CATALOG-BASED SINGLE-CHANNEL SPEECH-MUSIC SEPARATION FOR AUTOMATIC SPEECH RECOGNITION

Cemil Demir[1,3], A. Taylan Cemgil[2], Murat Saraçlar[3]

[1]TÜBİTAK-BİLGEM, Kocaeli, Turkey
[2]Computer Engineering Department, Boğaziçi University, Istanbul,Turkey
[3]Electrical and Electronics Engineering Department, Boğaziçi University, İstanbul,Turkey
cdemir@tubitak.uekae.gov.tr, (taylan.cemgil|murat.saraclar)@boun.edu.tr

ABSTRACT

In this study, we analyze the effect of the catalog-based single-channel speech-music separation method, which we proposed previously, on speech recognition performance. In the proposed method, assuming that we know a catalog of the background music, we developed a generative model for the superposed speech and music spectrograms. We represent the speech spectrogram by a Non-negative Matrix Factorization (NMF) model and the music spectrogram by a conditional Poisson Mixture Model (PMM). In this paper, we propose to recover the speech signals from the mixed signal in time-domain by detecting the active catalog frames using the catalog-based method. We compare the performances of 3 different signal reconstruction techniques; Expectation-Based, Posterior-Based and Time-Domain reconstruction. Moreover, we compare the performance of our system with the performance of the traditional NMF model. Our method outperforms the NMF method in ASR performance and separation performance in most experimental conditions.

1. INTRODUCTION

Recently automatic speech recognition (ASR) applications have become popular in broadcast news transcription systems. One major problem is the serious drop in the performance with the presence of background music, that is often present in radio and television broadcasts [1, 2]. Therefore, removing the background music is important for developing robust ASR systems. A real-world ASR solution should contain a front-end system capable of segmenting and separating music and speech from incoming audio signals. The aim of this study is to analyze the performance of the catalog-based speech-music separation method, that we proposed previously, when it is used as a front-end for an ASR system.

Many researchers studied single-channel source separation for mixture of speech from two speakers [3] but there are a few studies on single-channel speech-music separation [4, 5, 6]. Model-based approaches are used to separate sound mixtures that contain the same class of sources such as speech from different people [7, 8] or music from different instruments [9, 10].

In a previous study [11] , we introduced a simple probabilistic model-based approach to separate speech from music. Unlike other probabilistic approaches, we do not model the speech in great detail, but instead focus on a model for the music. The motivation behind our approach is that, especially in broadcast news, most of the time, the background music is composed of some repetitive piece of music, called a 'jingle'. Therefore, we can assume that we can learn a catalog of these jingles and hope to improve separation performance.

In our model, the catalog corresponds to a conditional mixture model. Each spectrogram frame of the music is generated by a single mixture component, i.e., a catalog element. The speech spectrogram is generated from an NMF model. The observed spectrogram is the sum of the speech and music. Separation is achieved by joint estimation of the unknown parameters and hidden variables of this hierarchical model.

We assume that, although we do not have any prior information about the speech part of the mixture, we can assume that the magnitude spectrogram of the speech signal is generated by an Non-Negative Matrix Factorization (NMF) model. This way, by finding the parameters of the NMF model, we can recover the speech signal from the mixture. We use the probabilistic interpretation of the NMF to develop the separation algorithm [12]. The reason for using the probabilistic approach is that we can easily extend the model so that it contains the prior information about the sources. For the time being, we assume that the music is created by playing a random part of a known clip and applying a frequency and volume adjustment filters to change the character of the music. Our aim is to find out which part of the clip is played when, while figuring out the values of parameters of adjustment filters. This corresponds to a Poisson Mixture Model (PMM) for the magnitude spectrogram of the music signal. The overall model consists of the combination of the NMF model for the speech part and the mixture model for the music part of the audio signal. In this study, we developed the inference method for this overall probabilistic model and apply this separation method to increase the ASR performance.

