Blind Separation of Audio Mixtures Through Nonnegative Tensor Factorisation of Modulation Spectograms

Tom Barker, Tuomas Virtanen

***Abstract*—This paper presents an algorithm for unsupervised single-channel source separation of audio mixtures. The approach specifically addresses the challenging case of separation where no training data is available.**

By representing mixtures in the modulation spectrogram (MS) domain we exploit underlying similarities in patterns present across frequency. A 3-dimensional tensor factorisation is able to take advantage of these redundant patterns, and is used to separate a mixture into an approximated sum-of-components by minimising a divergence cost. Furthermore, we show that the basic tensor factorisation can be extended with convolution in time being used to improve separation results and provide update rules to learn components in such a manner. Following factorisation, sources are reconstructed in the audio domain from estimated components using a novel approach based on reconstruction masks which are learned using MS activations, and then applied to a mixture spectrogram.

We demonstrate that the proposed method produces superior separation performance to a spectrally-based nonnegative matrix factorisation approach (NMF), in terms of source to distortion ratio. We also compare separation with the perceptually-motivated *IPS* metric and identify cases with higher performance.

***Index Terms*—NMF, Source Separation, Factorization, Speech Enhancement**

I. Introduction

REAL audio recordings usually consist of contributions from multiple sound sources, for which it is often useful to have access to each separately. The separation of mixtures into constituent sources is known as sound source separation. There are multiple applications of such a process, including speech enhancement [1], musical transcription [2], de-noising and increasing robustness in automatic speech recognition [3], [4], and improving quality in hearing-aid applications [5], [6].

Many current source separation techniques rely on decomposition of a mixture signal into a linear combination of components; so-called *compositional models* (CM) [7]. Generally, the most effective of these utilise a representation which expresses the signal as a matrix describing the energy in frequency bins or bands at each time-frame. The frequency resolution varies in different representations, but the spectrogram (alternatively called short-time Fourier transform or STFT), is popular, along with the perceptually motivated mel-band [8] and constant-Q [9] scalings. These mixture matrices are typically factorised into spectral basis patterns (sometimes referred to as *atoms*), in one dimension and their time-varying activations in another [10], [11]. The basic paradigm can also be extended to include convolutional models which learn time-varying spectro-temporal patterns, as in [12], [13], [14]. These CM techniques are practical for separating multiple audio mixture types, since many naturally occurring sounds can be effectively represented using a fixed spectrum and time-varying gains. Most established CM approaches do not generally take advantage of structure present across frequency though. In the case of nonnegative matrix factorisation (NMF) of a mixture spectrogram, the frequency relationship between bins is not exploited in the factorisation model, and each DFT bin is independent of all others within the factorisation. For example, permuting the position of any matrix rows prior to factorisation will produce the same results for that row in either the new or original position; the values of a frequency bin in the matrix spectrogram are not considered *relative* to any others. However, extensions to NMF which are able to take advantage of dependence between frequencies in the factorisation model do exist. Convolutive NMF in frequency [15], for example, allows translation in frequency for specific spectral patterns, where harmonic atoms are used with a logarithmic frequency axis. With this technique, an underlying relationship between partials of a fundamental can be learned and used to represent sounds with similar spectral structure at varying pitches.

Source separation can be generally divided into supervised, semi-supervised or unsupervised processes. These describe the availability of a training data for all sources, some sources, or no sources present in the mixture, respectively. Neural-network based methods have recently started to be used for supervised and semi-supervised separation and speech enhancement [16], [17], whilst compositional models are an established technique across all approaches. Generally, use of prior knowledge about the constituent sources within a mixture will improve separation performance, and it should be expected that a well-matched supervised approach should outperform an unsupervised approach. Unsupervised separation, where very little or no prior knowledge is used is often referred to as 'blind' separation and where no training data is available a blind separation approach must be employed. Blind separation is highly challenging, and particularly where the problem is under-determined, meaning that there are fewer

Tom Barker and Tuomas Virtanen are with the Department of Signal Processing, Tampere University of Technology, Finland. E-mail: thomas.barker@tut.fi.

Part of 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.

Manuscript received April 19, 2005; revised September 17, 2014.

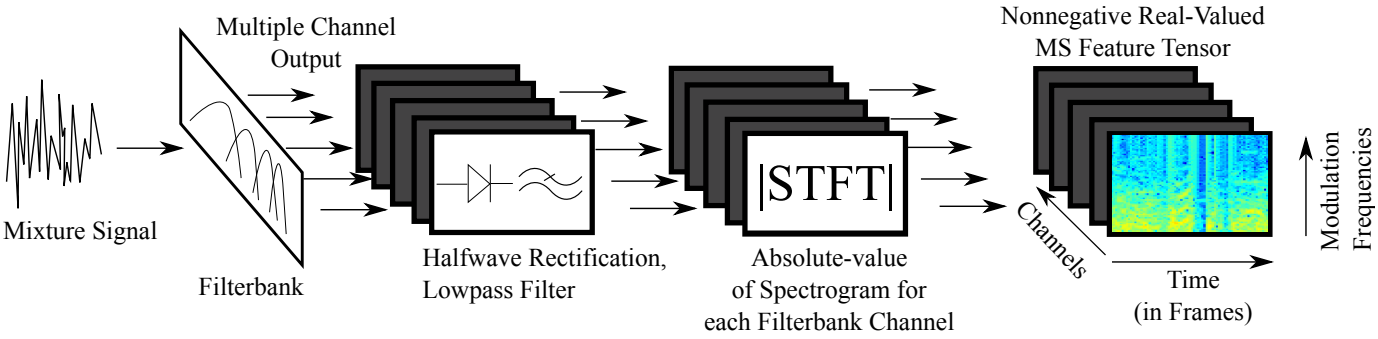

Fig. 1: Diagrammatic representation of modulation spectrogram feature tensor production from a time-domain audio mixture signal.

observations available than sources to be separated. Although less constrained in terms of requirement for *a priori* knowledge, blind separation does not suffer from over-fitting of training data, and is therefore useful as a general approach. It is with this in mind that we consider the challenging problem of single-channel blind separation of naturally occurring everyday sounds, and present our approach which relies only on the underlying sources having internal harmonicity, a common feature of sounds produced via natural physical processes.

The modulation spectrogram (MS) representation was proposed in [18], where it is argued that such a representation is somewhat analogous to that encoded by the human auditory system, and as such is robust to rapid temporal variations caused by effects such as reverberation. MS features have been successfully employed in automatic speech recognition (ASR) systems as described in [18], [19], [20], [21], [22] and in speech emotion recognition in [23]. Unlike in separation, signal reconstruction is not required for recognition uses. Reconstruction from the modulation domain is non-trivial, so introduces an additional challenge to source separation from modulation-based representations.

Mixture signals in the MS domain are represented as a 3-dimensional tensor. Nonnegative tensor factorisation (NTF) has been used previously to separate multichannel audio mixtures via decomposition in [24], [25], but until recently the application of NTF to single channel audio separation has not been widespread. The first uses of NTF for single-channel source separation were in [26], which this paper is a direct extension of, and [27]. Additionally, separation of unison musical sounds based on tensor factorisations of modulation patterns is presented in [28], whilst a complex-valued tensor factorisation for speech enhancement is shown in [29].

Unlike most of the compositional models that use a time-frequency representation, our sound-source separation approach is based on the decomposition of a modulation spectrogram (MS) representation. Such a representation captures the intrinsic redundancy in harmonic and modulation structure across frequency sub-bands. By separating signals in the 3-dimensional MS domain using an NTF model, a mixture is reduced to a sum of components. The aim is that each component models the activity of acoustic features grouped based on harmonic similarity.

This paper provides a thorough analysis of our modulation spectrogram based nonnegative tensor factorisation (MS-NTF) algorithm which we originally demonstrated in [26]. We extend this work by providing a set of convolutive update equations for the factorisation of MS tensors, which can provide increased separation performance under certain conditions and demonstrate the effectiveness on various material types. Additionally, we propose a novel reconstruction method, where activations learned with the MS-NTF model are used to initialise a reconstruction of sources from a spectrogram representation.

The structure of the rest of the paper is as follows: Section II introduces the modulation spectrogram and how it is obtained from a time-domain audio signal. In Section III, the tensor factorisation model is presented, alongside extended update rules for obtaining a decomposition which is convolutive in time. Toy separation examples and an analysis of the number of parameters of representations with varying rank are also provided. The novel method for reconstructing sources from factorised modulation spectrograms is presented in Section IV. In Section V, we describe the evaluation approach for the proposed MS-NTF source separation method, and compare its effectiveness to NMF-based separation. We also show the results of the simulation experiments and a discussion of the outcomes. Finally in Section VI we present conclusions and address the implications of the presented algorithm on speech separation.

