Semi-Supervised Non-Negative Tensor Factorisation of Modulation Spectrograms for Monaural Speech Separation

Tom Barker and Tuomas Virtanen

Abstract— **This paper details the use of a semi-supervised approach to audio source separation. Where only a single source model is available, the model for an unknown source must be estimated. A mixture signal is separated through factorisation of a feature-tensor representation, based on the modulation spectrogram. Harmonically related components tend to modulate in a similar fashion, and this redundancy of patterns can be isolated. This feature representation requires fewer parameters than spectrally based methods and so minimises overfitting.**

Following the tensor factorisation, the separated signals are reconstructed by learning appropriate Wiener-filter spectral parameters which have been constrained by activation parameters learned in the first stage.

Strong results were obtained for two-speaker mixtures where source separation performance exceeded those used as benchmarks. Specifically, the proposed semi-supervised method outperformed both semi-supervised non-negative matrix factorisation and blind non-negative modulation spectrum tensor factorisation.

I. Introduction

SOUND SOURCE SEPARATION is the process of decomposing a recorded audio mixture into the original constituent components from which it was formed. The process has many applications, including speech enhancement [1], noise reduction and robustness [2], musical re-mixing [3] and improvement of quality in hearing aid applications [4]. Depending on the type of signal to be separated, numerous techniques exist which may be well suited to the application at hand.

Non-negative matrix factorisation (NMF) and non-negative tensor factorisation (NTF) produce state-of-the-art blind single source separation [5], [6], [7], [8]. In these cases, prior knowledge about the mixture signal is not assumed. Where more information about the mixture is available, supervised separation approaches can be employed. Generally, the availability of a dictionary-based model for each source can produce good separation results. As shown in [9], [10], the established technique for supervised NMF-based separation is to assemble a dictionary of atoms from training material, relevant to each source, and learn a linear additive combination of these which approximate the mixture signal. The atoms active for each source approximate the sources that the mixture is comprised of in the magnitude spectrogram domain.

Tom Barker and Tuomas Virtanen are with the Department of Signal Processing, Tampere University of Technology, Finland (email: {thomas.barker, tuomas.virtanen}@tut.fi).

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

If only a model for one source is available, then semi-supervised separation approaches can be employed [11], [12], to provide superior performance compared to blind separation. In this scenario, both spectral and temporal activation parameters for one of the sources must be estimated, whilst another is modelled through existing atoms. It is quite possible for overfitting to occur in these cases, where the non-supervised portion of the signal which should ideally model missing data instead models the entire mixture. Where both sources in the mixture have similar properties, such as in speech-speech mixtures, this problem becomes exaggerated.

Our proposed method minimises the effect of overfitting by reducing the number of parameters required to represent the mixture signal. Instead of operating on the power spectrogram, as in conventional NMF-based approaches, we utilise a feature known as the modulation spectrogram (MS). Sources are separated based on the co-modulation across different frequency sub-bands, which exist in harmonic sounds [13]. It has also been shown that these modulation features are likely one of the cues utilised in higher-level human auditory system stream segregation [14], and are useful in representing speech signals [15]. The MS features are separated through a tensor factorisation model, which represents each component as modulation spectra being activated across different sub-bands at each time frame.

The rest of this paper is organized as follows: Section II details the existing NMF-based sound source separation approaches which currently exit. Section III introduces and explains the proposed method; first detailing the modulation spectrogram as a feature, then how this is incorporated into a semi-supervised tensor factorisation model and finally how the separated audio can be reconstructed from the factorised components. Evaluation of the proposed method is given in Section IV, and conclusions are presented in Section V.

II. Baseline NMF-Based Separation Approaches

Non-negative matrix factorisation (NMF) produces state-of-the-art single-channel source separation. It decomposes a magnitude spectrogram matrix $\mathbf{X}$ of a mixture signal into a sum of components which have a fixed magnitude spectrum and time varying weight. The magnitude spectrogram matrix $\mathbf{X}$ can then be modelled as matrix $\hat{\mathbf{X}}$, the product of spectral atom matrix, $\mathbf{B}$ with their corresponding weight matrix, $\mathbf{W}$ as:

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{BW}. \tag{1}$$

NMF's effectiveness in the blind case is based on its ability to isolate redundant patterns in an unsupervised manner. In

the case of Equation (1), both $\mathbf{B}$ and $\mathbf{W}$ would be initialised randomly, and structure of recurring elements would be learned. In the supervised case, the component spectra are learned from training material and only the weights, $\mathbf{W}$ are estimated.

