MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Convolutive Speech Bases and their Application to Supervised Speech Separation

Paris Smaragdis

TR2007-002 January 2007

Abstract

In this paper we present a convolutive basis decomposition method and its application on simultaneous speakers separation from monophonic recordings. The model we propose is a convolutive version of the non-negative matrix factorization algorithm. Due to the non-negativity constraint this type of coding is very well suited for intuitively and efficiently representing magnitude spectra. We present results that reveal the nature of these basis functions and we introduce their utility in separating monophonic mixtures of known speakers.

IEEE Transactions on Audio, Speech and Language Processing

Convolutive Speech Bases and their Application to Supervised Speech Separation

Paris Smaragdis, *Member, IEEE*

***Abstract*— In this paper we present a convolutive basis decomposition method and its application on simultaneous speakers separation from monophonic recordings. The model we propose is a convolutive version of the non-negative matrix factorization algorithm. Due to the non-negativity constraint this type of coding is very well suited for intuitively and efficiently representing magnitude spectra. We present results that reveal the nature of these basis functions and we introduce their utility in separating monophonic mixtures of known speakers.**

***Index Terms*— non-negative matrix factorization, source separation, convolutive bases**

I. Introduction

BASIS decompositions have long been an important tool in signal processing. The use of basis decompositions spans a wide variety of applications, equally rich as the variety of approaches to obtain bases. Most basis decomposition methods are deeply rooted in statistics and transform data so as to have desirable properties. Well known examples of these are the Principal Component Analysis (PCA) [4] or the Independent Component Analysis (ICA) algorithms [3]. Other types of basis decompositions are more algebraic in nature such as the Singular Value Decomposition (SVD), various higher order generalizations of it [5], or the Non-negative Matrix Factorization [8]. Many of these decompositions have been used in many ways for source separation tasks. A particular use of these decompositions is on the magnitude spectra of monophonic recordings. Results relating this approach to source separation have been reported in multiple publications [2][12][1][15] and have been a promising field of research for some time. The use of basis functions in the just referenced work has been in the context of unsupervised learning. Basis decomposition and dimensionality reduction were used to obtain a small set of components that usually resembles the various sounds contained in the original input. The basis functions describe the spectral characters of the components, whereas their weights provide their temporal evolution. Although this approach can be successful in specific contexts it suffers from two problems, a rigid spectral form and the fact that we often expect results from very little data. In this paper we propose two extensions to this approach that address these problems. We extend the expressive power of basis decompositions by specifying a convolutive model and we propose a supervised learning approach which can benefit from knowledge extracted outside the input samples. The supervised approach to separation has been used in the past in the context of various types of statistical models ([10][11][9]). Our convolutive model is based on a recently introduced decomposition [13] based on Non-negative Matrix Factorization (NMF). We have previously shown how this decomposition can be used to extract meaningful components out of spectrograms in an unsupervised manner. In this paper we introduce the use of this basis decomposition on speech and show how it discovers meaningful features which are very useful in the context of supervised source separation.

The remainder of this paper is organized as follows, section II introduces the convolutive basis decomposition approach we will use, section III presents a methodology of extracting these bases from speech signals and highlights their nature and some interesting properties. In section IV we introduce a methodology to perform speaker separation and we thoroughly evaluate it in section V. Finally in section VI we briefly consider some post-processing enhancements to further boost the quality of separation.

II. Convolutive NMF

In this section we describe the basis model we will employ and the appropriate adaptation procedures. We will start by reviewing the Non-Negative Matrix Factorization algorithm and then extend it to a convolutive form which we will employ for our simulations.

A. Non-negative matrix factorization

Non-Negative Matrix Factorization is a linear basis decomposition approach that assumes non-negativity on both the basis and the data to be approximated. We present it briefly in this section.

NMF was first introduced by Lee and Seung [8]. Simply stated having a non-negative matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$

the goal is to approximate it as a product of two non-negative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ where $R \leq M$. The objective is to minimize the error of reconstruction of $\mathbf{V}$ by $\mathbf{W} \cdot \mathbf{H}$ and to that extent Lee and Seung [6] provided two cost functions by which we can measure it. One of these cost functions is the Frobenius norm of the difference between the input and the reconstruction, and the other, which we will be employing in this paper, is an adaptation of the Kullback-Leibler divergence which was defined as:

$$D = \left\| \mathbf{V} \odot \ln\left(\frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}}\right) - \mathbf{V} + \mathbf{W} \cdot \mathbf{H} \right\|_F \quad (1)$$

where $\| \cdot \|_F$ is the Frobenius norm and $\odot$ is the Hadamard product (element-wise multiplication). The division between the matrices is also an element-wise division operation, and the logarithm is applied on all the matrix elements separately. Optimizing this function can be pursued by conventional means using constrained gradient descent, however Lee and Seung [6] provide multiplicative update rules for the two factors $\mathbf{W}$ and $\mathbf{H}$ which elegantly bypass the need for a non-negativity constraint (assuming non-negative initial values), and provide rapid convergence. These updates for the matrices $\mathbf{W}$ and $\mathbf{H}$ were defined as:

$$\mathbf{H} = \mathbf{H} \odot \frac{\mathbf{W}^\top \cdot \frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}}}{\mathbf{W}^\top \cdot \mathbf{1}} \quad (2a)$$

$$\mathbf{W} = \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\mathbf{W} \cdot \mathbf{H}} \cdot \mathbf{H}^\top}{\mathbf{1} \cdot \mathbf{H}^\top} \quad (2b)$$

where $\mathbf{1}$ is a $M \times N$ matrix with all its elements set to 1, and the matrix divisions are as before performed in an element-wise manner. Both of these updates are applied iteratively in an alternating manner until the two factors converge. The variable R, which is the number of columns of $\mathbf{W}$ and the rows of $\mathbf{H}$, determines the rank of the approximation. If $R = M$ we can achieve a perfect reconstruction of the input, as R is reduced we start obtaining low-rank approximations. In the low rank case if we have some structure in the input $\mathbf{V}$ we notice that the elements of $\mathbf{W}$ and $\mathbf{H}$ start to reveal it. The R columns of $\mathbf{W}$ tend to reveal the vertical structure of the input, and their corresponding rows in $\mathbf{H}$ reveal their horizontal structure. In terms of a basis decomposition we can view $\mathbf{W}$ as a set of R basis functions and $\mathbf{H}$ as their corresponding weights required to approximate $\mathbf{V}$.

Applications of NMF on audio data are presented in [7]. In these cases a magnitude spectrogram is presented as the input $\mathbf{V}$ and the resulting bases $\mathbf{W}$ end up representing dominant spectral patterns contained in the input whereas their weights $\mathbf{H}$ correspond to their temporal profiles.

B. Convolutive extensions to NMF

NMF provides a useful tool for analyzing data, it is however ignoring potential dependencies across successive columns of its input $\mathbf{V}$. A regularly repeating pattern that spans multiple columns of $\mathbf{V}$ would have to be represented by NMF using multiple bases that describe the entire sequence. The fact that there is a sequence would not be apparent by examination of the bases, but would be only discovered by tedious analysis of the basis weights. Since this is a regularly repeating pattern it would be more satisfying if it was represented by a single basis function that could span the pattern length. Such dependencies across columns are very frequently seen in time-frequency representations when analyzing audio signals and the expressive ability to capture these temporal dependencies within bases is a desirable feature. In this section we introduce a convolutive extension to NMF which can allow us to extract cross-column patterns as single bases.