II. Modulation Spectrogram Representation

In this section we provide an overview of the analysis of the effects and contributions of the various processing steps required to produce the MS domain representation.

The modulation spectrogram is the spectrogram of the low frequency amplitude envelope of the signal present in each MS-channel. We use the term *channel* to denote a certain sub-portion or sub-band of the spectrum. Audio data in the time domain is transformed into the modulation spectrogram domain through the application of the following steps:

1) Passing the signal through a filterbank.
2) Obtaining a modulation envelope for each filterbank channel via halfwave rectification and lowpass filtering.

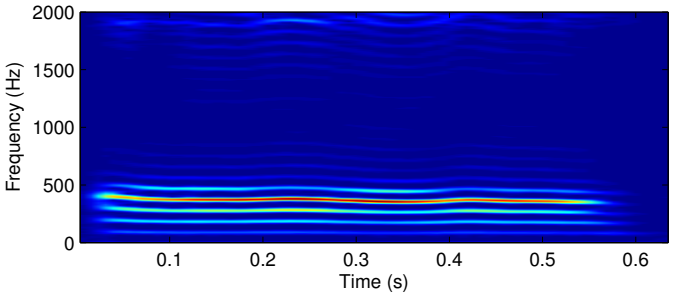

Fig. 2: Spectrogram of a male spoken /e/ sound. Similar frequency modulation is present in each partial.

3) Generation of the spectrogram of each modulation envelope via short-time-Fourier transform (STFT) and taking the absolute value of each bin.
4) Removal of unnecessary frequency bins, for frequencies much higher than the lowpass filter cutoff, to reduce model and factorisation complexity.

This processing (see Figure 1) produces a 3-dimensional data representation, with filterbank channel, STFT bin, and STFT frame being represented across each dimension.

The MS representation of a signal captures the structure present in the low-frequency modulation patterns present across frequency sub-bands, but not rapidly-varying fine temporal structure. Harmonically related sounds such as the partials present in voiced speech, or pitched musical instruments, have similar modulation envelopes within different sub-bands (see [26]), and the MS-NTF separation is able to utilise this by capturing the resulting spectral similarities within each sub-band.

When harmonicity exists within a signal, as is common in speech, for example, the fundamental f_0 generally co-modulates along with the harmonics (Figure 2). Each individual harmonic will have a similar modulation frequency, and therefore envelope. This similarity of envelopes produces similar spectra, whereas the spectral content of each sub-band will only reflect content at in-band frequency bins. This similarity in cross-channel patterns allows the use of a single representative component in the factorisation model. As the activity of a particular source varies, the cross channel gains for a harmonic relationship stay constant, but will co-modulate over time. The application of half wave rectification (HWR) and lowpass filtering captures the low-frequency modulating envelopes of the signal in each channel. The spectral shape of these exhibits more similarity than direct filterbank channel outputs (Figure 3).

Rectification of a narrowband signal such as produced by a bandpass filter, introduces spectral components centred at 0 Hz. An approximation to the power spectral density (PSD) $\Psi_y(f)$ of the output $y(t)$ of the HWR operation applied to a signal $x(t)$ with zero-mean has been shown in [30] to be:

$$\Psi_y(f) \approx \frac{\sigma_x^2}{2\pi}\delta(f)+\frac{1}{4}\Psi_x(f)+\frac{1}{4\pi\sigma_x^2}\int_{-\infty}^{\infty}\Psi_x(f')\Psi_x(f-f')df' \quad (1)$$

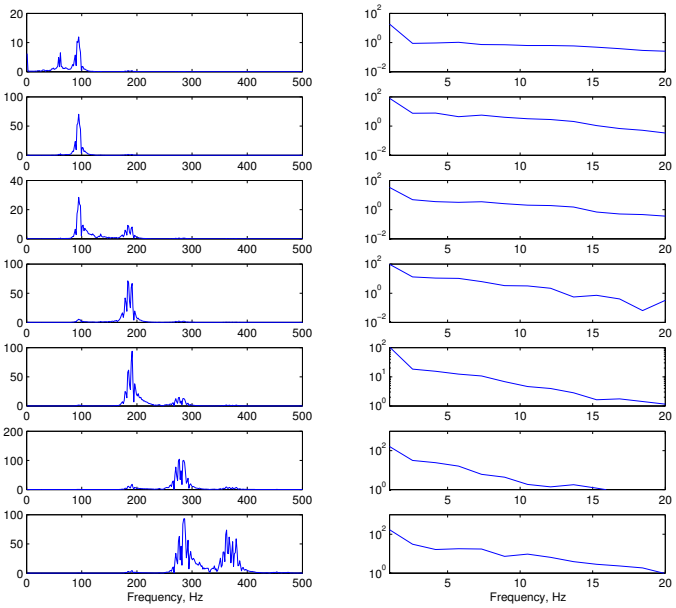

Fig. 3: Lowest 7 channels of magnitude spectra of filterbank outputs for a spoken /e/ vowel sound. Left column, prior to rectification and lowpass-filtering, right column, as modulation envelope spectra (log amplitude for clarity).

where σ_x^2 is the variance of the signal, $\Psi_x(f)$ is the input PSD and and $\delta(f)$ denotes a unit impulse function. As in [31], we equally consider the output of the gammatone filter as approximately a narrowband signal with bandwidth B, centred at f_c. The rectification of a signal with such a power spectrum additionally produces an amplitude scaled DC-component equivalent to the autoconvolution of the original power spectrum (see the third term in Equation 1), as well as reduced-amplitude versions of the DC-term at multiples of f_c (Figure 5). Lowpass filtering can be used to remove the original spectrum and higher frequency terms leaving only the signal centred around DC. Considering a single filterbank channel in our MS model as an approximation to the narrowband filter described in [31], similarities in spectral modulations across channels then begin to become apparent as a result of the HWR operation. Where the shape of the PSD within a particular band is similar to those in other bands, (e.g. as with the regular spacing of the harmonic peaks in speech or other harmonic sounds), it follows that the result of autoconvolution and shape of spectral patterns present at baseband will be similar.

III. Tensor Factorisation Model

The factorisation model approximates a 3-dimensional tensor as a sum of rank-1 components; this factorisation model [32] is known as the PARAFAC decomposition (also canonical polyadic decomposition (CPD) or CANDECOMP factorisation). Components are learned such that they minimise a divergence cost between the target and estimated components. The 3-dimensional structure ensures that for a single component, there exists similarity of modulation spectra across channels with variation only in activation magnitude. Cross-channel

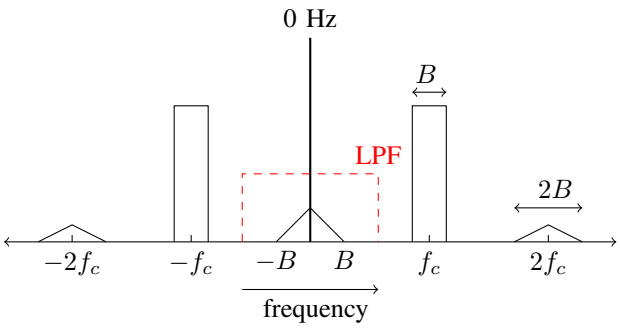

Fig. 4: Power spectrum of a half-wave rectified narrowband signal centred at f_c with bandwidth B. Dashed line 'LPF' shows how the use of a lowpass filter can be used to consider only the portion of the spectrum centred at 0Hz, as in the modulation envelope representation. Modified from [31] and based on [30].

similarity existing in simple signals in the MS-domain can therefore be efficiently encoded by a single component within the tensor model.

A. Factorisation Model

The 3-dimensional tensor representing the MS has dimensions of size of number of filterbank channels, DFT samples, and observation frames. This mixture tensor is denoted as $\mathcal{X}$, and the factors which approximate this are stored in matrices $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$. The outer product of each column in the matrices form the components which sum to form $\hat{\mathcal{X}}$, the approximation of $\mathcal{X}$.

