Discovering Speech Phones Using Convolutive Non-negative Matrix Factorisation with a Sparseness Constraint

Paul D. O'Grady Barak A. Pearlmutter

January 16, 2008

Abstract

Discovering a representation that allows auditory data to be parsimoniously represented is useful for many machine learning and signal processing tasks. Such a representation can be constructed by Non-negative Matrix Factorisation (NMF), a method for finding parts-based representations of non-negative data. Here, we present an extension to convolutive NMF that includes a sparseness constraint, where the resultant algorithm has multiplicative updates and utilises the beta divergence as its reconstruction objective. In combination with a spectral magnitude transform of speech, this method discovers auditory objects that resemble speech phones along with their associated sparse activation patterns. We use these in a supervised separation scheme for monophonic mixtures, finding improved separation performance in comparison to classic convolutive NMF.

Keywords: Non-negative matrix factorisation; Sparse representations; Convolutive dictionaries; Speech phone analysis.

1 Introduction

A preliminary step in many data analysis tasks is to find a suitable representation of the data. Typically, methods exploit the latent structure in the data. For example, ICA (Comon, 1994) reduces the redundancy of the data by projecting the data onto its independent components, which can be discovered by maximising a statistical measure such as independence (Bell and Sejnowski, 1995) or non-Gaussianity (Hyvärinen and Oja, 1997).

Non-negative Matrix Factorisation (NMF) is a *parts-based* approach that does not make a statistical assumption about the data. Instead, it assumes that for the domain at hand—for example grey-scale images—negative numbers are physically meaningless: The ICA decomposition of a grey-scale image may result in basis vectors that have both positive and negative components. The image is represented by a linear combination of these ICA basis vectors weighted by both positive and negative coefficients, with some basis vectors being cancelled out by others. Negative basis components have no real-world representation in a grey-scale image context, which has led researchers to argue that the decomposition should be confined to a non-negative space. Formally,

this idea can be interpreted as decomposing a non-negative matrix $\mathbf{V}$ into two non-negative factors $\mathbf{W}$ and $\mathbf{H}$. The lack of statistical assumptions makes it difficult to prove that NMF will give correct decompositions, although it has been shown geometrically that NMF provides a correct decomposition for some classes of images (Donoho and Stodden, 2004).

Data that contains negative components, for example audio, must be transformed into a non-negative domain before NMF can be applied. Here, we use a magnitude spectrogram. Spectrograms have been used in audio analysis for many years (Potter et al., 1947) and in combination with NMF have been applied to a variety of problems such as speech separation (Virtanen, 2003; FitzGerald et al., 2006; Smaragdis, 2004) and automatic transcription of music (Abdallah and Plumbley, 2004).

For some tasks it may be advantageous to perform NMF with additional constraints placed on either $\mathbf{W}$ or $\mathbf{H}$. One increasingly popular and powerful constraint is that the rows of $\mathbf{H}$ have a parsimonious activation pattern for the basis contained in the columns of $\mathbf{W}$. This is the so called *sparseness constraint* (Field, 1994; Olshausen and Field, 2004), which enables the discovery of an *over-complete* basis.

Although convolutive NMF produces activation patterns that tend to be sparse, the addition of the sparseness constraint on $\mathbf{H}$ provides a means of trading off the sparseness of the representation against accurate reconstruction. Previous algorithms for sparse NMF (Hoyer, 2002; Virtanen, 2003; O'Grady and Pearlmutter, 2006) have suffered from the scaling problem associated with the addition of a sparse constraint on $\mathbf{H}$. In order for the algorithm to behave as required, an additional normalisation step on $\mathbf{W}$ is needed (discussed in Section 3.2), which may result in $\mathbf{W}$ having an additive update rule. We overcome this restriction by using a normalised version of $\mathbf{W}$ explicitly in the reconstruction objective, and present an algorithm that has multiplicative updates for both $\mathbf{W}$ and $\mathbf{H}$.

We apply our algorithm to the analysis of speech spectrograms, it is therefore necessary to appropriately define the constituent *parts* of speech: At a conceptual level, the theoretical representation of a sound is called a *phoneme*, which is a sound in the most neutral form. Different Phonemes distinguish different words. Furthermore, Phonemes that are spoken by different speakers may be identical conceptually but differ physically, *e.g.*, Phonemes may differ in pitch and duration. A segment of speech that exhibits distinct physical or perceptual properties is called a *phone*. Phones occur frequently within speech and are the constituent components that create a speech spectrogram. In this context, the auditory objects that are extracted by convolutive NMF are expected to resemble phones, and will be referred to as such throughout the paper.

This paper is organised as follows: In Section 2 we discuss convolutive NMF (cNMF) and present an algorithm called sparse convolutive NMF (scNMF) in Section 3, which includes an additional sparseness constraint on $\mathbf{H}$. In Section 4 we apply sparse convolutive NMF to speech data, and demonstrate its utility in the extraction of speech phones. We apply such phone sets to a monophonic mixture separation task in Section 5, and discuss their utility in a speech coding task in Section 6. We finish, in Sections 7 and 8, with a discussion of related algorithms and a summarisation.

2 Convolutive NMF

NMF (Lee and Seung, 2001) is a bilinear non-negative approximate factorisation, and is formulated as follows: Given a non-negative matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ the goal is to approximate $\mathbf{V}$ as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$, $\mathbf{V} \approx \mathbf{WH}$, where $R \leq M$, such that the reconstruction error is minimised. In the context of speech spectrogram analysis, such a model produces R auditory objects that are composed of a single spectrum, calculated over all time. For our purposes, we require auditory objects that capture the time-varying nature of speech, which necessitates a convolutive NMF model. Such a convolutive NMF model has been previously proposed (Non-negative Matrix Deconvolution, Smaragdis (2004)), which we review in this section.

In conventional NMF each object is described by its spectrum and corresponding activation in time, while for convolutive NMF each object has a sequence of successive spectra and corresponding activation pattern across time. The conventional NMF model is extended to the convolutive case:

$$\mathbf{V} \approx \sum_{t=0}^{T_o-1} \mathbf{W}_t \overset{t\rightarrow}{\mathbf{H}}, \qquad v_{ik} \approx \sum_{t=0}^{T_o-1} \sum_{j=1}^{R} w_{ijt} (\overset{t\rightarrow}{h_{jk}}), \tag{1}$$

where T_o is the length of each spectrum sequence and the j-th column of $\mathbf{W}_t$ describes the spectrum of the j-th object t time steps after the object has begun.

The function $\overset{i\rightarrow}{(\cdot)}$ denotes a column shift operator that moves its argument i places to the right; as each column is shifted off to the right the leftmost columns are zero filled. Conversely, the $\overset{\leftarrow i}{(\cdot)}$ operator shifts columns off to the left, with zero filling on the right;

$$\mathbf{D} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad \overset{0\rightarrow}{\mathbf{D}} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad \overset{1\rightarrow}{\mathbf{D}} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix}$$
$$\overset{3\rightarrow}{\mathbf{D}} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad \overset{\leftarrow 2}{\mathbf{D}} = \begin{bmatrix} 3 & 4 & 0 & 0 \\ 7 & 8 & 0 & 0 \end{bmatrix} \quad \overset{\leftarrow 3}{\mathbf{D}} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 \end{bmatrix}, etc...$$

An important consideration in the formulation of the NMF algorithm is the selection of an appropriate reconstruction objective. Here, we use the beta divergence (Kompass, 2005; Cichocki et al., 2006), which is a parameterisable divergence measure,

$$D_{\mathrm{BD}}(\mathbf{V} \| \mathbf{\Lambda}, \beta) = \sum_{ik} \Big(v_{ik} \frac{v_{ik}^{\beta-1} - [\mathbf{\Lambda}]_{ik}{}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{\Lambda}]_{ik}{}^{\beta-1} \frac{[\mathbf{\Lambda}]_{ik} - v_{ik}}{\beta} \Big), \tag{2}$$

where β controls the reconstruction penalty and $\mathbf{\Lambda}$ is the current estimate of $\mathbf{V}$, $\mathbf{\Lambda} = \sum_{t=0}^{T_o-1} \mathbf{W}_t \overset{t\rightarrow}{\mathbf{H}}$. The choice of the β parameter depends on the statistical distribution of the data, and its selection requires prior knowledge (see O'Grady (2007, Chapter 3)). For $\beta = 2$, the Squared Euclidean Distance is obtained; for $\beta \rightarrow 1$, the divergence tends to the Kullback-Leibler Divergence; and for $\beta \rightarrow 0$, the divergence tends to Itakura-Saito Divergence.

It is evident that Eq. 1 can be viewed as a summation of T_o conventional NMF operations. Consequently, as opposed to updating two matrices ($\mathbf{W}$ and

```
Obj=sum(sum((V.*((((V+1E-9).^(b-1))-((lambda+1E-9).^(b-1))))./(b*(b-1)+1E-9))
+(((lambda+1E-9).^(b-1)).*((lambda-V)./(b+1E-9))))));
```

```
%Current estimate of V
for t=1:To
    Vt(:,:,t)=(W(:,:,t)*padshift(H,t-1));
end
lambda=sum(Vt,3);

%Update W
for t=1:To
    Hs=padshift(H,t-1);
    W(:,:,t)=W(:,:,t).*((V./(lambda+1e-9).^(2-b))*Hs')./((lambda+1e-9).^(b-1)*Hs');
end

%Update H
for t=1:To
    Qs=padshift((V./(lambda).^(2-b)),-(t-1));
    Ps=padshift((lambda.^(b-1)),-(t-1));
    Ht(:,:,t)=Ht(:,:,t).*(W(:,:,t)'*Qs)./((W(:,:,t)'*Ps)+1e-9);
end
H=mean(Ht,3);
```

Figure 1: Matlab notations for convolutive NMF.

