Joint Sound Source Separation and Speaker Recognition

Jeroen Zegers, Hugo Van hamme

KU Leuven, Dept. ESAT, Belgium

Abstract

Non-negative Matrix Factorization (NMF) has already been applied to learn speaker characterizations from single or non-simultaneous speech for speaker recognition applications. It is also known for its good performance in (blind) source separation for simultaneous speech. This paper explains how NMF can be used to jointly solve the two problems in a multichannel speaker recognizer for simultaneous speech. It is shown how state-of-the-art multichannel NMF for blind source separation can be easily extended to incorporate speaker recognition. Experiments on the CHiME corpus show that this method outperforms the sequential approach of first applying source separation, followed by speaker recognition that uses state-of-the-art i-vector techniques.

Index Terms: speaker recognition, source separation, non-negative matrix factorization, multichannel

1. Introduction

Non-negative Matrix Factorization (NMF) is a frequently used method to identify patterns in data. It was originally developed to recognize parts-based representation in images [1] and has since been used in various fields such as bioinformatics [2], noise-robust automatic speech recognition [3] and age and gender estimation [4]. Recently it has been used for speaker recognition (SR) tasks and obtained comparable results to state-of-the-art i-vector based approaches [5, 6]. Moreover, NMF has been used in scenarios of source separation (SS) and music transcription [7, 8]. Multichannel extensions of NMF have been developed with applications to blind source separation [9, 10]. These approaches combine spatial cues from phase differences between microphones and the segmentation strength from NMF without any prior knowledge of the sources.

State-of-the-art speaker recognition systems are made robust to speaker and inter-channel variability by determining a total variability space [11]. Also, attempts have been made at robustness to noise and reverberation [12]. However, these models cannot cope with simultaneous speech, where multiple speakers are active at the same time. In such a case one has to resort to a solution where first source separation is applied, followed by a single speaker recognizer [13]. Overlapping speech occurs frequently in conversations. For example, in a telephone conversation, a meeting, a panel discussion or natural speech in general. Instead of solving the problems of source separation and speaker recognition sequentially, it is shown how NMF is inherently capable of solving these two problems jointly. In this paper synthetic convolutive mixtures are used to analyze the problem.

This paper is organized as follows. Section 2 explains the basics of NMF and how it can be used in both SR and (multichannel) SS. Section 3 shows how NMF can jointly solve these problems. Experiments are explained in section 4 and a conclusion is given in section 5.

2. Non-Negative Matrix Factorization

2.1. Principle

Non-negative matrix factorization is a factorization method that approximates a non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{F\times N}$ using a non-negative dictionary matrix $\mathbf{T} \in \mathbb{R}_+^{F\times K}$ and a non-negative activation matrix $\mathbf{V} \in \mathbb{R}_+^{K\times N}$, such that $\mathbf{X} \approx \hat{\mathbf{X}} \triangleq \mathbf{TV}$. In our application $\mathbf{X} = |\tilde{\mathbf{X}}|^{\cdot 2}$ is a speech power spectrogram, with $\tilde{\mathbf{X}}$ the complex valued short time Fourier transform (STFT) of the audio signal, $|.|$ the absolute value and $^{\cdot 2}$ the element wise square. $\mathbf{X}$ is a matrix with F frequency bins and N time frames. NMF tries to capture the most frequent patterns of the speech in K F-dimensional basis vectors that form a dictionary $\mathbf{T}$ for the speech. The matrix $\mathbf{V}$ contains the coefficients of the linear combination and thus indicates how the kth basis vector is activated in the nth time frame. Usually $K < min(F, N)$ such that NMF is a rank reduction operation. A discrepancy measure is chosen between the original $\mathbf{X}$ and the reconstruction $\hat{\mathbf{X}}$ and can be minimized by finding optimal dictionaries and activations. The Euclidian (EU) distance, the Kullback-Leibler (KL) divergence and the Itakura-Saito (IS) divergence are well known measures. In this paper the IS divergence will be used.