The model $\hat{\mathcal{X}}$ is described by:

$$\mathcal{X}_{r,n,m} \approx \hat{\mathcal{X}}_{r,n,m} = \sum_{k=1}^{K} \mathbf{G}_{r,k}\mathbf{A}_{n,k}\mathbf{S}_{m,k} \tag{2}$$

where $\mathbf{G}^{R\times K}$ (size $R \times K$) is a matrix containing the auditory channel dependent gain, $\mathbf{A}^{N\times K}$ the frequency basis functions which model the spectral content of a modulation envelope feature, and $\mathbf{S}^{M\times K}$ is the time-varying activation of the component. Subscripts r, n, m are the channel, modulation spectral bin, and time frame indices, respectively, whilst k denotes the index of a particular component. The model therefore essentially describes each component's fixed modulation spectrum existing at different levels across channels, being activated at various points in time.

The model parameters contained in $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$ are estimated by minimising the generalised Kullback-Leibler (KL) divergence between $\mathcal{X}$ and $\hat{\mathcal{X}}$, notated D,

$$D(\mathcal{X}\|\hat{\mathcal{X}}) = \sum_{r,n,m} \mathcal{X}_{r,n,m} \log \frac{\mathcal{X}_{r,n,m}}{\hat{\mathcal{X}}_{r,n,m}} - \mathcal{X}_{r,n,m} + \hat{\mathcal{X}}_{r,n,m}. \tag{3}$$

KL divergence is widely used to estimate the components in source separation by nonnegative matrix and tensor factorisation [11], and is more sensitive to low-energy observations than Euclidean distance, an alternative measure of reconstruction error proposed in [33].

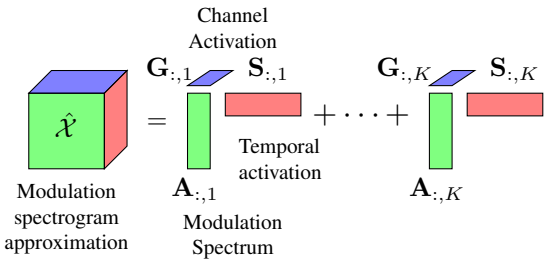

Fig. 5: An approximation to $\mathcal{X}$ the mixture tensor, $\hat{\mathcal{X}}$ is formed by the sum of outer products between rank-one tensors. Each rank-one tensor is a column of the component matrices $\mathbf{G}$, $\mathbf{A}$, and $\mathbf{S}$ and represents a different component in the separation. Update equations aim to minimise the divergence between $\mathcal{X}$ and $\hat{\mathcal{X}}$.

The divergence D can be minimised by applying update rules to $\mathbf{G}$, $\mathbf{A}$, and $\mathbf{S}$ which iteratively perform gradient descent with respect to each variable. The specific update rules given in this paper are derived in [24] and [34] although generalised multi-dimensional PARAFAC type updates such as presented in [35] can be applied, where the tensor is unfolded into a product of matrices and then updated via NMF matrix update rules.

The tensor factorisation algorithm applied is carried out as follows:

1) Generate modulation spectrogram tensor to be decomposed, $\mathcal{X}$.
2) Initialise matrices $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$ with random non-negative values. Matrix dimensions are defined by the corresponding dimensions of $\mathcal{X}$, and the number of components into which $\mathcal{X}$ should be decomposed.
3) Apply update rules to minimise the divergence between the sum of factors in $\mathbf{G}$, $\mathbf{A}$ and $\mathbf{S}$ and the tensor which they model.

The update rules applied in stage 3 of the algorithm are:

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

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

$$\mathbf{S}_{m,k} \leftarrow \mathbf{S}_{m,k} \frac{\sum_{r,n} \mathcal{C}_{r,n,m}\mathbf{G}_{r,k}\mathbf{A}_{n,k}}{\sum_{r,n} \mathbf{G}_{r,k}\mathbf{A}_{n,k}} \tag{6}$$

where $\mathcal{C} = \mathcal{X}/\hat{\mathcal{X}}$ elementwise and is recalculated after application of each update equation.

The update rules guarantee a reduction of the cost value, D, but do not ensure that the global minimum is reached. The update rules are applied until there is no longer significant reduction in D.

B. MS-NTD Model

Here we present a convolutive extension to the basic NTF-factorisation. By use of the convolutive factorisation, recurrent patterns across time or channel can be modelled within a single

factorisation component. We term this process modulation spectrogram nonnegative tensor deconvolution, or MS-NTD.

The use of a convolutive model is motivated by the assumption that a recurrent pattern present within a source may span more than a single time-frame or frequency channel. A convolutive factorisation model is able to represent such structure. In this way, a single component is able to represent more complex redundant structures than the non-convolutional case, and the lowest frequency changes which can be represented is covered by the context across multiple frames, rather than within a single frame.

Convolutive extensions to the basic NTF algorithm can span both/either time and/or frequency dimensions; we performed initial tests of separation performance with components which learn shifts over both channels and time. Temporal shifts produced most promising initial separation performance, and are also somewhat more intuitive in their data representation. For this reason we use and explain the model for shifts over time, although other cases can be covered by permuting the time and channel dimensions in the presented equations.

For spectral convolution over time, the basis functions containing spectra are estimated as a matrix, by summation over all convolutional time shifts. The algorithm is different compared to that presented in Section III-A in that the K spectral basis vectors are modified to become spectral basis matrices, and so increase their dimensionality.

The convolutive extension to the NTF factorisation model minimises the KL-divergence between the 3-dimensional MS tensor $\mathcal{X}$ and a linear combination of approximated factors, $\mathbf{G}'$ $\mathbf{A}'$, $\mathbf{S}'$ which form the approxiative model $\hat{\mathcal{X}}'$:

$$\hat{\mathcal{X}}'_{r,n,m} = \sum_{k=1}^{K} \sum_{d=0}^{D} \mathbf{G}'_{r,k} \mathbf{A}'_{n-d,k} \mathbf{S}'_{m,k,d}. \tag{7}$$

Update rules for a convolutive model with a maximum time shift of D frames are given as:

$$\mathbf{G}'_{r,k} \leftarrow \mathbf{G}'_{r,k} \frac{\sum_{n,m,d} \mathcal{C}'_{r,n,m} \mathbf{A}'_{(n-d),k} \mathbf{S}'_{m,k,d}}{\sum_{n,m,d} \mathbf{A}'_{(n-d),k} \mathbf{S}'_{m,k,d}} \tag{8}$$

$$\mathbf{A}'_{n,k} \leftarrow \mathbf{A}'_{n,k} \frac{\sum_{d,r,m} \mathcal{C}'_{r,(n+d),m} \mathbf{G}'_{r,k} \mathbf{S}'_{m,k,d}}{\sum_{d,r,m} \mathbf{G}'_{r,k} \mathbf{S}'_{m,k,d}} \tag{9}$$

$$\mathbf{S}'_{m,k,d} \leftarrow \mathbf{S}'_{m,k,d} \frac{\sum_{r,n} \mathcal{C}'_{r,n,m} \mathbf{G}'_{r,k} \mathbf{A}'_{(n-d),k}}{\sum_{r,n} \mathbf{G}'_{r,k} \mathbf{A}'_{(n-d),k}} \tag{10}$$

where $\mathcal{C}' = \mathcal{X}/\hat{\mathcal{X}}'$ element-wise, recalculated after each application of update equations.

C. Simulation Examples

In this section we provide an example to show how the MS-NTF factorisation is able to learn meaningful structure more effectively than NMF. In cases where the structure of individual sources in both time and frequency is well represented, good separation can be achieved. We illustrate the structure learned in matrix and tensor factorisation cases, and demonstrate via a toy example that it is the combination of the tensor model alongside the MS representation which is able to separate components.

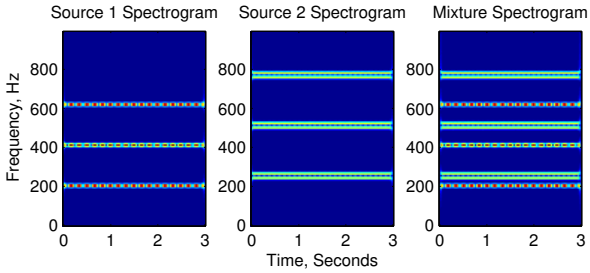

Fig. 6: Spectrograms for individual toy example sources and mixture.

Factors are learned by minimisation of a divergence function. We can evaluate the accuracy of the learned factors by comparing them with the *oracle factors*. Oracle factors are produced by rank-1 factorisation of the unmixed individual sources present in a simple mixture signal and allow us to gain intuition about the basic structure present in a signal. Inspection of the learned components relative to the oracle allows us to compare how each model captures source structure. The factors producing minimised divergence for a mixture approximation will not necessarily reflect the structure of individual sources, but in this toy example the NMF-derived factors show less similarity with the structure of each individual source than the NTF-derived factors.

