Gain Estimation Approaches in Catalog-Based Single-Channel Speech-Music Separation

Cemil Demir [1,3], Ali Taylan Cemgil [2], Murat Saraclar [3]

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

***Abstract*—In this study, we analyze the gain estimation problem of the catalog-based single-channel speech-music separation method, which we proposed previously. 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 model, we assume that the background music is generated by repeating and changing the gain of the jingle in the music catalog. Although the separation performance of the proposed method is satisfactory with known gain values, the performance decreases when the gain value of the jingle is unknown and has to be estimated. In this paper, we address the gain estimation problem of the catalog-based method and propose three different approaches to overcome this problem. One of these approaches is to use Gamma Markov Chain (GMC) probabilistic structure to impose the correlation between the gain parameters across the time frames. By using GMC, the gain parameter is estimated more accurately. The other approaches are maximum a posteriori (MAP) and piece-wise constant estimation (PCE) of the gain values. Although all three methods improve the separation performance as compared to the original method itself, GMC approach achieved the best performance.**

I. 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], [12], [13] , 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 a Non-Negative Matrix Factorization (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 NMF model. This way, by finding the parameters of the NMF model, we can recover the speech signal from the mixture.

Unlike the previous studies [11], [12], we address the gain estimation problem of the catalog-based method and propose three different techniques to enhance the gain estimation performance of the catalog-based method. With the analysis of the problem, we decided to use Gamma Markov Chain (GMC) probabilistic structure to impose the correlation between the gain parameters across the time frames. As alternative to this approach, we developed the maximum a posteriori (MAP) and piece-wise constant estimation (PCE) of the gain values. The separation performance of the method with gain estimation approaches is improved as compared to the method itself. However, the best improvement is obtained using the GMC approach.

This paper is organized as follows. In Section II, we overview the catalog-based speech-music separation method. In Section III, we analyze the gain estimation problem of the catalog-based speech-music separation method and propose three different gain estimation techniques. The experimental results and comparisons are provided in Section IV. Section V presents the discussion, conclusions and comments for further investigation.

978-1-4673-0367-5/11/$26.00 ©2011 IEEE

II. Catalog-Based Speech-Music Separation

In catalog-based 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. Moreover, the background music is composed of the jingles in the catalog. Which jingle is used to create the background music can be detected using the music parts of the audio. Although speech part of the segmented audio can be used in the separation phase, in this work we do not use the speech segment to separate speech from the mixture.

1) Model Description: In this model, we 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 [14], 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 where $M_{ut} = m_{ut}$ 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 the filtering parameter for frequency bin u and v_t represents the gain parameter for time frame t. The goal is here to model gain changes (fade-in, fade-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, the intensity parameters are multiplied by a gain factor and a filter.

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.

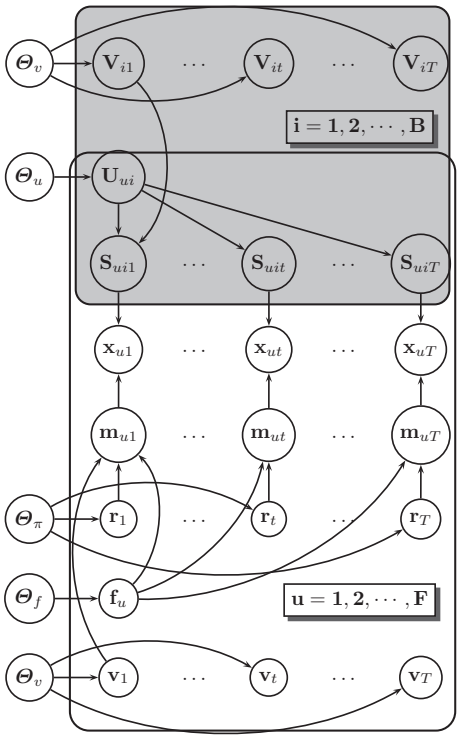

Fig. 1. Graphical Model For Speech-Music Mixture.