$$d_{IS}(x_{fn}, \hat{x}_{fn}) = \frac{x_{fn}}{\hat{x}_{fn}} - log\left(\frac{x_{fn}}{\hat{x}_{fn}}\right) - 1 \tag{1}$$

To minimize this divergence, multiplicative iterative update formulas with convergence guarantees have been derived [14].

$$t_{fk} \leftarrow t_{fk}\sqrt{\frac{\sum_n \frac{x_{fn}}{\hat{x}_{fn}}\frac{v_{kn}}{\hat{x}_{fn}}}{\sum_n \frac{v_{kn}}{\hat{x}_{fn}}}} \tag{2}$$

$$v_{kn} \leftarrow v_{kn}\sqrt{\frac{\sum_f \frac{x_{fn}}{\hat{x}_{fn}}\frac{t_{fk}}{\hat{x}_{fn}}}{\sum_f \frac{t_{fk}}{\hat{x}_{fn}}}} \tag{3}$$

where the sub-indices refer to the corresponding element in the matrix. To avoid scaling ambiguities the columns of $\mathbf{T}$ are to be normalized: $t_{fk} \leftarrow t_{fk}/\sum_{f^*} t_{f^*k}$.

2.2. NMF in Speaker Recognition

The use of NMF in SR applications for single speech is straightforward. In the training phase, training data of the jth target speaker $\mathbf{X}_{train}^j$ is factorized using equations 2 and 3. The obtained dictionaries $\mathbf{T}^j$, for each of the J target speakers, are assumed to be speaker dependent and are collected in the library $\mathbf{T}_{tot} = [\mathbf{T}^1, \mathbf{T}^2, \ldots, \mathbf{T}^J]$.

During testing the identity of a speaker s has to be found in a previously unseen $\mathbf{X}_{test}^s$. NMF is applied with a fixed library $\mathbf{T}_{tot}$ and the activations $\mathbf{V}_{tot,test}^s$ are found iteratively using equation 3. The activation matrix quantitatively indicates the activation of each basis vector for each target speaker in

each time frame. The combined activity of all basis vectors in a target speaker's dictionary is a measure of the activity of the target speaker in the test segment. It is possible to include Group Sparsity (GS-NMF) constraints on the activations $\mathbf{V}^s_{tot,test}$ to enforce solutions where it is unlikely that basis vectors from different target speakers are active at the same time frame [15, 16]. A simple way of estimating the speaker identity is to determine the target speaker for which the sum of the activations, over all its basis vectors and over all the time frames, is maximal. This way of classification can be seen as a per frame speaker probability estimation where the final estimation is a weighted average over all frames, giving more weight to frames with higher activation or more energy.

$$\hat{ID}_s = \arg\max_j \sum_{k\in\kappa^j} \sum_n v^s_{tot,kn} \tag{4}$$

where κ^j are the indices of the basis vectors belonging to the dictionary of the jth target speaker. It is possible to perform a more advanced classification of the activations to a speaker identity by using, for example, support vector machines.

2.3. NMF in Source separation

Aside from SR applications, NMF has also shown good results in source separation problems. When the different speakers are learned on single speech training data, the procedure is very similar to that of SR. However, in SS, the test data $\mathbf{X}_{test}$ contains speech of multiple sources that speak simultaneously. The task is not to determine the speaker identity, but the original source signal of each speaker.

After learning $\mathbf{T}_{tot}$ in the training phase, $\mathbf{V}_{tot,test}$ is calculated in the same way as in section 2.2. Using Wiener filtering and the phase of the observations, the original source signal $\hat{y}^s_{fn}$ can be estimated [8].

$$\hat{y}^s_{fn} = \left(\frac{\sum_{k\in\kappa^s} t_{fk}v_{kn}}{\sum_{s^*}\sum_{k\in\kappa^{s^*}} t_{fk}v_{kn}} |\tilde{x}_{fn}| \right) arg(\tilde{x}_{fn}) \tag{5}$$

where κ^s are the indices of the basis vectors belonging to the dictionary of the sth speaker and $arg(\tilde{x}_{fn})$ denotes the phase of $\tilde{x}_{fn}$.