The semi-supervised case consists of modelling one of the sources through a dictionary of spectral atoms, which remain unchanged, and estimating the spectra of the atoms for the unknown source(s). The mixture is therefore modelled by the concatenation of the matrices describing the known and unknown sources. That is,

$$\mathbf{B} = [\mathbf{B}^{S1}\mathbf{B}^{S2}] \tag{2}$$

$$\mathbf{W} = \begin{bmatrix}\mathbf{W}^{S1}\\ \mathbf{W}^{S2}\end{bmatrix} \tag{3}$$

where superscripts $S1$ and $S2$ denote spectral basis dictionaries for each source. Specifically, $\mathbf{B}^{S1}$ would be constructed from atoms obtained from training data consisting of material from $S1$ whilst $\mathbf{B}^{S2}$ would be estimated to represent $S2$ while processing the mixture. Likewise, weights matrix $\mathbf{W}$ is composed of the concatenation of weights for each source, such that the mixture model becomes:

$$\hat{\mathbf{X}} = \left[\mathbf{B}^{S1}\mathbf{B}^{S2}\right]\begin{bmatrix}\mathbf{W}^{S1}\\ \mathbf{W}^{S2}\end{bmatrix} \tag{4}$$

Matrices $\mathbf{B}^{S2}$, $\mathbf{W}^{S1}$ and $\mathbf{W}^{S2}$ are estimated by minimising the Kullback-Leibler (KL) divergence (shown to be effective in audio-based NMF applications [5]) between $\mathbf{X}$ and $\hat{\mathbf{X}}$ through iterative update equations which act across all weights in $\mathbf{W}$ but on only the untrained bases in $\mathbf{B}^{S2}$, as in [11].

III. Proposed NTF Separation Method

The proposed method makes use of a tensor factorisation model which represents the mixture signal as the sum of products in 3 dimensions, rather than 2 as in NMF. The signal is divided into sub-bands, and recurring low-frequency modulation patterns across bands approximate the mixture. The 3-dimensional tensor model and the limited number of spectral bins required to represent the modulations reduces the number of parameters required for the model, which in turn reduces the tendency for overfitting. The semi-supervised approach involves learning some of the modulation patterns from training material, whilst others are learned through update equations, similarly to in established NMF-based approaches.

A. Modulation Envelope Feature Representation

The use of the modulation spectrogram as a feature is inspired by the computational modelling of the human cochlea, where vibration in this inner-ear structure is transduced to electrically encoded signals. Spatial excitation response of the basilar membrane is dependent upon excitation frequency, and separate components must be sufficiently distinct in frequency to stimulate unique areas of the membrane. This

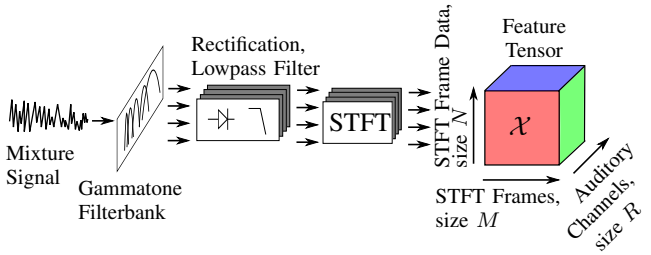

Fig. 2. Block diagram overview of the production of a modulation spectrogram tensor used in factorisation.

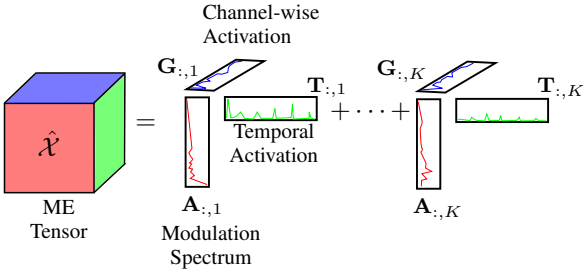

Fig. 3. An approximation to $\mathcal{X}$ the mixture tensor, $\hat{\mathcal{X}}$ is formed by the sum of outer products between rank-one tensors and an error term. Each rank-one tensor is a column of the component matrices $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{T}$ and represent a separate component in the separation. Update equations aim to minimise the difference between $\mathcal{X}$ and $\hat{\mathcal{X}}$.

motivates the idea that similar frequencies exist within the same auditory filter 'channels', and that the output of the cochlea can be divided into frequency bands. Each band's output approximates the instantaneous excitation energy present in that channel. A harmonic sound spanning many auditory channels will produce a similar modulation pattern in all channels. This redundancy between modulation envelope spectra does not exist in the conventional spectral representation of mixture signals, so can not be used as a feature when separating spectral components originating from the same source within a mixture.