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. For each j, the posterior distribution of the latent sources is a multinomial distribution as follows

$$\mathcal{M}(s^j_{u1t}, .., s^j_{uBt}, m^j_{ut}; X_{ut}, p^j_{u1t}, ..., p^j_{uBt}, p^j_{ut}) \tag{5}$$

where $\mathcal{M}$ represents the multinomial distribution. The parameters of the multinomial distribution which corresponds to the conditional posterior probability of i-th speech source and the j-th music source in frequency u and time t can be found as follows:

$$p^j_{uit} = \frac{U_{ui}V_{it}}{\sum_i U_{ui}V_{it} + C_{uj}f_u v_t} \tag{6}$$

$$p^j_{ut} = \frac{C_{uj}f_u v_t}{\sum_i U_{ui}V_{it} + C_{uj}f_u v_t} \tag{7}$$

The conditional marginal expectations of the latent sources in poisson model are:

$$\langle s^j_{uit} \rangle = X_{ut}p^j_{uit} \quad \text{and} \quad \langle m^j_{ut} \rangle = X_{ut}p^j_{ut}. \tag{8}$$

As a result, the hyper-parameters of the speech and music 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

$$U_{ui} = \frac{\sum_{t,j} \langle [r_t = j] \rangle \langle s^j_{uit} \rangle}{\sum_t V_{it}} \tag{9}$$

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

$$p(s_{uit}|r_t = j) = \sum_{m_{ut}} p(s_{uit}, m_{ut}|r_t = j). \tag{10}$$

Here, $\langle [r_t = j] \rangle$ represents expected value of active frame index $r(t)$ being equal to j at time frame t which is equal to the posterior probability of the active frame index and can be calculated as follows:

$$p(r_t = j|X) = \frac{\prod_{u,t} PO(X_{ut}; C_{uj} f_u v_t + \sum_i U_{ui} V_{it}) \pi_j}{\sum_j \prod_{u,t} PO(X_{ut}; C_{uj} f_u v_t + \sum_i U_{ui} V_{it}) \pi_j}. \tag{11}$$

Each entry of the excitation matrix of the speech spectrogram, V can be calculated using

$$V_{it} = \frac{\sum_{u,j} \langle [r_t = j] \rangle \langle s^j_{uit} \rangle}{\sum_u U_{ui}}. \tag{12}$$

The gain parameter at each time frame can be found by using

$$v_t = \frac{\sum_{u,j} \langle [r_t = j] \rangle \langle m^j_{ut} \rangle}{\sum_{u,j} \langle [r_t = j] \rangle C_{uj} f_u} \tag{13}$$

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

$$f_u = \frac{\sum_{t,j} \langle [r_t = j] \rangle \langle m^j_{ut} \rangle}{\sum_{t,j} \langle [r_t = j] \rangle C_{uj} v_t}. \tag{14}$$

III. Gain Estimation Problem

When we use the update in Equation 13 to estimate the gain parameter for each time frame, it is observed that the estimation error is very high at the mixture frames which either have

1) low input Music-to-Speech Ratio (MSR) values or
2) active catalog frames with low energy.

Figure 2 shows these two facts visually. In this example, the true gain parameter is constant at 1 for all frames. For example, for the first five frames though the input MSR values are high, the gain estimation error is very high due to the fact that the active catalog frame has low frame energy and so it is confused with other frames. This fact can be seen in Figure 3. When we analyze the gain parameter of the frames between the time indices twenty and twenty five, it can be seen that though the active catalog frames have high energy, the gain estimation error is high due to the fact that input MSR value is low and the speech signal suppresses the music signal. In fact, at these parts, the inference method cannot estimate the posterior probabilities of the catalog frames accurately. That is, the method cannot decide which part of the catalog is active at these parts. This fact is shown in Figure 3. Although most of

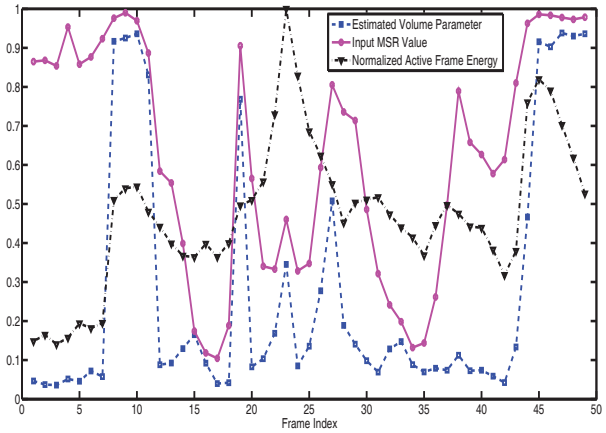

Fig. 2. Gain Estimation Problem Reasons: Low Input MSR and Low Active Frame Energy.

