Speech recognition based on Itakura-Saito divergence and dynamics / sparseness constraints from mixed sound of speech and music by non-negative matrix factorization

Naoaki Hashimoto, Shoichi Nakano, Kazumasa Yamamoto, Seiichi Nakagawa

Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

{hashimoto,snakano,kyama,nakagawa}@slp.cs.tut.ac.jp

10.21437/Interspeech.2014-160

Abstract

We considered a speech recognition method for mixed sound, which is composed of both speech and music, that only removes music based on non-negative matrix factorization (NMF). We used Itakura-Saito divergence instead of Kullback-Leibler divergence to compare the cost function, and the dynamics and sparseness constraints of a weight matrix to improve speech recognition. For isolated word recognition using the matched condition model, we reduced the word error rate of 52.1% relative from the case that didn't remove music (on average, from 69.3% to 85.3%).

Index Terms: speech recognition, mixed sound, music removal, vector quantization, non-negative matrix factorization

1. Introduction

Speech recognition performance is significantly degraded in noisy environments. In the presence of noise, we must reduce its effect. The spectral subtraction[1] and Wiener filter based methods [2] are general techniques for noise removal. Although both are valid for stationary noise, they are ineffective for non-stationary noise. In this paper, we consider speech recognition in background music that is comprised of non-stationary signals.

Several music removal methods have been proposed that separate speech and music using a single microphone, such as the binary masking [3] and non-negative matrix factorization (NMF) [4] methods. Independent component analysis (ICA) based methods have also been widely used [5] for sound source separation when multi-channel inputs are available from multiple microphones.

For mixed speech into a single channel, there is a monaural speech separation and recognition challenge. In [6], the keywords in sentences spoken by a target talker were identified with a background talker who made similar sentences. The main approaches for this task are based on missing feature theory, NMF, and Computational Auditory Scene Analysis (CASA).

We considered music removal for input speech with background music from a single microphone using vector quantization [7] and NMF, and applied these methods for speech recognition to mixed sounds consisting of speech and music. We improved the speech recognition performance by music removal through two methods [8]. However, since music removal based on NMF requires much computation, it is not practical. Therefore, we proposed a fast calculation technique of music removal based on NMF [9].

In this paper, for further improvement, we propose Itakura-Saito divergence (instead of Kullback-Leibler divergence) to compare the cost function, and the dynamics and sparseness constraints of the weight matrix to improve speech recognition.

2. Music removal by NMF

In recent years, NMF has been studied to solve the sound source separation problems dividing music into vocal and instrumental sounds [10], and mixed sound into music and speech [11].

2.1. Nonnegative Matrix Factorization

NMF decomposes $n \times m$ matrix Y into $n \times r$ matrix W and $r \times m$ matrix H:

$$Y \approx WH \tag{1}$$

where all the elements of matrices Y, W, and H are estimated by minimizing a cost function under the nonnegativity constraint. Kullback-Leibler divergence, which is usually used as the cost function, is defined as

$$D_{KL} = \sum_{i,j} \left(Y_{ij} \log \frac{Y_{ij}}{(WH)_{ij}} - Y_{ij} + (WH)_{ij} \right) \tag{2}$$

Using the following updating rules, W and H are updated until D_{KL} converges [4]:

$$H_{kj} \leftarrow H_{kj} \frac{\sum_i W_{ik} Y_{ij} / (WH)_{ij}}{\sum_i W_{ik}} \tag{3}$$

$$W_{ik} \leftarrow W_{ik} \frac{\sum_j H_{kj} Y_{ij} / (WH)_{ij}}{\sum_j H_{kj}} \tag{4}$$

The resulting matrices W and H are the results of decomposition.

We also use Itakura-Saito divergence to compare the cost functions, and define it as

$$D_{IS} = \sum_{i,j} \left(\frac{Y_{ij}}{(WH)_{ij}} - \log \frac{Y_{ij}}{(WH)_{ij}} - 1 \right) \tag{5}$$

Using the following updating rules, W and H are updated until D_{IS} converges [12]:

$$H_{kj} \leftarrow H_{kj} \sqrt{\frac{\sum_i \frac{Y_{ij}}{(WH)_{ij}} \frac{W_{ik}}{(WH)_{ij}}}{\sum_i \frac{W_{ik}}{(WH)_{ij}}}} \tag{6}$$

$$W_{ik} \leftarrow W_{ik} \sqrt{\frac{\sum_j \frac{Y_{ij}}{(WH)_{ij}} \frac{H_{kj}}{(WH)_{ij}}}{\sum_j \frac{H_{kj}}{(WH)_{ij}}}} \tag{7}$$

