Semi-supervised 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, İstanbul,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 propose a semi-supervised speech-music separation method which uses the speech, music and speech-music segments in a given segmented audio signal to separate speech and music signals from each other in the mixed speech-music segments. In this strategy, we assume, the background music of the mixed signal is partially composed of the repetition of the music segment in the audio. Therefore, we used a mixture model to represent the music signal. The speech signal is modeled using Non-negative Matrix Factorization (NMF) model. The prior model of the template matrix of the NMF model is estimated using the speech segment and updated using the mixed segment of the audio. The separation performance of the proposed method is evaluated in automatic speech recognition task.
Index Terms: speech-music separation, semi-supervised, speech recognition

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 develop a music-speech separation technique that can be used as a front-end for the ASR systems. In [1], it was shown experimentally that background music does not affect the ASR performance as seriously as the white noise at the same SNR values. However, standard noise reduction techniques are not applicable to music separation. Therefore, we approach the problem as a single-channel source separation task. The contribution of this study is to develop a semi-supervised probabilistic approach to single-channel speech-music separation problem and to analyze the performance improvement not only with source separation measures but also with ASR performance measures.

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]. Model-based approaches are used to separate sound mixtures that contain the same class of sources such as speech from different people [6, 7] or music from different instruments [8].

In a previous study [9] , we introduced a simple probabilistic model-based approach to separate speech and music signals. Unlike other probabilistic approaches, we did not model the speech in great detail, but instead focused 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 the same dull and repetitive piece of music, called a 'jingle'. Therefore, we can assume that we can learn a model of these jingles and hope to improve separation performance.

In this study in contrast to [9], we assume, we do not have any jingle as a priori. However, we assume, we have an audio signal which is manually or automatically segmented as speech, music and speech-music mixture. Using each segmented audio, the models for speech and music sources are trained and hence the mixed signal are separated as speech and music in this training phase. In other words, the training of the sources and the separation of the sources are done simultaneously. The main contribution of this study is to extend the previously proposed method [9] by incorporating the prior speech information to the separation task. We developed the inference method for incorporating the speech priors to the separation method. Moreover, unlike the previous separation methods, we propose a variational method to update the prior speech templates in the mixed part of the audio.

This paper is organized as follows: in Section 2, we overview the proposed semi-supervised speech-music separation methods. The experimental results and comparisons are provided in Section 3. Section 4 presents the discussion, conclusions and comments for further investigation.

2. Method

In the proposed semi-supervised speech-music separation framework, it is assumed that a speech-music segmentation system can partition an incoming audio as speech, music and speech-music mixture. However, it is not necessary to segment the entire audio signal in our approach. That is, the segmentation system can label some part of the audio and label the remaining part of the audio as the unsegmented part. The proposed separation method will segment the unlabeled part of the audio as speech, music and speech-music mixture. Moreover, in this approach we assume that the background music in the mixed part is partially composed of the repetition of the music part in the audio. This assumption is realistic especially for the broadcast news audio. Therefore, the music model, which is a mixture model as in the previous study [9], are trained only using the segmented music part in the audio itself. In the previous study [9], though an NMF model is used to represent the speech source, no training data was used to model the speech source and hence the speech signal is estimated in an unsuper-

vised manner. Although the previous approach is well-suited for the case that whole background music is composed of the repetition of the music segment in the audio, incorporation of the prior speech information is useful when some part of the background music is not included in the music segment.

In the current study, we consider three different speech-music separation methods and compare their performances. In the first method, the templates of the NMF model of the speech signal are trained using the speech part of the audio. Then using these fixed templates and the music model, the corresponding excitations in the mixed part of the audio are estimated to recover the speech signal. This method is called as NMF based separation. The second method updates the speech templates, which are estimated using the speech part as a prior, in the mixed part of the audio and estimates the corresponding excitations in the mixed part simultaneously to recover the speech signal. Since we use the variational technique to do inference of the sources, the second strategy is called Variational based separation. In the last method, the speech and the mixed parts of the audio are used jointly to estimate the speech templates and the corresponding excitations simultaneously. The last method is called as Joint Separation method.