Unlike the previous study [11], we use the catalog-based method as a front-end for the ASR task and measured the separation performance with ASR evaluation criteria, Word Error Rate (WER). Moreover, speech-music separation performance of the proposed method is compared with the separation performance of the traditional NMF based method. Furthermore, the effect of reconstruction techniques, Expectation-based and Posterior-based techniques, are examined and the superiority of Posterior-based approach is observed experimentally. Time-domain reconstruction technique which can be used in the case of the original version of the jingle is accessible in the separation phase is also proposed and evaluated in this study.

This paper is organized as follows: in Section 2, we overview the catalog-based and NMF based speech-music separation methods. In Section 3, we briefly explain 3 different speech reconstruction techniques using source sepa-

ration methods. The experimental results and comparisons are provided in Section 4. Section 5 presents the discussion, conclusions and comments for further investigation.

2. SEPARATION METHODS

2.1 Catalog-Based Speech-Music Separation

2.1.1 Model Description

In this model, we can express each time-frequency entry of the magnitude spectrogram of the mixture at time t and frequency bin u as

$$x_{ut} = S_{ut} + m_{ut} \tag{1}$$

where S and m represent the magnitude spectrogram of the speech and music signals, respectively. We assume an NMF based generative model, which uses a Poisson observation model [12], for the spectrogram of the speech. In this probabilistic model, each time-frequency entry of the spectrogram of the speech is generated by B Poisson sources as

$$S_{ut} = \sum_{i=1}^{B} s_{uit} \quad \text{where} \quad s_{uit} \sim PO(s_{uit}; U_{ui}V_{it}) \tag{2}$$

where U and V matrices contain the hyper-parameters of the spectrogram of the speech signal and also correspond to template and excitation matrices respectively in NMF model. We also use a Poisson observation model in the generative model of the magnitude spectrogram of the music part as

$$m_{ut}|r_t = j \sim PO(m_{ut}; C_{uj}f_u v_t)^{[r_t=j]} \tag{3}$$

where $[r_t = j]$ represents the indicator function, which is 1 when j-th frame of the catalog is used and its value is 0, otherwise. In Equation (3), C_{uj} represents the magnitude spectrogram corresponding to the u-th frequency bin and the j-th member of the jingle catalog, f_u represents frequency adjustment parameter for frequency bin u and v_t represents the volume adjustment parameter for time frame t. The goal is here to model volume changes (pan-in, pan-out) and filtering (equalization). Each active frame index is drawn independently from a set of catalog indexes as

$$r(t) = j \in \{1,2,..,N\} \text{ with probability } \pi_j \tag{4}$$

where π represents probability distribution on the catalog frame indexes. The difference from the speech model is that, the intensity parameter of the Poisson model is chosen from a magnitude spectrogram of a set of previously obtained catalog frames. Moreover, a frequency and volume adjustment is applied to that intensity.

The overall graphical model corresponding to the generation of the mixture of the speech and music signals is shown in Figure 1. Upper side of the graphical model generates the spectrogram of the speech part of the mixture whereas the lower side generates the spectrogram of the music part.

2.1.2 Multiplicative Update Rules

In the previous study [11], it was shown that the overall joint posterior distribution over hidden sources (speech, music sources and catalog indexes) is a mixture of multinomials. As a result, the hyper-parameters of the speech and music

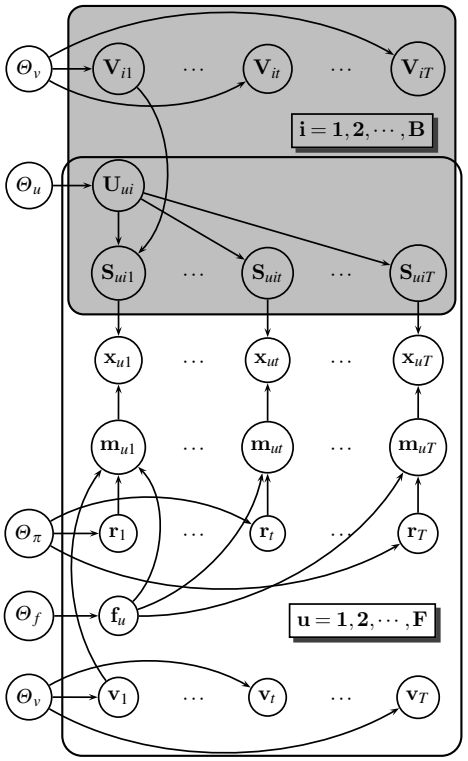

Figure 1: Graphical Model For Speech-Music Mixture.