Copyright © 2014 ISCA

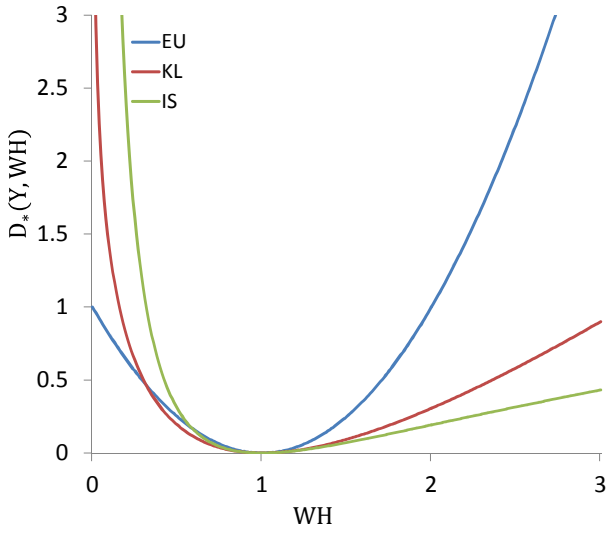

Figure 1: Comparison of the cost functions

2.2. Comparison of cost functions

Figure1 compares the Itakura-Saito divergence (IS), the Kullback-Leibler divergence (KL), and the Euclid distance (EU) [12, 13, 14]. The Euclid distance has axial symmetry, whereas the Itakura-Saito and Kullback-Leibler divergences have axial asymmetric, although the cost functions of all of them are 0 when $Y = WH$. If WH is less than Y, the Itakura-Saito and Kullback-Leibler divergences impose more excessive penalties. In addition, the Itakura-Saito divergence is independent of the scale of WH and Y for the function represented by the ratio of only WH and Y.

2.3. Sound source separation of NMF based on dynamic and sparseness constraints

2.3.1. *Basic procedure*

In this paper, we refer to the idea of phoneme recognition using NMF [15] to separate speech and music in mixed sound. Matrix Y is composed of an amplitude spectrogram, which is a sequentially arranged amplitude spectrum for each frame of the input sound as a column vector, and matrix Y is decomposed into matrices W and H. Matrix W is arranged as a set of column basis vectors of speech and music. Matrix H is arranged as row vectors for each input frame weight of each basis. The basis matrices of speech W_s and music W_m are determined beforehand; $W = [W_s W_m]$. H is obtained from W using the update rule in Eq.(3). In the experiment, we fixed W, because the VQ code vectors are considered to be representative basis vectors. We used VQ code vectors for the speech and music sounds as basis vectors for W_s and W_m, respectively. After this processing,

$$Y \approx W_s H_s + W_m H_m \tag{8}$$

can be separated into $W_s H_s$ and $W_m H_m$ that correspond to speech and music, respectively. To obtain the estimated spectrum of speech and music, we construct a filter from the decomposed results and, multiply the input signal as follows:

$$\hat{S} = Y \otimes \frac{W_s H_s + C_1}{W_s H_s + W_m H_m + C_2} \tag{9}$$

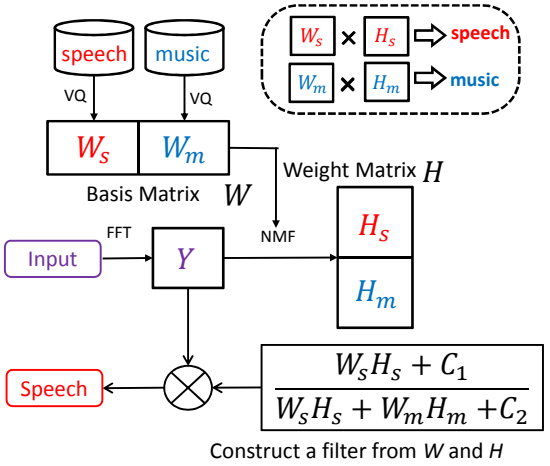

Figure 2: *Overview of music removal by NMFtest.*

$$\hat{M} = Y \otimes \frac{W_m H_m + C_1}{W_s H_s + W_m H_m + C_2} \tag{10}$$

where $\hat{S}$ is the estimated amplitude spectrogram of speech, $\hat{M}$ is the estimated amplitude spectrogram of music, C_1 and C_2 are constant values for smoothing, and operator $\otimes$ and all the divisions are element wise multiplications and divisions [9]. Figure 2 shows an overview of our NMF method.