Factorisation of simultaneous signals: Here we inspect the structure obtained by factorisation of two differently modulated tones. Consider the synthetic signal with the mixture spectrogram shown in Figure 6. Each source in the mixture is a 3-partial harmonic, amplitude modulated at either 3 Hz or 11 Hz.

Source 1 has an f_0 of 207 Hz and is modulated at a rate of 3 Hz, modulation depth 0.7. Source 2 has an f_0 of 257 Hz and is modulated at a rate of 11 Hz, modulation depth 0.7. The mixture is created by summing the time domain source 1 and source 2 signals.

We factorise the mixture into 2 factors in both the 2-dimensional spectrogram representation (NMF), and the 3-dimensional MS domain as well as a matrix factorisation of the unfolded MS mixture tensor. Unfolding, or tensor matricization (see [35]) is performed over the channel dimension, so that the tensor of dimensions $R \times N \times M$ becomes a matrix of size $(R \times N) \times M$.

Figure 7 shows components learned with the NTF model whilst Figure 8 shows the factors learned in the NMF separation. Figure 9 and 10 show the factors learned with the matrix factorisation of the MS tensor unfolded over the channel dimension.

The spectral basis functions obtained with NMF have significant contribution bleed from the interfering source, and components are not well separated from one another. The NTF model better learns the distinct components comparable with the oracle factors in this example, and peaks in the channel activation dimension are learned at the same location as in the oracle examples. It could also be argued that there is

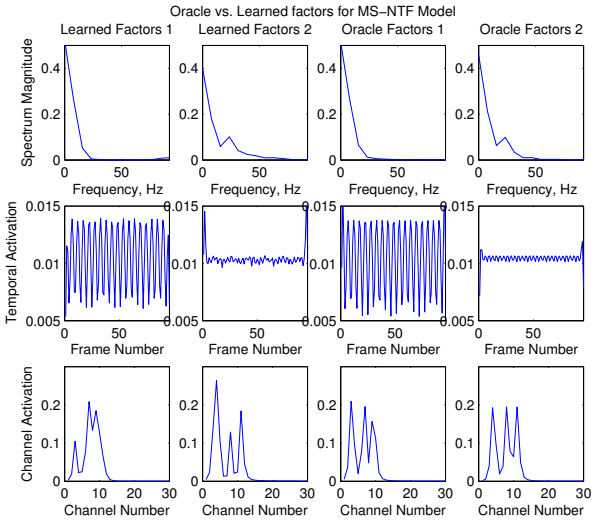

Fig. 7: NTF-derived mixture factors compared to oracle MS-NTF factors derived from constituent source modulation spectrograms.

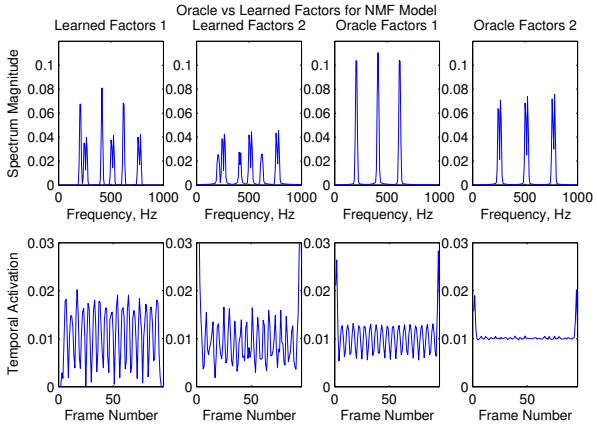

Fig. 8: NMF-derived mixture factors compared to oracle factors derived from constituent source spectrograms.

greater similarity of time activations. The source interference apparent with the NMF applied to the MS demonstrates that it is the combination of the tensor factorisation with the representation which make the proposed method effective at separating sources.

D. Model Complexity

The MS-NTD model is able to approximate much of the energy in the mixture representation using relatively fewer parameters than other approaches. Fewer parameters means less chance of over-fitting in production of the separated components, resulting in a more meaningful source separation. We can compare and describe the number of parameters in different factorisation approaches, for factorisation rank K. As rank increases, it should be expected that a better approximation to the mixture can be achieved.

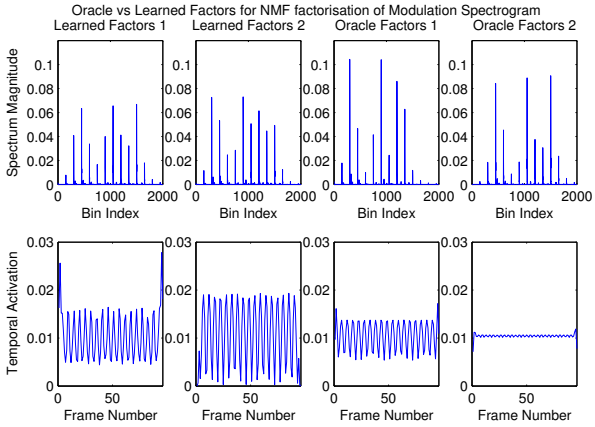

Fig. 9: Factors obtained with NMF applied to unfolded MS matrix representation, compared to oracle unfolded MS matrices. Spectral bins truncated for clarity (from 4500 bins) since very little energy is present within the bins relating to higher channels.

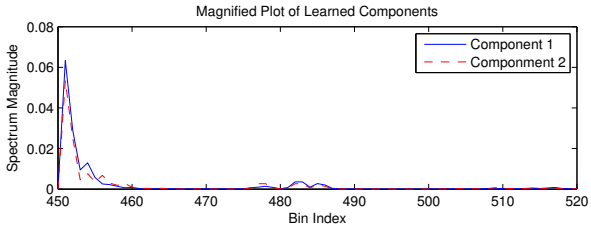

Fig. 10: Clarified view of portion of factors obtained through matrix factorisation of modulation spectrogram. Components have similar and overlapping shapes, resulting in poor separation.

In an NMF spectrogram factorisation, the number of entries in factorisation matrices, and hence parameters is $K \times (P + M)$. For the MS-NTF model (referring to dimension definitions in Section III-A), we have $K \times (R + N + M)$ parameters. If the MS is unfolded over frequency channels and factorised as a matrix, we introduce many more degrees of freedom in the spectral dimension, requiring $K \times (R \times N + M)$ parameters. Where the NTD model is used, for a shift of D frames, $K \times ((D \times N) + R + M)$ parameters are needed.

Since N is the length of a truncated spectrum based on the lowpass frequency used in producing the MS, in practice $R + N < P$ resulting in many fewer parameters in MS-NTF than NMF for equivalent factorisation rank.

In Figure 11 we show the normalised residual power calculated from subtraction of the factorisation approximation from the target in different factorisation models and summation over all dimensions. Normalisation was carried out by dividing the power (absolute value squared) of the residual by the initial power present in the representation. Values were calculated with $R = 30, N = 64, M = 256, P = 1024$.

The results of this experiment demonstrate the ability of the MS-NTF model to represent a signal more compactly, by taking advantage of redundancies. Even the convolutive factorisations, spanning several frames have fewer parameters

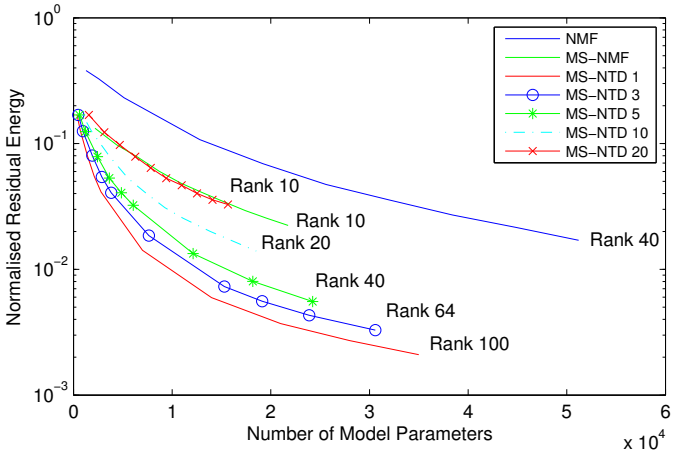

Fig. 11: Average residual energy present after factorisation of signals with 3 different approaches, NMF, MS-NMF and MS-NTD plotted against number of model parameters. MS-NTD is shown with varying shift lengths. For a given number of parameters, the proposed MS-NTD model has lower error in the approximation. For equivalent factorisation rank, the MS-NTD model has fewer parameters. Increasing convolution length within the MS-NTD approach increases number of parameters for a given rank but produces increased residual energy for a given number of parameters. Results shown averaged over 50 speech mixtures as used in later evaluation.

than the single frame NMF-based models. A compact representation does not necessarily ensure good separation capability though, however we address the separation performance of such models in more detailed evaluations in Section V.