The MS based tensor used in our factorisations is produced in the following way: First, the monaural audio signal is filtered with a 20-band gammatone filterbank, implemented with Slaney's *Auditory Toolbox* [16]. The output of each filterbank channel is half-wave rectified then low-pass filtered with a single-pole recursive filter with -3dB bandwidth of 26Hz, to produce the modulation envelope (ME). The modulation spectrogram for each filterbank channel is obtained by taking the short-time Fourier transform (STFT) of each channel with a Hamming analysis window. The frequency dimension of the STFT output is truncated to 150 positive frequency bins since much of the high frequency content is removed by filtering, resulting in no meaningful contribution from these bins during factorisation. A 3-dimensional tensor, $\mathcal{X}$ (Figure 2) is therefore produced, with dimensions of (number of filterbank channels x size of truncated STFT x number of observation frames) which we denote as R x N x M.

B. Tensor Model

Tensor $\mathcal{X}$ is approximated by a sum of K components. Each component is the product of three factors, one each from $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{T}$ where $\mathbf{G}^{R\times K}$ contains the auditory channel dependent gains, $\mathbf{A}^{N\times K}$ the frequency basis functions which models the spectral content of a modulation envelope feature, and $\mathbf{T}^{M\times K}$ is the time-varying activation of each component (Figure 3). The approximation, $\hat{\mathcal{X}}$, for $\mathcal{X}$ is given as:

$$\mathcal{X}_{r,n,m} \approx \hat{\mathcal{X}}_{r,n,m} = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{A}_{n,k}\mathbf{T}_{m,k} \tag{5}$$

The model therefore describes a component's ME existing at different levels across channels, being activated at particular points in time.

This tensor representation benefits from requiring fewer parameters than conventional NMF, making it less prone to over-fitting. The truncation of the discrete Fourier transform (DFT) results in our representation means that the total number of entries in the factor matrices $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{T}$ is $K \times (M + R + N)$. Conventional NMF approaches retain the full-length DFT result (defined here as length P), which totals $K \times (M + P)$ entries in the NMF representation. In our implementations, $R = 20$, $N = 150$ and $P = 513$ (redundancy in the 1024 bin DFT output allows removal of complex conjugates), so $R + N < P$.

C. Semi-Supervised Data Representation

In our semi-supervised application of MS-NTF, a model for a single known speaker, $S1$ is learned from training material. Model parameters describing channel-wise activations and modulation spectra basis functions for $S1$ are contained in matrices $\mathbf{G}^{S1}$ and $\mathbf{A}^{S1}$ respectively. Modulation spectra and channel gain data in $\mathbf{G}^{S1}$ and $\mathbf{A}^{S1}$ are produced from training material utterances. For each example, modulation tensors are produced as in Figure 2, and each single time frame of the modulation tensors are factorised into a single component as presented in [6]. Each component forms a column in $\mathbf{G}^{S1}$ and $\mathbf{A}^{S1}$. The spectra and channel-gains of the unknown speaker $S2$ are estimated through update equations given in Section III-D, and are contained in the randomly non-negative initialised matrices $\mathbf{G}^{S2}$ and $\mathbf{A}^{S2}$ (Figure 4). Time-wise activations for both sources are contained in the matrices $\mathbf{T}^{S1}$ and $\mathbf{T}^{S2}$.

For a 2-source mixture, with $K1$ the number of components approximating source 1 and $K2$ the number for source 2, the mixture signal $\hat{\mathcal{X}}$ can be described as the sum of components:

$$\hat{\mathcal{X}}_{r,n,m} = \sum_{k=1}^{K1} \mathbf{G}^{S1}_{r,k}\mathbf{A}^{S1}_{n,k}\mathbf{T}^{S1}_{m,k} + \sum_{k'=1}^{K2} \mathbf{G}^{S2}_{r,k'}\mathbf{A}^{S2}_{n,k'}\mathbf{T}^{S2}_{m,k'} \tag{6}$$

The left hand sum-of-products in Equation (6) is the approximation to source 1 (supervised) whilst the right-hand sum-of-products approximates source 2 (unsupervised).