$\mathbf{H}$) as in conventional NMF, T_o+1 matrices require an update ($\mathbf{W}_0, \ldots, \mathbf{W}_{T_o-1}$ and $\mathbf{H}$). The resultant convolutive NMF update equations are

$$w_{ij_t} \leftarrow w_{ij_t} \frac{\sum_{k=1}^{T} (v_{ik}/[\mathbf{\Lambda}]_{ik}^{2-\beta}) \overset{t\rightarrow}{h}_{jk}}{\sum_{k=1}^{T} [\mathbf{\Lambda}]_{ik}^{\beta-1} \overset{t\rightarrow}{h}_{jk}}, \qquad h_{jk} \leftarrow h_{jk} \frac{\sum_{i=1}^{M} w_{ij_t} (v_{ik}/[\mathbf{\Lambda}]_{ik}^{2-\beta}) \overset{\leftarrow t}{}}{\sum_{i=1}^{M} w_{ij_t} [\overset{\leftarrow t}{\mathbf{\Lambda}}]_{ik}^{\beta-1}}; \quad (3)$$

corresponding Matlab notations are presented in Figure 1. At every iteration, both $\mathbf{H}$ and $\mathbf{W}_t$ are updated for each t. It is worth noting that $\mathbf{W}_t$ for $t = 0, \ldots, T_o$ is a tensor and contains a separate $\mathbf{W}$ for each t, while a shifted version of $\mathbf{H}$ is shared across all t's. It is possible to update $\mathbf{W}_t$ and $\mathbf{H}$ at each t, however this is not advisable as it results in a biased estimate of $\mathbf{H}$, with the $t = T_o - 1$ update dominating over the others (Smaragdis, 2004). A more correct scheme is to update $\mathbf{H}$ to the average result of its updates for all t,

$$h_{jk} \leftarrow \left\langle h_{jk} \frac{\sum_{i=1}^{M} w_{ijt} (v_{ik} \overset{\leftarrow t}{/} [\mathbf{\Lambda}]_{ik})}{\sum_{i=1}^{M} w_{ijt}} \right\rangle, \forall t. \quad (4)$$

3 Sparse Convolutive NMF

Combining our reconstruction objective (Eq. 2) with a sparseness constraint on $\mathbf{H}$ results in the following objective function:

$$G(\mathbf{V} \| \mathbf{\Lambda}, \mathbf{H}, \beta, \lambda) = D_{\mathrm{BD}}(\mathbf{V} \| \mathbf{\Lambda}, \beta) + \lambda \sum_{jk} h_{jk}, \quad (5)$$

where the left term of the objective function corresponds to convolutive NMF, and the right term is an additional constraint on $\mathbf{H}$ that enforces sparsity by minimising the L_1-norm of its elements (Entropy minimisation has also been used (Shashanka et al., 2007)). The parameter λ controls the trade off between sparseness and accurate reconstruction.

3.1 Basis Normalisation

The objective of Eq. 5 creates a new problem: The right term is a strictly increasing function of the absolute value of its argument, so it is possible that the objective can be decreased by scaling $\mathbf{W}_t$ up and $\mathbf{H}$ down ($\mathbf{W}_t \mapsto \alpha\mathbf{W}_t$ and $\mathbf{H} \mapsto (1/\alpha)\mathbf{H}$, with $\alpha > 1$). This situation does not alter the left term in the objective function, but will cause the right term to decrease, resulting in the elements of $\mathbf{W}_t$ growing without bound and $\mathbf{H}$ tending toward zero. Consequently, the solution arrived at by the optimisation algorithm is not influenced by the sparseness constraint.

To avoid the scaling misbehaviour of Eq. 5 another constraint is needed; by normalising the convolutive bases we can control the scale of the elements in $\mathbf{W}_t$ and $\mathbf{H}$. Normalisation is performed for each object matrix, $\mathbf{W}_j$, by rescaling it to the unit L_2-norm,

$$\bar{\mathbf{W}}_j = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|}, \qquad j = 1, \ldots, R, \tag{6}$$

where the matrix $\mathbf{W}_j$ is constructed from the j-th column of $\mathbf{W}_t$ at each time step, $t = 0, 1, \ldots, T_o - 1$. Furthermore, normalisation of $\mathbf{W}_j$ has no adverse effects on the NMF algorithm, as the objective function (Eq. 5) does not depend on the norm of the object matrices.

3.2 Additive W Update

An NMF algorithm that uses Eq. 5 as its objective and performs the necessary basis normalisation results in the following multiplicative update for $\mathbf{H}$,

$$h_{jk} \leftarrow h_{jk} \frac{\sum_{i=1}^{M} w_{ij_t} (v_{ik} / [\overleftarrow{\mathbf{\Lambda}}^{t}]_{ik}^{2-\beta})}{\sum_{i=1}^{M} w_{ij_t} [\overleftarrow{\mathbf{\Lambda}}^{t}]_{ik}^{\beta-1} + \lambda}. \tag{7}$$

The additional unit norm constraint on each object, $\mathbf{W}_j$, complicates the $\mathbf{W}$ update rule and impedes the discovery of a suitable diagonally rescaled learning rate, $\eta_{w_{ijt}}$, which would result in a multiplicative update (Hoyer, 2002). Consequently, the following additive update is used,

$$w_{ijt} = w_{ijt} + \eta_{w_{ijt}} \left[\sum_{k=1}^{T} (v_{ik} / [\mathbf{\Lambda}]_{ik}^{2-\beta}) \overrightarrow{h}_{jk}^{t} - \sum_{k=1}^{T} [\mathbf{\Lambda}]_{ik}^{\beta-1} \overrightarrow{h}_{jk}^{t} \right]. \tag{8}$$

Subsequent to this update, any negative values in the set of matrices $\mathbf{W}_t$ are set to zero (non-negativity constraint), and each $\mathbf{W}_j$ is normalised (Eq. 6).

3.3 Multiplicative W Update

A multiplicative update can be obtained by including the normalisation requirement in the objective. Previously, this has been achieved for conventional NMF using the Squared Euclidean Distance reconstruction objective (Eggert and Körner, 2004). Here, we derive the multiplicative updates for a convolutive NMF algorithm utilising beta divergence. The classic NMF update rules (Lee and Seung, 2001) implement gradient descent, our new updates will also follow this approach. First, we introduce our new reconstruction objective, which is a modification of Eq. 2, where each object contained in $\mathbf{W}$ is normalised,

$$D_{\mathrm{BD}}(\mathbf{V}\|\mathbf{\Delta},\beta) = \sum_{ik}\Big(v_{ik}\frac{v_{ik}^{\beta-1} - [\mathbf{\Delta}]_{ik}{}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{\Delta}]_{ik}{}^{\beta-1}\frac{[\mathbf{\Delta}]_{ik} - v_{ik}}{\beta}\Big). \tag{9}$$

Here, $\mathbf{\Delta}$ is the current estimate of $\mathbf{V}$ following the normalisation of $\mathbf{W}_j$ (Eq. 6). A consequence of the normalisation requirement is that each $\mathbf{W}_j$ must be treated separately, resulting in column by column generative model,

$$\mathbf{\Delta} = \sum_{t=0}^{T_o-1}\sum_{j=1}^{R}\bar{\mathbf{w}}_{jt}(\overset{t\rightarrow}{\mathbf{h}_j}), \tag{10}$$

where $\bar{\mathbf{w}}_{jt}$ is a column vector and $\mathbf{h}_j$ is a row vector. By substituting Eq. 10 into Eq. 5 we obtain

$$G(\mathbf{V}\|\mathbf{\Delta},\mathbf{H},\beta,\lambda) = D_{\mathrm{BD}}(\mathbf{V}\|\mathbf{\Delta},\beta) + \lambda\sum_{jk}h_{jk}. \tag{11}$$

We can now derive the gradient descent update for $\mathbf{H}$,

$$h_{jk} \leftarrow h_{jk} + \eta_{h_{jk}}\frac{\partial G}{\partial h_{jk}}. \tag{12}$$

Taking the gradient of Eq. 11 with respect to $\mathbf{H}$ gives

$$\frac{\partial G}{\partial h_{jk}} = \sum_{i=1}^{M}\bar{w}_{ijt}(v_{ik}/\overset{\longleftarrow t}{[\mathbf{\Delta}]_{ik}^{2-\beta}}) - \sum_{i=1}^{M}\bar{w}_{ijt}\overset{\leftarrow t}{[\mathbf{\Delta}]_{ik}^{\beta-1}} + \lambda. \tag{13}$$

Diagonally rescaling the variables and setting the learning rate to

$$\eta_{h_{jk}} = \frac{h_{jk}}{\sum_{i=1}^{M}\bar{w}_{ijt}\overset{\leftarrow t}{[\mathbf{\Delta}]_{ik}^{\beta-1}} + \lambda} \tag{14}$$

guarantees to decrease the reconstruction objective (Lee and Seung, 2001), and gives the following multiplicative update rule for $\mathbf{H}$

$$h_{jk} \leftarrow h_{jk}\frac{\sum_{i=1}^{M}\bar{w}_{ijt}(v_{ik}/\overset{\longleftarrow t}{[\mathbf{\Delta}]_{ik}^{2-\beta}})}{\sum_{i=1}^{M}\bar{w}_{ijt}\overset{\leftarrow t}{[\mathbf{\Delta}]_{ik}^{\beta-1}} + \lambda}, \tag{15}$$

which is the same as the update of Eq. 7.

```
Obj=sum(sum((V.*((((V+1E-9).^(b-1))-((delta+1E-9).^(b-1))./(b*(b-1)+1E-9))
+(((delta+1E-9).^(b-1)).*(((delta)-V)./(b+1E-9))))))+lambda*sum(sum(H));
```

```
%Current estimate of V
for t=1:To
    Vt(:,:,t)=W(:,:,t)*padshift(H,t-1);
end
delta=sum(Vt,3);

%Update W
for t=1:To
    Hs=padshift(H,t-1);
    for j=1:R
        NumMatW=((V./delta.^(2-b))+(W(:,j,t)*W(:,j,t)'*(delta.^(b-1))))*Hs(j,:)';
        DenMatW=(delta.^(b-1)+(W(:,j,t)*W(:,j,t)'*(V./(delta.^(2-b)))))*Hs(j,:)';
        W(:,j,t)=W(:,j,t).*(NumMatW)./(DenMatW+1e-9);
    end
end
%Normalise W
for j=1:R
    scaling=sqrt(sum(sum(W(:,j,:).^2)));
    W(:,j,:)=squeeze(W(:,j,:))./((ones(M,To)*scaling)+1e-9);
end

%Update H
for t=1:To
    Qs=padshift((V./(delta).^(2-b)),-(t-1));
    Ps=padshift(delta.^(b-1),-(t-1));
    Ht(:,:,t)=H.*(W(:,:,t)'*Qs)./((W(:,:,t)'*Ps)+lambda+1e-9);
end
H=mean(Ht,3);
```

Figure 2: Matlab notations for sparse convolutive NMF.

