CATALOG-BASED SINGLE-CHANNEL SPEECH-MUSIC SEPARATION WITH THE ITAKURA-SAITO DIVERGENCE

Cemil Demir[1,3], A. Taylan Cemgil[2], Murat Saraclar[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 introduce a catalog-based single-channel speech-music separation method with the Itakura-Saito (IS) divergence measure. Previously, we have developed the catalog-based separation method with the Kullback-Leibler (KL) divergence. In the probabilistic point of view, IS divergence corresponds to a complex Gaussian observation model. Comparison of divergence measures or observation models in speech-music separation task is carried out with both of catalog-based and traditional Non-Negative Matrix Factorization (NMF) methods. The separation performance is compared using Speech-to-Music Ratio (SMR), Speech-to-Artifact Ratio (SAR) and speech recognition performance measure via the Word Error Rate (WER). We showed that, using IS divergence in both of catalog-based or NMF based speech-music separation methods yields better separation performance than KL divergence. Moreover, in this study, it is shown that catalog-based approaches with both divergence measures outperform traditional NMF based approaches in speech recognition experiments.

1. INTRODUCTION

Recently automatic speech recognition (ASR) applications have become popular in broadcast news transcription systems. One major problem in this sytems 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]. 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]. Raj [5] used the NMF method for compensating of the music signal for an ASR system for the first time. They showed that NMF-based approaches are capable of generating enhanced signals that significantly improve the speech recognition performance.

In previous studies [8, 9, 10], we have 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 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.

Unlike the previous studies, we introduced the catalog-based approach with Itakura-Saito (IS) Divergence and developed the inference method for this approach. Moreover, we compare the separation performance with catalog-based approach with Kullback-Leibler (KL) Divergence, which we proposed previously [8], in speech-music separation task. We also compared the separation performance of catalog-based method with traditional NMF based methods for both of IS and KL divergences [3, 11]. We evaluate the separation performances of the methods not only by using the signal separation measures such as the amount of music suppression or artifact ratios in the recovered speech signal. But also, we evaluate the separation performance of the methods by analyzing the effect of the separation in ASR task.

This paper is organized as follows: in Section 2, we overview the catalog-based separation method with IS divergence. In section 3, we briefly summarize the NMF based speech-music separation method. The experimental results and comparisons are provided in Section 4. Section 5 presents the discussion, conclusions and comments for further investigation.

2. CATALOG-BASED SPEECH-MUSIC SEPARATION WITH IS DIVERGENCE

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. 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.

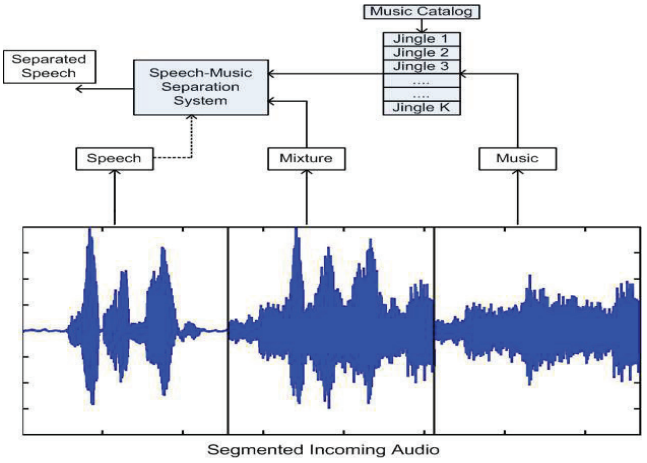

Figure 1: Catalog-Based Speech-Music Separation System Framework

The framework for this scenario is shown in Figure 1. Although the 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. Since we describe the catalog-based method with KL divergence in the previous studies [8], in this study we will describe the method for IS divergence.

2.1 Model Description

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

$$X_{ut} = S_{ut} + M_{ut}$$

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. It is known that the maximization of the likelihood of the complex spectrum of the signal with complex Gaussian observation model corresponds to minimization of the Itakura-Saito (IS) divergence between the power spectrogram of the signal with its NMF approximation [11].

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

$$S_{ut} = \sum_{i=1}^{B} s_{uit}.$$

Each complex Gaussian source is defined as follows:

$$s_{uit} \sim \mathcal{N}_c(s_{uit}; 0, U_{ui}V_{it})$$

where $\mathcal{N}_c$ represents the complex Gaussian distribution and U and V matrices contain the hyper-parameters of the complex spectrum of the speech signal and also correspond to template and excitation matrices respectively in NMF model.