The matrices to be factorised can also be formed as per the NMF examples presented in Section II thus:

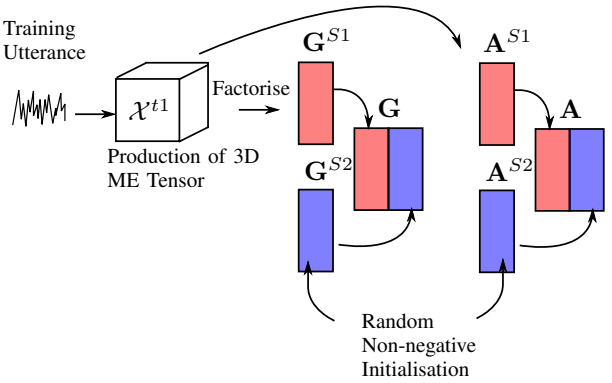

Fig. 4. Stylised overview of the production of channel activation and modulations spectra data matrices $\mathbf{G}^{t3}$ and $\mathbf{A}^{t3}$ used in semi-supervised separation of the mixture signal. Concatenation of matrices obtained from training material and those randomly initialised matrices to model the unknown source form the matrices used in tensor factorisation.

$$\mathbf{G} = [\mathbf{G}^{S1}\mathbf{G}^{S2}] \tag{7}$$

$$\mathbf{A} = [\mathbf{A}^{S1}\mathbf{A}^{S2}] \tag{8}$$

$$\mathbf{T} = [\mathbf{T}^{S1}\mathbf{T}^{S2}] \tag{9}$$

This allows writing the the model in Equation (6) using Equation (5).

D. Factorisation Algorithm

The model parameters $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{T}$ are estimated by minimising the generalised Kullback-Leibler (KL) divergence D,

$$D(\mathcal{X}\|\hat{\mathcal{X}}) = \sum_{r,n,m} \mathcal{X}_{r,n,m} \log \frac{\mathcal{X}_{r,n,m}}{\hat{\mathcal{X}}_{r,n,m}} - \mathcal{X}_{r,n,m} + \hat{\mathcal{X}}_{r,n,m} \tag{10}$$

between X and $\hat{\mathcal{X}}$. This divergence measure is commonly used in non-negative tensor and matrix factorisation, [17], producing effective performance with NMF based source separation [5], [18].

The divergence is minimised by iteratively applying update equations derived as in [19], [20], to $\mathbf{G}^{S2}$, $\mathbf{A}^{S2}$, $\mathbf{T}^{S1}$ and $\mathbf{T}^{S2}$. The update equations use the definition of $\mathcal{C} = \mathcal{X}/\hat{\mathcal{X}}$, element-wise. The update rule for $\mathbf{G}^{S2}$ is:

$$\mathbf{G}^{S2}_{r,k'} \leftarrow \mathbf{G}^{S2}_{r,k'} \frac{\sum_{n,m} \mathcal{C}_{r,n,m}\mathbf{A}^{S2}_{n,k'}\mathbf{T}^{S2}_{m,k'}}{\sum_{n',m'} \mathbf{A}^{S2}_{n',k'}\mathbf{T}^{S2}_{m',k'}} \tag{11}$$

The update rule for $\mathbf{A}^{S2}$ is:

$$\mathbf{A}^{S2}_{n,k'} \leftarrow \mathbf{A}^{S2}_{n,k'} \frac{\sum_{r,m} \mathcal{C}_{r,n,m}\mathbf{G}^{S2}_{r,k'}\mathbf{T}^{S2}_{m,k'}}{\sum_{r',m'} \mathbf{G}^{S2}_{r',k'}\mathbf{T}^{S2}_{m',k}} \tag{12}$$

The update rule for $\mathbf{T}$ is:

$$\mathbf{T}_{m,k} \leftarrow \mathbf{T}_{m,k} \frac{\sum_{r,n} \mathcal{C}_{r,n,m}\mathbf{G}_{r,k}\mathbf{A}_{n,k}}{\sum_{r',n'} \mathbf{G}_{r',k}\mathbf{A}_{n',k}} \tag{13}$$

which updates weights for both estimated sources in $\mathbf{T}^{S1}$ and $\mathbf{T}^{S2}$, as per Equation (9). $\mathcal{C}$ is recalculated following

each update of $\mathbf{G}$, $\mathbf{A}$ or $\mathbf{T}$. The equations should be applied until convergence is reached.

