Effect of speech priors in single-channel speech-music separation for ASR

Cemil Demir1,3, A. Taylan Cemgil2, Murat Saraçlar3

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

Abstract

In this study, we extend the catalog-based single-channel speech-music separation method such that it incorporates prior speech information to enhance the separation performance of the method. We develop an inference method that enables us to use a prior speech model. The method uses a complex Gaussian observation model and an inverse-gamma prior model. We compare the separation performance of the catalog-based method with and without the prior speech model in both complex Gaussian and Poisson observation models. It is shown that for both observation models incorporating prior speech information improves the separation performance of the catalog-based method.

Index Terms: speech-music separation, prior speech model, speech recognition

1. Introduction

Recently automatic speech recognition (ASR) applications have become popular in broadcast news transcription systems. One major problem in such systems 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. The aim of this study is to extend the catalog-based speech-music separation method, that we proposed previously, such that it can improve the separation performance by using a prior speech model.

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] or music from different instruments [7]. In [5] it was shown that NMF-based approaches are capable of generating enhanced signals that significantly improve the speech recognition performance.

In previous studies [8, 9], we have introduced a 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 contains the jingles. By using the music segment of the audio, the jingle identity can be detected. For this study, we assume, the identity of the jingle is known as a prior. Each spectrum frame of the music is generated by a single mixture component, i.e., a jingle frame. The speech spectrum is generated by an Non-negative Matrix Factorization (NMF) model. The observed spectrum is the sum of the speech and music. Separation is achieved by joint estimation of the unknown parameters and latent variables of this hierarchical model.

In these studies [8, 9], we used the magnitude spectrogram of the signals which are assumed to be generated by a Poisson observation model. Using Poisson observation model in the generation of the magnitude spectrograms corresponds to minimizing Kullback-Leibler (KL) Divergence between the true and estimated values of the magnitude spectrograms of the signals. In a previous study [10], we extend the catalog-based approach to incorporate prior speech information into the separation framework. However, in [10], we proposed using Gamma prior models for the speech signal due to the conjugacy property of Poisson observation model.

As an extension to the previous studies [8, 9], in this study, we use a prior speech model to increase the performance of the catalog-based speech-music separation method. Unlike the previous studies [8, 9, 10], we use a complex Gaussian observation model which corresponds to minimizing Itakura-Saito (IS) Divergence between the true and estimated values of the power spectrograms of the signals. As the prior model, we use an Inverse-Gamma prior for the speech signal due to the conjugacy property of the variance parameter of the complex Gaussian distribution.

2. Catalog-Based Separation With Speech Priors

In this section, we develop the inference method for incorporating prior speech model into the separation framework which uses complex Gaussian model to generate the spectrum of the signals.

2.1. Model Description

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

\[ X_{tu} = S_{tu} + M_{tu} \]

where \( S \) and \( M \) represents the complex spectrum of the speech and music signals, respectively. We assume an NMF based generative model, which uses a complex Gaussian observation model [11], for the complex spectrum of the speech.

In this probabilistic model, each time-frequency entry of the complex spectrum of the speech signal is generated by \( B \) latent
complex Gaussian sources as

\[ S_{uit} = \sum_{i=1}^{B} s_{uit}, \quad s_{uit} \sim \mathcal{N}_c(s_{uit}; 0, U_{ui}V_{it}) \]

where \( \mathcal{N}_c \) represents the complex Gaussian distribution. The template matrix, \( U \), and the excitation matrix, \( V \), contain the hyper-parameters of the complex spectrum of the speech signal.

In complex Gaussian model, the latent sources are complex Gaussians and they generate the complex spectrum of the speech signal. Moreover, maximization of the likelihood of the complex spectrum of the signal with complex Gaussian sources corresponds to minimizing the Itakura-Saito (IS) divergence between the power spectrogram of the signal and its NMF approximation [11]. Complex Gaussian density of the random variable \( s \) is given as

\[ \mathcal{N}_c(s; \mu, \Sigma) = |\pi\Sigma|^{-1/2} \exp(-\frac{1}{2} (s - \mu)^H \Sigma^{-1} (s - \mu)) \]

In this study, we assume a Inverse-Gamma prior on the template matrix as follows:

\[ U_{ui} \sim \mathcal{IG}(U_{ui}; a^u_{ui}, b^u_{ui}) \]

where \( a^u_{ui}, b^u_{ui} \) are the hyper-parameters of the template matrix. Inverse-Gamma distribution is defined as:

\[ \mathcal{IG}(x; a, b) = \exp(- (a+1) \log x - \frac{1}{2} \log \Gamma(a)) \]