As just described NMF attempts to reconstruct a matrix $\mathbf{V}$ using a matrix product by $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$. In the convolutive Non-Negative Matrix Factorization we extend this expression to:

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \overset{t\rightarrow}{\mathbf{H}} \quad (3)$$

where $\mathbf{V} \in \mathbb{R}^{\geq 0, M \times N}$ is the input we wish to decompose, $\mathbf{W}(t) \in \mathbb{R}^{\geq 0, M \times R}$ is a set of bases, and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$ contains their weights. The $\overset{i\rightarrow}{(\cdot)}$ operator is a shift operator that moves the columns of its argument by i spots to the right, and consequently $\overset{\leftarrow i}{(\cdot)}$ shifts to the left, such that:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad \overset{0\rightarrow}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}$$

$$\overset{1\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{bmatrix} \quad \overset{2\rightarrow}{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{bmatrix}$$

$$\overset{\leftarrow 0}{\mathbf{A}} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix} \quad \overset{\leftarrow 1}{\mathbf{A}} = \begin{bmatrix} 2 & 3 & 4 & 0 \\ 6 & 7 & 8 & 0 \end{bmatrix}$$

$$\overset{\leftarrow 2}{\mathbf{A}} = \begin{bmatrix} 3 & 4 & 0 & 0 \\ 7 & 8 & 0 & 0 \end{bmatrix} \quad \overset{\leftarrow 3}{\mathbf{A}} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 \end{bmatrix}, etc...$$

The columns that are shifted in from outside the matrix are set to zero.

Equation 3 is essentially a summation of convolution operations between corresponding elements from a set

of two-dimensional bases $\mathbf{W}$ and a set of weights $\mathbf{H}$. Effectively what happens is that the set of ith columns of $\mathbf{W}(t)$ defines a two-dimensional structure (one which we will refer to as a basis). This basis will be shifted and scaled by convolution across the axis of t with the ith row of $\mathbf{H}$. The resulting reconstruction will be a summation of all the basis convolution results for each of the R bases.

In order to estimate the appropriate set of matrices $\mathbf{W}(t)$ and $\mathbf{H}$ to approximate $\mathbf{V}$ we can use the already existing framework of NMF. In accordance to the NMF cost function, we define the convolutive NMF cost function as:

$$D = \left\| \mathbf{V} \odot \ln\left(\frac{\mathbf{V}}{\hat{\mathbf{V}}}\right) - \mathbf{V} + \hat{\mathbf{V}} \right\|_F \tag{4}$$

Where $\hat{\mathbf{V}}$ is the approximation of $\mathbf{V}$ defined as:

$$\hat{\mathbf{V}} = \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \overset{t\rightarrow}{\mathbf{H}} \tag{5}$$

Due to the linearity we can decompose the above cost function to a collection of simultaneous NMF approximations, one for each value of t. Noting this fact we can now optimize the above cost function by optimizing this set of T NMF approximations. For each NMF approximation we have to update the equivalent $\mathbf{W}(t)$ and the appropriately shifted $\mathbf{H}$. This results into the convolutive NMF update equations which are:

$$\mathbf{H} = \mathbf{H} \odot \frac{\mathbf{W}(t)^\top \cdot \overset{\leftarrow t}{\left[\frac{\mathbf{V}}{\hat{\mathbf{V}}}\right]}}{\mathbf{W}(t)^\top \cdot \mathbf{1}} \tag{6a}$$

$$\mathbf{W}(t) = \mathbf{W}(t) \odot \frac{\frac{\mathbf{V}}{\hat{\mathbf{V}}} \cdot \overset{t\rightarrow}{\mathbf{H}}^\top}{\mathbf{1} \cdot \overset{t\rightarrow}{\mathbf{H}}^\top} \tag{6b}$$

In every updating iteration, for each t we update $\mathbf{H}$ and $\mathbf{W}(t)$. Note that for each t the corresponding NMF problem has its own $\mathbf{W}(t)$, but $\mathbf{H}$ is shared (albeit shifted) across all t's. It is possible to update $\mathbf{W}(t)$ and $\mathbf{H}$ for each t, however this will result in a biased estimate of $\mathbf{H}$ with the update for $t = T-1$ dominating over others. Therefore it is best to update all $\mathbf{W}(t)$ first and then assign to $\mathbf{H}$ the average of all the NMF subproblems:

$$\mathbf{H} = \left\langle \mathbf{H} \odot \frac{\mathbf{W}(t)^\top \cdot \overset{\leftarrow t}{\left[\frac{\mathbf{V}}{\hat{\mathbf{V}}}\right]}}{\mathbf{W}(t)^\top \cdot \mathbf{1}} \right\rangle, \forall t \tag{7}$$

In terms of computational complexity this technique depends mostly on T. If $T = 1$ then it reduces to standard NMF, otherwise it is burdened with extra matrix updates equivalent to one NMF per unit of T.

Some examples of convolutive NMF analysis are presented in [13] and [15]. In these papers the appropriateness of a convolutive model for the analysis of sounds is demonstrated using a variety of audio signals. It is shown that this type of analysis is good at finding the salient spectral sequences contained in auditory scenes and can be further employed to extract them. In the following section we will examine the results of convolutive NMF analysis as applied on speech signals. Unlike previous attempts we will not attempt to extract large sequences like words or entire sounds as reported before, but rather smaller segments which can represent the building blocks of a speech.

III. Convolutive NMF on speech spectra

In this section we will be presenting some results on speech signals which reveal the nature of the convolutive NMF components. We will show that the extracted bases are akin to speech phones with various pitch inflections. We will discover that qualitatively similar bases are extracted whether the input is a single speaker, or a mixture of multiple speakers (an important observation which we will take advantage of later on), and that the bases encode a lot of information about the speakers and naturally reflect the speakers' particular speech patterns.

Due to its non-negative nature this type of basis approximation is best suited for representing magnitude spectra. Therefore to apply this analysis on speech signals we will be operating in the magnitude frequency domain. Starting with a finite length monophonic speech signal $s(t)$ we denote its short-time magnitude spectrum as $F(\omega, t)$, containing at each element (ω, t) the energy of frequency ω at time t. Viewing each instance of $F(\omega, t)$ as an element of a matrix $\mathbf{F}$ we now have a non-negative set of data on which we can apply convolutive NMF.

To illustrate the nature of the convolutive NMF bases we performed this process on a 28 second speech signal from the TIMIT database (speaker `mwbt0`) sampled at $16kHz$. We used an $L = 1024$ point spectrum which resulted into 513 distinct frequency magnitudes, t was advanced by 256 samples at a time and before the DFT we applied a hanning window on the time-domain signal to reduce the presence of sidelobes. We extracted $R = 40$ components with a time span $T = 8$, which roughly amounts to a 0.17 second time span for the bases. The results after 200 iterations are shown in figure 1.