In complex Gaussian model, the latent sources are complex Gaussian 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 minimize the Itakura-Saito (IS) divergence between the power spectrogram of the signal with its NMF approximation [11] which can be defined as follows:

$$D_{IS}(|S|^2|U,V) = \sum_{ut}\left(\frac{|S_{ut}|^2}{\sum_i U_{ui}V_{it}} - \log\frac{|S_{ut}|^2}{\sum_i U_{ui}V_{it}} - 1\right)$$

where $|S|^2$ represents the power spectrogram of the speech signal.

Complex Gaussian density of the random variable s is given as

$$\mathcal{N}_c(s;\mu,\Sigma) = |\pi\Sigma|^{-1}\exp(-(s-\mu)^H\Sigma^{-1}(s-\mu)).$$

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

$$M_{ut} = m_{ut}|r_t \sim \mathcal{N}_c(m_{ut}; 0, C_{uj}f_u v_t)^{[r_t=j]} \quad (1)$$

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_u 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 frame index is drawn independently from a set of jingle indexes as

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

where π represents probability distribution on the jingle frame indexes 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 2. Upper side of the graphical model generates the complex spectrum of the speech part of the mixture whereas the lower side generates the complex spectrum of the music part.

2.2 Inference

After describing the probabilistic model, the appropriate inference methodology must be developed to estimate the hyper-parameters of the latent speech and music sources to be reconstructed. Since the probabilistic model contains the latent sources and hyper-parameters, Expectation-Maximization approach can be used as an inference method. Firstly, in E-step, the expectation of the joint log-likelihood of the latent sources and data under the posterior distribution of the latent sources must be calculated.

We know, if the observation is the sum of the values of complex Gaussian sources, the posterior distribution over the sources given that observation is a complex Gaussian distribution [11]. Since we have a different gaussian for each jingle index, j, the overall posterior distribution over hidden sources is a mixture of gaussian. For each j, the conditional

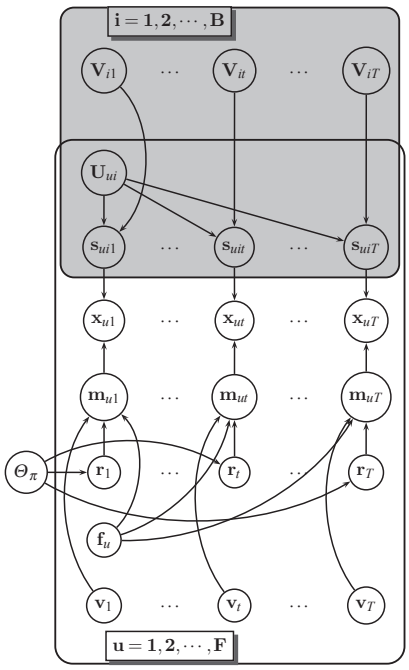

Figure 2: Graphical Model For Speech-Music Mixture.

posterior of the latent speech and music sources can be written as

$$\begin{aligned} p(s_{uit}|X,r_t) &= \mathcal{N}_c(s^j_{uit};\mu^j_{uit},\Sigma^j_{uit}) \\ p(m_{ut}|X,r_t) &= \mathcal{N}_c(m^j_{ut};\mu^j_{ut},\Sigma^j_{ut}). \end{aligned}$$

The conditional posterior mean and variance of i-th speech source and the j-th music source in frequency bin u and time frame t can be found as

$$\begin{aligned} \mu^j_{uit} &= \frac{U_{ui}V_{it}}{\sum_h U_{uh}V_{ht}+C_{uj}f_u v_t}X_{ut} \\ \Sigma^j_{uit} &= \frac{U_{ui}V_{it}}{\sum_h U_{uh}V_{ht}+C_{uj}f_u v_t}(\sum_{h\neq i} U_{uh}V_{ht}+C_{uj}f_u v_t) \\ \mu^j_{ut} &= \frac{C_{uj}f_u v_t}{\sum_h U_{uh}V_{ht}+C_{uj}f_u v_t}X_{ut} \\ \Sigma^j_{ut} &= \frac{C_{uj}f_u v_t}{\sum_h U_{uh}V_{ht}+C_{uj}f_u v_t}(\sum_h U_{uh}V_{ht}) \end{aligned}$$

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

$$\langle|s^j_{uit}|^2\rangle = \Sigma^j_{uit}+|\mu^j_{uit}|^2$$

$$\langle|m^j_{ut}|^2\rangle = \Sigma^j_{ut}+|\mu^j_{ut}|^2.$$

The posterior probability of the active jingle index, j, at time t in gaussian model is:

$$p(r_t=j|X)=\frac{\prod_{ut}\mathcal{N}_c(X_{ut};0,C_{uj}f_u v_t+\sum_i U_{ui}V_{it})\pi_j}{\sum_j\prod_{ut}\mathcal{N}_c(X_{ut};0,C_{uj}f_u v_t+\sum_i U_{ui}V_{it})\pi_j}.$$