E. Signal Reconstruction

Following factorisation of the mixture signal into $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{T}$ through update equations, components of the mixture signal are separated in the modulation envelope domain. The basis functions in $\mathbf{A}$ exist in the (band-limited) MS domain, but are required in the STFT domain for reconstruction. The full-bandwidth spectral basis functions for reconstruction are therefore learned by factorisation of a component synthesis tensor, $\mathcal{V}$, using channel and temporal activations $\mathbf{G}$ and $\mathbf{T}$ derived in the semi-supervised factorisation. $\mathcal{V}$ is created by filtering the mixture signal with the auditory filterbank described in Section III-A and taking the STFT of each channel output. This process is conceptually like that demonstrated in Figure 2, omitting the rectification and filtering stage. The complex-valued STFT output is retained, but only the magnitudes are used during factorisation. No truncation of DFT results is performed, so $\mathcal{V}$ has dimensions of R x P x M. Signal reconstruction spectral basis functions are generated in a matrix $\mathbf{B}$ (dimensions: P x K) which 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 $\mathbf{T}$ as:

$$|\hat{\mathcal{V}}|_{r,p,m} = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{T}_{m,k}. \tag{14}$$

B is calculated by continued application of the update rule until convergence:

$$\mathbf{B}_{p,k} \leftarrow \mathbf{B}_{p,k} \frac{\sum_{r,m} \mathcal{E}_{r,p,m}\mathbf{G}_{r,k}\mathbf{T}_{m,k}}{\sum_{r,m} \mathbf{G}_{r,k}\mathbf{T}_{m,k}}, \tag{15}$$

where $\mathcal{E} = |\mathcal{V}|/|\hat{\mathcal{V}}|$ and $\mathbf{B}$ is randomly initialised with positive values.

Summation of all components either trained as $S1$ or generated from update equations as in $S2$ produces the Wiener filter which is applied to (complex-valued) $\mathcal{V}$ as in [10]. The spectrograms for each source, $\mathbf{V}^{P\times M}$ are, then, generated with the following operations:

$$\mathbf{V}^{S1}_{p,m} = \sum^{R} \mathcal{V}_{r,p,m} \frac{\sum_{k=1}^{K1} \mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{T}_{m,k}}{\sum_{k'=1}^{K} \mathbf{G}_{r,k'}\mathbf{B}_{p,k'}\mathbf{T}_{m,k'}} \tag{16}$$

$$\mathbf{V}^{S2}_{p,m} = \sum^{R} \mathcal{V}_{r,p,m} \frac{\sum_{k=1}^{K2} \mathbf{G}_{r,k}\mathbf{B}_{p,k}\mathbf{T}_{m,k}}{\sum_{k'=1}^{K} \mathbf{G}_{r,k'}\mathbf{B}_{p,k'}\mathbf{T}_{m,k'}} \tag{17}$$

which involve summation over the sub-band channels to produce a 2-dimensional complex valued matrix of time-frequency points in each case. Conversion back to the time-domain is achieved by applying an inverse STFT and overlap-add combination of the frames.

IV. Evaluation

The performance of the proposed semi-supervised NTF algorithm was evaluated against both semi-supervised NMF, and blind NTF separation. One-hundred test mixtures were produced between pairs of talkers, where one talker was consistently present in all mixtures. The consistent talker was used to inform the pre-learned for supervised speaker model, $S1$. The separation performance across all mixtures was evaluated using the BSS Eval Matlab toolbox [21].

A. Generation of Test and Training Material

Mixture material was produced by summing pairs of speech utterances from the CMU Arctic speech database. The database consists of 7 talkers, from which one formed the 'supervised' model (US English, Scottish Accent) in all cases, whilst the 'unknown' source was selected from the other available speakers. 10 training utterances from the supervised speaker were randomly selected as training material, and were excluded from the test set. Pairs of utterances from the test set were randomly selected, one from the supervised speaker model and another from the other speakers. Each pair was RMS normalised and summed to form the test mixtures.