signals can be updated using the EM algorithm which is described in [11] in detail. These update equations correspond to the Multiplicative Update Rules of the NMF method. Each entry of the template matrix, U, can be calculated as

$$U_{ui} = \frac{\sum_{t,j} \langle [r_t = j] \rangle \left\langle s^j_{uit} \right\rangle}{\sum_t V_{it}} \tag{5}$$

where $\langle s^j_{uit} \rangle$ represents the expected value of hidden speech source w.r.t the conditional posterior which can be calculated as

$$p(s_{uit}|r_t = j) = \sum_{m_{ut}} p(s_{uit}, m_{ut}|r_t = j) \tag{6}$$

and $\langle [r_t = j] \rangle$ represents the posterior probability of the active frame index, j at time t. Each entry of the excitation matrix of the speech spectrogram, V can be calculated using

$$V_{it} = \frac{\sum_{u,j} \langle [r_t = j] \rangle \left\langle s^j_{uit} \right\rangle}{\sum_u U_{ui}}. \tag{7}$$

The volume adjustment parameter at each time can be found by using the next formula

$$v_t = \frac{\sum_{u,j} \langle [r_t = j] \rangle \langle m^j_{ut} \rangle}{\sum_{u,j} \langle [r_t = j] \rangle C_{uj} f_u} \tag{8}$$

where $\langle m^j_{ut} \rangle$ similarly represents the expected value of hidden music source. The frequency adjustment parameters for each frequency can be found using

$$f_u = \frac{\sum_{t,j} \langle [r_t = j] \rangle \langle m^j_{ut} \rangle}{\sum_{t,j} \langle [r_t = j] \rangle C_{uj} v_t}. \tag{9}$$

2.2 NMF Based Speech-Music Separation

In NMF based speech-music separation systems, during training phase,the magnitude spectrogram of the speech and music signals are used to train an NMF model for each source as

$$S = U_s V_s \quad \text{and} \quad M = U_m V_m. \tag{10}$$

The template and excitation matrices can be calculated via Multiplicative Update Rules [13] efficiently. In the separation phase, using the template matrices, an overall template matrix is constructed. Using the magnitude spectrogram of the mixed signal and the overall template matrix, the excitation matrix for each source is calculated by solving the equation

$$X = [U_s U_m][W_s' W_m'] \tag{11}$$

where W_s and W_m represents the excitation matrix for speech and music sources in the mixture respectively. After finding the excitation matrix for each source, the reconstruction of the speech and music signals can be done using the techniques described in Section 3.

3. SIGNAL RECONSTRUCTION

3.1 Expectation-Based Reconstruction

By using the proposed method or a traditional NMF method, template and excitation matrices are estimated using

$$(U^*, V^*, R^*, f^*, v^*) = \arg \max_{U,V,R,f,v} p(X|U,V,R,f,v).$$

where R represents posterior probability of the catalog frames for each time t. The magnitude spectrogram of the speech and music signals are estimated as the expectations of the hidden speech and music sources which corresponds to the intensity parameters of these sources in Poisson observation model. The estimated magnitude spectrograms of the sources are

$$\widehat{S} = \langle S|U^*, V^*\rangle = U^* V^* \tag{12}$$

$$\widehat{M} = \langle M|R^*, f^*, v^*\rangle = (CR^*) \otimes (f^* v'^*) \tag{13}$$

where $\otimes$ represents the element-wise multiplication. Using estimated magnitude spectrograms and phase of the mixed signal, we can reconstruct the time-domain signal.