Similarly, we derive a new update for $\mathbf{w}_{jt}$,

$$w_{ijt} \leftarrow w_{ijt} + \eta_{w_{ijt}} \frac{\partial G}{\partial w_{ijt}}. \tag{16}$$

To calculate the gradient of Eq. 11 with respect to $\mathbf{w}_{jt}$, we first need to calculate the gradient of $\mathbf{\Delta}$ using the quotient rule,

$$\frac{\partial [\mathbf{\Delta}]_{ak}}{\partial w_{ijt}} = \frac{\partial \left(\sum_{p=0}^{T_o-1} \sum_{q=1}^{R} \frac{w_{aqp}}{\|\mathbf{W}_q\|} \overset{p\rightarrow}{h}_{qk} \right)}{\partial w_{ijt}} \tag{17}$$

$$= \frac{\|\mathbf{W}_j\| \overset{t\rightarrow}{h}_{jk} - (w_{ijt} \overset{t\rightarrow}{h}_{jk}) \frac{\partial \|\mathbf{W}_j\|}{\partial w_{ijt}}}{\|\mathbf{W}_j\|^2}, \tag{18}$$

where $a = i; p = t; q = j$, and $\frac{\partial \|\mathbf{W}_j\|}{\partial w_{ijt}} = \bar{\mathbf{W}}_j$ for the L_2-norm. The gradient of Eq. 11 can now be expressed as

$$\frac{\partial G}{\partial w_{ijt}} = \sum_{k=1}^{T} \left[\frac{v_{ik}}{[\mathbf{\Delta}]_{ik}^{2-\beta}} - [\mathbf{\Delta}]_{ik}^{\beta-1} \right] \frac{\partial [\mathbf{\Delta}]_{ik}}{\partial w_{ijt}}. \tag{19}$$

Setting the learning rate to

$$\eta_{w_{ijt}} = \frac{w_{ijt}\|\mathbf{W}_j\|^2}{\sum_{k=1}^{T} \overrightarrow{h}_{jk}^{t} \left[\|\mathbf{W}_j\|[\mathbf{\Delta}]_{ik}^{\beta-1} + \bar{w}_{ijt}(w_{ijt}(v_{ik}/[\mathbf{\Delta}]_{ik}^{2-\beta}))\right]}, \tag{20}$$

then rearranging Eq. 16 and scaling by $\|\mathbf{W}_j\|/\|\mathbf{W}_j\|$ results in the following element-wise update,

$$w_{ijt} \leftarrow w_{ijt} \frac{\sum_{k=1}^{T} \overrightarrow{h}_{jk}^{t} \left[(v_{ik}/[\mathbf{\Delta}]_{ik}^{2-\beta}) + \bar{w}_{ijt}(\bar{w}_{ijt}[\mathbf{\Delta}]_{ik}^{\beta-1})\right]}{\sum_{k=1}^{T} \overrightarrow{h}_{jk}^{t} \left[[\mathbf{\Delta}]_{ik}^{\beta-1} + \bar{w}_{ijt}(\bar{w}_{ijt}(v_{ik}/[\mathbf{\Delta}]_{ik}^{2-\beta}))\right]}, \tag{21}$$

and column-wise update,

$$\mathbf{w}_{jt} \leftarrow \mathbf{w}_{jt} \otimes \frac{\left[(\mathbf{V}/\mathbf{\Delta}^{2-\beta}) + (\bar{\mathbf{w}}_{jt}\bar{\mathbf{w}}_{jt}^{\mathsf{T}}\mathbf{\Delta}^{\beta-1})\right]\overrightarrow{\mathbf{h}}_{j}^{t}}{\left[\mathbf{\Delta}^{\beta-1} + (\bar{\mathbf{w}}_{jt}\bar{\mathbf{w}}_{jt}^{\mathsf{T}}(\mathbf{V}/\mathbf{\Delta}^{2-\beta}))\right]\overrightarrow{\mathbf{h}}_{j}^{t}}, \tag{22}$$

where $\otimes$ denotes an element-wise (also known as Hadamard or Schur product) multiplication, and division is also element-wise. The update for $\mathbf{w}_{jt}$ is now in terms of its normalised version, $\bar{\mathbf{w}}_{jt}$, which is calculated (Eq. 6) subsequent to the update. As long as $\eta_{w_{ijt}}$ and $\eta_{h_{jk}}$ are sufficiently small, these updates should reduce Eq. 11. Matlab notations for sparse convolutive NMF are presented in Figure 2.

3.4 Sparse Convolutive NMF Applied to Audio Spectra

An interesting property of the sparseness constraint is that it enables the discovery of an over-complete basis, *i.e.*, a basis that contains more basis functions than are necessary to span the projection space. To illustrate the performance of convolutive NMF on data generated from an over-complete basis, consider the example presented in Figure 3. A signal that is composed of three auditory objects, each occurring at least twice, is presented: The first object is an exponentially decreasing then increasing frequency sweep centred around 5 kHz, the second object has a frequency sweep that is the reverse of the first centred at 3 kHz, and the third object is a combination of the first two. Convolutive NMF is applied to the signal with $R = 3$ and $T = 2$ seconds; the discovered auditory object are presented along with their activations, which indicate the start time of each auditory object. It is evident from the discovered objects that only the first two auditory objects are identified. The reason being that the third object can be expressed in terms of the first two, and the signal can be adequetly described by using the first two objects. Therefore, convolutive NMF achieves its optimum with just the first two linearly independent objects, without the need for an over-complete representation.

When a sparseness constraint is introduced, the existence of an over-complete representation helps minimise the objective, allowing for a sparser description of the signal. Sparse convolutive NMF applied to the same signal (Figure 4) identifies all three objects and their associated activation patterns, successfully revealing the over-complete basis used to generate the signal. Furthermore, sparse convolutive NMF produces ten activations while convolutive NMF produces twelve ($R = 2$).

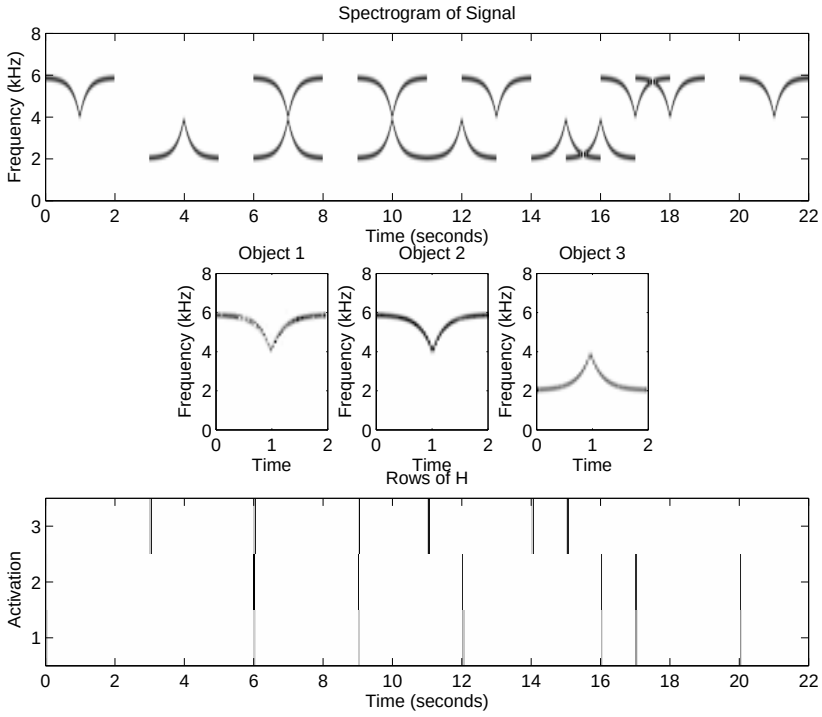

Figure 3: Spectrogram of a signal composed of an over-complete basis, and its factors obtained by convolutive NMF. It is evident that the first two objects are discovered, while the third object is represented in terms of the first two.

4 Sparse Convolutive NMF on Speech Spectra

We have demonstrated the properties of sparse convolutive NMF when applied to synthetic audio data, we will now turn our attention to real-world data. We apply sparse convolutive NMF to speech, and present a learned basis for the sparse representation of speech using the TIMIT (Garofolo et al., 1993) database. Recently, such work has been presented for convolutive NMF (Smaragdis, 2007).