Our procedure can be summarized by the following steps in the separation of speech and music using NMF:

1. Obtain the basis matrices for speech and music, and then combine them to form W.
2. Create matrix Y from the amplitude spectrogram of the input sound.
3. Obtain weight matrix H by the iterative updating rule (W is fixed).
4. Construct a filter from W and H obtained by NMF.
5. Separate speech and music by multiplying the filter to amplitude spectrogram of input signal.

2.3.2. *Constraints of dynamic (NMF+Δ)*

Since matrix decomposition in NMF deals independently with each column, NMF treats each frame independently of the spectrogram in the mixed sound. However, dynamic alteration in the spectrum's time direction is an important feature constraint. Therefore, to obtain better decomposition results, our framework considers dynamic features as constraints in the following [16]:

$$C_t(H) = \sum_{k=1}^{K} \frac{1}{\sqrt{(1/J)\sum_{j=1}^{J} H_{k,j}^2}^2} \sum_{j=2}^{J} (H_{k,j} - H_{k,j-1})^2 \tag{11}$$

$$\left[\nabla C_t^+(H)\right]_{k,j} = 4J \frac{H_{k,j}}{\sum_{t=1}^{J} H_{k,t}^2} \tag{12}$$

$$\left[\nabla C_t^-(H)\right]_{k,j} = 2J \frac{H_{k,j-1} + H_{k,j+1}}{\sum_{t=1}^{J} H_{k,t}^2} + 2J H_{k,j} \frac{\sum_{t=2}^{J} (H_{k,t} - H_{k,t-1})^2}{(\sum_{t=1}^{J} H_{k,t}^2)^2} \tag{13}$$

where $C_t(H)$ denotes a constraint cost, that is added to Eq.(2) or (5). The new updating rule for Eq.(2) is given by

$$H_{k,j} \leftarrow H_{k,j} \frac{\sum_i \frac{W_{ik}Y_{ij}}{(WH)_{ij}} + \lambda \nabla C_t^-(H)}{\sum_i W_{i,k} + \lambda \nabla C_t^+(H)} \quad (14)$$

For Eq.(5),

$$H_{k,j} \leftarrow \sqrt{H_{k,j} \frac{\sum_i \frac{Y_{ij}}{(WH)_{ij}} \frac{W_{i,k}}{(WH)_{ij}} + \lambda \nabla C_t^-(H)}{\sum_i \frac{W_{i,k}}{(WH)_{ij}} + \lambda \nabla C_t^+(H)}} \quad (15)$$

where λ is a parameter that controls the weight of the criterion.

2.3.3. Constraints of sparseness (NMF+sp) [17]

An overfitting problem exists when the number of basis vectors becomes too large for representing speech and music. Therefore, cost function $L1$ norm term is added to reduce the components of the basis vectors:

$$C_s(H) = \sum_{k,j} H_{k,j} \quad (16)$$

$$\nabla C_s^+(H) = 1 \quad (17)$$

$$\nabla C_s^-(H) = 0 \quad (18)$$

The following new updating rule H for Eq.(2) is given by

$$H_{k,j} \leftarrow H_{k,j} \frac{\sum_i \frac{W_{ik}Y_{ij}}{(WH)_{ij}} + \lambda \nabla C_s^-(H)}{\sum_i W_{i,k} + \lambda \nabla C_s^+(H)} \quad (19)$$

For Eq.(5),

$$H_{k,j} \leftarrow H_{k,j} \sqrt{\frac{\sum_i \frac{Y_{ij}}{(WH)_{ij}} \frac{W_{i,k}}{(WH)_{ij}} + \lambda \nabla C_s^-(H)}{\sum_i \frac{W_{i,k}}{(WH)_{ij}} + \lambda \nabla C_s^+(H)}} \quad (20)$$

where λ is a parameter that controls the weight of the criterion.

3. Fast calculation technique of NMF based approach (fastNMF)

The standard NMF method described in Section 2.3 requires matrix decomposition to be performed for each input speech, so it is not practical due to the large amount of calculation. We previously proposed a fast calculation technique of an NMF based approach [9]. This technique achieves an approximate separation based on NMF by creating a VQ codebook from the mixed sound of the training data and decomposing the matrix of the VQ code vectors a priori, and then using the decomposition results that correspond to the input speech. This fastNMF runs about 20 times faster than the standard NMF.