3.2 Posterior-Based Reconstruction

In Posterior-based approach, using estimated intensity parameters of the speech and music sources and the observation values, we can estimate the source values as joint posterior of the sources as

$$(\widehat{S}, \widehat{M}) = \arg \max_{S,M} p(S, M|X, U^*, V^*, R^*, f^*, v^*).$$

This corresponds to the estimation of the magnitude spectrogram of the speech and music sources as

$$\widehat{S} = X \otimes \frac{U^* V^*}{(U^* V^* + (CR^*) \otimes (f^* v'^*))}. \tag{14}$$

$$\widehat{M} = X \otimes \frac{(CR^*) \otimes (f^* v'^*)}{(U^* V^* + (CR^*) \otimes (f^* v'^*))}. \tag{15}$$

This is also known as the Wiener Filtering approach and was used in NMF based speech-music separation in [5]. Since NMF methods find an approximation to the magnitude spectrogram of the mixed signal, the error term between the approximation and the real value is not assigned to any source. This problem can be solved by estimating the source values jointly using the mixed signal spectrogram. This enables the perfect reconstruction of the target sources.

3.3 Time-Domain Reconstruction

Since, in catalog-based approach the posterior probability of the catalog frames at each time frame are estimated, the music signals can be recovered using the frames which have the maximum posterior probability at each time frame. Mathematically, for each time frame, the music signal is estimated as

$$\widehat{m}(t) = m(\widehat{r}(t)) \quad \text{where} \quad \widehat{r}(t) = \arg \max_j \langle [r_t = j] \rangle$$

where $\widehat{m}$ represents the reconstructed music signal and $\widehat{r}(t)$ represents the catalog frame which has the maximum posterior probability at time frame t. Afterwards, the speech signal can be found by subtracting the recovered music signal from the mixed signal.

4. EXPERIMENTAL RESULTS

Since the ultimate goal of the speech-music separation is to increase the ASR performance, we analyze the performance of the method using ASR performance measure, Word Error Rate (WER). However, in order to relate the performances of the methods in source-separation and ASR tasks, we also calculated Speech-to-Music Ratio (SMR) and Source-to-Artifact Ratio (SAR) values [14].

4.1 Speech Recognition System and Test Set

For speech recognition tests, we used a CMU-Sphinx HMM-based continuous density speech recognizer which is trained to recognize Turkish Broadcast News speech. The gender-dependent acoustic models are trained using MFCCs and their deltas and double-deltas calculated in 25ms frames of the clean speech data. The vocabulary size of the recognition system is about 30k. The test set contains 1232 utterances distributed approximately uniformly across 8 speakers. The total length of the test set is about 2 hours. The test utterances are mixed with a 4 sec. length jingle at different SMR levels to create the test set. The background music signal is generated by repeating the jingle up to the length of the speech. The average length of the speech sentences is 6 sec. The jingle is taken from the broadcast news jingles. The magnitude spectrogram is computed using 1024-point length frames and 512 point frame shift is used. The number of speech bases is fixed at 30.

4.2 Experimental Analysis

For the catalog-based separation, we apply the 3 different reconstruction techniques, Expectation-based Reconstruction (ER), Posterior-based Reconstruction (PR) and Time-domain Reconstruction (TR), to compute the recovered speech sources. For the NMF based separation, we used ER and PR techniques to recover speech signals because of the fact that in NMF case, there is no direct relation between

each element of the template matrix and time-domain signal. We compare the performances of the separation methods and reconstruction techniques (RT) in Table 1 with SMR values, Table 2 with SAR values and Table 4 with ASR results.

Table 1: Average Output SMR values (in dB) obtained using the original jingle

SMR (dB)		Input SMR Values				
Method	RT	0dB	5dB	10dB	15dB	20dB
PMM	TR	**28.1**	**34.8**	**37.4**	41.5	47.2
	ER	16.7	23.2	30.1	37.5	45.4
	PR	17.5	24.1	30.9	38.2	46.1
NMF	ER	23.2	27.1	34.9	42.3	50.3
	PR	23.4	27.2	34.9	**42.4**	**50.3**

Table 2: Average Output SAR values (in dB) obtained using the original jingle

SAR (dB)		Input SMR Values				
Method	RT	0dB	5dB	10dB	15dB	20dB
PMM	TR	**15.6**	**17.2**	**18.4**	**20.4**	23.1
	ER	8.5	10.3	11.6	12.2	12.6
	PR	10.9	14.2	17.2	20.2	**23.2**
NMF	ER	9.6	10.2	11.6	12.4	12.8
	PR	11.5	13.3	16.1	18.6	20.8

Table 3: Baseline WER values (in %)