We also use a complex Gaussian observation model in the generative model of the complex spectrum of the music signal as

\[ M_{uit} = m_{uit} | r_t \sim \mathcal{N}_c(m_{uit}; 0, C_{uj}f_{ut}v_t)^{[r_t=m]} \]

where \([r_t = j]\) represents the indicator function, which is 1 when \( j \)-th frame of the jingle is used and its value is 0, otherwise. In Equation (1), \( C_{uj} \) represents the power spectrogram corresponding to the \( u \)-th frequency bin and the \( j \)-th frame of the jingle, \( f_{ut} \) represents frequency filtering parameter for frequency bin \( u \) and \( v_t \) represents the gain parameter for time frame \( t \). The goal is here to model volume changes (fade-in, fade-out) and filtering (equalization). Each active jingle frame is drawn independently from a set of jingle frames as

\[ r_t = j \in \{1, 2, ..., N\} \text{ with probability } \pi_j \]

where \( \pi \) represents probability distribution on the jingle frames and \( N \) represent the number of frames in the jingle. The difference from the speech model is that, the variance parameter of the complex Gaussian model is chosen from a power spectrogram of a set of previously obtained jingle frames. Moreover, a filtering and gain adjustment is applied to that variance parameter.

The overall graphical model corresponding to the generation of the mixture of the speech and music signals is shown in Figure 1.

### 2.2. Variational Inference

In this section, we describe the inference technique used to derive the update equations of the posterior distributions of the latent sources and parameters of the speech and music signals in the probabilistic model. Since the posterior distributions of the template matrix, \( U \) and the latent speech, music and active frame sources, \( s, m, r \) are coupled, we cannot compute the overall posterior distribution exactly. Therefore, we use the variational inference technique that factorizes the posterior distribution into the posteriors of the decoupled random variables as follows:

\[ q(s, m, r) \propto \exp((\log \phi)_{q(s)}) \]

\[ q(U) \propto \exp((\log \phi)_{q(s, m, r)}) \]

\[ V^* \propto \arg \max_{V} (\log \phi)_{q(s, m, r)}(U) \]

where \( \phi = p(X, s, m, U, \Theta) \) and \( \Theta \) represents the \( a^u_{ui}, b^u_{ui}, \pi, f, v \). The joint posterior distribution of the latent speech and music sources and jingle indexes, \( q(s, m, r) \), is a complex Gaussian mixture model (CGMM) as shown in [8]. The overall joint posterior distribution of the latent speech and music sources can be decomposed conditioned on the jingle frame index, \( j \), as

\[ q(s, m, r) = q(s, m|r)q(r) \]

Conditional posterior distribution of the latent speech and music sources are complex Gaussian distributed as:

\[ p(s_{uit}|X, r_t) = \mathcal{N}_c(s_{uit}; \mu_{uit}^{(j)}, \Sigma_{uit}^{(j)}) \]

\[ p(m_{uit}|X, r_t) = \mathcal{N}_c(m_{uit}; \mu_{uit}^{(k)}, \Sigma_{uit}^{(k)}) \]

The parameters of the complex Gaussian distributions of the latent speech and music sources can be computed using:

\[ \mu_{uit} = \nabla_{\mu_{uit}} \log p(s_{uit}|X, r_t) \]

\[ \Sigma_{uit} = \nabla_{\Sigma_{uit}} \log p(s_{uit}|X, r_t) \]

\[ p_{uit} = p_{uit}^{(j)} X_{uit} \]

\[ p_{uit}^{(k)} = p_{uit}^{(k)} X_{uit} \]

\[ \mu_{uit} = \mu_{uit}^{(j)} X_{uit} \]

\[ \Sigma_{uit}^{(j)} = \Sigma_{uit}^{(j)} X_{uit} \]
where \( p_{ui}^{\alpha} \) and \( p_{ui}^{\beta} \) are the auxiliary variables used to shorten the equations. The conditional posterior distribution of the jingle indexes can be computed as follows:

\[
q(r_t | X) = \prod_{c=1}^{C} \frac{N_c(X_{ui}; 0, C_{ui} f_u v_t + \sum_j \frac{1}{\pi_{ij}^{c}} V_{it}) \pi_{ij}^{c}}{\sum_j \prod_{c=1}^{C} N_c(X_{ui}; 0, C_{ui} f_u v_t + \sum_j \frac{1}{\pi_{ij}^{c}} V_{it}) \pi_{ij}^{c}}.
\]