The expected value of active jingle frame index r_t being equal to j at time frame t is

$$\langle[r_t=j]\rangle = p(r_t=j|X).$$

After calculating the expectations, we can find out the model parameters that maximize the likelihood of the data. Firstly, we compute the hyper-parameters of the speech spectrogram, U and V matrices. Each entry of the template vector matrix in complex Gaussian model, U, and corresponding excitation matrix of the speech spectrogram, V, can be calculated using the following equations:

$$\begin{aligned} U_{ui} &= \frac{1}{T}\sum_{t,j}\langle[r_t=j]\rangle\frac{\langle|s^j_{uit}|^2\rangle}{V_{it}} \\ V_{it} &= \frac{1}{F}\sum_{u,j}\langle[r_t=j]\rangle\frac{\langle|s^j_{uit}|^2\rangle}{U_{ui}}. \end{aligned}$$

The filtering parameter for each frequency bin, f_u, and gain parameter for each time frame, v_t can be found using

$$\begin{aligned} f_u &= \frac{1}{T}\sum_{t,j}\langle[r_t=j]\rangle\frac{\langle|m^j_{ut}|^2\rangle}{C_{uj}v_t} \\ v_t &= \frac{1}{F}\sum_{u,j}\langle[r_t=j]\rangle\frac{\langle|m^j_{ut}|^2\rangle}{C_{uj}f_u} \end{aligned}$$

where $\langle|m^j_{ut}|\rangle$ similarly represents the expected value of latent music source. After finding the hyper-parameters of the sources, we can reconstruct the complex spectrum of the sources using the following equations:

$$\widehat{S}=X\otimes\frac{UV}{UV+CR\otimes(fv')}$$

$$\widehat{M}=X\otimes\frac{CR\otimes(fv')}{UV+CR\otimes(fv')}$$

where R contains the posterior probabilities of each active frame for each time frame t and $\otimes$ represents the element-wise multiplication.

3. NMF BASED SPEECH-MUSIC SEPARATION

In NMF based speech-music separation systems, during training phase, the power or magnitude spectrogram of the speech and music signals are used to train an NMF model for each source. For this study, although we assume, we can obtain the music template as a prior information, we assume that no training data for the speech signal is available.

In this section, we briefly summarize IS divergence based speech-music separation in the case of known jingle which is used template matrix for the music signal. For KL divergence case, instead of power spectrograms of the sources, magnitude spectrograms are used for the separation.

The template and excitation matrices can be calculated via Multiplicative Update Rules [11] efficiently. In the separation phase, using the template matrices, an overall template matrix is constructed. Using the power spectrogram of the mixed signal and the overall template matrix, the excitation matrix for each source is calculated by solving the equation

$$|X|^2=[UC][V'W']$$

In our case, the template matrix for the music signal (C) is assumed to be known. After finding the excitation matrix

Table 1: Average Output SMR values (in dB)

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-NMF	**22.1**	**28.5**	**35.3**	**42.7**	**50.6**
IS-NMF	16.5	23.8	31.4	39.3	47.5
KL-Catalog	17.6	24.2	30.9	38.2	46.2
IS-Catalog	15.9	23.4	31.1	38.9	46.8

Table 2: Average Output SAR values (in dB)

Separation Method	Input SMR Values				
	0dB	5dB	10dB	15dB	20dB
KL-NMF	10.8	13.4	15.9	18.3	20.4
IS-NMF	11.3	14.6	17.6	20.6	23.4
KL-Catalog	10.9	14.2	17.2	20.2	23.2
IS-Catalog	**12.1**	**15.2**	**18.1**	**21.1**	**24.1**

for each source, the reconstruction of the speech and music signals can be done using the following equations:

$$\widehat{S} = X \otimes \frac{UV}{UV + CW}$$

$$\widehat{M} = X \otimes \frac{CW}{UV + CW}$$

Since we used the IS divergence, we estimated the complex spectrum of the sources. In other words, we estimated both of magnitude and phase of the sources directly.

Table 3: Average Output WER values (in %)

Separation Method	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
KL-NMF	74.3	57.2	43.2	36.2	31.5
IS-NMF	66.2	46.1	35.6	28.9	27.6
KL-Catalog	69.4	52.5	39.5	32.6	29.4
IS-Catalog	**63.2**	**44.5**	**34.6**	**28.8**	**27.5**

4. EXPERIMENTAL RESULTS

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 separation quality which characterize the separation performance to ASR tasks, we also calculated Speech-to-Music Ratio (SMR) and Source-to-Artifact Ratio (SAR) values. In this study, for simplicity, the gain and frequency filtering parameter are assumed to be constant.

4.1 Speech Recognition System and Test Set

For speech recognition tests, we have used the 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. 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 4 sec. length jingles 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 jingles are taken from the broadcast news jingles. The spectrum 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 than speech recognition setup is to decrease the computational complexity of the separation algorithm. The number of speech bases is fixed at 30.