Baseline Results	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
Clean	24.9	24.9	24.9	24.9	24.9
Mixed	99.6	97.4	84.7	59.1	39.6

In our experiments, it is shown that TR method has the best ASR performance among the reconstruction techniques. However, since relation of the clustered frames and time-domain signal are not one-to-one, we do not apply TR method in clustered jingle case. The TR method has the highest SMR and SAR values and obtains the best performance. The question about the TR method is that why the separation performance of the TR method is slightly getting worse as the input SMR value is increasing. The reason is that as the input SMR value is increasing, the percentage of correctly identified number of frames is decreasing due to the fact that the speech signal is suppressing the background music as expected. Therefore, the estimation performance of the TR method is slightly worse in high input SMR values. In order to compare the ASR performances of the separation methods, it is not enough to examine only the output SMR values, SAR values must also be considered. When we compare the ER and PR performances, it is seen that although output SMR values of the methods are very close to each other, WERs of the ER method are very high compared to PR WERs. This is due to the fact that estimating the sources using joint posteriors results higher output SAR values. For example, average SAR value of PR method over all experiments is 16.61dB whereas average SAR value of ER method is 10.53dB.

Table 4: Average WER values (in %) obtained using the original jingle

WER (%)		Input SMR Values				
Method	RT	0dB	5dB	10dB	15dB	20dB
PMM	TR	**26.1**	**26.6**	**27.2**	**27.5**	**27.1**
	ER	88.9	78.9	67.3	56.5	46.1
	PR	70.8	53.3	40.6	33.1	29.6
NMF	ER	89.0	78.6	67.6	58.0	50.8
	PR	75.0	58.1	44.3	36.7	31.9

When original version of the jingle is used in the separation task, it corresponds to an example-based separation method. In the experiments, it is observed that modeling the jingle with a mixture model produces better ASR results than modeling with NMF model in the example-based separation framework. Actually, this result is not surprising that we use the prior information about the music generation process from the jingle. Each frame of the music is generated by choosing a frame of the jingle, that is, at each time frame, only one frame of the jingle is used. Therefore, using a mixture model for the music signal is more appropriate than using an NMF model which generates the music signal as an additive combination of the jingle frames. Instead of using the jingle itself in the separation task, we reduced the size of the catalog by half using PMM and NMF methods. In Table 5,6 the separation performance results with the clustered jingle are presented. In Table 7, we analyzed the performance of the ASR system in the case of using the clustered jingle. For all input SMR values, clustering with PMM method improves the ASR performance more than NMF method. This shows the advantage of using PMM clustering over NMF clustering. The ASR performances of the method with or without clustering are compared in Figure 2 and the advantage of using PMM method is shown in this figure. In this figure, 'O' and 'C' represent original and clustered versions of the jingle respectively.

Table 5: Average Output SMR values (in dB). The results are obtained using the clustered version jingle

SMR (dB)		Input SMR Values				
Method	RT	0dB	5dB	10dB	15dB	20dB
PMM	ER	17.9	24.4	31.8	39.1	44.9
	PR	**19.8**	**24.6**	**31.9**	**39.2**	**45.2**
NMF	ER	5.2	16.3	26.5	35.8	44.9
	PR	4.9	15.4	25.8	35.1	44.3

Table 6: Average Output SAR values (in dB). The results are obtained using the clustered version jingle

SAR (dB)		Input SMR Values				
Method	RT	0dB	5dB	10dB	15dB	20dB
PMM	ER	8.5	10.5	11.7	12.2	11.5
	PR	**11.6**	**14.1**	**17.1**	**19.9**	21.9
NMF	ER	7.8	9.6	11.2	12.2	12.9
	PR	11.1	13.2	16.1	19.1	**22.3**

Table 7: Average WER values (in %) obtained using the clustered version jingle

WER (%)		Input SMR Values				
Method	RT	0dB	5dB	10dB	15dB	20dB
PMM	ER	92.3	82.1	71.1	56.4	47.4
	PR	**73.2**	**62.4**	**46.5**	**35.1**	**32.7**
NMF	ER	99.3	94.8	81.5	62.2	47.8
	PR	98.4	87.6	61.7	42.8	32.7