To produce the training data for the proposed method from the training utterances, each was transformed to the modulation spectrogram domain. Each temporal frame was factorised to produce a frame spectrum and channel gains. From the 10 training utterances, approximately 12,000 frames of atoms were initially produced for both channel-wise gains and spectra which were then normalised based on their L1-norm. Each atom spectrum was concatenated with its corresponding channel-wise gains. The concatenated frames were k-means clustered to form 100 vector atoms using the KL-divergence between cluster centres and observations, as in [9]. The resulting dictionary contains a more generalised version of the speaker model, and reduces complexity for factorisation. The 100 vectors were then separated into channel-gain and spectral matrices $\mathbf{G}^{S1}$ and $\mathbf{A}^{S1}$ which were used in factorisation.

For the semi-supervised NMF factorisation, exemplar atoms were selected from the same training material. The atoms were k-means clustered to produce 1,000 atoms for the supervised model, an approach which produced stronger source separation performance than randomly sampled dictionaries in [9] and where performance does not increase significantly with dictionary size larger than 1,000 atoms.

B. Evaluation Procedure

Each test mixture was separated using the proposed method, and the source-distortion-ratio (SDR) metric used to evaluate separation performance. The mean SDR over the 100 test mixtures was compared for separation methods on trial.

The proposed semi-supervised factorisation, training and reconstruction methods were used to separate each of the test mixtures. The same test mixtures, with the same training data were also separated using the semi-supervised NMF algorithm [11], with and without sparsity constraints. Blind NTF separation as in [6] was evaluated, where the mixture was separated directly into 2 components, without the reliance on clustering, since this provided greatest separation performance in the blind case.

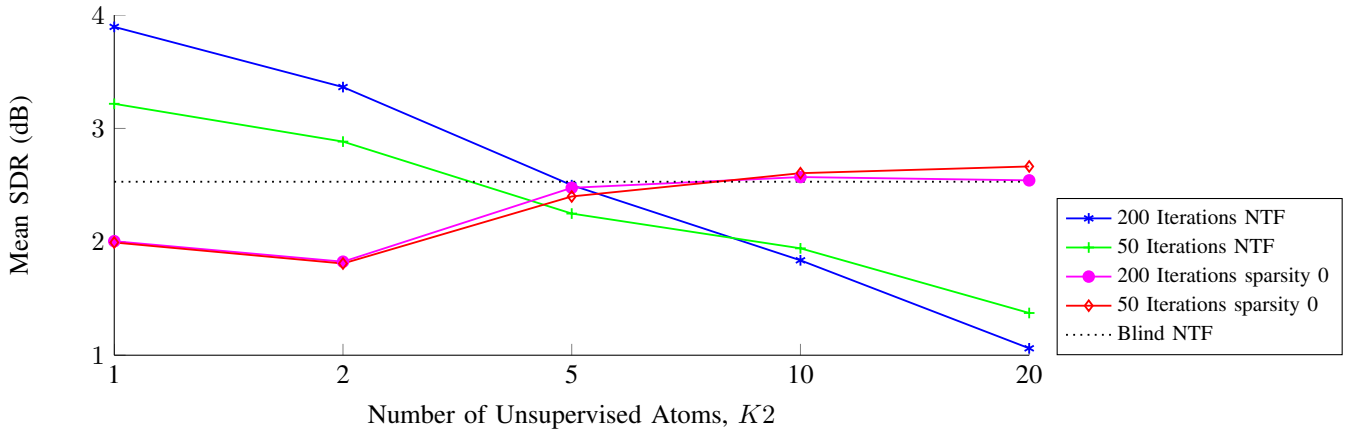

Fig. 5. Mean SDR figures for semi-supervised NTF (proposed) compared to semi-supervised NMF and blind NTF, against varying number of unknown speaker atoms, $K2$.

For both NMF and NTF-based separation methods, a 1024 sample (64 ms) window with 50% overlap was used in the analysis, as this had provided satisfactory separation in previous experiments [10], [6]. The Wiener-filtering reconstruction method (Equations 16, 17) was used to reconstruct each source.

Update equations were applied for different numbers of iterations. It has been noted in [11] that overfitting (and hence decreased separation performance) can occur as a greater number of iterations are applied in the NMF case. Separation performance was evaluated after applying 50, 100, and 200 update iterations. 200 iterations was the point where generally the cost function ceased to decrease by a non-negligible amount with each additional iteration. In the unsupervised NTF case, 200 iterations of updates were used. NMF separation was trialled both with and without sparsity constraints, and also for differing numbers of iterations. Both semi-supervised methods were trialled with 1,2,5,10 and 20 components modelling the unknown speaker.