IV. Source Reconstruction

Reconstruction of audio from the modulation spectrogram is an inherently challenging problem, due the MS not being directly invertible. The filterbank (FB) stage can be inverted if an appropriate function is used (an oversampled analysis FB allows perfect reconstruction with the correct synthesis reconstruction FB [36]). Lowpass-filtering discards high-frequency information however, which is difficult to recover, as does the non-linearity resulting from halfwave rectification, and taking the absolute value of STFT frames. Inversion of modulation envelopes (not spectra) has been addressed in [37] via efficient optimisation of a cost function. Such an approach assumes that the signal-representation for inversion was derived from a real signal rather than being estimated from a mixture signal. Inversion of arbitrary signals such as those derived from estimated separation may not produce meaningful time-domain waveforms though. Informal testing of such an approach produced worse separation performance than our existing and proposed methods for sources reconstructed from estimations obtained via factorisation and so was not explored further.

In [26] we presented a method for source synthesis based on the activations learned in the NTF model. Using learned temporal activation, full bandwidth basis functions were obtained through factorisation of a reconstruction tensor. In this work, we propose a new method for reconstruction of sources separated in the modulation spectrogram domain. A similar approach of maintaining initial source activation values is used, but instead of factorisation of a 3-dimensional MS-derived tensor, a less-complex data representation based on a simple spectrogram is used in the second stage reconstruction factorisation. The use of this 2-dimensional spectrogram allows for less computation and a more intuitive method. The new approach also seems to produce better source-to-distortion values for reconstructed sources compared with the approach in [26] (see Section V-B).

Keeping the time-varying activations obtained during the MS-NTF stage fixed, a matrix factorisation is subsequently used to produce spectra bases to approximate a reconstruction matrix. The reconstruction matrix, $\mathbf{V}$ is produced by taking the magnitude spectrogram of the time-domain mixture signal. $\mathbf{V}$ is subsequently decomposed into approximative factors in $\mathbf{B}$ which are estimated using fixed activations $\mathbf{A}$, from the initial MS-NTF and MS-NTD model factorisations.

Matrix $\mathbf{B}$ contains factors which produce minimal KL-divergence for a given set of activations and the structure of these will vary depending on the structure of sources within the mixture. Where source spectra have structure which is inherently low rank e.g. for harmonic sounds such as the example shown in Figure 6, $\mathbf{B}$ is able to learn components which have frequency content at those bins present in the sources. Where a low-rank representation can not accurately model the sources, such as with speech, the components in $\mathbf{B}$ just represent the bins with most activity for that source estimate.

A. Non-convolutive reconstruction

In the non-convolutive case, reconstruction is performed using NMF update rules. Spectral bases in the matrix $\mathbf{B}$ are estimated according to the model

$$\mathbf{V} \approx \hat{\mathbf{V}} = \mathbf{B}\mathbf{A}^{\top} \tag{11}$$

by minimising the KL-divergence

$$KL(\mathbf{V}|\hat{\mathbf{V}}) = ||\mathbf{V} \otimes \log \frac{\mathbf{V}}{\hat{\mathbf{V}}} - \mathbf{V} + \hat{\mathbf{V}}|| \tag{12}$$

via NMF updates of $\mathbf{B}$ for the fixed activations in $\mathbf{A}$ learned during initial factorisation. Wiener filters are derived from factors which minimise Equation (12). These filters are applied to the mixture spectrogram as in [38], before inversion to obtain time-domain waveforms.

B. Convolutive reconstruction

Reconstruction of mixtures factorised with MS-NTD, makes use of the convolutive NMF model, as in [12] but updates only the spectra, forming an approximation to the reconstruction matrix $\hat{\mathbf{V}}'$:

$$\hat{\mathbf{V}}' = \sum_{d=0}^{D-1} \mathbf{B}_d \overset{d\rightarrow}{\mathbf{H}} \tag{13}$$

where we use activations obtained from MS-NTD,

$$\mathbf{H} = \mathbf{A}'^{\top}, \tag{14}$$

and $d \rightarrow$ is a non-circular shift of the matrix d columns to the right. Where $D = 1$ the model reduces to the non-convolutive case.

To minimise $KL(\mathbf{V}|\hat{\mathbf{V}}')$ as in Eq. 12, $\mathbf{B}$ is updated via:

$$\mathbf{B}_d \leftarrow \mathbf{B}_d \otimes \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}'} \overset{d\rightarrow}{\mathbf{H}}^\top}{\mathbf{1} \overset{d\rightarrow}{\mathbf{H}}^\top} \tag{15}$$

as in [12].

Following convergence of the cost function, each of the K sources are reconstructed by generating a Wiener filter soft-mask from each base at the k'th index. Filters are applied to the complex mixture spectrogram, $\mathbf{Y}$ so that source k's spectrogram is:

$$Source_k = \mathbf{Y} \otimes \frac{\sum_{d=0}^{D-1} \mathbf{B}_{:,k,d} \overset{d\rightarrow}{\mathbf{H}}_{k,:}}{\sum_{k=1}^{K} \sum_{d=0}^{D-1} \mathbf{B}_{:,k,d} \overset{d\rightarrow}{\mathbf{H}}_{k,:}} \tag{16}$$

where $\mathbf{Y}$ has a frequency resolution defined by the analysis frame length.

Time domain reconstruction for each source is performed by inversion of the resulting spectrogram via the inverse DFT of each frame followed by the overlap-add operation.

V. Simulation Experiments

We compare our blind single-channel MS-NTF approach to blind single-channel NMF, in both non-convolutive and convolutive implementations. The separation performance of the methods is demonstrated on 4 classes of mixture signal, each containing two sources, which are common in everyday life and often used in source separation evaluations. The four mixture classes we evaluate on are *Speech-Speech*, *Speech-Musical Instrument*, *Speech-Noise* and *Music-Music* mixtures.

Speech-Speech mixtures provide a challenging separation task, since the properties of each source tend to be more similar to each other. The musical-instrument mixtures generally contain highly harmonic content (although unpitched percussive test material is also part of the evaluation). Where single musical notes are present for each source, the underlying structure is not complicated and lends itself well to low-rank models. Speech-noise mixtures are a common separation task, for which unsupervised separation approaches are highly appropriate due to the non-deterministic nature of real world noise.

A. Test material

For each class of mixture, 500 test examples were created. Speech-speech mixtures were generated by summing a single utterance from each of two different randomly selected speakers from the CMU-Arctic database [39]. Speech-noise test mixtures were generated again using speech and noise mixtures from a single microphone channel in the CHiME3 database [40]. For each mixture, a noise type was selected at random from CHiME3 and a 3-second section was summed with a 3-second speech segment from a randomly-selected talker. Speech-music mixtures were generated with a randomly-chosen 3-second speech sample from the CHiME3 database summed with a randomly selected 3-second monophonic sample of different musical instruments from the RWC musical instrument database [41]. Music-music mixtures were generated by summing two randomly-selected 3-second monophonic samples from the same RWC database. The fixed 3-second length across all mixtures allows for meaningful comparison of algorithm performance on each mixture type. Sources were RMS normalised prior to mixing so that each source contributed equal power to the mixture. Test mixtures were re-sampled to 16 kHz in cases where original material was at a different samplerate.

B. Evaluating separation performance

The proposed convolutive MS-NTD method was used to separate the test mixtures and the results were compared with those produced using unsupervised convolutive NMF [12]. For the case of a single convolutive frame shift, the model is equivalent to MS-NTF. For MS-NTD, two reconstruction methods were preliminarily tested; both the novel reconstruction method (with respect to modulation spectrogram based source separation) described in Section IV and the method in [26], modified to make use of the convolutive update rules in Section III-B. Following the results of these tests, the novel method was considered to produce better performance and so used in all further evaluations. In all experiments, test-mixtures were separated directly into 2 components. In [26], the blind 2-factor separation cases outperformed naive clustering approaches using more components prior source assignment. This additionally detaches the effect of clustering algorithms from any analysis, and allows comparison of solely a method's separation performance for simple additive mixtures.