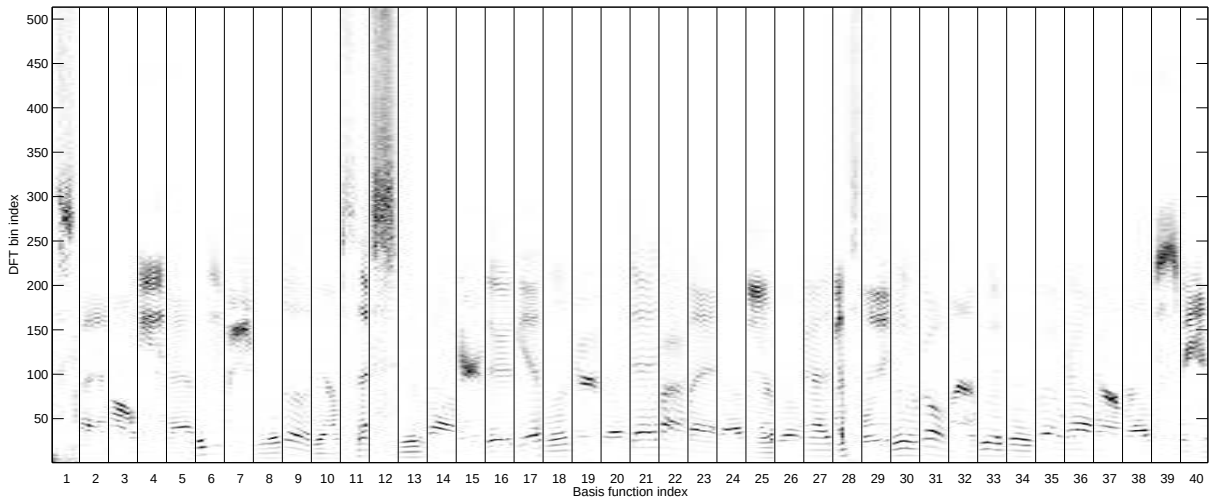

Fig. 1. *Basis functions derived from the magnitude spectra of a single speaker. Each basis function resembles phone-like components of the analyzed magnitude spectrogram.*

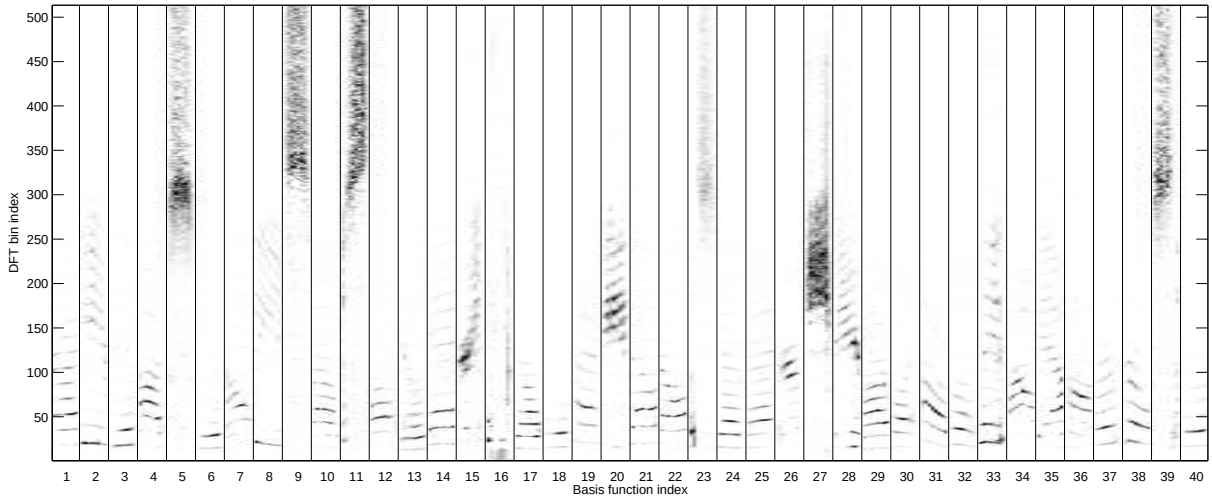

Fig. 2. *Basis functions derived from the magnitude spectra of a single speaker. Unlike the bases in figure 1, these were produced from a female speaker. Note how the higher fundamental of the female voice is reflected in the bases when compared to the male speaker in figure 1.*

Note how the bases are roughly corresponding to speech phone instances. Most bases are representing harmonic series with various pitch inflections, and a smaller subset contains wideband elements that correspond to consonant sounds. Audible reconstructions can be done by modulating the phase of the original input by the magnitude of a selected a selected basis. Doing so and listening at the results verifies that most bases sound like short speech phones. There are some bases left with the burden of representing signal portions that the rest of the bases do not reflect, these few usually have some compound nature combining various speech elements occasionally with some noise to approximate background or spurious portions of the signal.

As should be expected these basis functions are reflecting the acoustical characteristics of the speaker that was analyzed. To illustrate this consider figure 1's equivalent basis set from another TIMIT speaker (speaker `faks0`). All other training parameters are the same as in the previous example. This basis set is shown in figure 2. Qualitatively the bases are similar, however after close inspection it is quite evident that they reflect key differences between the two analyzed speakers. Most notably we can see the harmonics in figure 1 being spaced closer indicating a lower pitched voice, as compared to the bases in figure 2 where the harmonics are farther apart from each other indicating a higher pitched voice. A keen observer can also pick up some formant differences between the two speakers. Noting that the two speakers that were used were a male and a female speaker, we see that the extracted bases are indeed coding speaker dependent characteristics.

Due to the linearity of the algorithm and the fact that the magnitude spectra are roughly added in the case of monophonic mixtures, we should expect to get qualitatively the same results when the input is a mixture of multiple speakers instead of a single speaker. Unless there are spurious correlations between the speakers in a mixture it is natural to theorize that the set of bases extracted from the mixture will contain bases describing each isolated speaker.

We repeat the experiment with a mixture input made by summing the previously used speakers. The results are as we predicted and are shown in figure 3. We can see some of the mixture basis functions resembling the bases of the male speaker (figure 1), whereas other resemble the bases of the female speaker (figure 2). Naturally our assumption that magnitude spectra mix linearly is not precisely correct, but at a more general and qualitative level this assumption is approximately true for most sound mixtures. As reported in [10] a binary mask is often sufficient to maintain source information in sound mixtures, and our assumption is a generalization of that which allows some degree of approximate additivity between the mixed spectra.

The model that this approach imposes on the data is that of a convolutive basis function. In usual basis function expansions, like NMF, we have a set of bases (corresponding to spectra in our case) being scaled by a set of weights. In the convolutive case we have a set of bases that correspond to patches of a spectrogram which are convolved along the time axis according to their weights in order to reconstruct the input. The underlying assumption is that the inputs can be adequately described by a set of these patches. This is the case in speech where repeating patterns are often reused, but it is also the case for other types of sounds that exhibit a regular temporal structure in their spectrograms.

The number of bases R that we request is not particularly important in this context. If R is too small then the basis functions will be forced to approximate simultaneous clumps of speech phones resulting in worse reconstruction performance and a more blurry distinction between the bases. For a large value of R we can see certain bases adapting to individual harmonics as opposed to entire phonemic structures. In general a value between 100 to 500 bases is usually a good estimate for a rich in phonetic content speech input.

IV. Separation of known speakers

In this section we will introduce a way to take advantage of the basis functions we just introduced to perform separation from monophonic mixtures of known speakers. We will show that once the basis functions of a speaker are known they can be used to reconstruct only that speaker's signal from a monophonic mixture. We will first describe the methodology and then present results from our experiments.

A. Extracting speaker dependent bases from a mixture

As we showed in the previous section the basis functions we extract from speech are dependent on the timbral characteristics of the speaker who provided the training data. We would expect the learned bases to characterize that speaker best. Furthermore when we analyze a magnitude spectrogram which is generated from a mixture of speakers the basis functions are still resembling individual phones from all included speakers. Groups of these bases can be attributed to only one speaker respectively. If we could reconstruct the mixture magnitude spectrogram using only the bases that correspond to one speaker in the mixture we could be effectively performing separation. Performing this in an unsupervised manner is a rather complicated process, however if we have a sufficient set of learned basis functions from a specific speaker we can use these bases to extract that speaker's voice from sound mixtures.