C. Results

The average SDR for each separated sources produced a single separation performance figure for each test case. For each test condition, the mean over 100 randomly generated trials is presented in Figure 5. Only the NMF results without sparsity constraints are shown, as these provided appreciably better separation SDR across all test conditions than those with sparsity constraints. Also omitted are the results for 100 iterations in both the NMF and NTF approaches, which lie between the 50 and 200 iteration results at both extremes of $K2$.

The semi-supervised MS-NTF produces improved separation performance over blind MS-NTF separation for low numbers of noise components, and also removes the requirement to classify each speaker or cluster separated components. Better SDRs are also produced compared to semi-supervised NMF, where the unknown talker model can easily overfit to approximate the trained talker model. Overfitting however also appears to occur with increasing number of $K2$ atoms in the NTF case, since performance decreases with this parameter and becomes quite poor for $K2 > 5$. For higher values of $K2$, fewer update equation iterations produce stronger results for both NTF and NMF approaches, which suggests that overfitting is occuring in these cases.

V. Conclusions

We have presented a semi-supervised sound-source separation method which has been shown to be effective in separating two-speaker mixtures. Using training material, we were able to model a single speaker, whilst the unknown speaker parameters where learned through update equations applied to a tensor model. The representation of the mixture in the modulation spectrogram domain enabled redundancy arising from harmonic relationships across sub-bands to be exploited in identifying and modelling parts of a mixture arising from the same original source. Following semi-supervised factorisation in the modulation-spectrum domain, full bandwidth spectral models were learned for each source using the activations learned in the modulation domain. For test signals consisting of two speakers (one unknown), the proposed method was shown to be able to produce greater SDR-based separation performance than both blind-NTF and semi-supervised NMF methods.

References

[1] B. Raj, R. Singh, and T. Virtanen, "Phoneme-Dependent NMF for Speech Enhancement in Monaural Mixtures," in *proceedings of INTERSPEECH*. ISCA, 2011, pp. 1217–1220.

[2] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[3] D. FitzGerald, "Upmixing From Mono - A Source Separation Approach," in *proceedings of the 17th International Conference on Digital Signal Processing (DSP)*, 2011.

[4] M. S. Pedersen, "Source Separation for Hearing Aid Applications," Ph.D. dissertation, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2006.

[5] T. 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, March 2007.

[6] T. Barker and T. Virtanen, "Non-negative Tensor Factorisation of Modulation Spectrograms for Monaural Sound Source Separation," in *proceedings of INTERSPEECH*, 2013, pp. 827–831.

[7] D. FitzGerald and E. Coyle, "Shifted 2D Non-negative Tensor Factorisation," in *proceedings of the Irish Signals and Systems Conference*, 2006.

[8] P. Smaragdis, "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," in *Independent Component Analysis and Blind Signal Separation*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2004, vol. 3195, pp. 494–499.

[9] T. Virtanen, J. Gemmeke, and B. Raj, "Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2277–2289, 2013.

[10] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition," in *proceedings of INTERSPEECH*, 2010, pp. 717–720.

[11] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization," in *Latent Variable Analysis and Signal Separation*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7191, pp. 322–329.

[12] G. J. Mysore and P. Smaragdis, "A Non-negative Approach to Semi-supervised Separation of Speech from Noise with the Use of Temporal Dynamics," in *proceedings of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 17–20.

[13] A. Klapuri, "Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 255 –266, Feb. 2008.

[14] 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.

[15] S. Greenberg and B. Kingsbury, "The Modulation Spectrogram: in Pursuit of an Invariant Representation of Speech," in *proceedings of IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 1647 –1650.

[16] 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

[17] D. FitzGerald, E. Coyle, and M. Cranitch, "Using Tensor Factorisation Models to Separate Drums from Polyphonic Music," in *proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2009, pp. 294–298.

[18] 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. [Online]. Available: http://dx.doi.org/10.1007/s11265-012-0673-7

[19] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative Tensor Factorisation for Sound Source Separation," in *proceedings of Irish Signals and Systems Conference*, 2005.

[20] D. FitzGerald, E. Coyle, and M. Cranitch, "Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation," *Computational Intelligence and Neuroscience*, 2008.

[21] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, July 2006.