In many situations, however, there is no possibility for supervised source separations. In blind source separation (BSS) no $\mathbf{X}^s_{train}$ is available to learn the library $\mathbf{T}_{tot}$. Instead the library will be created during the separation itself. Usually one resorts to multichannel techniques in such cases, where Time Direction Of Arrival (TDOA) techniques can be used to assist the source separation. The mixing matrix $\mathbf{M} \in \mathbb{C}^{F\times I\times S}$ is assumed static and thus independent on n.

$$\tilde{x}_{i,fn} = \sum_s m_{is,f} y_{s,fn} \tag{6}$$

where I is the amount of microphones and $m_{is,f}$ indicates the frequency domain representation of the room impulse response (RIR) between the sth speaker and the ith microphone for the fth frequency bin. $\tilde{x}_i$ is the received STFT spectrogram in the ith microphone and y_s is the STFT spectrogram of the original signal of the sth speaker. Because of the scaling ambiguity in equation 6, only the relative RIRs between the microphones can be estimated. The notation for the combined microphone signals is as follows.

$$\tilde{\mathbf{x}}_{fn} = [\tilde{x}_{1,fn}, \tilde{x}_{2,fn}, \ldots, \tilde{x}_{I,fn}] \tag{7}$$

Sawada et al. proposed a multichannel IS divergence [9].

$$\begin{aligned} D_{IS}(\mathbf{X}, \{\mathbf{T}, \mathbf{V}, \mathbf{H}, \mathbf{Z}\}) &= \sum_f \sum_n d_{IS}(X_{fk}, \hat{X}_{fk}) \\ d_{IS}(X_{fk}, \hat{X}_{fk}) &= tr(X_{fk}\hat{X}_{fk}) - logdet(X_{fk}\hat{X}_{fk}) \end{aligned} \tag{8}$$

where $tr(.)$ is the trace of a matrix, $logdet(.)$ is the natural logarithm of the determinant of a matrix, $X_{fn} = \tilde{\mathbf{x}}_{fn}\tilde{\mathbf{x}}^H_{fn}$ with $.^H$ the Hermitian transpose of a matrix and $\hat{X}_{fn} = \sum_k \left(\sum_s H_{fs}z_{sk}\right) t_{fk}v_{kn}$. The same interpretations are given to t_{fk} and v_{kn} as in single channel NMF. z_{sk} is a latent speaker-indicator that indicates the certainty that the kth basis vector belongs to the dictionary of the sth speaker under the constraints $z_{sk} \geq 0$ and $\sum_s z_{sk} = 1$. H_{fs} is an $I \times I$ Hermitian positive semi-definite matrix with on its diagonals the power gain of the sth speaker at the fth frequency bin to each microphone. The off-diagonal elements include the phase differences between microphones and thus contain spatial information of the speaker. Multiplicative update formulas have been found in [9, eq. 42-47] that minimize the divergence in equation 8. The separated signals are then obtained through Wiener filtering.

$$\hat{y}^s_{fn} = \left(\sum_k z_{sk}t_{fk}v_{kn} \right) H_{fs}\hat{X}^{-1}_{fn}\tilde{\mathbf{x}}_{fn} \tag{9}$$