4.1 Discovering a Phone-like Basis

To illustrate the differences between the phones extracted by convolutive NMF and sparse convolutive NMF we perform the following three experiments for each algorithm: We take around 30 seconds of speech from a single male speaker (`DMT0`), a single female speaker (`SMA0`), and around 15 seconds from both to create a contiguous mixture. The data is normalised to unit variance, downsampled from 16 kHz to 8 kHz and a magnitude spectrogram of the data is constructed. We use a FFT frame size of 512, a frame overlap of 384 and a hamming window to reduce the presence of sidelobes. We extract 40 bases, $R = 40$, with a temporal extent of 0.176 seconds, $T_o = 8$, and run convolutive NMF with $\beta = 1$ for 200 iterations. The extracted bases for male, female and mixed speech are presented in Figures 5, 6 & 7 respectively. The experiments

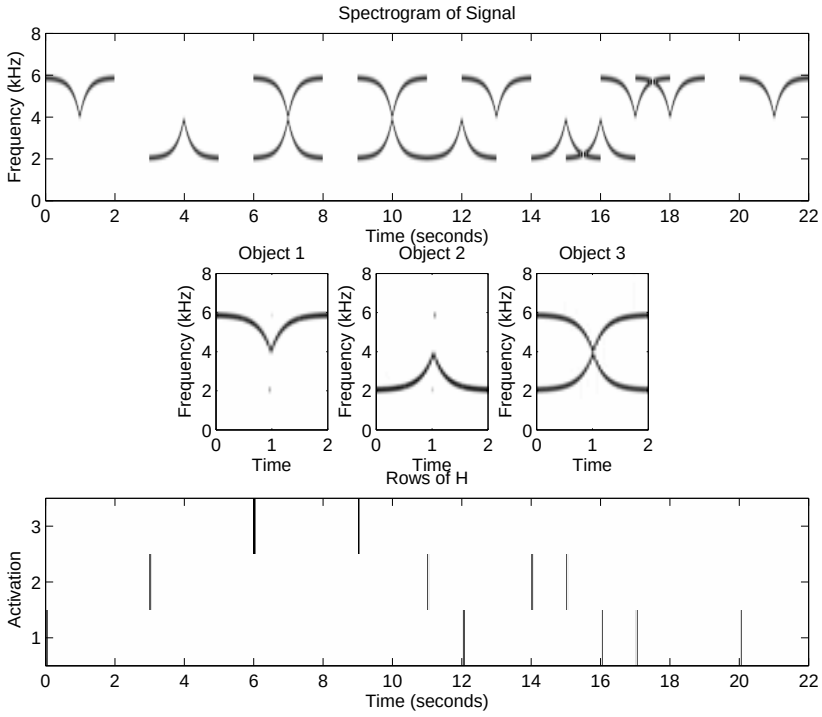

Figure 4: Spectrogram of a signal composed of an over-complete basis, and its factors obtained by sparse convolutive NMF. It is evident that the algorithm successfully reveals an over-complete basis for the data.

are repeated for sparse convolutive NMF with $\lambda = 15$, and the corresponding bases are presented in Figures 8, 9 & 10

4.1.1 Convolutive NMF Basis

For convolutive NMF, the harmonic nature of the extracted bases suggest that they correspond to speech phones. The verification of which, can be achieved by listening to an audible reconstruction, which produces sounds that resemble small segments of speech *i.e.*, speech phones. An audible reconstruction can be created by combining the magnitude spectrum of the NMF estimate with the phase of the original input, which represents a Polar form of the complex FFT coefficients, and returning to Cartesian form where an inverse FFT transformation can be performed. The resultant waveform exhibits perfect phase, and its quality is uniquely dependant on the magnitude spectrum estimated by NMF.

It is evident that most of the phones represent harmonic series with differing pitch inflections, while a smaller subset of phones contain wideband components that correspond to consonant sounds. The form of the extracted basis functions are very dependent on the data, and reflect the timbral characteristics of each speaker's voice. Comparison of the male and female phone sets reveal that the most important difference between the two is the spacing between the harmonics of the phones. For the male speaker the harmonics are spaced much closer together, which is indicative of a lower pitched voice, while the fe-

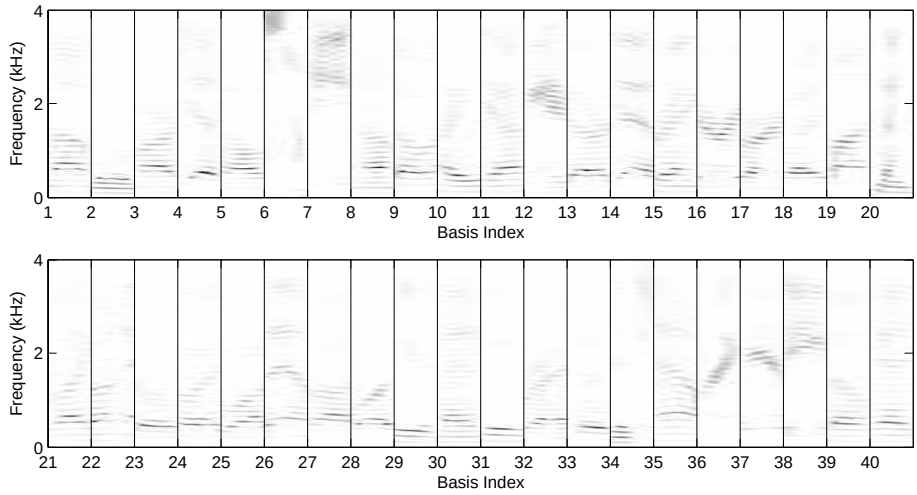

Figure 5: A collection of 40 phone-like basis functions discovered by cNMF for a single male speaker (DMT0) taken from the TIMIT speech database.

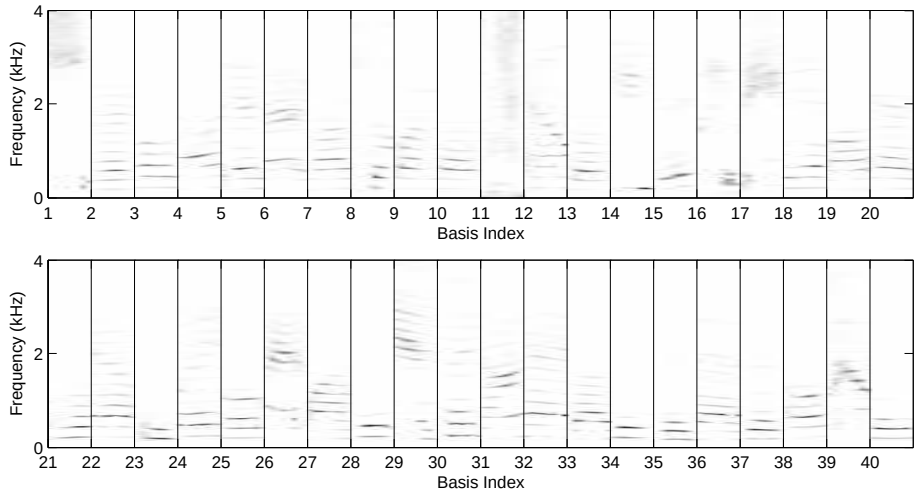

Figure 6: A collection of 40 phone-like basis functions discovered by cNMF for a single female speaker (SMA0) taken from the TIMIT speech database.

male speaker phone set contains harmonics which are farther apart, indicating a higher pitched voice. Otherwise, both phone sets are quite similar. For the mixture phone set, it is evident that the extracted phones correspond to either the male or female phone set. This indicates that the timbral characteristics of the male and female speaker are sufficiently different, such that phones that are representative of both cannot be extracted. Although, this may not be true for the consonant phones.

Due to the approximative nature of NMF, the number of bases, R, and the temporal extent of each basis, T_o, affects the ability of the algorithm to repre-

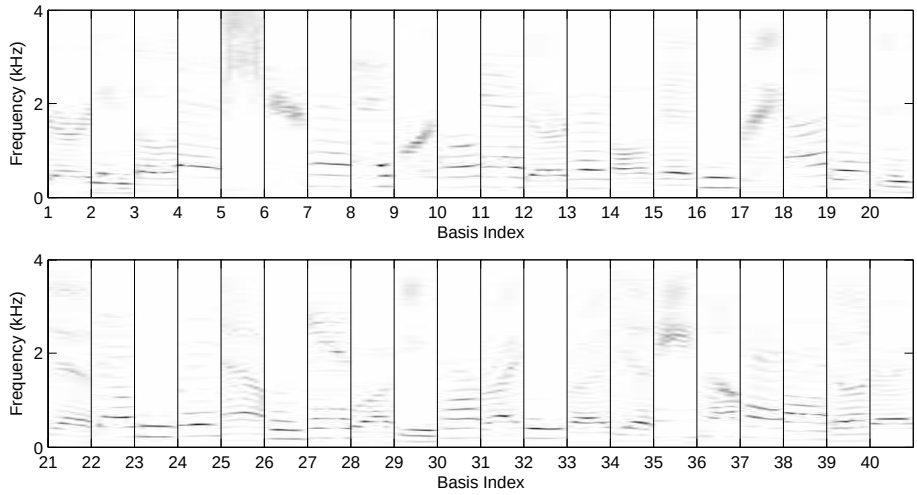

Figure 7: A collection of 40 phone-like basis functions discovered by cNMF for a mixture of a male (DMT0) and female speaker (SMA0) taken from the TIMIT speech database.

sent phonetic content in a speech spectrogram. This is reflected in the *Signal-to-Noise Ratio* (SNR) of the original spectrogram and its NMF reconstruction. For a large value of R, convolutive NMF can more accurately represent individual phones as individual basis functions, resulting in better reconstruction quality. For small values of R the resultant bases are forced to simultaneously represent multiple phones in each individual basis function, resulting in a blurry distinction between the bases, and poor reconstruction quality. For the purposes of our illustrative examples, the chosen algorithm parameters suffice.

4.1.2 Convolutive NMF Basis with a Sparseness Constraint

By placing a sparseness constraint on the activations of the basis functions, we specify that the expressive power of each basis be extended such that it is capable of representing speech parsimoniously, much like an over-complete dictionary. The result is that the extracted phones exhibit a structure that is rich in phonetic features, where harmonics at higher frequencies have a much greater intensity than seen in the phones extracted by convolutive NMF. This reflects the requirement that the basis functions in our new sparse phone set exhibit enough features to produce a parsimonious activation pattern.

Analysis of the male and female sparse phone set reveals another important difference between the two speakers. In addition to difference in harmonic spacing, it is evident that the structure of the male phones are of a more complex nature, where changes over time are much more varied than for the female phone set. Furthermore, for the male sparse phone set, basis functions that contain both harmonic series and wideband components are extracted. For the mixture phone set, the effects are the same as those previously observed, where extracted phones correspond to either the male or female sparse phone set.

It is worth noting the effects of the selection of the weighting parameter λ.