The conditional marginal expectation of the latent sources can be calculated using the parameters as:

\[
\langle r_t = j \rangle = q(r_t = j | r, \theta)
\]

\[
\langle s_{ui}^{j} \rangle = \frac{\sum_{t=1}^{T} \langle r_t = j \rangle \langle s_{ui}^{j} \rangle}{\sum_{t=1}^{T} \langle r_t = j \rangle}
\]

\[
\langle m_{ui}^{j} \rangle = \frac{\sum_{t=1}^{T} \langle r_t = j \rangle \langle m_{ui}^{j} \rangle}{\sum_{t=1}^{T} \langle r_t = j \rangle}
\]

The posterior distribution of the each entry of the template matrix is Inverse-Gamma distribution due to the conjugacy property of the complex Gaussian and Inverse-Gamma distributions with parameters:

\[
q(U_{ui}) \propto \frac{1}{\Gamma(\alpha_{ui}, \beta_{ui})} \alpha_{ui}^{\alpha_{ui}} \beta_{ui}^{\beta_{ui}}
\]

\[
a_{ui} = a_{ui}^{\alpha_{ui}} + T
\]

\[
\beta_{ui}^{\beta_{ui}} = \left( \frac{1}{b_{ui}} + \sum_{j=1}^{T} \langle r_t = j \rangle \langle s_{ui}^{j} \rangle \langle s_{ui}^{j} \rangle \right)^{-1}
\]

The excitation matrix which maximizes the expectation of the joint log-likelihood of the data is calculated using the following equation:

\[
V_{it} = \frac{1}{P} \sum_{u,j} \langle r_t = j \rangle \langle m_{ui}^{j} \rangle \frac{1}{\langle s_{ui}^{j} \rangle}\n\]

The gain parameter for each time frame, \( v_t \) can be found using

\[
v_t = \frac{1}{P} \sum_{u,j} \langle r_t = j \rangle \langle m_{ui}^{j} \rangle \frac{1}{\langle s_{ui}^{j} \rangle}
\]

2.3. Gain Estimation Strategies

In our analysis, it was realized that the accuracy of the estimated gain values in complex Gaussian model with Equation 2 is very low. In a previous study [9], we analyzed the gain estimation problem of catalog-based method with Poisson model. In [9], we proposed using a Gamma Markov Chain (GMC) which is a prior structure for a chain of positive variables, where the correlation between consecutive variables is positive. Since the gain value of each frame is estimated independently, the abrupt change of gain values of consecutive frames is possible. In order to prevent the abrupt changes in the gain values, Inverse-Gamma-Markov-Chain (IGMC) is used to impose the correlation between consecutive variables for the catalog method with complex Gaussian model.

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 240 utterances distributed approximately uniformly across 8 speakers. The total length of the test set is about 70 minutes and the average length of the utterances is about 18 seconds.

The test utterances are mixed synthetically with 10 different jingles at 10dB Speech-to-Music Ratio (SMR) level to create the test set. The average length of the jingles is 7 seconds. The background music signal is generated by repeating the jingles up to the length of the speech. The jingles are taken from real broadcast news jingles. In this study, we assume, the which jingle is used to generate the background music is known as a prior. While Word Error Rate (WER) of the clean speech data is %26.1, WER of the mixed data without any separation method is %38.2. The magnitude or power spectrum are computed using 1024-point length frames and 512 point frame shift is used. 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. In our approach, the prior speech model contains the hyper-parameters of gamma distributions (Gamma distribution in Poisson model or Inverse-Gamma distribution in complex Gaussian model). It is assumed that each frequency bin of the template vector of the speech signal has a different gamma distribution.

Table 1: Speech Training Data Set Properties

<table>
<thead>
<tr>
<th>Data Set</th>
<th># of Speakers</th>
<th>Definition of the set</th>
<th>Length (min.)</th>
<th># of Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self</td>
<td>1</td>
<td>The same speaker</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>All</td>
<td>4</td>
<td>Including Speaker</td>
<td>8</td>
<td>30</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>Excluding Speaker</td>
<td>6</td>
<td>30</td>
</tr>
</tbody>
</table>

In our experimental study, the effects of three major factors on the catalog-based separation method are tested. These factors are can be listed as follows. Divergence Measure (KL or IS): The aim is to compare the effect of divergence measures on the separation performance. Prior Speech Data Type (None (N), Self (S), All (A) and Other (O): The aim is to analyze the effect of imposing different types of speech training data on the separation performance. None type refers to the catalog-based method without prior speech model which is used in [8].