3. Joint approach

When performing speaker recognition in simultaneous speech scenarios, one could opt for a sequential approach. First apply blind source separation to obtain multiple, supposedly single speech, segments from simultaneous speech. Proceed as if those segments do not contain any crosstalk and apply single speaker recognition as explained in section 2.2. However in this paper a joint approach is chosen, where speakers are characterized through dictionaries while separating the sources. During training, source separation is performed as explained in section 2.3. The kth basis vector is then assigned to the sth speaker for which z_{sk} is maximal. The dictionaries are collected in the library $\mathbf{T_{tot}}$. While testing, a similar source separation algorithm is used but $\mathbf{T_{tot}}$ remains fixed. Since every basis vector is contained in a dictionary, the meaning of the $\mathbf{Z}$ variable is changed. It now maps a complete dictionary j, and its corresponding target speaker identity, to a test speaker s. A new variable indicator c_{jk} is introduced that assigns a basis vector k to a dictionary j if $c_{jk} = 1$ under the constraints $c_{j^*k} \geq 0$ and $\sum_{j^*} c_{j^*k} = 1$. The variable $\hat{X}_{fn}$ is then reformulated as follows.

$$\hat{X}_{fn} = \sum_k \sum_j \sum_s H_{fs}z_{sj}c_{jk}t_{fk}v_{kn} \tag{10}$$

It can be easily shown that the update formulas below then extend [9, eq. 42-47].

$$t_{fk} \leftarrow t_{fk} \sqrt{\frac{\sum_j c_{jk} \sum_s z_{sj} \sum_n v_{kn} tr(\hat{X}^{-1}_{fn} X_{fn} \hat{X}^{-1}_{fn} H_{fs})}{\sum_j c_{jk} \sum_s z_{sj} \sum_n v_{kn} tr(\hat{X}^{-1}_{fn} H_{fs})}} \tag{11}$$

$$v_{kn} \leftarrow v_{kn} \sqrt{\frac{\sum_j c_{jk} \sum_s z_{sj} \sum_f t_{fk} tr(\hat{X}^{-1}_{fn} X_{fn} \hat{X}^{-1}_{fn} H_{fs})}{\sum_j c_{jk} \sum_s z_{sj} \sum_f t_{fk} tr(\hat{X}^{-1}_{fn} H_{fs})}} \tag{12}$$

$$z_{sj} \leftarrow z_{sj} \sqrt{\frac{\sum_k c_{jk} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}^{-1}_{fn} X_{fn} \hat{X}^{-1}_{fn} H_{fs})}{\sum_k c_{jk} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}^{-1}_{fn} H_{fs})}} \tag{13}$$

$$c_{jk} \leftarrow c_{jk} \sqrt{\frac{\sum_s z_{sj} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} H_{fs})}{\sum_s z_{sj} \sum_f \sum_n t_{fk} v_{kn} tr(\hat{X}_{fn}^{-1} H_{fs})}} \tag{14}$$

To update H_{fs} an algebraic Riccati equation is solved.

$$H_{fs} A H_{fs} = B \tag{15}$$

$$A = \sum_j \sum_k c_{jk} z_{sj} t_{fk} \sum_n v_{kn} \hat{X}_{fn}^{-1} \tag{16}$$

$$B = H_{fs}^{'} \left(\sum_j \sum_k c_{jk} z_{sj} t_{fk} \sum_n v_{kn} \hat{X}_{fn}^{-1} X_{fn} \hat{X}_{fn}^{-1} \right) H_{fs}^{'} \tag{17}$$

where $H_{fs}^{'}$ is the H_{fs} of the previous update. To avoid scale ambiguity these normalizations should follow: $H_{fs} \leftarrow H_{fs}/tr(H_{fs})$, $t_{fk} \leftarrow t_{fk}/\sum_{f^*} t_{f^*k}$, $z_{sj} \leftarrow z_{sj}/\sum_{j^*} z_{sj^*}$ and $c_{jk} \leftarrow c_{jk}/\sum_{j^*} c_{j^*k}$. In the test phase a basis vector is kept fixed to a dictionary. Therefore, $c_{jk} = 1$ only if the kth basis vector belongs to the jth dictionary, otherwise $c_{jk} = 0$. Using equation 14 and the normalization, one can see that the values for c_{jk} are then fixed for the whole iterative process.