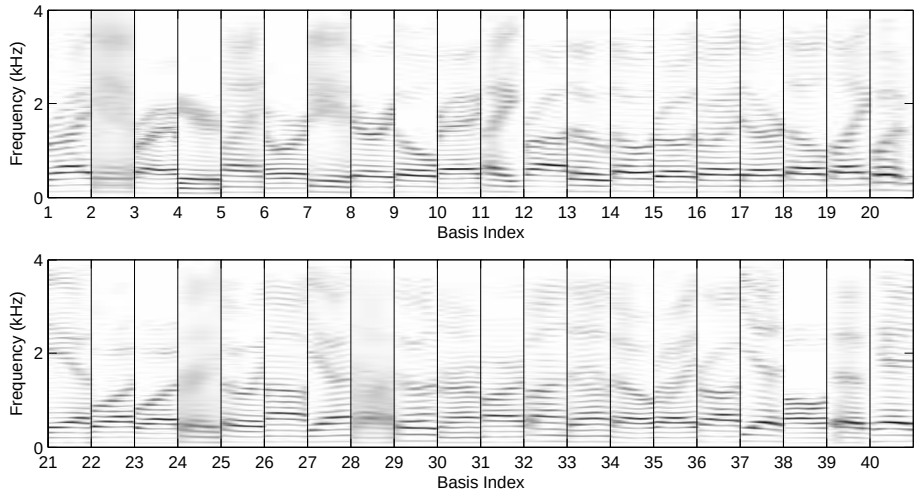

Figure 8: A collection of 40 phone-like basis functions for a single male speaker (DMT0) taken from the TIMIT speech database. The basis is extracted using scNMF with $\lambda = 15$.

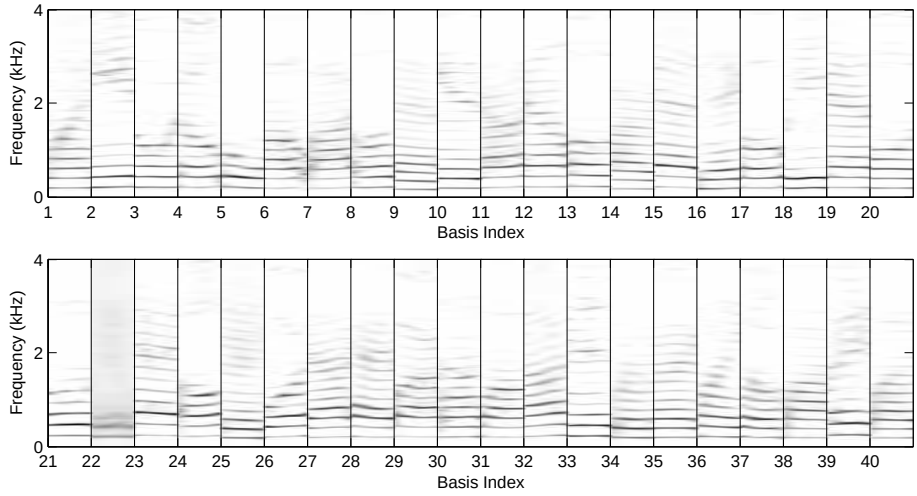

Figure 9: A collection of 40 phone-like basis functions for a single female speaker (SMA0) taken from the TIMIT speech database. The basis is extracted using scNMF with $\lambda = 15$.

Since λ controls the tradeoff between accurate reconstruction and sparseness of the activations, larger values for λ will result in degradation of the quality of the approximation. This effect can be ameliorated by increasing R or reducing λ.

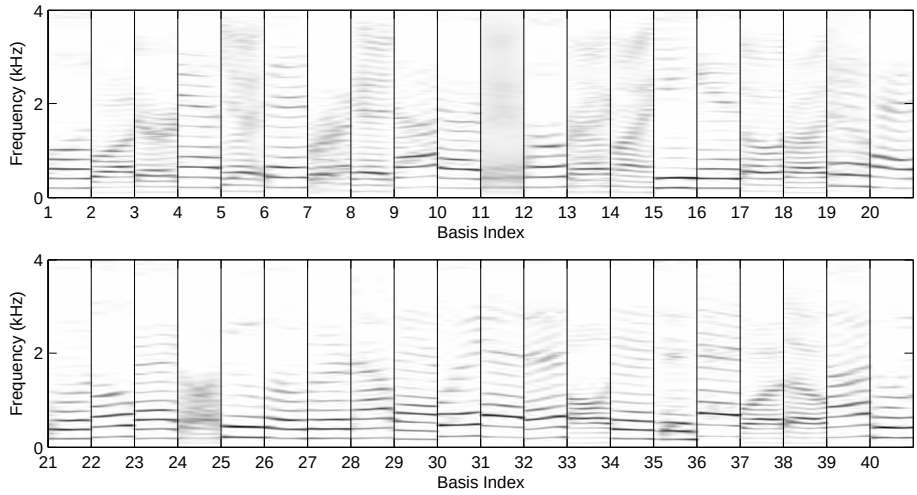

Figure 10: A collection of 40 phone-like basis functions for a a mixture of a male (DMT0) and female speaker (SMA0) taken from the TIMIT speech database. The basis is extracted using scNMF with $\lambda = 15$.

4.1.3 Sparseness of Activations

The sparseness of the activations produced by convolutive NMF, $\mathbf{H}^{\mathrm{c}}$, and sparse convolutive NMF, $\mathbf{H}^{\mathrm{sc}}$, can be compared using the *Kurtosis Ratio* (KR),

$$\mathrm{KR}(\mathbf{H}^{\mathrm{sc}}, \mathbf{H}^{\mathrm{c}}) = \frac{\frac{1}{R}\sum_{j=1}^{R} \mathrm{kurt}(\mathbf{h}_j^{\mathrm{sc}})}{\frac{1}{R}\sum_{j=1}^{R} \mathrm{kurt}(\mathbf{h}_j^{\mathrm{c}})}, \tag{23}$$

where $\mathrm{kurt}(c) = \frac{\langle (c-\mu)^4 \rangle}{\sigma^4} - 3$; $\mathrm{KR} > 1$ indicates that the $\mathbf{H}^{\mathrm{sc}}$ is sparser than $\mathbf{H}^{\mathrm{c}}$, and vice versa. The KR values for our male, female and mixed representations are 2.03, 1.74 and 1.98 respectively. Indicating that sparse convolutive NMF has indeed discovered a sparse representation for each.

5 Supervised Known Speaker Separation

To demonstrate the utility of the extracted phone sets, we apply them to the separation of speakers from a monophonic mixture. From inspection of the NMF generative model, we can see that the estimate for $\mathbf{V}$ is constructed by taking the outer product of each column of $\mathbf{W}$ and row of $\mathbf{H}$, then summing the resultant matrices,

$$\mathbf{V} \approx \sum_{j=1}^{R} \mathbf{w}_j \mathbf{h}_j.$$

This reconstruction scheme together with a magnitude spectrogram representation, where overlapping spectra sum approximately, constitute a scheme whereby different sounds, represented by different basis functions, can be separated from the mixture (this scheme can also be extended to the convolutive NMF). In

Table 1: Information on the Training Data for Each Speaker, Including Duration of Training Data and Phone Information (39 Phoneme Set, Lee and Hon (1989)).

Speakers		Training Len. (sec.)	Phone Information			
			Total No.	Time Len. (ms) Min.	Avg.	Max.
Male	ABC0	23	322	17	70	206
	BJV0	24	331	23	73	175
	DWM0	27	328	15	82	186
Female	EXM0	32	350	17	96	213
	KLH0	26	327	24	87	179
	REH0	25	361	16	61	161

contrast, ICA separates signals by making statistical assumptions about the signals, i.e, statistical independence. Therefore, the thoerem that proves identifiability of sources in ICA, *i.e.*, Darmois' Theorem (Darmois, 1953), cannot be extended to NMF as it has no solid statistical notions behind it. Furthermore, ICA can not be applied to the separation of monophonic signals, as more than one observation is required to separate independent sources.

As illustrated in our previous experiments, the structure of the bases that are extracted from the speech data are uniquely dependent on the speaker (given the same algorithm parameters). In the context of speech separation, it is not unreasonable to expect that the bases extracted for a specific speaker adequately characterise the speaker, such that they can be used to discriminate them from other speakers. For a monophonic mixture where a number speakers are added together, it is possible to separate the speakers in the mixture by constructing an individual magnitude spectrogram for each speaker, using the phones specific to that speaker.

It is evident that this scheme requires that the bases be categorised into individual phone sets. If the speakers are known in advance, a phone set can be extracted for each speaker and used in this scheme in a supervised manner. For example, consider a mixture of a known male and female speaker. The set of male bases, $\mathbf{W}_t^m$, and female bases, $\mathbf{W}_t^f$, are learned from the training data, and it is assumed that they will roughly correspond to bases extracted from any unknown sentences spoken by that speaker. By arranging the respective bases contiguously to form a combined basis, $\mathbf{W}_t^{mf} = [\mathbf{W}_t^m | \mathbf{W}_t^f]$, we can fit the mixture to the combined basis by fixing $\mathbf{W}_t = \mathbf{W}_t^{mf}$ and updating $\mathbf{H}$. Separation can be achieved by constructing an individual magnitude spectrogram using each speaker's bases and associated activations. The separation performance of such an approach is highly dependant on the *similarity* of each speaker's phone set. For a typical male and female mixture, the respective phone sets will be sufficiently different to achieve good results.

We use the following procedure for the separation of a known male and female speaker from a monophonic mixture:

1. Obtain training data for the male, $s_m(t)$, and female, $s_f(t)$, speaker; create a magnitude spectrogram for both, and extract corresponding phone sets, $\mathbf{W}_t^m$ and $\mathbf{W}_t^f$, using sparse convolutive NMF.

Table 2: The Speakers and Sentences Used for Each Male and Female Mixture, Including Information About Sentence Duration and Phone Content (39 Phoneme Set, Lee and Hon (1989)).