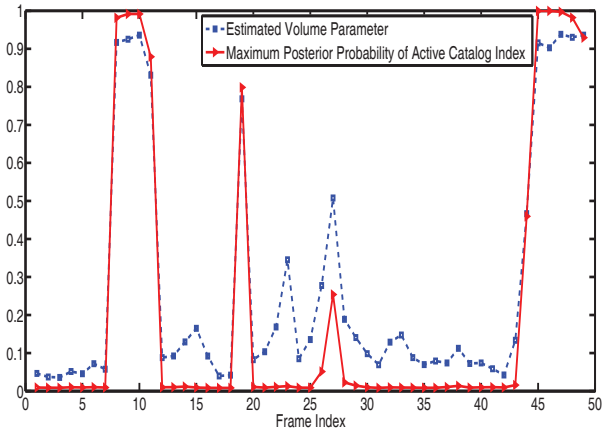

Fig. 3. Relation between the Gain Parameter and the MAP.

the maximum a posterior probabilities (MAP) of the catalog frames are very low, 67% of the frames with MAP are the actual active frames for this example.

Using this analysis about the gain estimation problem, we propose two different correction methods so as to enhance the gain estimation performance of the inference method. The methods are called "MAP Estimation Method" and "Piece-wise Constant Estimation Method".

A. MAP-Estimation Method

Experimentally, we observe that although MAP for most of the mixture frames are very low, the frames which have MAP are indeed the active frames. Therefore, after some iterations with the original posterior update Equation 11, the frames with the MAP can be chosen as the active frames. Then the posterior probability of these MAP frames are assigned to 1 so as to estimate the gain parameter more accurately. After this assignment, even though the posterior probabilities are not updated, other update rules are applied via reassigned posterior probabilities. This approach can be shown mathematically as

follows:

$$r_t^* = \arg\max_{r_t} p(r_t|X,\theta) \tag{15}$$

$$p(r_t = j) = \begin{cases} 1 & \text{if } j = r_t^*, \\ 0 & \text{Otherwise} \end{cases} \tag{16}$$

B. Piece-wise Constant Estimation

When we analyze the gain estimation results obtained using the original update Equation 13 in Figure 3, it is observed that when the MAP of a frame is high enough, the gain parameter for this frame is estimated correctly. Therefore, we can use the gain parameter of the closest frame which has high MAP values as gain parameter for the frame which has low MAP value. The question here is, what will be used as the threshold to decide whether the MAP of a frame is low or high? We decide on this threshold value using a development set which maximizes the separation performance. We call this estimation method, given in Algorithm 1, as 'Piece-wise Constant Estimation' because of the fact that the resultant gain parameter is a piece-wise constant version of the originally estimated gain parameter.

Algorithm 1 Piece-wise Constant Estimation Algorithm

```
f=empty
for t = 1 to T do
  if MAP(t) ≥ Threshold then
    Add t to f
  end if
end for
L = length(f)
i = 1
for t = 1 to T do
  if i < L and |t − f(i+1)| < |t − f(i)| then
    i = i + 1;
  end if
  v_es(t) = v_es(i)
end for
```

C. Gamma Markov Chain For Gain Estimation

A Gamma Markov chain (GMC) [15], which is shown in Figure 4, is a prior structure for a chain of positive variables, where the correlation between consecutive variables is positive. In addition, each variable is conditionally conjugate, i.e., their prior and full conditional distributions are Gamma. A GMC of $v_{1:T}$ can be defined as

$$v_1 \sim \mathcal{G}(v_1; a_v, b_v/a_v) \tag{17}$$

$$z_t|v_t \sim \mathcal{IG}(z_t; a_z, a_z v_t) \tag{18}$$

$$v_{t+1}|z_t \sim \mathcal{G}(v_{t+1}; a_v, z_t/a_v) \tag{19}$$

where a_v, a_z, b_v are the hyper-parameters of the chain and $z_{1:T-1}$ are auxiliary variables introduced to have positive correlation and conjugacy properties simultaneously. a_v and a_z are the coupling hyper-parameters and they determine the degree of correlation between variables. $\mathcal{G}$ and $\mathcal{IG}$ represent Gamma and Inverse-Gamma distributions respectively.