The estimated ID for speaker s is j for which z_{sj} is maximal.

$$\hat{ID}_s = \arg\max_j z_{sj} \tag{18}$$

Through H_{fs} and z_{sj}, the speaker recognition can be interpreted as assigning the jth dictionary, and thus the target speaker identity of the jth dictionary , to the position of the sth speaker in the test utterance. Notice that S, the amount of speakers in the test mixture, can be equal to or smaller than J, the amount of speakers in the library. Not all known speakers must appear in the test mixture. However, for the experiments in this paper $S = J$.

4. Experiments

4.1. Data, set-up and parameters

The CHiME corpus has been used to perform our experiments [17]. It contains 34 speakers with 500 spoken utterances per speaker and a vocabulary of 52 words. Each utterance is about 1.5 seconds long. A scenario in time domain is simulated with a microphone array of $I = 2$ microphones and three randomly chosen speakers ($S = 3$) on randomly chosen spatial positions that are at least 20° apart. The RIR for each speaker to the microphones is determined via its spatial angle to the microphone array and some mild reverberation (RT_{60} = 280 ms) [18]. The microphones are placed 15 cm apart and sample at 16kHz. A Generalized Cross Correlation with Phase Transform (GCC-PHAT) is used to estimate the Direction of Arrival (DOA) and to perform a source count [19]. Aside from speaker recognition and source separation, also information on the number of sources and their location will be known. It has already been shown that such a GCC-PHAT is useful to initialize the multichannel NMF [20]. Here it will be used to set up the initial off-diagonal elements of $\mathbf{H}_{fs}$.

For mixtures of 1.5 seconds, the GCC-PHAT finds the correct number of sources in about 85% of the cases. This goes to 100% for longer mixtures. Since this paper is not about source counting, mixtures with a false estimate of the number of sources will not be included in the results of the experiments in this paper. A spectrogram for each microphone is calculated using an STFT with a window length of 64 ms and an overlap of 32 ms.

During the training phase each person speaks U_{tr} utterances, without moving. The library is learned according to section 2.3 in 1000 iterations. For each speaker $K_s = K$ basis vectors are assigned, giving a total of $K_{tot} = KS$ basis vectors. Random initializations are used for $\mathbf{T}$ and $\mathbf{V}$. Basis vectors are fixed in a dictionary, belonging to a speaker on a certain location, by setting the speaker-indicators $\mathbf{Z}$ as follows.

$$z_{sk} = \begin{cases} 1 & \text{if } k \in \kappa^s \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Through z_{sk} and the spatial information in H_{fs}, a basis vector is then fixed to a speaker on a certain location.

In the test phase the same speakers as in the training phase are used ($J = S = 3$), but placed on different virtual locations. Each person speaks $U_{test} = 1$ utterance, without moving. The algorithm in section 3 is applied and 1000 iterations are used. The library is taken from the training phase, $\mathbf{V}$ is initialized randomly and $\mathbf{H}$ is initialized using GCC-PHAT. The estimated IDs are determined by using equation 18. 50 such test mixtures are created and for each three speakers have to be estimated. This gives a total of 150 recognition tasks per training set. In total twenty independent training sets are created and their recognition accuracies are averaged to cope with the training variability.

Pure SR performance is analyzed in the single speaker scenario, where no SS takes place and dictionaries are learned from non-reverberated speech. The *joint* approach (section 3), which can be applied during either or both, testing and training, is compared with a single speaker scenario and a sequential scenario. In the *single speaker* scenario, the dictionaries are trained or tested on non-reverberated speech of a single speaker, without applying SS. In the *sequential* approach, SS is first applied on speaker mixture data, then dictionaries are trained or tested on the segregated streams. In the remainder of the paper, the scenario where both training and testing are performed on simultaneous speech, will be referred to as the *full simultaneous* scenario. In the *full single* scenario both training and testing are performed on a single speaker.