Mix.	Speaker		Sentence		Len. (sec.)		Total Phones	
	Male	Female	Male	Female	Male	Female	Male	Female
1	ABC0	EXM0	SX331	SX291	2.45	3.48	32	36
2	ABC0	KLH0	SX331	SX357	2.45	3.69	32	43
3	ABC0	REH0	SX331	SX325	2.45	1.93	32	25
4	BJV0	EXM0	SX347	SX291	3.62	3.48	59	36
5	BJV0	KLH0	SX347	SX357	3.62	3.69	59	43
6	BJV0	REH0	SX347	SX325	3.62	1.93	59	25
7	DWM0	EXM0	SX286	SX291	3.66	3.48	52	36
8	DWM0	KLH0	SX286	SX357	3.66	3.69	52	43
9	DWM0	REH0	SX286	SX325	3.66	1.93	52	25

2. Construct a combined basis set $\mathbf{W}_t^{mf}$. This results in a basis that is twice as big as R.

3. Take a mixture that is composed of two unknown sentences spoken by our selected speakers, and create a magnitude spectrogram of the mixture. Fit the mixture to $\mathbf{W}_t^{mf}$ by performing sparse convolutive NMF with $\mathbf{W}_t$ fixed to $\mathbf{W}_t^{mf}$, and learn only the associated activations $\mathbf{H}$.

4. Partition $\mathbf{H}$ such that the activations are split into male, $\mathbf{H}^m$, and female, $\mathbf{H}^f$, parts that correspond to their associated bases, $\mathbf{H} = [\mathbf{H}^m|\mathbf{H}^f]^\mathsf{T}$.

5. Construct a magnitude spectrogram for both speakers, using their respective bases and activations: $\mathbf{S}^m = \sum_{t=o}^{T_o-1} \mathbf{W}_t^m \mathbf{H}^m$; $\mathbf{S}^f = \sum_{t=o}^{T_o-1} \mathbf{W}_t^f \mathbf{H}^f$.

6. Use the phase information from the mixture to create an audible reconstruction for both speakers, $\hat{s}_m(t)$ & $\hat{s}_f(t)$.

This procedure may also be used for convolutive NMF, and can be generalised for more than two speakers, and speakers of the same gender.

5.1 Separation Experiments

Here, we compare the separation performance of convolutive NMF and sparse convolutive NMF. Our interest lies in how the algorithms perform for the same algorithm parameters, which may not necessarily be the optimal choice for each algorithm. For an extensive study of the relationship between parameter selection and separation performance for convolutive NMF, see Smaragdis (2007).

We select three male and three female speakers from the TIMIT database, and create a training set for each that includes all but one sentence spoken by that speaker. We artificially generate a monophonic mixture by summing the remaining sentences for a selected male female pair, generating a total of nine mixtures in this way. More formally, each sentence pair is normalised to unit variance, down-sampled from 16 kHz to 8 kHz, and summed together. A magnitude spectrogram of each mixture is constructed using a FFT frame size

of 512, a frame overlap of 256 and a hamming window. Information pertaining to the speakers and their training data is presented in Table 1, while information on the mixtures is presented in Table 2.

The separation performance for both algorithms is evaluated for each mixture over a selection of values for R ($R = [40\ 80\ 140\ 220]$). For both algorithms the temporal extent of each phone is set to 0.224 seconds ($T_o = 6$), the number of iterations is 150, β is set to 1 and each experiment is repeated for 10 Monte Carlo runs. For convolutive NMF, a total of 24 speaker phone sets are extracted and used in 360 ($9 \times 4 \times 10$) separation experiments. For sparse convolutive NMF separation performance is tested for $\lambda = [0.01\ 0.1\ 0.3\ 1.0\ 2.0]$; resulting in 120 ($6 \times 4 \times 5$) speaker phone sets and 1800 ($9 \times 4 \times 5 \times 10$) separation experiments.

For the purposes of ease of comparison with existing separation methods, we evaluate the separation performance of the sparse convolutive NMF algorithm using the measures provided by the `BSS_EVAL` toolbox (Févotte et al., 2005). The performance measures are based on the principle that a given source estimate, $\hat{s}$, is composed as a sum that includes the original source and different classes of noise,

$$\hat{s}(t) = s(t) + \epsilon_i(t) + \epsilon_n(t) + \epsilon_a(t), \tag{24}$$

where $\epsilon_i(t)$ is noise due to interference from other sources, $\epsilon_n(t)$ is perturbating noise (such as Gaussian noise) and $\epsilon_a(t)$ is the noise due to artifacts (such as musical noise). The noise introduced by each class is estimated by the toolbox functions and used in the following global performance measures:

- *Source-to-Artifact Ratio* (SAR): Measures the level of artifacts in the source estimate,
$$\text{SAR} = \frac{\|s + \epsilon_i + \epsilon_n\|^2}{\|\epsilon_a\|^2}. \tag{25}$$

- *Source-to-Interferences Ratio* (SIR): Measures the level of interference from the other sources in each source estimate,
$$\text{SIR} = \frac{\|s\|^2}{\|\epsilon_i\|^2}. \tag{26}$$

- *Source-to-Distortion Ratio* (SDR): Provides an overall separation performance criterion,
$$\text{SDR} = \frac{\|s\|^2}{\|\epsilon_i + \epsilon_n + \epsilon_a\|^2}. \tag{27}$$

All performance measures are expressed in dB, with higher performance values indicating better quality estimates.

5.1.1 Convolutive NMF Separation Performance

In this section, we examine the separation performance of convolutive NMF when applied to our generated mixtures. The results for each experiment are averaged over all runs and are presented in Figure 11. Each separation measure is illustrated as a bar chart, where mixtures are plotted against the number of bases used, and bar height indicates performance. For illustrative clarity, the 9 mixtures are arranged in ascending order.

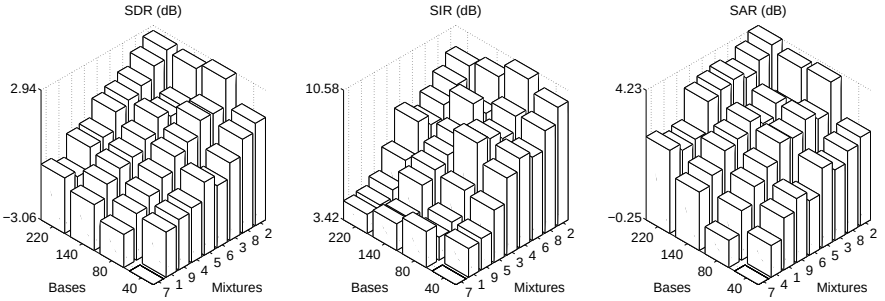

Figure 11: Separation performance for cNMF: A bar chart for each performance measure (SDR, SIR and SAR) is presented, where the performance for each mixture, in ascending order, is plotted against the number of bases. Note that the scales for the z-axis are expressed in dB and change for each plot.

The resultant performance values are very dependant on the mixture under consideration, this may reflect similarity in the timbral characteristics of the speakers in each mixture. On average, mixture 7 performed worst for all performance measures, while mixture 2 performed best. The SDR results, which indicate overall performance, improve for most mixtures as the number of bases gets larger. The average SDR over all mixtures range from -0.18 dB for 40 bases to 0.96 dB for 220 bases. The same is also true for SAR, where performance rises from 1.75 dB at 40 bases to 3.37 at 220 bases. For the SIR results, best performance is achieved when 80 bases are used.

5.1.2 Sparse Convolutive NMF Separation Performance

The results in Figure 11 can be compared with the corresponding results for sparse convolutive NMF in Figure 12, in which 4 sets of results pertaining to different values of λ are presented.

For added clarity, we statistically analyse the performance of convolutive NMF and sparse convolutive NMF by collating the results from all experiments (Figure 11 & Figure 12), and represent the results using box plots, where SDR, SIR and SAR are presented in Figures 13, 14 & 15 respectively. Each box presents information about the median and the statistical dispersion of the results. The top and bottom of each box represents the upper and lower quartiles, while the length between them is the interquartile range; the whiskers represent the extent of the rest of the data, and outliers are represented by +.

The SDR results indicate that for $\lambda = [0.1, 0.3]$, the median performance obtained (0.66 dB, 0.62 dB) exceeds convolutive NMF (0.44 dB) for our given algorithm parameters and data. It is also evident that a better spread of results is produced for sparse convolutive NMF; demonstrating that when λ is chosen appropriately, sparse convolutive NMF achieves superior overall performance. For SIR, $\lambda = 0.3$ produces the best spread of results, which indicates that sparse convolutive NMF is more resilient to interference from other sources. However, for SAR, convolutive NMF produces the best results; this may reflect the fact that each sparse phone set exhibits phones that are rich in phonetic content, which may manifest as artifacts in the resultant source estimates. It is also evident that the performance of the sparse convolutive algorithm degrades

