbox version 2," Interval Research Corporation Technical Report No. 10, 1998. [Online]. Available: https://engineering.purdue.edu/~malcolm/interval/1998-010/AuditoryToolboxTechReport.pdf

[12] D. FitzGerald, E. Coyle, and M. Cranitch, "Using tensor factorisation models to separate drums from polyphonic music," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, 2009, pp. 294–298.

[13] F. Weninger and B. Schuller, "Optimization and parallelization of monaural source separation algorithms in the openblissart toolkit," *Journal of Signal Processing Systems*, vol. 69, pp. 267–277, 2012.

[14] D. FitzGerald, E. Coyle, and M. Cranitch, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, 2008.

[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[16] J. Kominek and A. W. Black, "*CMU Arctic Databases for Speech Synthesis*," [Online]., Available: http://www.festvox.org/cmu_arctic/

[17] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *In Proc. 3rd International Conference on Music Information Retrieval*, 2002, pp. 287–288.

[18] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *INTERSPEECH*, 2010, pp. 717–720.