Consider a mixture of the male and female speaker we used in the preceding section. Based on the observations from the previous section we can assume that learning a set of bases $\mathbf{W}_m(t)$ from the male speaker and a set of bases $\mathbf{W}_f(t)$ from the female speaker will roughly resemble the set of bases learned from a mixture of their voices. This means that $\mathbf{W}_m(t)$ and $\mathbf{W}_f(t)$ can be used to reconstruct the mixture. If this is the case then we can assume that the part of the mixture that is reconstructed by $\mathbf{W}_m(t)$ will predominantly rebuild the male voice and $\mathbf{W}_f(t)$ the female voice, thereby providing the spectrograms from each speaker that we can easily invert back to time signals.

So to formalize and generalize for N speakers we take the following steps:

1) Obtain training data $x_i(t)$ for each speaker and separately derive convolutive NMF bases $\mathbf{W}_i(t)$ from their magnitude spectrograms $\mathbf{X}_i$ using the methodology in section III.
2) Construct a union of all the bases by combining them: $\mathbf{W}(t) = \mathbf{W}_1(t) \bigcup \mathbf{W}_2(t) \bigcup ... \bigcup \mathbf{W}_N(t)$. This will result in a basis set N times bigger than the individual speaker sets.
3) Take a mixture $y(t)$ containing the learned speakers uttering an unknown phrase $y_i(t)$. Obtain its magnitude spectrogram $\mathbf{Y}$ and perform convolutive NMF training on it. During training keep the

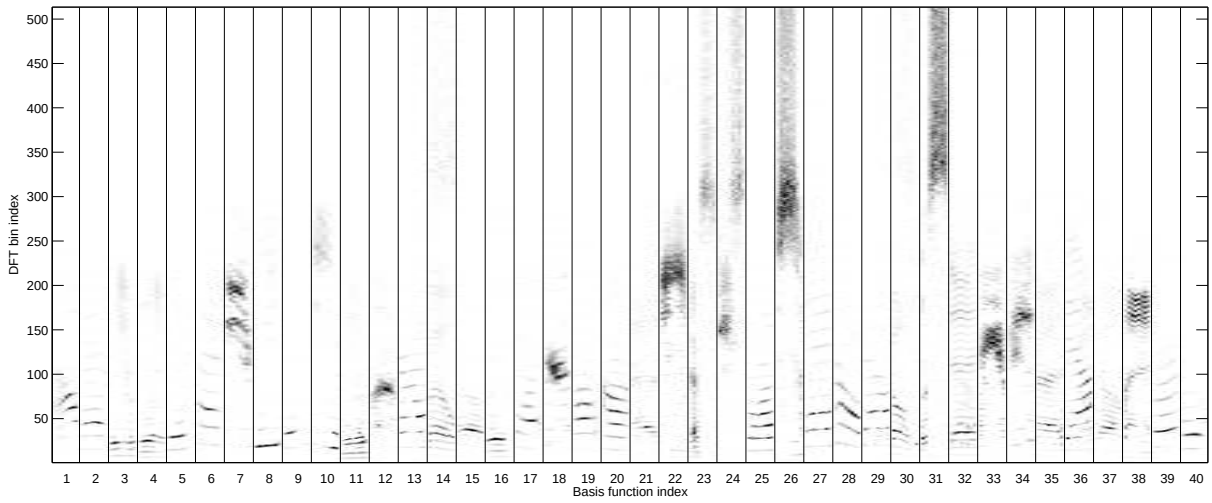

Fig. 3. *Basis functions derived from the magnitude spectra of a two-speaker mixture. Note how some of the bases seem to fit best the female speaker whereas others fit the male.*

bases fixed to $\mathbf{W}(t)$ and learn only their weights $\mathbf{H}$.

4) Break up $\mathbf{H}$ into N parts $\mathbf{H}_i$ each corresponding to the weights that belong to a single speaker's bases. This will result into N sets of weights.
5) Reconstruct the magnitude spectrogram $\mathbf{Y}$ of the analyzed mixture using only an individual speaker bases and weights: $\mathbf{Z}_i = \sum_{t=0}^{T-1} \mathbf{W}_i(t) \cdot \overset{t\rightarrow}{\mathbf{H}_i}(t)$. Each $\mathbf{Z}_i$ will be a magnitude spectrogram containing parts of the mixture that were best explained by bases from speaker i.
6) Use the phase data from the original mixture and modulate it by $\mathbf{Z}_i$ to obtain N spectrograms for each speaker.
7) Using the inverse short-time Fourier transform transform the speaker spectrograms to the time domain and obtain the extracted speech signals $z_i(t)$.

The signals $z_i(t)$ will be approximations of $y_i(t)$ since they are constructed from bases belonging only to these speakers. This of course presupposes that the speakers have discernible voices and somewhat different timbral character and pitch inflections, which seemed to be usually the case in our experience.

B. Mixing and evaluating methodology

In this section we will describe the steps we took while conducting our experiments on speech separation. We describe the construction of the mixing cases and their evaluation.

To test this approach monophonic mixtures were synthetically generated by summing two different but roughly equal length sentences from different speakers from the TIMIT database. These sentences were normalized to unit variance so that when added would produce a $0dB$ mixture. The remaining sentences of the two used speakers in the TIMIT database were used as training data from which we derived a basis set for each speaker. The training data were in the order of 30 seconds of continuous speech per speaker, whereas the evaluation sentences were 2 to 3 seconds long. As in the above examples the sample rate was $16kHz$.

Evaluation of the quality of speech separation algorithms is always a very hard task and the non-linear unmixing procedure we propose is especially hard to evaluate reliably. In order to provide a comparable measure with existing literature we will be using standard correlation-based measurements. For each separation example we will provide three types of performance indexes, the signal to noise ratio for each extracted speaker, the log correlation of the extracted source with the original, and the amount of unaccounted energy in the original and extracted signals.

Using the notation introduced in the preceding section, once we have the separated speaker sounds $z_i(t)$ we compare them with the original mixed sources $y_i(t)$ to see how good the separation is. We derive three measures to measure performance, the speaker ratio in dB, the similarity of the output with the target, and the residual noise. The speaker ratio is computed by comparing the correlations of the original sources with the extracted sounds:

$$c_{i,j} = cor\left(z_i(t), y_i(t)\right) \tag{8}$$

Where $cor(\cdot)$ denotes correlation. We define the speaker ratio for each output as a log ratio of its correla-

tion with the desired sentence divided by the correlation with the other sentences, i.e.:

$$SR_i = 10 \log_{10} \frac{c_{i,i}}{\sum_{\forall j \neq i} c_{i,j}} \quad (9)$$

This measure will tell us how much the signals of the undesired speakers have been suppressed. Higher values will reveal better extraction of the desired speaker.

The similarity index measures how much the output resembles the desired output. We measure it by taking the correlation of the extracted source with the desired output:

$$SI_i = 10 \log_{10} cor\left(z_i(t), y_i(t)\right) \quad (10)$$

We will have $SI_i <= 0$, with $SI_i = 0$ being the most desired case. Lower values indicate that the result is not too similar to the desired sentence. Note that this measure also is influenced by the quality of separation since traits of the undesired speakers would lower its value.

Finally the residual energy is the variance of the difference between the input signal and the sum of the extracted signals:

$$RE = var\left(\sum_i y_i(t) - \sum_i z_i(t)\right) \quad (11)$$

RE will reveal how much of the output signal is not accountable by any of the original sounds, and how much is an artifact of the separation procedure. Values closer to zero are best, indicating good accountability of the input signal and little or no residual noise.

One thing to note is that the SNR measurements, although sufficient, are not directly indicative of the performance of this algorithm. The separation process is very complex and non-linear and the SNR measurement will only provide a standardized way of evaluating it. More appropriate performance measures would be in terms of the cost function being optimized by the convolutive NMF procedure. However interpreting these metrics would be obscure at best and would not provide values amenable to comparisons with other approaches.