4.2. i-vector baseline

A second baseline for a *sequential* approach is also presented which uses i-vectors for the speaker recognition part. The MSR Toolbox is used to set up the speaker recognizer [21]. First a Universal Background Model (UBM) and the total variability space are determined using 25 speakers not used in training phase. For every universal background speaker 500 single speech utterances are used. Source separation was performed on mixtures using the multichannel NMF with dictionary size $K = 10$. The STFT features of the separated signals are transformed to MFCC features with differentials (MFCCΔ) and accelerations (MFCC$\Delta\Delta$). i-Vectors are determined from the separated signals, the total variability space and the UBM. A test segment is classified using the cosine distance between the test i-vector and the training i-vectors. Table 1 shows how the speaker recognition accuracies in the *full simultaneous* scenario depend on the amount of UBM components N_{comp} and the dimensionality of the i-vector D_{ivec}. The optimal values were found to be $N_{comp} = 1024$ and $D_{ivec} = 15$. This is comparable to the values found in [6].

N_{comp}	D_{ivec}					
	5	10	15	20	50	100
256	68.3	73.7	74.6	74.5	62.5	62.4
512	66.1	74.8	74.8	73.4	68.8	66.5
1024	60.9	75.2	**75.4**	75.2	71.9	68.4
2048	55.4	68.5	70.6	72.7	73.8	72.2

Table 1: i-vector based SR accuracies (in %) for the *full simultaneous* scenario.

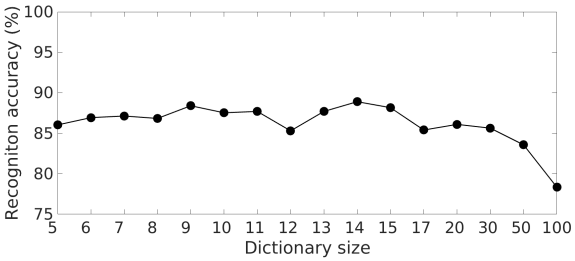

Figure 1: Average speaker recognition accuracy (in %), depending on the dictionary size, using the *joint* approach in the *full simultaneous* scenario.

4.3. Results

In figure 1 speaker recognition accuracies relative to the dictionary size K are plotted for the *full simultaneous* scenario. The training size U_{tr} was taken at 20 utterances per speaker. Results were obtained using the *joint* solution. There is little influence of the dictionary size to the recognition accuracy if K is between 8 and 30. In the remainder of the paper the dictionary size is taken at $K = 10$.

Speaker recognition accuracies for the different scenarios and used methods, using the above found optimal parameters ($K = 10$, $N_{comp} = 1024$ and $D_{ivec} = 15$), are shown in table 2(a) for NMF-based speaker recognition and table 2(b) for i-vector based speaker recognition. In the *full single* scenario i-vector based SR performs slightly better then NMF based SR. However, in *simultaneous* scenarios, i-vectors struggle with the crosstalk after SS. The same problem occurs for NMF when a sequential approach is used in, either or both, training and testing. This problem is circumvented when the *joint* approach for NMF in the *full simultaneous* scenario is used. The Speaker Error Rate (SER) for the *joint* approach (12%) is significantly lower than a *sequential* approach using i-vectors for the SR (24%).

Figure 2 shows how the recognition accuracy increases with the amount of training data. The optimal parameters above are used again. Only for very low amount of training utterances, the recognition accuracy decreases. The sufficiency of a limited amount of training data is probably due to the limited vocabulary size in the CHiME corpus. Notice that increasing the amount of training is beneficial twice to the recognition accuracy. SS is improved (see table 3) and the amount of patterns seen in training data is increased which allows the speakers to be better characterized and more easily recognized. To cope better with the increasing amount of seen patterns, the dictionary size could also be increased. This way the extra detected patterns can be stored in the dictionary. A system is chosen that can cope with any training size and thus the dictionary size is fixed.

Train	Test		
	single	seq	joint
single	98.0	**86.4**	73.2
seq	88.8	76.1	62.5
joint	80.2	61.3	**87.9**

(a) NMF based SR.

Train	Test	
	single	seq
single	**98.4**	78.5
seq	**91.2**	76.0

(b) i-vector based SR.