Gain Estimation Strategy (The original method (O) and Gamma Chains (G)): As a complete example of naming, 'IS-G-A' represents separation with IS divergence with IGMC is used as the gain estimation method and 'All' type speech training data is used to get the prior speech model.

3.2. Experimental Analysis

In our experiments, it is seen that the separation performance of Other type model is as good as the Self and All type models in terms of SMR, Source-to-Artifact Ratio (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 separation results obtained using the Self model are not better than All and Other models. The reason for that can be the insufficiency of the training data in Self model. When we analyze the SMR results in Table 2, it is shown that, for both of divergence measures (KL or IS), prior speech model (All, Self and Other) increases the SMR values with Original gain estimation strategy. However, in GMC case, the SMR values are almost the same as with no speech prior
Table 2: Average SMR values (in dB) vs. Separation Methods

<table>
<thead>
<tr>
<th>Prior Speech Data</th>
<th>Divergence Measure</th>
<th>KL</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>GMC</td>
<td>Original</td>
</tr>
<tr>
<td>None</td>
<td>25.9</td>
<td>33.5</td>
<td>29.6</td>
</tr>
<tr>
<td>Self</td>
<td>29.5</td>
<td>33.2</td>
<td>31.5</td>
</tr>
<tr>
<td>All</td>
<td>28.9</td>
<td>33.6</td>
<td>31.3</td>
</tr>
<tr>
<td>Other</td>
<td>28.5</td>
<td>33.4</td>
<td>31.2</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Prior Speech Data</th>
<th>Divergence Measure</th>
<th>KL</th>
<th>IS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
<td>GMC</td>
<td>Original</td>
</tr>
<tr>
<td>None</td>
<td>15.6</td>
<td>17.2</td>
<td>11.5</td>
</tr>
<tr>
<td>Self</td>
<td>16.1</td>
<td>17.6</td>
<td>18.3</td>
</tr>
<tr>
<td>All</td>
<td>16.6</td>
<td>17.6</td>
<td>18.2</td>
</tr>
<tr>
<td>Other</td>
<td>16.3</td>
<td>17.5</td>
<td>18.2</td>
</tr>
</tbody>
</table>

model (None). For SAR values in Table 3, it is observed that for both of observation models (KL or IS) and Gain estimation strategies (Original or GMC), incorporating prior speech information increases the SAR values. Since the speech recognition performance of a separation method is affected by both of the SMR and the SAR values, using prior speech model in the separation for all conditions (KL or IS, Original or GMC) improves the speech recognition performance. This fact can be seen in Table 4.

When we compare the effects of prior speech models and the gain estimation strategies for divergence measures, it is really interesting that using prior speech models in IS case improves the separation performance more than using GMC in gain estimation. However, for KL case, using GMC in gain estimation makes more improvement than using prior speech models in the separation. While the relative improvement in WER with prior speech model in KL case (KL-O-A compared to KL-O-N) is %21.8, with GMC in KL case (KL-G-N compared to KL-O-N) is %32.7. For IS case, relative improvement with prior speech model (IS-O-A compared to IS-O-N) and GMC (IS-G-N compared to IS-O-N) are %16.3 and %10.8, respectively. The reason why using the prior speech model or GMC in gain estimation strategy makes more relative improvement in KL case than IS case is that the baseline separation performance of IS (IS-O-N) is better than KL (KL-O-N) case. As a result, incorporating prior speech information in the separation method enhances the separation performance of both of divergence measures with both of gain estimation strategies (Original or GMC).

4. Conclusion

As a conclusion, in this study we propose using Inverse-Gamma prior speech models in the catalog-based single-channel speech-music separation method which we proposed previously. We developed the variational inference method for the complex Gaussian model (IS) which uses the prior speech model in the separation method. We evaluated the proposed separation method in speech recognition test and show the advantage of using prior speech model in the separation method. Moreover, we compare the effect of prior speech model in Poisson (KL) and complex Gaussian (IS) models with different gain estimation strategies. As a result we showed that using prior speech information improves the separation performance.

5. Acknowledgements

This research is supported in part by TUBITAK (Scientific and Technological Research Council of Turkey) (Project code: 105E102). Murat Saracılı is supported by the TUBA-GEBIP award. ATC is funded by TUBITAK grant 110E292.

6. References