To determine performance we computationally assess the separation for a large number of mixtures. Separation was evaluated according to widely-used metrics from the BSS and PEASS toolkits [42], [43] which provides objective measurements for source separation quality. Source-to-distortion ratio (SDR) is a measurement of energy contributions from the desired source compared to unwanted energy from interference, noise, and artefacts and so is a good and widely used evaluation of separation quality. A high SDR could be expected to lead to good enhancement results in a computational speech-recognition test, for example. Since the separated sources are also often used in human evaluations, their subjective quality should also be considered. A lot of energy in a low frequency region may not be highly-audible to a human listener, but may have a large effect on SDR ratings. For this reason, it is also beneficial to consider perceptual separation metrics. Interference-related perceptual score (IPS) is a measure from the PEASS toolkit, where an overall score is calculated based on a model created by the toolkit's authors and obtained from listening test ratings. We considered this the most appropriate PEASS metric in terms of quantifying source separation algorithm performance, although other PEASS measures were also calculated and displayed similar general trends.

It should be stated that it can be problematic to measure meaningful separation performance of truly 'blind' separation

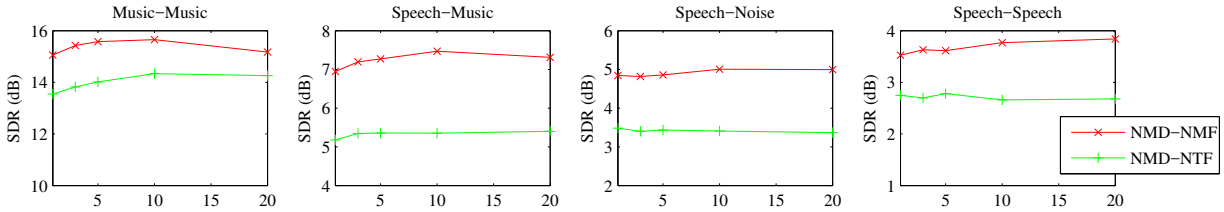

Fig. 12: Source-distortion-ratio performance with reconstruction based on Wiener-like filters derived from NMF of a reconstruction spectrogram-matrix (NMD-NMF, proposed method) vs. from a reconstruction tensor (NMD-NTF, from [26]). Analysis window 1024 samples (64 ms). Subplots show results for different material type averaged over 500 test mixtures.

approaches. Since in practice the sources are not defined, evaluation procedures are inevitably constrained to a particular type of material, which may not describe the performance on other types of source. Even in these so-called blind separation cases then, some assumption tends to be made about the mixture to be separated. For example, that the mixture contains speech or that sources are harmonic, or will have a certain level of statistical independence. We attempt to give an accurate description of comparative real-world source separation performance using the stated metrics.

C. NTF Reconstruction Results

Figure 12 shows the SDR performance of sources reconstructed with each method across various material types and convolution lengths, averaged over 500 test mixtures. Performance with a 1024 sample (64 ms) analysis window is shown, but a similar result was obtained for window lengths of 512, 2048 and 4096 samples. Superior SDR values are obtained using the MS-NTD derived activations within a reconstruction matrix spectrogram as proposed in Section IV as opposed to the use of a reconstruction tensor in [26]. The proposed method also provided higher perceptual IPS scores across all window lengths and material types. For all subsequent evaluation, sources separated with the MS-NTD model are reconstructed using the method in Section IV.

D. Algorithm Parameters

The choice of parameters, such as window length and function for generating representations for non-negative decomposition will clearly have an effect on separation performance. Depending on the specifics of a particular mixture signal, one particular analysis function may outperform another.

We perform our experiments using a range of parameters, although exhaustive trials of all implementation variations are impossible. We present the results with the aim of using the MS-NTD approach as a general separation approach, and attempt to provide intuitions and explanations about how and why parameter variation influences separation performance.

1) Window size: We evaluate our approach with analysis window sizes of 32 ms, 64 ms, 128 ms and 256 ms (512, 1024, 2048, 4096 samples). NMF-based methods typically use analysis frames in the order of 30-100 ms [44], and previous work [6], [44], [45] has shown that this range works well in both the NMF and MS-NTF algorithms. The window length limits the minimum *within window* frequency which can be meaningfully represented, according to the relationship $f_{min} = 1/T$. The minimum frequency within a 32 ms window is 31.25 Hz whilst for a 256 ms window is 3.91 Hz. However, low frequency temporal structure variation information can still be encoded by such an approach, as the sliding window analysis allows the convolutive factorisation model to represent changes spanning multiple overlapping frames.

2) Hop size: In conjunction with window size, analysis-hop size will affect the temporal context represented by a single component in the convolutive implementation. a 20-frame convolution with a short hop might represent less context than a 10-frame convolution at a longer hop length. For all frame lengths, we evaluated hop sizes of 64, 128 and 256 frames, as well as hop sizes relative to window length, by using a hop of 50% of window length.

3) Filterbank choice: An FIR gammatone filterbank with 30 channels of equivalent rectangular bandwidth (ERB) [46] was used as the analysis filterbank in the creation of MS-domain mixtures, and was implemented with the LTFAT toolbox [47]. We do not make the assertion that a gammatone filterbank will produce the absolute best performance, however this filterbank has some properties (as do others) which produce useful structure in the production of the MS. Its extensive use in auditory modelling, for example in $F0$ estimation [48], influence our use of such a filterbank here however. As Bregman points out in [49], the ability to estimate $F0$ in the presence of other sounds means the correct assignment of spectral components to sound sources, and gammatone-based methods have been successful in achieving this.

Increasing bandwidth with centre frequency means that multiple harmonics can be covered in a single band even as frequency increases. Overlapping filters provide mutual information across channels, which aid in a single component representing redundant information across channels in the factorisation.

An insight into the effects of various filterbank parameters can be observed in Figure 13, where the results of preliminary performance tests are shown. We compare SDR and IPS for separated sources with a variety of filterbanks in the generation of the MS tensor. The number of channels in a gammatone filterbank is varied, and the effect on separation performance shown. Also, a different filterbank spacing, constant-Q transform (CQT) spacing is compared. There is less overlap between channels with this filterbank.

From these initial results, it can be seen that there is a performance disadvantage to using CQT filters,

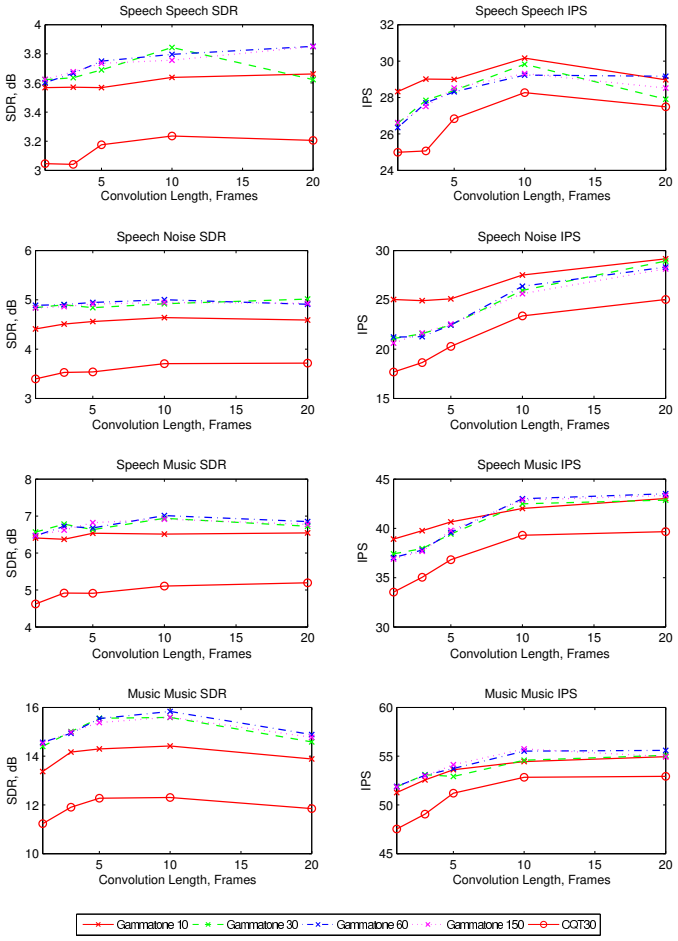

Fig. 13: Separation performance for a fixed window and hop size, and different MS filterbank functions. Filter type and number of channels shown in legend. Note different y-axis scale across plots.