Table 2: SR accuracies (in %) for different training and testing scenarios. In *seq* and *joint* 3 speakers are active simultaneously. *Single* uses data from one speaker at a time.

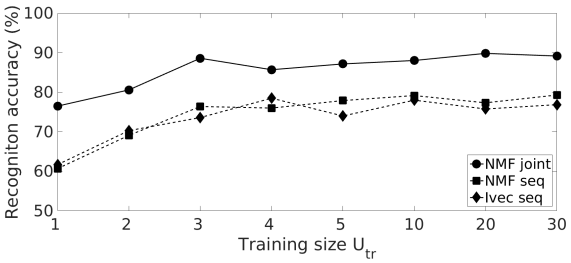

Figure 2: Average speaker recognition accuracy, depending on the training size U_{tr}, in the *full simultaneous case*. Results are shown for the *joint* approach (circles) and for a *sequential* approach using NMF (squares) or i-vectors (diamonds).

As this paper analyzes joint SR and SS, an evaluation metric for the SS quality is considered. Signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) between the original source signal and the signal separated from the mixture, are calculated and are shown in table 3 [22]. Three different situations are shown: a mixture where each person speaks 20 utterances, a mixture where each person speaks 1 utterance and a mixture where each person speaks 1 utterance that is separated using previously learned dictionaries. 20 independent runs are used to calculate the ratios and the average is shown. As expected, the SS results are better when the mixture is longer. NMF has more examples to recognize recurring patterns and build distinctive dictionaries per speaker which enhances the separation quality. However, performing SS using learned and fixed dictionaries decreases performance. If a pattern is not seen in the training stage or is not frequent enough to fit in the dictionary, it cannot be used during testing for the reconstruction of the signal and will thus degrade the reconstructed signal. A solution is to allow flexibility in the trained dictionaries when testing, but this would interfere with the speaker recognition.

	SDR	SIR	SAR
SS on 20 utts	6.93	12.50	9.36
SS on 1 utt	4.28	7.99	8.26
SS on 1 utt using learned dicts	3.08	5.62	8.81

Table 3: SS performance for different scenarios measured in SDR, SIR and SAR (dB).

5. Conclusion

In this paper it is shown that in simultaneous speech environments for three overlapping speakers, a joint approach for SS and SR outperforms sequential approaches for both NMF and i-vector based SR. This benefit is inherent to the multichannel NMF as patterns per speaker are learned anyway to perform a segregation. These patterns can be used to recognize speakers during the test phase.

6. References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[3] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[4] M. H. Bahari and H. Van hamme, "Speaker age estimation and gender detection based on supervised non-negative matrix factorization," in *In proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*, 2011.

[5] C. Joder and B. Schuller, "Exploring nonnegative matrix factorization for audio classification: Application to speaker recognition," in *Speech Communication*. VDE, 2012.

[6] S. Drgas and T. Virtanen, "Speaker verification using adaptive dictionaries in non-negative spectrogram deconvolution," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 462–469.

[7] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

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

[9] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[10] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 836–845, 2014.

[13] T. May, S. van de Par, and A. Kohlrausch, "A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, 2012.

[14] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," *In Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), Kittila, Finland, 29 August – 1 September 2010*, pp. 283–288.

[15] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-saito nonnegative matrix factorization with group sparsity," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 21–24.

[16] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group sparsity for speaker identity discrimination in factorisation-based speech recognition." in *INTERSPEECH*, 2012, pp. 2138–2141.

[17] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *INTERSPEECH*, 2010, pp. 1918–1912.

[18] D. Campbell, K. Palomäki, and G. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," *Computing and Information Systems*, vol. 9, no. 3, p. 48, 2005.

[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[20] S. Mirzaei, H. Van hamme, and Y. Norouzi, "Blind audio source separation of stereo mixtures using bayesian non-negative matrix factorization," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 621–625.

[21] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, November 2013. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=205119

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