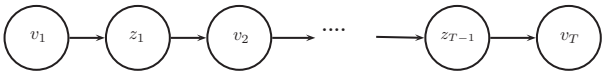

Fig. 4. GMC Graphical Model For Gain Parameter

The full joint distribution of the catalog-based model with GMC can be decomposed as:

$$\begin{aligned} \log\phi &= \log p(X,s,m,r,v,z|\Theta) \\ &= \log p(X|s,m) + \log p(s|\Theta^{u,v}) + \log p(m|r,v) + \\ &\quad \log p(v,z|\Theta^v) + \log p(r|\Theta) \end{aligned}$$

where Θ represents the hyper-parameters of the latent speech and music sources. Since the posterior distributions of the gain parameters, v, z and the hidden sources are coupled, we cannot compute the overall joint 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:

$$\begin{aligned} q(s,m,r) &\propto \exp(\langle \log p(X,s,m,r,v,z|\Theta)\rangle_{q(v)q(z)}) \\ q(v) &\propto \exp(\langle \log p(X,s,m,r,v,z|\Theta)\rangle_{q(s,m,r)q(z)}) \\ q(z) &\propto \exp(\langle \log p(X,s,m,r,v,z|\Theta)\rangle_{q(s,m,r)q(v)}) \end{aligned}$$

The joint posterior distribution of the latent speech and music sources and the catalog indexes is also a multinomial mixture model (MMM). However, the calculation of the parameters of the distribution differ from the original model which is described in Section II. The overall posterior distribution can be decomposed conditioned on the catalog 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:

$$\begin{aligned} p^j_{uit} &= \frac{U_{ui}V_{it}}{(\sum_i U_{ui}V_{it}) + C_{uj}f_u \exp(\langle \log v_t\rangle)} \\ p^j_{ut} &= \frac{C_{uj}f_u \exp(\langle \log v_t\rangle)}{(\sum_i U_{ui}V_{it}) + C_{uj}f_u \exp(\langle \log v_t\rangle)} \\ q(r_t = j) &= \frac{\prod_{u,t} \mathcal{PO}(X_{ut}; \sum_i U_{ui}V_{it} + C_{uj}f_u\langle v_t\rangle)\pi_j}{\sum_j \prod_{u,t} \mathcal{PO}(X_{ut}; \sum_i U_{ui}V_{it} + C_{uj}f_u\langle v_t\rangle)\pi_j} \\ &= \langle [r_t = j]\rangle \end{aligned}$$

The only difference from Equations 6 and 11 is that instead of using v_t , its expectations are used to calculate the posteriors. The marginal expectation of the latent sources under the posterior distribution can found using:

$$\langle s_{uit}\rangle = X_{ut}(\sum_j \langle [r_t = j]\rangle p^j_{uit}) \tag{20}$$

$$\langle m_{ut}\rangle = X_{ut}(\sum_j \langle [r_t = j]\rangle p^j_{ut}) \tag{21}$$

$$\tag{22}$$

Now, the posterior distribution of the gain parameter, v_t and the auxiliary variable, z_t, are calculated. The posterior of the gain parameter, v_t, is also Gamma-distributed due to the conjugacy of Poisson and Gamma distributions. The posterior distribution of v_{t+1} conditioned on the auxiliary variable z_t is:

$$q(v_{t+1}) \propto \mathcal{G}(v_{t+1}; \alpha^v_{t+1}, \beta^v_{t+1}) \quad (23)$$

$$\alpha^v_{t+1} = a_v + \sum_u \langle m_{u(t+1)} \rangle \quad (24)$$

$$\beta^v_{t+1} = (a_v \langle \frac{1}{z_t} \rangle + \sum_{u,j} C_{uj} f_u)^{-1} \quad (25)$$