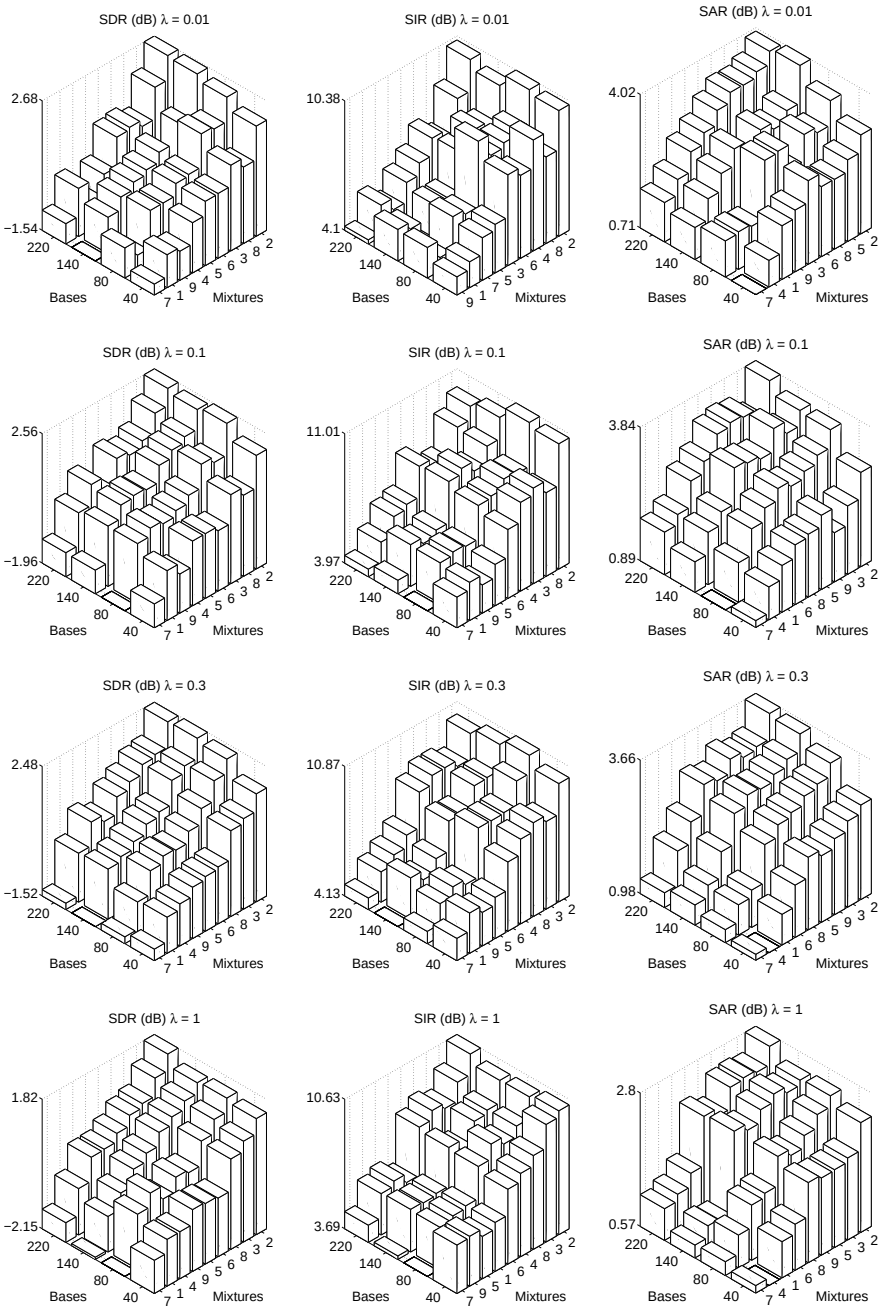

Figure 12: Separation performance for sparse convolutive NMF: A bar chart for each performance measure (SDR, SIR and SAR) is presented for a selection of λ values, where the performance for each mixture, in ascending order, is plotted against the number of bases. Note that the scales for the z-axis change for each plot and are expressed in dB.

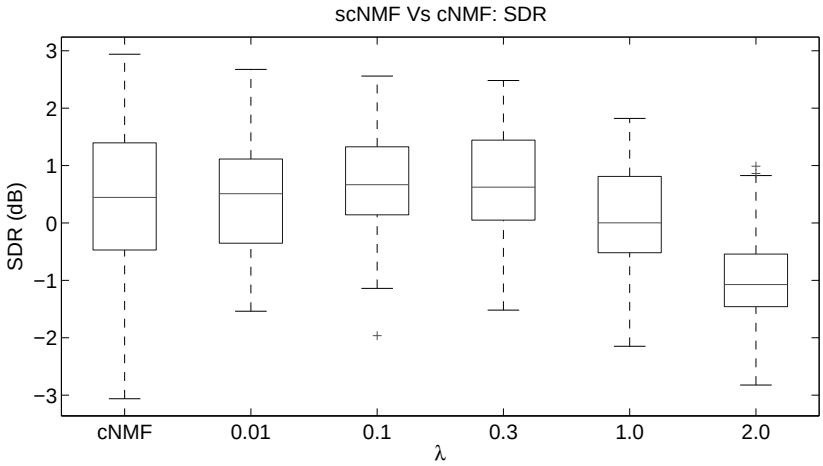

Figure 13: A comparison of the SDR results obtained by convolutive and sparse convolutive NMF: Box plots are used to illustrate the performance results, with each box representing the median and the interquartile range of the results. It is evident that for $\lambda = 0.1$, a better spread of results is obtained, indicating that sparse convolutive NMF achieves superior overall performance.

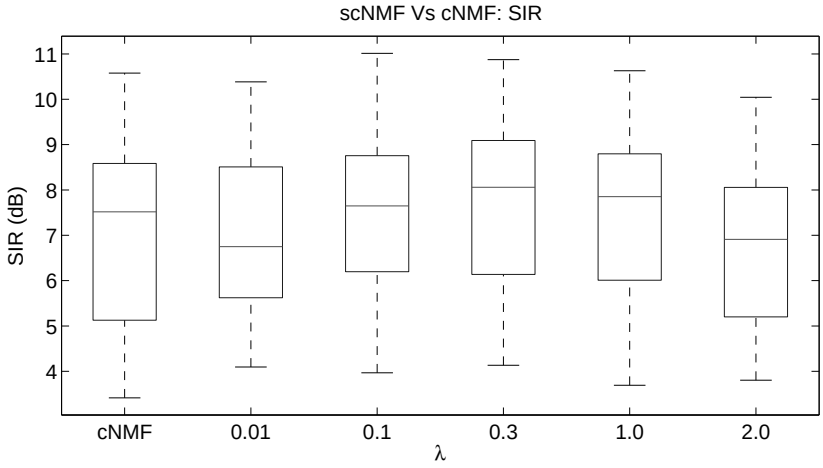

Figure 14: A comparison of the SIR results obtained by convolutive and sparse convolutive NMF: Box plots are used to illustrate the performance results. For $\lambda = 0.1$, a better spread of results is obtained, indicating that sparse convolutive NMF produces estimates that are more resilient to interference from other sources.

significantly for large λ values, so much so, that it renders the results useless, for our data this is especially evident when $\lambda \geq 1$.

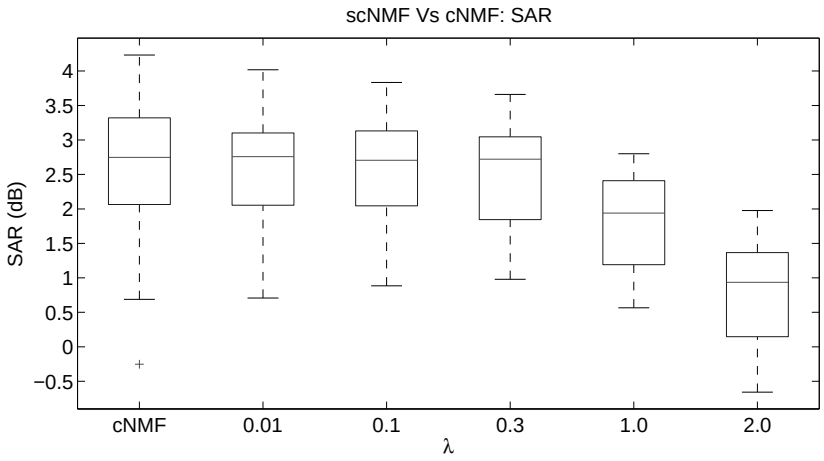

Figure 15: A comparison of the SDR results obtained by convolutive and sparse convolutive NMF: Box plots are used to illustrate the performance results, with each box representing the median and the interquartile range of the results, the whiskers represent the spread of the results. Here, convolutive NMF achieves the best results. This may reflect the fact that sparse phone sets exhibit phones that are rich in features, which may produce artifacts in the resultant source estimates.

6 Coding Efficiency of Bases

We demonstrate the utility of sparse convolutive NMF in the information coding of speech data, we employ a rudimentary scheme whereby the K largest coefficients in each column of $\mathbf{H}$, along with their positions, are used to reconstruct the data,

$$\mathbf{\Lambda}_K = \sum_{t=0}^{T_o-1} \mathbf{W}_t \ \max(K, \overset{t\rightarrow}{\mathbf{H}}) \qquad 0 < K \leq R, \tag{28}$$

where the max operator creates a matrix the same dimensions as $\mathbf{H}$, with all but the K largest coefficients in each column being zeroed. Here, we consider only the reconstruction of the magnitude spectrogram and do not address how to encode phase information.

We use an experimental procedure similar to that used in our separation experiments, whereby we fix $\mathbf{W}$ to the basis for our speaker and fit an unknown sentence to it by updating $\mathbf{H}$. We then reconstruct $\mathbf{\Lambda}_K$ by using its K largest coefficients, over $K = 1, \ldots, R$, and measure reconstruction quality using SNR. We select a male (`ABC0`) and female speaker (`EXM0`), and use the 220 basis set learned for the experiments in the previous section. The reconstruction quality for a range of λ values is investigated and each experiment is repeated for 10 Monte Carlo runs; the results are presented in Figure 16.

The curves in Figure 16 illustrate the trade-off between the fidelity of the reconstruction and the coding cost, expressed in coefficients. We are interested in the transitional phase leading to quiescent value for SNR; the quicker the convergence the fewer coefficients needed to reconstruct $\mathbf{\Lambda}$. The coding efficiency for convolutive NMF ($\circ$) can be easily compared with the other curves, which

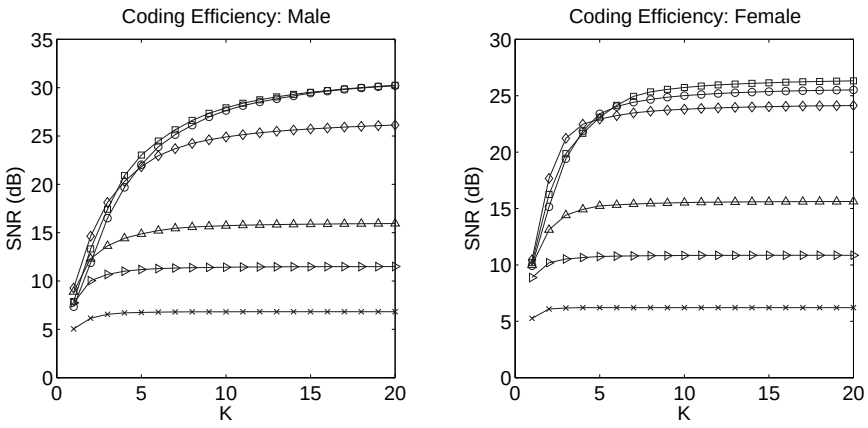

Figure 16: Coding efficiency curves for sparse convolutive NMF, for both a male (left: `ABC0`) and female (right: `EXM0`) speaker. The curve for convolutive NMF ($\circ$) can be contrasted with sparse convolutive NMF ($\square$: $\lambda = 0.001$, $\Diamond$: $\lambda = 0.01$, $\triangle$: $\lambda = 0.1$, $\triangleright$: $\lambda = 0.3$, $\times$: $\lambda = 1$). It is evident that sparse convolutive NMF provides a faster rate of convergence to a quiescent SNR. Furthermore, for large λ, a degenerative effect on reconstruction quality is evident, which is indicative of the tradeoff between sparseness and reconstruction quality.

represent the results for sparse convolutive NMF.