2.1. Model Description

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

$$x_{ut} = k_{ut} + n_{ut} \tag{1}$$

where K and N represent the magnitude spectrograms of the speech and music signals, respectively. We assume an NMF based generative model, which uses a Poisson observation model [10], 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

$$k_{ut} = \sum_{i=1}^{B} s_{uit} \tag{2}$$

and each Poisson source model is given by

$$s_{uit} \sim PO(s_{uit}; d_{ui}e_{it}) \tag{3}$$

where D and E matrices contain the hyper-parameters of the spectrogram of the speech signal. D contains template vectors for the magnitude spectrogram of the speech signal and E contains the corresponding excitations for the template vectors. In this study, we assume a Gamma prior on the dictionary and excitation matrices as follows:

$$d_{ui} \sim \mathcal{G}(d_{ui}; a^d_{ui}, b^d_{ui}) \quad \text{and} \quad e_{it} \sim \mathcal{G}(e_{ui}; a^e_{it}, b^e_{it}) \tag{4}$$

where $a^d_{ui}, b^d_{ui}, a^e_{it}, b^e_{it}$ are hyper-parameters of the template and excitation matrices respectively. We also use a Poisson observation model in the generative model of the magnitude spectrogram of the music part, $n_{ut} = m_{ut}$, as

$$m_{ut}|r_t = j \sim PO(m_{ut}; C_{uj}f_u v_t)^{[r_t=j]} \tag{5}$$

where $[r_t = j]$ represents the indicator function, which is 1 when j-th frame of the jingle component is used and its value is 0, otherwise. In Equation (5), 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 filtering parameter for frequency bin u and v_t represents the gain parameter for time frame t. The goal here is to model volume changes (fade-in, fade-out) and filtering (equalization). Each active frame index is drawn independently from a set of jingle indexes as

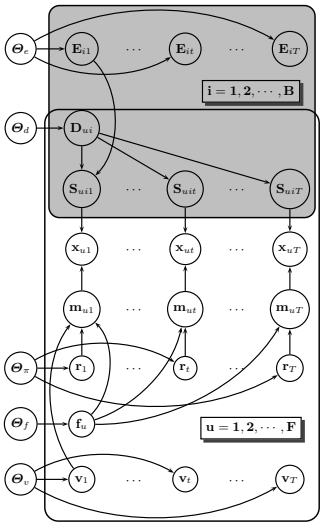

Figure 1: Graphical Model For Speech-Music Mixture.

$$r(t) = j \in \{1, 2, .., I\} \text{ with probability } \pi_j \tag{6}$$

where π represents probability distribution on the jingle frame indexes and I represents the number of jingle frames. 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 jingle frames. Moreover, a filtering and gain is applied to that intensity parameter. 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.

2.2. Inference Method

In this section, we describe the inference technique that are used in the mixed segment of the audio. We derive the update equations of the posterior distributions of the latent sources and parameters of the speech and music signals in the previously described probabilistic model. Since the posterior distributions of the template and excitation parameters, d_{ui}, e_{it} and the latent speech, music and active frame sources, S, M and R are coupled, we cannot compute the overall posterior distribution exactly. In this case, we use the variational technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

$$q(S, M, R) \propto \exp(\langle \log \phi \rangle_{q(D)q(E)}) \tag{7}$$

$$q(D) \propto \exp(\langle \log \phi \rangle_{q(S,M,R)q(E)}) \tag{8}$$

$$q(E) \propto \exp(\langle \log \phi \rangle_{q(S,M,R)q(D)}) \tag{9}$$

where $\phi = p(X, S, M, D, E, R|\Theta)$ and Θ represents the $a^d_{ui}, b^d_{ui}, a^e_{it}, b^e_{it}, \pi, f, v$. The joint posterior distribution of the latent speech and music sources and the jingle indexes, $q(S, M, R)$, is a multinomial mixture model (MMM) as shown in [9]. The overall joint posterior distribution of the latent sources can be decomposed conditioned on the jingle frame,

j, as

$$\begin{aligned} q(S,M,R) &= q(S,M|R)q(R) \\ q(S,M|R) &= \mathcal{M}(s_{u1t},.,s_{uBt},m_{ut};x_{ut},p^j_{u1t},.,p^j_{uBt},p^j_{ut}) \end{aligned}$$

The parameters of this MMM can be computed using:

$$\begin{aligned} p^j_{uit} &= \frac{\exp(\langle \log d_{ui}\rangle + \langle \log e_{it}\rangle)}{(\sum_i \exp(\langle \log d_{ui}\rangle + \langle \log e_{it}\rangle)) + (C_{uj} f_u v_t)} \\ p^j_{ut} &= \frac{C_{uj} f_u v_t}{(\sum_i \exp(\langle \log d_{ui}\rangle + \langle \log e_{it}\rangle)) + (C_{uj} f_u v_t)} \\ q(r_t = j) &= \frac{\mathcal{PO}(x_{ut};\sum_i \langle d_{ui}\rangle\langle e_{it}\rangle + C_{uj} f_u v_t)\pi_j}{\sum_j \mathcal{PO}(x_{ut};\sum_i \langle d_{ui}\rangle\langle e_{it}\rangle + C_{uj} f_u v_t)\pi_j}. \end{aligned}$$