V. Separation Results

In this section we present some results from speech mixtures and shed some light on the importance of various parameters involved in this process. We averaged the results from a set of twelve runs from each of eight randomly selected male/female pairs of speakers from the TIMIT database and attempted separation using the aforementioned process. The parameters that were used can be divided in two groups. One group included the parameters relating to the short time Fourier transform: the FFT size, the transform hop size, the zero padding for the FFT and the analysis window. The other group of parameters were the ones relating to the convolutive NMF training: the number of bases R, their extent in time T, and the training iterations. Although there are plenty of parameters to compare, the most important ones were the size of the FFT used, the number of basis functions R and their temporal extent T. At first we will examine these three and then present some additional results exhibiting the effect of the rest of the parameters.

A. Most Important Parameters

For this set of results we will assume that the STFT hop size was set to one fourth of the FFT size, that zero padding was not used, before the FFT the data was scaled according to a Hanning window, and that we estimated the bases and their weights for two hundred iterations. For the remaining parameters we used the following values, FFT size = [128 256 512 1024 2048 4096], Number of Bases = [20 40 80 120 200], Length of Bases = [1 2 4 6 8 10]. The sampling rate of the inputs was $16kHz$. We performed separation using all combinations of these parameter on our data set which amounted to 180 experiments for each of the eight speaker pairs (1440 runs over all speakers, repeated 12 times for a total of 17280 experiments). We averaged the performance measures for all these experiments and analyzed the effect of various parameters. We present our findings in this section.

Of major importance is the size of the FFT we use to analyze our inputs. If we average the results over all other parameters and speakers for each FFT size we can then observe its effect on the speaker energy ratio (figure 4 left). This value fluctuates from about $2.5dB$ for an FFT size of 128 points (8 ms) at worse, to about $4.8dB$ for 1024 points (64 ms). FFT sizes outside 1024 points tend to produce progressively worse results indicating that this is a good value for this parameter. The similarity index behaves in a similar way, the optimal average comes out for an FFT size of 512 points (32 ms) with an approximate value of -0.7. The similarity index progressively deteriorates for diverging FFT sizes with the recorded worst being 4096 points with a value of about -1.2. Deterioration seems to be more rapid for larger FFT frame values. Residual energy is roughly increasing with the FFT size. This is to be expected, short time windows provide small building blocks which can fit the data well without extending their errors to

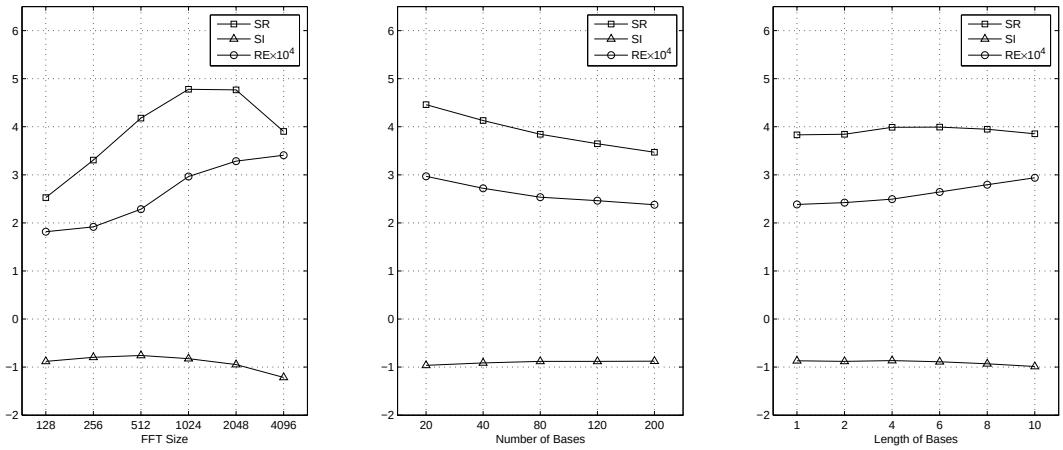

Fig. 4. Effects of some parameters to the performance indexes. Each point is the average of all other cases when fixing a single parameter value. Note that for readability reasons the value of the residual noise energy has been scaled by 10^4. The left plot displays the effect of the FFT size, the middle plot the effect of the number of bases R and the right one the effect of the length of the bases T.

too wide a time window, wider analysis windows are extending over larger time periods having to fit much more information. Using bigger analysis windows results into a coarser approximation. Values ranged from an average of about 1.8×10^{-4} for 128 point FFTs to about 3.4×10^{-4} for 4096 point FFTs.

The number of bases is obviously also a major parameter. Regarding its effect on the speaker energy ratio, we generally observe that fewer bases provide a better result (figure 4 center). The learning of more bases for each speaker allows a greater expressive power which can model utterances of other speakers as well. Having too many bases will result in some reconstruction of the interfering signal which can negatively affect signal separation. Best results were obtained for 20 bases per speaker with about $4.5dB$ separation, and after a steady decline the worst value was at $3.5dB$ for 200 bases per speaker. However using less bases comes at a price since reconstruction results in a more coarse fit which then exhibits poor values for the similarity index and residual noise. We see the residual noise dropped monotonically as we moved from 20 bases to 200 bases per speaker taking values ranging from 3×10^{-4} to 2.4×10^{-4}. Inversely the similarity index rose from about -1 to $-.9$ So we see a tradeoff, at the expense of a worsening speaker energy ratio, adding more bases resulted in an increase in the similarity index and a decrease in the residual noise.

The length of the bases is another important factor. On average a value of 4 to 6 produced the best separation results, although only by a minor margin with worst values at around $3.8dB$ and best at around $4dB$ (figure 4 right). Residual noise energy tended to increase with longer basis lengths ranging from 2.6×10^{-4} for a length of 1 to 3×10^{-4} for a length of 10. That was expected because it will be more difficult to use longer bases and still have them precisely fit on the evaluation data. The similarity was at its best around a basis length of 4 with a value of -.9, and tended to fall for greater basis lengths down to -1 for a basis length of 10. Note that when the length of the bases is 1 then we are essentially performing NMF.

Of more importance than the individual parameters is the interaction between them. Figure 5 is instrumental in pointing this out. We briefly describe some of the major interactions here. As we pointed out in the previous paragraph, the length of the bases was not a parameter that varied the performance measures significantly. However we can see that its effect was heavily dependent on the FFT size. In general for long FFTs we saw better separation for short bases regardless of how many we use, and for shorter FFTs we saw better performance for few bases regardless of their length. For small FFT sizes and long bases we obtained better separation and similarity, however the amount of residual energy increased. As the FFT size grows, longer bases acted as a detriment to the separation quality as well as to the similarity index, therefore in this case shorter bases are to be preferred. The residual energy was heavily dependent on the number of bases. In general more bases introduced more residual noise, although that effect is not as pronounced for larger FFT sizes. For larger FFT sizes though the residual energy increased significantly for longer bases. Finally as the FFT size increased we noted that the number of bases become more important to the effect of the length of the bases with regard to similarity.

The results shown in this section are suppressed due to

averaging with poor parameter selection. In most speaker set cases boosting up to $6dB$ was achieved for each speaker, and proper post-processing (section VI), can boost this to double digit dB improvements.

B. Remaining Parameters

In this section we will examine the remaining parameters, such the number of iterations for the convolutive NMF training, the STFT hop, zero padding size and analysis window. Their interplay is not significant so we examine them independently. For our experiments we used an FFT size of 1024 and 40 basis functions which extended for 4 time points. The results shown for each parameter are averages over 12 independent runs on each of 8 pairs of speakers.