4.2 Experimental Analysis

In this section, we compare the separation performances of the proposed catalog-based approaches, which are called as 'IS-Catalog' and 'KL-Catalog' methods. As a reference, the separation performances of the traditional NMF method approaches, which are called as'IS-NMF' and 'KL-NMF', are also measured. In this part, we use the jingle itself as the Catalog or NMF model for the music signal. However, it should be noted that any prior speech information is not used in any experiments. The SMR, SAR and WER values are shown in Tables 1, 2 and 3, respectively. The separation results are obtained using each frame of the magnitude or power spectrogram of the jingle as a mixture component in catalog based approaches or a template vector in NMF based approaches.

In [8], it was shown that the ASR results with KL-Catalog method is better than KL-NMF method. When we examine the results in Tables 1, 2 and 3 and Figure 3, we can draw the same conclusion for the IS-Catalog and IS-NMF methods. Average SMR values of KL-NMF, KL-Catalog, IS-NMF and IS-Catalog on all input SMRs are 35.9, 31.3, 31.7 and 31.2 dB, respectively. Similarly, Average SAR values of KL-NMF, KL-Catalog, IS-NMF and IS-Catalog on all input SMRs are 15.8, 17.2, 17.5 and 18.1 dB, respectively.

Although the SMR values of NMF methods are higher than SMR values of the Catalog methods, since SAR values of Catalog methods are better than SAR values of NMF methods, the speech recognition performance of the Catalog method outperforms the NMF-methods'. From these results, it can be understood that in speech-music separation, preserving the speech signal is more important than suppressing the music signal in speech recognition point of view.

With the analysis of the experimental results, using IS divergence or complex Gaussian observation model in speech-music separation task yields better separation results than KL divergence or poisson observation model. Using IS divergence in separation decreases the suppression ratio of the music signal. However, since the reconstruction of the speech signal with IS divergence results in higher SAR values, the speech recognition performances of IS methods are better than KL methods' performances. From these results, it can be concluded that using IS divergence or complex Gaussian observation model is more appropriate than KL divergence or poisson observation model for speech-music separation task.

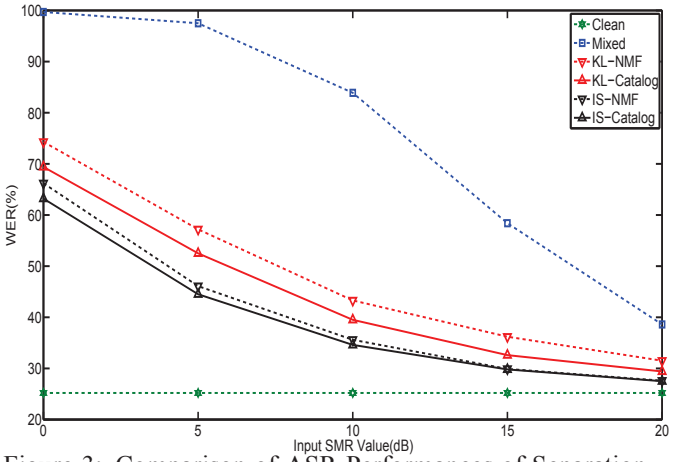

Figure 3: Comparison of ASR Performances of Separation Methods

5. CONCLUSIONS

The aim of this study is to develop the previously proposed catalog based speech music separation method for complex Gaussian observation model and make comparison. The inference method for complex Gaussian model is derived in this study. We have evaluated the separation performance of the proposed IS-Catalog method and compare its performance with previously proposed KL-Catalog method.

Moreover, traditional NMF methods are used in separation tests as a baseline systems. As similar to KL case, IS-Catalog method gets better results than IS-NMF method. In this study, we showed that using IS divergence based methods (Catalog or NMF) in speech-music separation outperforms KL-divergence based methods. 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 Markov 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.

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 TUBA-GEBIP award. Taylan Cemgil is supported by the Bogazici University research grant BAP 5723 and TUBITAK (Project code: 110E292).

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] 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] 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.

[8] C. Demir, A.T. Cemgil, and M. Saraçlar, "Catalog-Based Single-Channel Speech-Music Separation For Automatic Speech Recognition," in *Proc. of EUSIPCO*, 2011.

[9] C. Demir, A.T. Cemgil, and M. Saraçlar, "Semi-supervised Single-Channel Speech-Music Separation For Automatic Speech Recognition," in *Proc. of Interspeech*, 2011.

[10] C. Demir, A.T. Cemgil, and M. Saraçlar, "Gain Estimation Approaches in Catalog-Based Single-Channel Speech-Music Separation," in *Proc. of ASRU*, 2011.

[11] C. Févotte, N. Bertin, and J.L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.