where p^j_{uit} and p^j_{ut} represent the conditional posterior probability of i-th speech source and the j-th music source in frequency bin u and time frame t. $q(r_t = j)$ represents the posterior probability of the jingle frame index, j, at time t. The marginal expectation of the latent sources can be calculated using the parameters as:

$$\langle [r_t = j]\rangle = q(r_t = j) \quad (10)$$

$$\langle s_{uit}\rangle = x_{ut}(\sum_j \langle [r_t = j]\rangle p^j_{uit}) \quad (11)$$

$$\langle m_{ut}\rangle = x_{ut}(\sum_j \langle [r_t = j]\rangle p^j_{ut}) \quad (12)$$

The posterior distribution of the parameters of the template and excitation matrices are Gamma distributions due to the conjugacy property of the Poisson and Gamma distributions with parameters:

$$q(d_{ui}) \propto \mathcal{G}(d_{ui};\alpha^d_{ui},\beta^d_{ui}) \qquad q(e_{it}) \propto \mathcal{G}(e_{it};\alpha^e_{it},\beta^e_{it}) \quad (13)$$

$$\alpha^d_{ui} = a^d_{ui} + \sum_t \langle s_{uit}\rangle \qquad \alpha^e_{it} = a^e_{it} + \sum_u \langle s_{uit}\rangle \quad (14)$$

$$\beta^d_{ui} = (\frac{1}{b^d_{ui}} + \sum_t \langle e_{it}\rangle)^{-1} \qquad \beta^e_{it} = (\frac{1}{b^e_{it}} + \sum_u \langle d_{ui}\rangle)^{-1} \quad (15)$$

The sufficient statistics of these distribution can be calculated using the following equations:

$$\exp(\langle \log d_{ui}\rangle) = \exp(\Psi(\alpha^d_{ui}))\beta^d_{ui} \quad (16)$$

$$\exp(\langle \log e_{it}\rangle) = \exp(\Psi(\alpha^e_{it}))\beta^e_{it} \quad (17)$$

$$\langle d_{ui}\rangle = \alpha^d_{ui}\beta^d_{ui} \qquad \langle e_{it}\rangle = \alpha^e_{it}\beta^e_{it} \quad (18)$$

2.3. Speech-Music Separation Methods

2.3.1. NMF Based Separation

In this method, we use the fixed templates, which are trained using the speech segment of the audio, while separating speech from the music in the mixed part of the audio. Likewise the traditional NMF based approaches, the hierarchical prior on the template and excitation matrices are applied and the prior model for the template matrix is trained using the speech segment of the audio. The estimation of the prior model from the speech segment corresponds to the training phase of the traditional NMF method and described in [10] in great detail. In this method, the template matrix model, which represents the corresponding speech signal model, is fixed at the separation step and the excitation matrix is estimated in the mixed part of the audio using Equations 13-15.

2.3.2. Variational Based Separation

This separation strategy requires to update the speech model in the mixed part of the audio. In the first stage, using the speech segments of the audio, the prior model for the template matrix is estimated. In the second stage, using Variational Inference Method, which is described in Section 2.2, the posterior distribution of the template matrix and the corresponding excitation matrices are estimated and hence, the speech-music separation is performed. Instead of using the posterior distribution of the template and excitation matrices, the maximum a-posteriori (MAP) estimation of the matrices can be carried out using an iterative conditional modes (ICM) algorithm. The MAP estimation of the matrices can be carried out using the following update equations instead of Equations 7-9

$$\begin{aligned} q(S,M,R) &\propto p(X,S,M,R,D^*,E^*|\Theta) \\ D^* &\propto \arg\max_D(\exp(\langle \log \phi\rangle_{q(S,M,R)})) \\ E^* &\propto \arg\max_E(\exp(\langle \log \phi\rangle_{q(S,M,R)})) \end{aligned}$$

where $\phi = p(X,S,M,R,D^*,E^*|\Theta)$.

2.3.3. Joint Separation

Unlike the previous two methods, in this method, the speech and mixed segments of the audio are used simultaneously to train the speech models and to separate speech from the music. That is, the speech model, which corresponds to the template matrix of the speech signal, is estimated jointly with excitation matrices and the music signal parameters using both of the speech and music segments. The update equations for joint separation method are derived in Section 2.2.