The performance indexes for various values of these parameters are shown in figures 6 and 7. We note that the in general denser packed FFT windows facilitated better separation. This is because such denser sampling of time frames helped develop more time invariance in the basis set, provided a richer data set and bypassed alignment problems. This seemed to come at a cost though since it introduced more computational requirements due to a larger training set and it also resulted into poor fitting for extreme values. We can see that in the case where the STFT window hop size was $1/8$th that of the FFT size where performance seems to degrade. The excessive amount of data to learn and fit introduced a computational complexity which presumably required more training to reach equal results as larger hop sizes. In the case of zero padding, we note that in general it is a bad idea since it increased the dimensionality of the input dramatically and it didn't seem to offer any particular performance advantage.

In figure 7 we show the performance effects of the adaptation iterations of convolutive NMF. These can be separated into two groups, the training iterations and the approximation iterations. Training iterations are the number of iterations we train on each speaker, whereas approximation iterations are the number of iterations we perform to adapt the speaker bases to a mixture. The effect of training iterations displayed an interesting trend. Early on we got the best dB improvement in separation, albeit at a cost of high residual noise. As we kept iterating the separation index dropped as did the residual noise whereas the similarity index was more or less stabilized. We saw that around 100 iterations performance stabilized and further training was unnecessary (that is on training on 30 seconds of speech, larger training sets would require more training to reach that state). These effects can be explained by the fact that due to the rapid

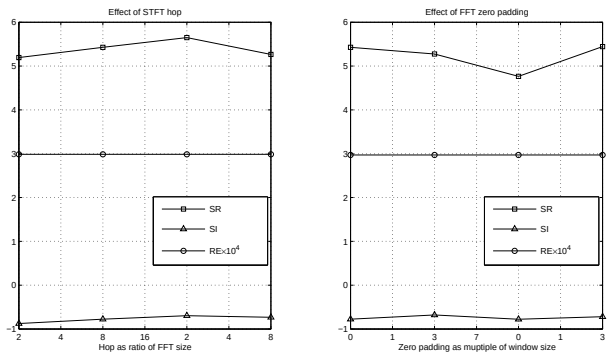

Fig. 6. The effect of some FFT parameters on the performance of separation. In these plots we examine parameters that do not create a significant performance changes. On the left plot we see the effects of the STFT hop size, in the right plot the effects of the FFT zero padding.

convergence of NMF training the most salient features of each speaker were discovered very early whereas later on they were refined to include more generic speech features that all speakers might exhibit (such as consonants). As training progressed the bases became more refined so that reconstruction was more effective and that drove down the amount of residual noise.

The approximation iterations are more predictable in their performance effect. The more we trained the better the separation and the less the residual noise. We also note an interesting trend for the similarity index. We note that it peaked at around 60 iterations (that is on a 3 seconds mixture, lengthier mixtures would delay this), and then asymptotically decreased. However the peak of the similarity index didn't coincide with the peak in separation quality. The justification for the separation and residual noise trend is obvious, the more we iterated the better the fit. The trend of the similarity index can be explained by the fact that prolonged training would result into more cross-pollination of speaker bases which can alter the characteristics of each speaker.

The effect of the window type is rather negligible, it seemed best to not use a rectangular window since it resulted into noisier basis functions and about $2dB$ to $3dB$ worst separation. Aside from this the selection of window type is not an issue in performance.

C. Denoising Examples

Given that the separation procedure we introduce is based on finding and extracting sound elements that compose the input sources, separation between multiple speakers is obviously a harder example since the sources are very similar. In this section we briefly consider the case of denoising where the types of sources are quite distinct and can be seen as composed from a non-

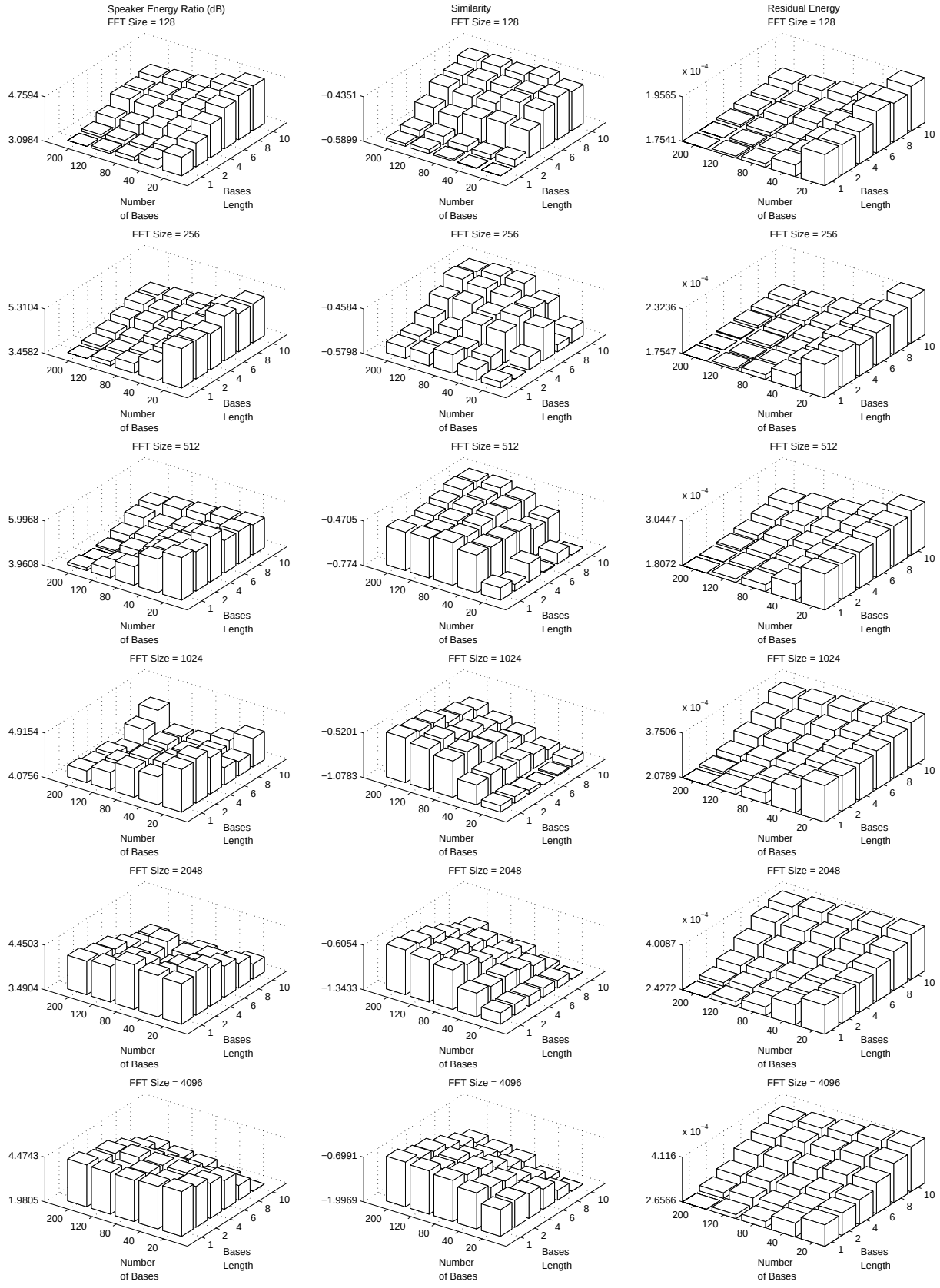

Fig. 5. Performance measures for combinations of major parameters. Each row of plots is for a different FFT size as denoted over each plot plots. Each column is a different performance measure. First column is the speaker ratio in dB, second column the similarity index, and third column is the residual energy. Note that all plots are plotted across different scales indicated by their maximum and minimum values along the vertical axis.