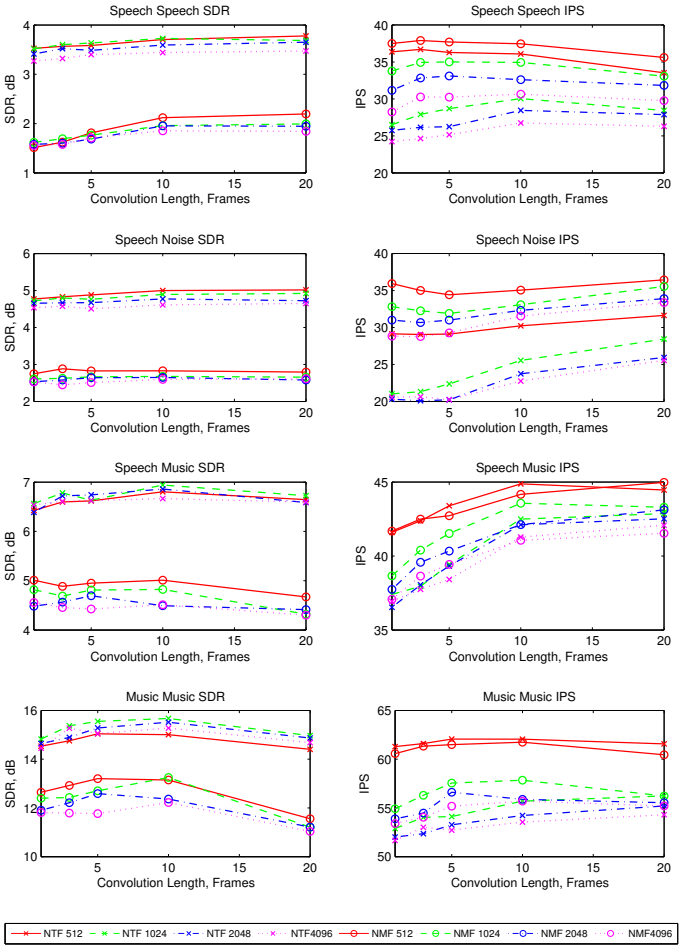

Fig. 14: Hop length 256 separation source-to-distortion-ratio (SDR) and interfering source perceptual suppression (IPS) for convolutional modulation spectrogram NTF (MS-NTD) and NMF (NMD) for different material types. Different analysis window lengths are compared. Note different y-axis scale across plots.

4) Truncation length: The MS signals are lowpass filtered at a fixed frequency during their generation. With different analysis frame lengths, the DFT bin relating to cutoff frequency varies. The truncation length can be changed accordingly. We vary truncation length with frame size to remove information above a fixed frequency, and truncate at $1/16^{\text{th}}$ of frame size.

5) Convolution Length: Convolution lengths of 1, 3, 5, 10, 20 frames were used in the MS-NTD factorisation. This, in combination with the hop size, determines how much context (and resulting variation) is captured in a single component.

E. Results

Separation results for each mixture type are presented in Figures 14 and 15. Figure 14 shows results with a fixed hop size of 256 samples, whilst Figure 15 shows results with a hop size proportional to the analysis window length at 50% overlap and allows comparison for larger hop sizes. Hops of 64 and 128 frames (across all analysis frame lengths) were also tested but on average produced inferior performance compared with a 256 frame hop, so are not shown here.

The MS-NTD separation approach gives consistently higher separation performance than NMF in terms of SDR for all analysis window lengths and across all material types. For the proposed MS-based representation, convolutive factorisation (MS-NTD) increases SDR performance over non-convolutive (MS-NTF) for at least one convolution length in each case when a 256 frame hop is used. However, for longer analysis frames, as with a 50% hop for frame lengths $\geq$ 1024, convolutive shifts tend to reduce separation SDR. In these cases, the overall context time covered by multiple frames is enough that a single component can not properly model the changes present.

For MS-NTD, a window length of 1024 samples produces the best within-method separation quality, for all material types except speech-noise mixtures, where a window of 512 samples produces better separation.

Although the plotted results show a difference in mean separation performance, the statistical significance of differences between mean separation across methods should also

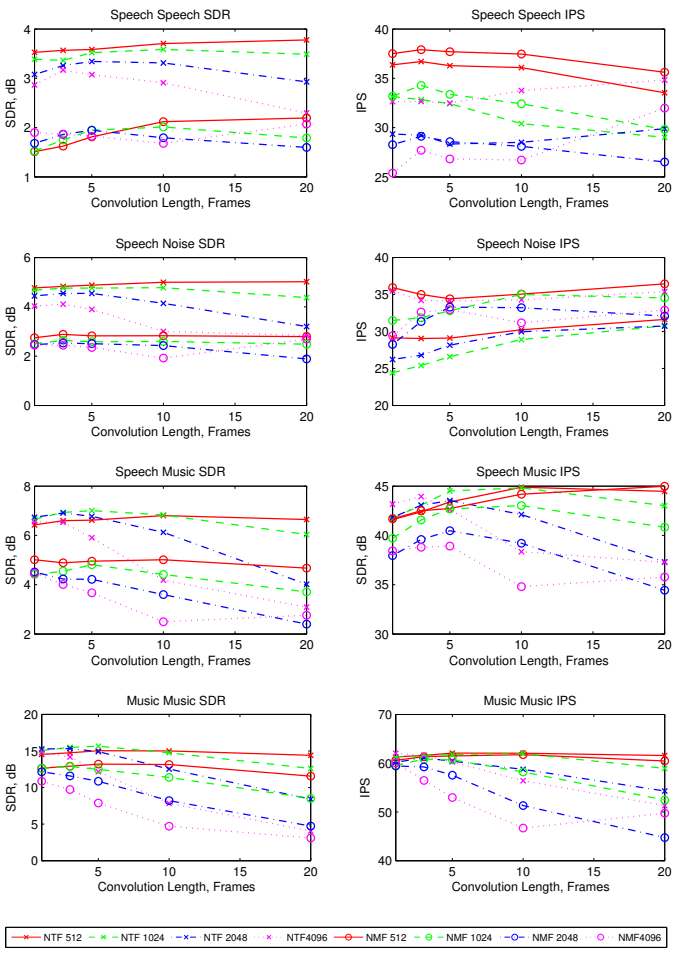

Fig. 15: HOP 50% Separation source-to-distortion-ratio (SDR) and interfering source perceptual suppression (IPS) for convolutional modulation spectrogram NTF (MS-NTD) and NMF (NMD) for different material types. Different analysis window lengths are compared. Note different y-axis scale across plots.

be considered. A paired t-test was used to determine whether the differences in mean performance measured over 500 test mixture samples was statistically significant. Significance was assessed across separation method (MS-NTD vs. NMF) and in terms of effect from convolution length. For SDR, for each window length and convolution length, the mean improvement observed with MS-NTD over NMF was highly statistically significant $(P < 0.001)$. Within the MS-NTD results, the statistical significance of improvement using convolutional factorisation compared to single-frame factorisation was also tested. For the best performing window length of each material type, the key results and statistical significance can be seen in Table I. Such results validate the use of the convolutional MS-NTD model over MS-NTF.

Within the perceptual separation metrics, neither the non-negative matrix deconvolution (NMD) or MS-NTD is consistently producing better performance. Superiority of one method over another depends more on material type. Mixtures containing speech have a higher mean IPS score for NMD separation than the proposed MS-NTD. For mixtures containing

TABLE I: Statistical significance of SDR improvements achieved with convolutive factorisation model and 256 sample hop

Material+ Window Length	NTF SDR Mean (dB)	Convolution Length (Frames)	NTD SDR Mean (dB)	P-Value
Speech-Speech 512	3.52	20	3.78	1.5e-07
Speech-Noise 512	4.77	20	5.01	8.20e-07
Speech-Music 1024	6.56	10	6.94	0.0011
Music-Music 1024	14.82	10	15.67	2.9e-05

music, for similar analysis window lengths we observe similar performance across both methods. A window of 512 produces the best IPS scores across all material types.

F. Discussion

There is a clear variation in performance for different types of material. The differences are likely due to the differences in structural complexity (underlying rank) of each signal type. Representing complex signals accurately using a only a single component will never be totally effective if the inherent rank of an individual signal is much greater than one. This is true regardless of the domain in which signals are represented. This shortcoming can be addressed by factorising mixtures using higher-rank models, but this introduces the need to assign factors to specific sources, a challenging problem in its own right [50], [51].

A large amount of overlap in time-frequency points also makes separation of sources more challenging. For speech-speech mixtures, each source will tend to have greater statistical similarity than other material types since speech tends to occupy specific frequency ranges, whereas noise and music have a much looser expectation in terms of frequency range, so have lower expectation of overlap between sources. In comparison of IPS score with SDR, we notice that for material types which produce higher mean SDR values also produce higher mean IPS.