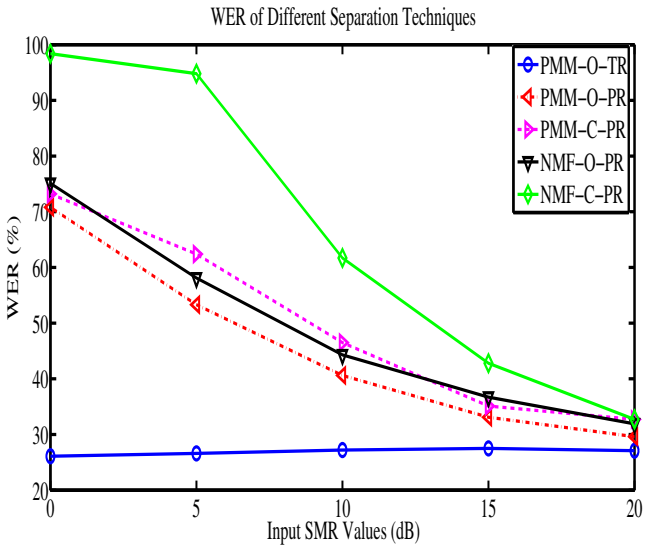

Figure 2: WERs for different methods.

5. CONCLUSIONS

The aim of this study is to evaluate the speech recognition performance of the previously proposed method. The performance comparison with NMF method in source separation and ASR tasks is carried out and it was shown that PMM based method yields better ASR performance for all experimental conditions. Moreover, we proposed to reconstruct the source signals using TR method and it is shown that TR method outperforms PR and ER methods in ASR performance. In the future, we will focus on methods that enable to use TR method in clustered jingle case. Advantage of using PR method over ER method is also shown experimentally. We used the two different clustering techniques to decrease the size of the jingle catalog, PMM and NMF. We showed that the clustered versions of the jingles can also be used for source separation. We also show that ASR performance of PMM clustering is better than the NMF clustering. In this study, we assumed a mixture model on the catalog frames, however, in the case of a known catalog, it is more realistic to assume a Markovian structure on the catalog frame indexes. In the future, we are planning to use a Markov Model instead of using the mixture model on the catalog frames. Moreover, in this study, any prior information about the speech signal is not used in the separation stage. Incorporating prior speech information can enhance the separation performance. We are planning to developed the proposed method such that the model can use the prior information about the speech signal.

6. ACKNOWLEDGEMENTS

This research is supported in part by TUBITAK (Scientific and Technological Research Council of Turkey) (Project code: 105E102). Murat Saraçlar is supported by the TUBAGEBIP award. Taylan Cemgil is supported by the Bogazici University research grant BAP 09A105P.

REFERENCES

[1] B. Raj, V.N. Parikh, and R.M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. of ICASSP*, 1997.

[2] E. Arisoy, D. Can, S. Parlak, H. Sak, and M. Saraclar, "Turkish broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 5, pp. 874–883, 2009.

[3] M.N. Schmidt and R.K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of ICSLP*, 2006.

[4] R. Blouet, G. Rapaport, and C. Fevotte, "Evaluation of several strategies for single sensor speech/music separation," in *Proc. of ICASSP*, 2008, pp. 37–40.

[5] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-Negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition," in *Proc. of Interspeech*, 2010.

[6] L. Benaroya, F. Bimbot, G. Gravier, and R. Gribonval, "Experiments in audio source separation with one sensor for robust speech recognition," *Speech Communication*, vol. 48, no. 7, pp. 848–854, 2006.

[7] P. Smaragdis, M. Shashanka, M. Inc, and B. Raj, "A Sparse Non-Parametric Approach for Single Channel Separation of Known Sounds," Proc. of NIPS, 2009.

[8] R.J. Weiss and D.P.W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Computer Speech & Language*, 2008.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[10] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[11] C. Demir, A.T. Cemgil, and M. Saraçlar, "Catalog-Based Single-Channel Speech-Music Separation," in *Proc. of Interspeech*, 2010.

[12] A.T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009.

[13] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, 1999.

[14] C. Févotte, R. Gribonval, and E. Vincent, "A toolbox for performance measurement in (blind) source separation," .