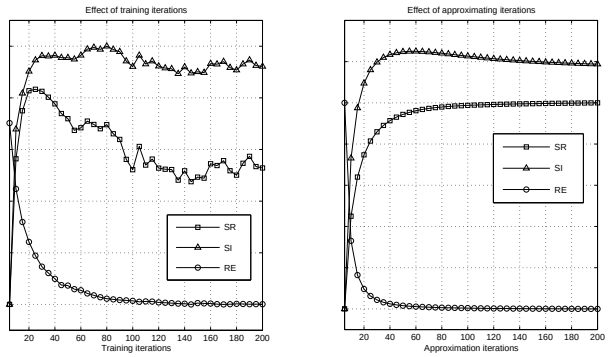

Fig. 7. The effect of training and approximating iterations on the performance of the separation. The left panel displays the effect of the training iterations, whereas the right one displays the effect of the mixture approximation iterations. Best separation results are achieved early on in training, however at the expense of high residual noise. Note that the three performance measures are not drawn on the same scale and the plots are only informative of the measures progression during iterating.

overlapping set of bases. We briefly discuss a couple of examples in this section.

We generated two mixtures that were composed out of one speaker (`mwbt0`) and either ambient street noise or chimes. For both cases we used an FFT size of 1024 points and 40 bases of length 4. In order to perform the separation we learned models of both the speaker and the interfering sounds. In the first example speech was boosted to $5.2dB$ over the street noise. The noise signal was composed out of background speech, a street performed playing accordion and high frequency ambience. The accordion and the ambience were suppressed to inaudible levels, speech babble was suppressed but not as much. This is due to the fact that speech babble had a lot of common spectral content with the speaker and some of it could also be explained by the speaker's bases. In the second example the separation was up to $16.3dB$. The much better result is attributed to the fact that the chimes had a spectral character that had almost no spectral overlap from the speaker's voice which facilitated separation. Although not a primary objective, for both mixtures the street noise and the chimes were also well separated with minimal traces of the speaker. Figure 8 displays the spectrograms of the two mixtures and the separated speech. In these spectrograms we can clearly see elements of the street noise as well as the chimes being suppressed. Audible reconstructions of the extracted sounds exhibit an excellent degree of separation and a minimal degradation of quality due to the low rank encoding of the signal.

D. General remarks

On average the resulting separated sounds sounded very much like linear mixtures of various mixing proportions. Depending on the success of the separation these ranged from slightly perceptible to significant interference. There were no echo residues or spectral coloring imposed by the algorithm, and thus the speaker ratio proved to be an adequate indicator to measure separation performance. Results that were obtained using too few bases often had missing speech phones (usually wideband ones) or a muffled quality to them, this was not the case when using more bases. This effect was measured effectively by the sound similarity indicator which correlated well with subjective listening evaluations. The use of too few bases in addition to a large T also contributed to some hissing or scratching noise which is most likely a product of a poor tapering across frames in the frequency magnitude domain which in turn produced subtle discontinuities in the time domain signal. These were usually reflected in the residual noise measure. Once again depending to the settings used this effect ranged from imperceptible to noticeable. These are all the audible artifacts that we encountered during testing and driven by that we designed the appropriate performance evaluation measures. Although there is some correlation between some of the performance measures, in average (and in our subjective opinion) they did a fairly good job in describing the audible result quantitatively.

From figures 4 and 5 one can note that a larger T is not significantly beneficial. Although that is true in the sense conveyed by the performance parameters, it is important to note that the extracted features are much more informative when $T > 1$. Consider the case of a fixed analysis window. In the case where $T = 1$ we lose a lot of temporal information which is carried by the phase of the spectra. We essentially represent the data using a single magnitude spectrum. For the same length window when $T > 1$ the previously lost temporal evolution information is now conveyed by the extracted bases. The resulting trade-off is that we can obtain a single long spectrum, or a series of shorter spectra that describes common patterns of spectral evolutions. Although this is not an important factor for separation, it is valuable when we subsequently need to perform speech or sound recognition and the extracted features need to be maximally informative and unique. The value of the extracted bases is of course context specific, but in extreme cases we can even extract entire words per basis using similar processing [14].

As mentioned before there needs to be some spectral

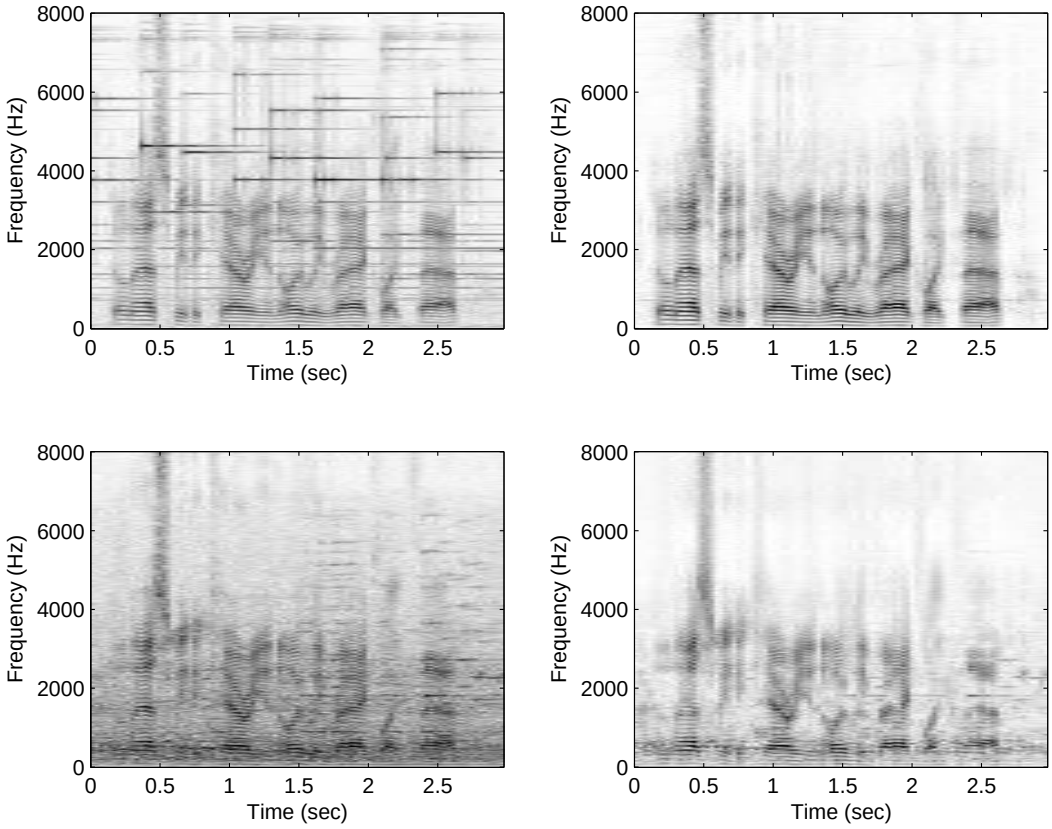

Fig. 8. Spectrograms of the denoising examples. The two left spectrograms display the two mixtures, the top one with chimes and the bottom one with the street noise. The two right spectrograms show the respective speech separation. On the top spectrogram one can see the chimes harmonics being virtually non-existent. On the bottom separation spectrogram one can see much of the ambience and the accordion harmonics being suppressed.

difference between the sounds that need to be separated in order to obtain good performance. In our speech experiments we used female/male sets of speakers to loosely ensure some spectral variation. Separation between male/male sets produces in general slightly worse results since the spectra to be extracted have more similarity (although that is highly dependent on the speaker character and does not in general mean that all male/male pairs will be harder problems than female/male). To further stress the importance of spectral dissimilarity note the dramatic improvement of separation quality when using non-spectrally similar sounds as we have done in section V-C.