3. Experimental Results

3.1. 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. The test set contains 704 utterances distributed approximately uniformly across 8 speakers. The total length of the test set is about 1 hour. The test utterances are mixed synthetically with a 4 sec. length jingle at 15dB SMR level to create the test set. The background music signal is generated by repeating the jingle up to the length of the speech. The jingle is taken from real broadcast news jingles. While WER of the clean speech data is %23, WER of the mixed data without any separation method is %59. The magnitude spectrogram is computed using 1024-point length frames and 512 point frame shift is used. In this study, we assume, only half of the jingle is labeled as music segment. That is, unlike the previous study [9], we do not have the whole of the jingle to separate speech from the music. In order to train the speech model, three types of speech data set are used and the properties of these sets are listed in Table 1.

Table 1: Speech Training Data Set Properties

Data Set	# of Speakers	Definition of the set	Length (min.)	# of Bases
Self	1	The same speaker	2	300
All	4	Including Speaker	8	600
Other	3	Excluding Speaker	6	600

3.2. Experimental Analysis

In our experiments, it is pointed out that the separation performance of Other type model is as good as the Self and All type models in terms of SMR, SAR and WER performance measures as shown in Tables 2, 3 and 4. This is a good result for the speech-music separation systems due to the fact that it is not always possible to make sure that the speaker in the mixed segment of the audio are in the training data of the speech model. It is surprising that the worst separation results are obtained using the Self model. The reason for that can be the insufficiency of the training data. However, Variational method improves the separation performance in terms of SAR and WER values as shown in Tables 3 and 4.

We observe that Joint method cannot increase the separation performance as compared to the other methods. This can be due to the fact that when we update the templates of the speech signal at the speech and mixed segments synchronously, the negative effect of the noisy observations, the mixed segment data, is more than their contribution to the training of the speech templates. In our experiments, we observe that although updating the templates with Self type increases the separation performance, it does not increase the performance with All and Other type models. The reason for that can be the length of the speech segment we used in the separation performance. That is, since the average length of the speech segment is about 5 seconds, this amount of the data is not enough to update the All and Other types of the models. Lastly, it is observed that SAR value of a separation method is more indicative than SMR values to show the effect of the separation method to ASR performance. For example, although the SMR value of the Self type with Joint method is the highest SMR value over all experiments, its ASR performance is the worst over all experiments.

We can use the previously proposed method [9] as a baseline for these experiments. SMR and SAR values of the previous method by using the half of the jingle is measured as 31.8 and 18.2 dB respectively. WER of the method is $\%48.1$. Although most of the proposed method improves the ASR performance as compared to the previous method, the improvement is not as high as expected. The reason for that in the previous method speech signal is recovered using only the mixed segment itself and this causes to decrease the artifacts of the separation method as compared to the currently proposed framework.

Table 2: Average SMR values (in dB) vs. Separation Methods

Prior Speech Data	Separation Methods			
	NMF	Variational	ICM	Joint
Self	34.2	34.0	33.3	**35.7**
All	34.6	34.6	33.8	33.4
Other	34.4	34.4	33.8	35.4

Table 3: Average SAR values (in dB) vs. Separation Methods

Prior Speech Data	Separation Methods			
	NMF	Variational	ICM	Joint
Self	17.2	17.6	18.3	16.7
All	18.5	18.6	**19.1**	17.2
Other	18.2	18.2	18.9	17.1

4. Conclusion

In this study, we extend the method which we proposed previously by incorporating the prior speech information to the speech-music separation task. Moreover, we also propose a Variational method to update the prior speech templates, which is estimated using the speech segment of the audio, in the mixed part of the audio. Furthermore, Joint estimation of the speech templates using both of the speech and mixed segment of audio is proposed. However, Joint estimation method does not increase the separation performance. We are planning to test the separation methods in a large database to show the performances. Moreover, we will try to use the weighted effect of the speech and mixed segments of the audio to estimate the templates in Joint method.

Table 4: Average WER values (in %) vs. Separation Methods

Prior Speech Data	Separation Methods			
	NMF	Variational	ICM	Joint
Self	48.6	44.9	47.3	48.3
All	42.6	43.6	**42**	47.5
Other	42.7	45.8	42.7	42.5

5. 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 TUBA-GEBIP award. ATC is funded by TUBITAK grant 110E292.

6. References

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

[2] E. Arısoy, H. Sak, and M. Saraçlar, "Language modeling for automatic Turkish broadcast news transcription," *Proc. of Interspeech*, 2007.

[3] M. Schmidt and R. 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] 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.

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

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

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

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