For both speakers, it is evident that sparse convolutive NMF needs fewer coefficients to reach a quiescent SNR value, the SNR achieved is very dependent on λ, which is indicative of the trade-off between the sparseness of $\mathbf{H}$ and accuracy of reconstruction, this effect is particularly evident for large λ. For the male speaker, $\lambda = 0.001$ provides an increase in SNR over convolutive NMF when $K \leq 15$, while $\lambda = 0.01$ achieves the best performance when $K \leq 3$. Furthermore, for $\lambda = 0.01$ with $K = 3$ the SNR achieved (18 dB) produces a level of reconstruction quality such that the encoded sentence is intelligible. For the female speaker, $\lambda = 0.01$ produces superior quality when $K \leq 4$, while $\lambda = 0.001$ produces superior quality for $K \geq 6$. For both speakers $\lambda = [0.1, 0.3, 1]$ never exceeds the performance of convolutive NMF at any point along the curve. Therefore, these values are an inappropriate choice for our data and produce results that are of no use. It is also evident that there is a faster convergence rate for the female speaker's coding efficiency curves, which may be due to the lower SNRs achieved for these reconstructions. Furthermore, the female coding efficiency curves reveal that $\lambda = 0.001$ is superior for all $K \geq 6$, which indicates that for a carefully selected λ the sparseness constraint may also improve reconstruction. Although, in this case the improvement in SNR is marginal.

7 Discussion

The advantage obtained by combining convolutive NMF with a sparseness constraint on the activations, is due to the requirement that a parsimonious representation must be found in order to satisfy sparseness. Such representations

extract bases that are rich in phonetic structure, and exhibit superior separation properties.

In contrast to previously proposed algorithms, which have additive updates (Hoyer, 2002; Virtanen, 2003; O'Grady and Pearlmutter, 2006), our NMF algorithm retains its advantages of parameter-independent gradient descent and fast convergence. Moreover, multiplicative updates ensure that the algorithm arrives at some solution, which from our experience, has not always been the case for additive update algorithms. Similar multiplicative algorithms have been proposed (Virtanen, 2007), although the proposed updates do not necessarily decrease the value of the cost function. From our experience with the updates derived for our algorithm, the value of the cost function always decreases, which is possibly due to our careful treatment of the scaling ambiguity caused by using the L_1-norm sparseness constraint, as demonstrated by Eggert and Körner (2004).

An additional benefit to our algorithm is that it utilises the beta divergence, which enables different reconstruction penalty schemes to be selected depending on some additional requirement, *e.g.*, the perceptual quality of the NMF reconstruction (O'Grady, 2007, Chapter 3); although, such benefits are not discussed in this paper. Furthermore, since the beta divergence encompasses both the Square Euclidean Distance and the Kullback-Leibler Divergence—the NMF reconstruction objectives originally proposed by Lee and Seung (2001), and which have remained the most popular choice for implementing the algorithm—derivation of updates for each is unified, which differs to the general approach taken in the literature where algorithms for both are derived and presented individually.

Normalisation of the objects in $\mathbf{W}$ introduces an asymmetry between $\mathbf{W}$ and $\mathbf{H}$, which makes it difficult to prove convergence properties of Eq. 22 as discussed in Lee and Seung (2001). Nonetheless, we have performed many experiments with our algorithm and it converges to sensible solutions every time. Eggert and Körner (2004) propose that convergence can be explained by the fact that the rescaling of the gradient introduced by the multiplicative update rule, results in a gradient step that has a positive projection on the true gradient, due to the non-negativity constraint. Furthermore, as long as the gradient step size is sufficiently small (this is true when $\mathbf{\Lambda}$ approaches $\mathbf{V}$), convergence is achieved—we believe this to be true in our case also.

Finally, due to the fact that our algorithm is implemented using column-wise updates for $\mathbf{W}_t$ (because of the normalisation of the objects, $\mathbf{W}_j$, contained in $\mathbf{W}$), the run time of the algorithm increases greatly: Consider speaker `ABC0` from Table 1, to extract 40 bases (as per our experiments) on a 2.53 GHz Intel Pentium 4 computer with 256 Mb of RAM, takes four minutes for convolutive NMF, while the same experiment takes 50 minutes for sparse convolutive NMF. Furthermore, sparse convolutive NMF algorithms with additive updates may run faster too. However, our multiplicative algorithm will always arrive at a solution with better quality results, and removes the requirement to select both an appropriate learning rate and λ, which can sometimes be painfully difficult to achieve.

8 Conclusion

In this paper, we presented a sparse convolutive NMF algorithm, which effectively discovers a sparse parts-based representation for non-negative data. This method extends the convolutive NMF objective by including a sparseness constraint on the activations, enabling the discovery of over-complete representations. Moreover, in contrast to previously proposed algorithms, normalisation of the basis vectors is explicitly included in the reconstruction objective, resulting in multiplicative updates and more stable convergence properties. We have applied the algorithm to speech data, and have demonstrated its superiority to convolutive NMF, when applied to the separation of monophonic speech mixtures and speech coding.

8.1 Acknowledgements

Supported by Higher Education Authority of Ireland (An tÚdarás Um Ard-Oideachas), and Science Foundation Ireland grant 00/PI.1/C067.

References

S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 318–25, 2004.

A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neu. Comp.*, 7(6):1129–59, 1995.

A. Cichocki, R. Zdunek, and S.-i. Amari. Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. In J. P. Rosca, D. Erdogmus, J. C. Príncipe, and S. Haykin, editors, *Independent Component Analysis and Blind Signal Separation, 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006, Proceedings*, volume 3889 of *Lecture Notes in Computer Science*, pages 32–39. Springer, 2006. ISBN 3-540-32630-8. doi: doi:10.1007/11679363_5.

P. Comon. Independent component analysis: A new concept. *Signal Processing*, 36:287–314, 1994.

G. Darmois. Analyse générale des liaisons stochastiques. *Rev. Inst. Internat. Stat.*, 21:2–8, 1953.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Adv. in Neu. Info. Proc. Sys. 16*. MIT Press, 2004. URL `http://books.nips.cc/papers/files/nips16/NIPS2003_LT10.pdf`.

J. Eggert and E. Körner. Sparse coding and NMF. In *IEEE International Joint Conference on Neural Networks, 2004. Proceedings*, volume 4, pages 2529–33. IEEE, July 2004.

C. Févotte, R. Gribonval, and E. Vincent. BSS_EVAL toolbox user guide. Technical Report 1706, IRISA, 2005.

D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6: 559–601, 1994.

D. FitzGerald, M. Cranitch, and E. Coyle. Sound source separation using shifted non-negative tensor factorisation. In *Proceedings, IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing, 2006. URL `http://homepage.eircom.net/~derryfitzgerald/ICASSP06.pdf`.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR-4930, U.S. Department of Commerce, National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA, Feb. 1993. Available on CD-ROM, NIST Speech Disc 1-1.1.

P. O. Hoyer. Non-negative sparse coding. In *IEEE Workshop on Neural Networks for Signal Processing*, 2002.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neu. Comp.*, 9(7):1483–92, Oct. 1997.

R. Kompass. A generalized divergence measure for non-negative matrix factorization. In *Neuroinformatics workshop*, Torun, Poland, Sept. 2005.

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Adv. in Neu. Info. Proc. Sys. 13*, pages 556–62. MIT Press, 2001. URL `citeseer.ist.psu.edu/lee00algorithms.html`.

K. F. Lee and H. W. Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-37(11):1641–1648, 1989.

P. D. O'Grady. *Sparse Separation of Under-Determined Speech Mixtures*. PhD thesis, National University of Ireland Maynooth, 2007. URL `http://ee.ucd.ie/~pogrady/ogrady2007_phd.pdf`.

P. D. O'Grady and B. A. Pearlmutter. Convolutive non-negative matrix factorisation with sparseness constraint. In *International Workshop on Machine Learning for Signal Processing*, pages 427–432, Maynooth, Ireland, Sept. 6–8 2006. IEEE Press.

B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Curr Opin Neurobiol*, 14(4):481–7, 2004.

R. K. Potter, G. A. Kopp, and H. C. Green. *Visible Speech*. D. Van Nostrand Company, 1947.

M. Shashanka, B. Raj, and P. Smaragdis. Sparse overcomplete decomposition for single channel speaker separation. In *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transaction on Audio, Speech and Language Processing*, 2007.

P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *Fifth International Conference on Independent Component Analysis*, LNCS 3195, pages 494–9, Granada, Spain, Sept. 22–24 2004. Springer-Verlag.

T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 15(3):1066–1074, 2007. URL `http://ieeexplore.ieee.org/xpls/abs\_all.jsp?arnumber=4100700`.

T. Virtanen. Sound source separation using sparse coding with temporal continuity objective. In *in Proceedings of the International Computer Music Conference (ICMC 2003)*, 2003.

Paul O'Grady received the Ph.D. degree in computer science from the National University of Ireland, Maynooth in 2007. He is currently a post-doctoral researcher at the Complex and Adaptive Systems Laboratory, University College Dublin. His research interests include audio signal processing, blind source separation, non-negative matrix factorisation, and machine learning algorithms.

Prof. Barak A. Pearlmutter received a Ph.D. in Computer Science from Carnegie Mellon University, and has held positions at Yale, the Oregon Graduate Institute, Siemens Corporate Research, UCSD, and the University of New Mexico. He is currently on the faculty at the Hamilton Institute and in the Computer Science Department at the National University of Ireland Maynooth. His main research interest is in theoretical neurobiology: attempting to understand how the brain achieves its remarkable feats of perception.