The sufficient statistics of the gain parameter, which are used to estimate the posteriors of the other parameters are:

$$\exp(\langle \log v_{t+1} \rangle) = \exp(\Psi(\alpha^v_{t+1}))\beta^v_{t+1} \quad (26)$$

$$\langle v_{t+1} \rangle = \alpha^v_{t+1}\beta^v_{t+1} \quad (27)$$

where Ψ denotes the digamma function defined as $\Psi(\alpha) \equiv d \log \Gamma(a)/da$. We also need to compute the posterior and the sufficient statistics of the inverse of the gain parameter which has an Inverse-Gamma distribution as follows:

$$\frac{1}{v_{t+1}} \sim \mathcal{IG}(\frac{1}{v_{t+1}}; \alpha^v_{t+1}, \frac{1}{\beta^v_{t+1}}) \quad (28)$$

$$\langle \frac{1}{v_{t+1}} \rangle = \frac{1}{(\alpha^v_{t+1} - 1)\beta^v_{t+1}} \quad (29)$$

The posterior of auxiliary variable, z_t, is also Inverse Gamma-distributed due to the conjugacy of Poisson and Inverse Gamma distributions. The posterior distribution of z_t conditioned on the gain parameter, v_t is:

$$q(z_t) \propto \mathcal{IG}(z_t; \alpha^z_t, \beta^z_t) \quad (30)$$

$$\alpha^z_t = a_z \quad \text{and} \quad \beta^z_t = (\frac{1}{a_z}\langle \frac{1}{v_t} \rangle)^{-1} \quad (31)$$

The sufficient statistics of the auxiliary variable, which are used to estimate the posterior of the gain parameter are:

$$\langle z_t \rangle = \frac{\beta^z_t}{\alpha^z_t - 1} \quad (32)$$

We also need to compute the posterior and the sufficient statistics of the inverse of the auxiliary variable which has a Gamma distribution as follows:

$$\frac{1}{z_t} \sim \mathcal{G}(\frac{1}{z_t}; \alpha^z_t, \frac{1}{\beta^z_t}) \quad \text{and} \quad \langle \frac{1}{z_t} \rangle = \frac{\alpha^z_t}{\beta^z_t}. \quad (33)$$

IV. Experimental Results

A. 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 with 10ms shift of the clean speech data. For each gender , 40 hours of speech data is used to train context dependent phone models. 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 Speech-to-Music Ratio (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 reason why we use a larger window and shift size is to decrease the computational complexity of the separation algorithm. The number of speech bases is fixed at 30. The speech recognition performance is measured using Word Accuracy Ratio (WAcc).

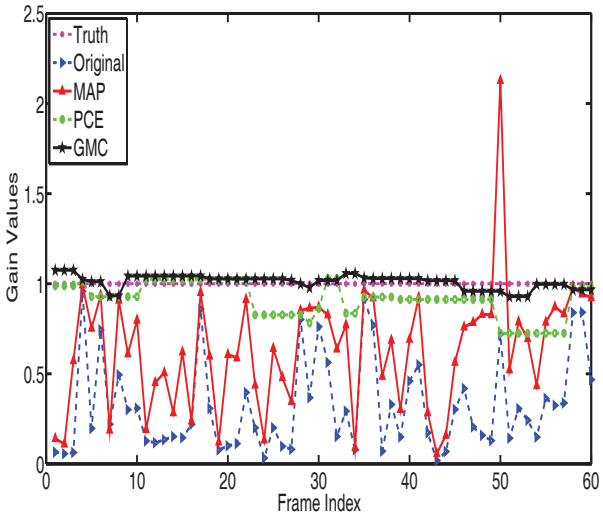

Fig. 5. Estimation of constant gain parameter.

TABLE I
AVERAGE SMR VALUES (DB)

Output SMR (dB)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
Truth	17.6	24.2	32.1	38.2	46.2
Original	6.5	15.3	24.3	33.5	42.8
MAP	12.5	20.3	28.4	36.8	45.3
PCE	15.3	22.5	30.1	37.9	46.2
GMC	18.5	24.6	31.4	38.8	46.7

B. Experimental Analysis

In this section, the effects of proposed gain estimation techniques to the separation performance are analyzed and compared. As an example for comparing estimation performances of the methods, the estimated gain parameter values for constant and fading gain cases are shown in Figure 5 and 6.