Generally an improvement in performance with convolutive mixtures could be attributed to a higher number of parameters compared to the single frame factorisation. In the results presented, all frames overlap by 256 samples (16 ms) effective convolution over 10 frames captures temporal variations of the order of 160 ms. For mixtures containing speech, temporal variation is higher than the music-music mixtures, which would explain why the frame length 512 (32 ms) gives better results than longer context. It can also be expected that certain other factorisation constraints which have been shown to help separation performance in NMF-based separation, such as the introduction of enforced sparsity may also improve separation performance.

The described evaluations and comparisons should be considered as measure of each technique's general separation performance but will not ensure superiority in all cases. In

practical applications, one can use the results presented to make informed decisions about implementation parameters of a particular separation approach, based on expected source material.

VI. Conclusion

This paper has presented a sound separation technique based on the factorisation of mixture signals in the modulation spectrogram representation. Non-negative factors are estimated for each source by minimisation of the Kullback-Leibler divergence between factors and a mixture tensor. Through use of iterative update rules, a single component is learned for each source within a mixture, from which individual source estimations can be reconstructed.

We have proposed a convolutive extension to our original MS-NTF algorithm, termed MS-NTD, and shown that it can produce a statistically-significant mean improvement in SDR for separated signals. Furthermore, we presented a novel reconstruction method for audio signals separated using MS-NTD factorisation, which makes use of the estimated source activities in order to learn reconstruction masks in the STFT domain.

Computational tests across many mixtures on various real world mixture types show that the proposed methods outperform spectrogram based NMF, in terms of SDR. For the perceptually-derived IPS metric, NMF produces better performance on mixtures containing speech, although we consider this evaluation criterion less relevant.

The results suggest that a large advantage can be gained by the use of blind MS-NTF compared to NMF in producing higher mean separation metrics in terms of SDR, but do not necessarily produce an expected improvement in terms of perceptually estimated IPS.

References

[1] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Oct 2013.

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2003*, Oct 2003, pp. 177–180.

[3] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 4029–4032.

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

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

[6] T. Barker, T. Virtanen, Pontoppidan, and N. H., "Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries," in *In proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[7] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing," *IEEE Signal Processing Magazine*, March 2015.

[8] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[9] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *in proceedings of 7th Sound and Music Computing Conference, Barcelona, Spain*, 2010.

[10] J. C. Brown and P. Smaragdis, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing, Audio and Acoustics (WASPAA)*, New Paltz, NY, 2003, pp. 177–180.

[11] 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.

[12] 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.

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

[14] A. Hurmalainen, J. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[15] M. Mørup, M. N. Schmidt, and L. K. Hansen", "Shift invariant sparse coding of image and music data"," Technical University of Denmark, Tech. Rep., 2008.

[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 1562–1566.

[17] ——, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 12, pp. 2136–2147, Dec 2015.

[18] 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.

[19] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, no. 13, pp. 117 – 132, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639398000326

[20] N. Moritz, J. Anemuller, and B. Kollmeier, "Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5492–5495.

[21] D. Baby, T. Virtanen, J. Gemmeke, T. Barker, and H. Van hamme, "Exemplar-based noise robust automatic speech recognition using modulation spectrogram features," in *Spoken Language Technology Workshop (SLT)*, 2014.

[22] S. Ahmadi, S. Ahadi, B. Cranen, and L. Boves, "Sparse coding of the modulation spectrum for noise-robust automatic speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, 2014. [Online]. Available: http://dx.doi.org/10.1186/s13636-014-0036-3

[23] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768 – 785, 2011, perceptual and Statistical Audition.

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

[25] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: Statistical insights and towards self-clustering of the spatial cues," in *Exploring Music Contents*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6684, pp. 102–115.

[26] 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.

[27] S. Kırbız and B. Günsel, "A multiresolution non-negative tensor factorization approach for single channel sound source separation," *Signal Processing*, vol. 105, no. 0, pp. 56 – 69, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168414002370

[28] F. Stöter, S. Bayer, and B. Edler, "Unison source separation," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx)*, 2014.

[29] S. Masaya and M. Unoki, "Complex Tensor Factorization in Modulation Frequency Domain for Single-Channel Speech Enhancement," in *proceedings of INTERSPEECH*, 2015.
[30] W. Davenport and W. Root, *An Introduction to the Theory of Random Signals and Noise*. IEEE Press, 1987.
[31] A. Klapuri, "Signal processing methods for the automatic transription of music," Ph.D. dissertation, Tampere University of Technology, 2004.
[32] F. L. Hitchcock, "The expression of a tensor or a polyadic as a sum of products," *Journal of Mathematics and Physics*, vol. 6, no. 1-4, pp. 164–189, 1927. [Online]. Available: http://dx.doi.org/10.1002/sapm192761164
[33] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*. MIT Press, 2000, pp. 556–562.
[34] D. FitzGerald, E. Coyle, and M. Cranitch, "Extended Nonnegative Tensor Factorisation Models for Musical Sound Source Separation," *Computational Intelligence and Neuroscience*, 2008.
[35] A. Chichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. Wiley, 2009.
[36] H. Bolcskei, F. Hlawatsch, and H. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *Signal Processing, IEEE Transactions on*, vol. 46, no. 12, pp. 3256–3268, Dec 1998.
[37] R. Decorsière, "Spectrogram inversion and potential applications to hearing research," Ph.D. dissertation, Technical University of Denmark, 2013.
[38] 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.
[39] J. Kominek and A. W. Black, "*CMU Arctic Databases for Speech Synthesis*," [Online]. Available: http://www.festvox.org/cmu_arctic/, http://www.festvox.org/cmu_arctic/, 2003. [Online]. Available: http://www.festvox.org/cmu_arctic/
[40] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *submitted to IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, December 2015.
[41] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *5th International Symposium on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 229–230.
[42] 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.
[43] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046–2057, Sept 2011.
[44] 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.
[45] 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.
[46] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.
[47] Z. Průša, P. L. Søndergaard, N. Holighaus, C. Wiesmeyr, and P. Balazs, "The Large Time-Frequency Analysis Toolbox 2.0," in *Sound, Music, and Motion*, ser. Lecture Notes in Computer Science, M. Aramaki, O. Derrien, R. Kronland-Martinet, and S. Ystad, Eds. Springer International Publishing, 2014, pp. 419–442.
[48] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.
[49] A. Bregman, *Auditory Scene Analysis*. MIT Press, 1994.
[50] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx)*, September 2009, pp. 36–42.
[51] Z. Yang, B. Tan, G. Zhou, and J. Zhang, "Source number estimation and separation algorithms of underdetermined blind separation," *Science in China Series F: Information Sciences*, vol. 51, no. 10, pp. 1623–1632, 2008. [Online]. Available: http://dx.doi.org/10.1007/s11432-008-0138-6

Tom Barker Tom Barker is a Doctoral Student and Researcher within the Audio Research Group in the Department of Signal Processing, Tampere University of Technology (TUT), Finland. He received the M.Eng. Degree in Electronic Engineering from the University of York, UK in 2011. From 2011 to 2012 he was a researcher at the University of Aveiro, Portugal, and between 2013 and 2015 was the recipients of a Marie-Curie Fellowship as part of the EU-funded INSPIRE (Investigating Speech Processing In Realistic Environments) project.

Tuomas Virtanen Tuomas Virtanen is an Academy Research Fellow and Associate Professor (tenure track) at the Department of Signal Processing, Tampere University of Technology (TUT), Finland, where he is leading the Audio Research Group. He received the M.Sc. and Doctor of Science degrees in information technology from TUT in 2001 and 2006, respectively. He has also been working as a research associate at Cambridge University Engineering Department, UK. He is known for his pioneering work on single-channel sound source separation using non-negative matrix factorization based techniques, and their application to noise-robust speech recognition, music content analysis and audio event detection. In addition to the above topics, his research interests include content analysis of audio signals in general and machine learning. He has authored more than 100 scientific publications on the above topics, which have been cited more than 3000 times. He has received the IEEE Signal Processing Society 2012 best paper award for his article "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria" as well as three other best paper awards. He is an IEEE Senior Member, member of the Audio and Acoustic Signal Processing Technical Committee of IEEE Signal Processing Society, Associate Editor of IEEE/ACM Transaction on Audio, Speech, and Language Processing, and recipient of the ERC 2014 Starting Grant.