The issue of spectral similarity between sources is an important one which needs to be studied in more depth. The side effect of dealing with spectrally similar sounds is that their resulting bases are most likely linearly dependent and thereby inhibit good separation. Various ad-hoc steps can be taken to reduce that effect, such as discarding very similar bases across the different sounds, or reassignment of bases to remove dependencies, which can result in better sounding results but are not satisfying solutions. The fact that the bases need to be non-negative complicates any straightforward processing that can to be done to ensure maximal linear independence and points to a non-trivial solution. Employing longer bases so that $T > 1$ provides a minor relief in this respect since the issue becomes one of a 'spectral sequence similarity' which is a less likely possibility between two sounds.

Finally a note about reverberation. Because of this type of analysis, factors such as echoes can be safely ignored since their effect will be undone by the implicit deconvolution in training (long echoes will result in repeating sections of the magnitude spectrogram which is what this algorithm is designed to discover). Shorter echoes will not be as present in the magnitude spectrum, but more so in the phase spectrum which we discard, and thus does not pose an issue. If these echoes are strong enough to color the magnitude spectrum then they will be learned as part of the characteristics of the input sound, but will not interfere with the learning. In extreme cases we can have an unnatural amount of spectral smearing due to reverberation which can make separation impossible. Such cases are rarely encountered

though and even then the most acute listeners or systems can have a hard time telling different sounds apart.

VI. Post-Processing

Although the process described so far can obtain reasonable results in separating sounds, it does not need to be the end of the separating process. We can use various post-processing techniques on the outputs to boost the quality of the results even more. In this section we briefly describe a couple of possible approaches.

A trivial step that we can employ to improve the quality of the results is to modify the reconstruction step to account for all the energy in our testing signals. Recall that we used the phase of the mixture and modulated it with the magnitudes dictated by the basis approximation. Energy from the mixture that was not approximated well will be missing from the reconstructions resulting in a choppy sounding output. This can be remedied by longer training in the approximation step, but we have observed that this can sometimes be a detriment in separation quality and also a computationally intensive process. Instead we can compute the spectrogram of each speaker as:

$$\mathbf{F}_i = (\angle \mathbf{H}) \cdot \frac{\mathbf{Z}_i}{\sum_j \mathbf{Z}_j} \tag{12}$$

where $\angle \mathbf{H}$ is the phase of the mixture spectrogram and $\mathbf{Z}_i$ is the approximated magnitude spectrogram of each speaker. This is essentially a spectral filter that ensures that the unaccounted energy in the input mixture is redistributed to the resulting speakers' spectrograms in proportion. This results in definitely better sounding reconstructions in terms of quality, something we can note in the performance indexes. The separation quality remained the same, however the similarity was improved a lot and residual energy was not only reduced but also stabilized to a fixed low value for all other parameters.

A notable point of this approach is that it starts with a monophonic mixture and results into a multi-channel output. Had the separation been perfect there would be no need for post-processing, however the separation is not always satisfactory and in that case we can view the outputs as a multi-channel mixture. This transition from monophonic to multi-channel opens the possibility of employing multi-channel separation techniques to further separate the sources. Although this is not a linear mixture anymore, an application of straightforward unmixing algorithms like Independent Component Analysis can be applied to it and can provide on average a boost of about $5dB$ to $7dB$ (we employed the JADE algorithm [16] in our simulations). Given the non-linear relationship of the two resulting sources this is not necessarily as notable an improvement as the numbers indicate, it is however noticeable, and could be improved with a specialized unmixing approach based on this type of mixture. It should be pointed that this is quite an ad-hoc step and there is no guarantee that an application of ICA will work at all, since the resulting mixtures will be non-linearly mixed.

Either of the above approaches are fairly straightforward and generic and are not meant as ultimate solutions. They show however that there can be a considerable improvement of results after the convolutive NMF approximation and opens up an interesting avenue of research.

VII. Conclusions

In this paper we have presented a supervised method for separating known types of sounds from monophonic mixtures. We introduced the concept of a convolutive non-negative basis set, demonstrated how it maps to meaningful features in the case of audio spectra and demonstrated how we can use it in the context of supervised source separation. We also provided simulation material which can provide some intuition about the importance of various parameters and suggested a couple of ways this process can be enhanced using post-processing. Depending on the nature of the inputs and the post-processing employed we obtained interference suppression ranging from $5dB$ in the worst cases to up to $20dB$. As described there are numerous options to trade separation performance for better audio quality results, and this is a choice which to our experience has been application dependent. There seem to be many ways this approach can be enhanced and this article only attempted to present a basic implementation and its operational characteristics, it is our hope that future work will address additional performance enhancing extensions.

Acknowledgment

The author would like to thank Bhiksha Raj of Mitsubishi Electric Research Laboratories for fruitful discussions while preparing this paper, as well as the assigned reviewers for pointing out areas that needed development.

REFERENCES

[1] Abdallah S.A. and Plumbley, M.D. ”Polyphonic transcription by non-negative sparse coding of power spectra”, in Proceedings of the 5th International Conference on Music Information Retrieval 2004.

[2] Casey, M. and Westner, A. ”Separation of Mixed Audio Sources by Independent Subspace Analysis”, in the proceedings of the International Computer Music Conference 2000.

[3] Hyvärinen, A. ”Survey on Independent Component Analysis”, in Neural Computing Surveys 2:94–128, 1999.

[4] Jackson, J.E. ”A User’s Guide to Principal Components”, Wiley-Interscience paperback series, ISBN: 0-471-47134-8.

[5] De Lathauwer L., De Moor B., Vandewalle J. ”A multilinear singular value decomposition”, in SIAM Journal on Matrix Analysis and Applications, vol. 21, no. 4, Apr. 2000, pp. 1253-1278.

[6] Lee, D.D., Seung, H.S. ”Algorithms for non-negative matrix factorization”, in Advances in Neural Information Processing Systems 13, 2000.

[7] Brown J.C. and Smaragdis, P., ”Non-negative Matrix Factorization for Polyphonic Music Transcription”, in Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 177-180, October 2003

[8] Lee, D.D., Seung, H.S. ”Learning the Parts of Objects by Non-negative Matrix Factorization”, in Nature 1999 (401):788.

[9] Reyes-Gomez, M.J., Raj, B. and Ellis, D.P.W. ”Multi-channel source separation by factorial HMMs”,in the Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[10] Roweis, S.T. ”One Microphone Source Separation”, in Neural Information Processing Systems 13, 2000.

[11] Roweis, S.T. ”Factorial Models and Refiltering for Speech Separation and Denoising”, in Proceedings of Eurospeech, 2003.

[12] Smaragdis, P. ”Redundancy Reduction for Computational Audition, a Unifying Approach”, Doctoral dissertation, Massachusetts Institute of Technology, Media Laboratory 2001.

[13] Smaragdis, P. ”Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs”, in International Symposium on Independent Component Analysis and Blind Signal Separation, September 2004.

[14] Smaragdis, P., ”Discovering Auditory Objects Through Non-Negativity Constraints”, in the proceedings of Statistical and Perceptual Audio Processing (SAPA), SAPA 2004, October 2004.

[15] Virtanen, T. ”Sound Source Separation Using Sparse Coding with Temporal Continuity Objective”, in the proceedings of the International Computer Music Conference 2003.

[16] Cardoso, J-F. ”High-order contrasts for independent component analysis”, in Neural Computation 11(1):157–192, 1999.