When we examine the gain estimation results with the original method in Table I, II and corresponding estimated gain parameter in Figure 5, it is observed that when the speech signal suppresses the music signal, the gain parameter is under-estimated. Therefore, music signal is contaminated in speech signal and so SMR value of the original method is

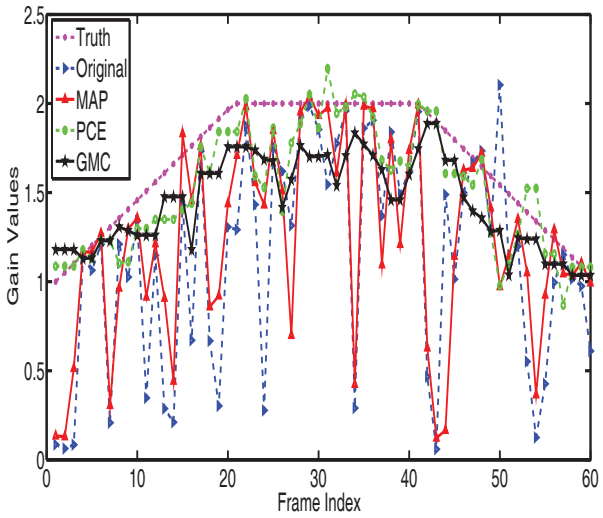

Fig. 6. Estimation of fading gain parameter.

TABLE II
AVERAGE SAR VALUES (DB)

Output SAR (dB)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
Truth	10.9	14.2	17.2	20.2	23.2
Original	7.5	11.4	14.9	18.4	21.9
MAP	10.5	13.8	16.8	19.6	22.3
PCE	11.7	14.5	17.3	20.1	22.9
GMC	11.7	14.4	17.1	20.1	23.1

very low compared to "Truth" case. In this part, the original method corresponds to estimating the gain parameter with Equation 13. When we use MAP, since some of the frames with low MAP are actual active frames, the estimated gain parameters for these frames are increased, so the SMR and Speech-to-Artifact Ratio (SAR) values are higher compared to the original case. However, for the frames whose active frames are not correctly identified, the gain parameter is over or under estimated. In PCE case, the gain parameter is estimated using the frames which have high MAP, so the gain parameter of the frames are smoothed over these frames. As a result, the gain estimation performance increases as compared to the original case. In GMC method, by imposing correlation between the gain parameter along the frames, it is not allowed to have abrupt changes in the estimated values. This scenario is more realistic because of the fact that the gain parameter is not changed instantaneously in real life. ASR results with different gain estimation techniques are presented in Table III and it is experimentally shown that the proposed gain estimation techniques enhance the speech recognition results. It is very promising that by using the proposed techniques the speech recognition performance can be improved to a level that is very close to the speech recognition performance with the true gain values as can be seen in Table III.

V. CONCLUSIONS

As a conclusion, we address the gain estimation problem of the catalog-based method and propose three different solutions to this problem. MAP and Piece-wise constant estimation

TABLE III
AVERAGE WACC VALUES (%)

WAcc (%)	Input SMR Values				
Method	0dB	5dB	10dB	15dB	20dB
Clean	75.1	75.1	75.1	75.1	75.1
Mixed	0.4	2.6	15.3	40.9	60.4
Truth	29.2	46.7	59.4	66.9	70.4
Original	3.5	14.8	41.7	52.4	66.6
MAP	13.7	35.1	48.8	61.2	69.5
PCE	17.6	35.9	55.6	63.6	70.0
GMC	26.1	41.5	57.3	63.4	70.9

methods are ad-hoc methods which we developed by analyzing the reasons of estimation errors. Also we applied GMC structure to overcome this gain estimation issue. It is shown that all of these enhancement techniques improves the gain estimation performance of the catalog-based method. Moreover, by using the proposed approaches the separation performance can be improved as if the truth gain values are used in the separation process.

VI. 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. Taylan Cemgil is supported by the Bogazici University research grant BAP 09A105P.

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. 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. 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] 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. Weiss and D. Ellis, "Speech separation using speaker-adapted eigen-voice 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. Cemgil, and M. Saraçlar, "Catalog-Based Single-Channel Speech-Music Separation," in *Proc. of Interspeech*, 2010.
[12] ——, "Catalog-Based Single-Channel Speech-Music Separation For Automatic Speech Recognition," in *Proc. of EUSIPCO*, 2011.
[13] ——, "Semi-supervised Single-Channel Speech-Music Separation For Automatic Speech Recognition," in *Proc. of Interspeech*, 2011.
[14] A. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009.
[15] A. Cemgil and O. Dikmen, "Conjugate gamma Markov random fields for modelling nonstationary sources," *Independent Component Analysis and Signal Separation*, pp. 697–705, 2007.