4. Experiments

4.1. Experimental setup

We experimentally conducted a recognition evaluation using 200 isolated words from 20 speakers in the Tohoku University and Matsushita word speech database [18]. We used 15 speakers for the training data and 5 speakers for the test data. We used a piano trio in G minor Op.8 as the music data. The audio data were sampled at 12 kHz in the mono-mode. In a representative vector set of mixed sound in the fastNMF, the code vector is the amplitude spectrum of 256 dimensions, and the codebook size is 4096. The speech analysis conditions in the NMF method

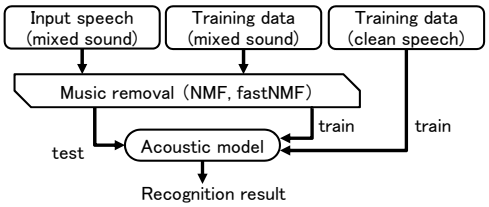

Figure 3: *Matched condition model*

were a 512 point Hanning window and a 256 point frame shift. Matrix W, which is the base vectors, was composed by both the speech and music code vectors of size 512 constructed using the VQ technique. In addition, the constant values for smoothing were set to $C_1 = C_2 = 1.0$ for the D_{KL} cost function, and $C_1 = C_2 = 2.0$ for the D_{IS} cost function from the result of a preliminary experiment.

We constructed acoustic models for speech recognition by entire word based HMMs, with 14 states and 8 (clean model) or 16 (matched condition model) mixtures of Gaussians (diagonal covariance matrix). As features, we used 12 dimensions of MFCCs, their deltas, double-deltas, delta power, and double-delta power (38 dimensions) obtained with a 25ms window size and 10ms frame shift. Music was added to the 1000 (200 words $\times$ 5 speakers) words in the test data at 20, 10, 0, and -5 dB SNRs. We conducted recognition experiments using two models: clean and matched condition (so colled "feature-based noise adaptive training"[19]) models. Figure 3 illustrates the learning procedure in the matched condition.

4.2. Speech Recognition

4.2.1. Clean Model

Table 1 gives the recognition results for HMMs trained with the clean speech data. The NMF using cost functions D_{KL} and D_{IS} improved the recognition performance from "no processing." The recognition performance also improved to 64.3% from 56.3% on average by NMF using the D_{KL} cost function that added sparse constraints and smoothing . On the other hand, the D_{IS} cost function that added the sparseness or dynamic constraints did not improve the rate. For the fastNMF method using the D_{KL} cost function, improvement was obtained from the "no processing" (46.1% vs 37.4% on average).

4.2.2. Matched Condition Model

Table 2 gives the recognition results for the matched condition model. "Mixed sound" shows the results when such mixed sound is recognized using a mixed sound HMM, which was learned by the condition of ∞, 20, 10, 0, and -5 dB SNRs. The NMF using the D_{IS} cost function improved the recognition rates of all the SNRs more than using the D_{KL} cost function (81.6% vs 74.4% on average). The recognition rate also improved to 85.3% from 81.6% on average by the NMF using the D_{IS} cost function to which smoothing was added. However, by adding the sparseness or dynamic constraints to cost function (D_{IS}), the recognition rate was not improved. On the other hand, for the cost function D_{KL}, the rate improved from 79.9% to 81.2% or 81.1%.

For the fastNMF method using the D_{IS} cost function, improvement was obtained from the mixed sound (71.1% vs 69.3% on average).

Table 1: Recognition rate for clean model[%]

method	λ	SNR				
		-5dB	0dB	10dB	20dB	ave
no processing	-	2.2	7.8	53.4	86.1	37.4
NMF_D_{KL}	-	17.4	37.8	77.6	92.2	56.3
NMF_D_{IS}	-	15.9	39.0	80.0	93.5	57.1
$NMF\_D_{KL}+C$	-	21.1	43.3	83.2	93.2	60.2
$NMF\_D_{KL}+C+\Delta$	0.5	20.5	43.5	82.9	93.0	60.0
$NMF\_D_{KL}+C+sp$	1.0	25.5	51.5	85.7	94.5	64.3
$NMF\_D_{IS}+C$	-	12.7	33.0	77.8	92.0	53.9
$NMF\_D_{IS}+C+\Delta$	0.1	8.3	20.6	70.1	88.1	46.8
$NMF\_D_{IS}+C+sp$	0.2	9.7	26.7	76.5	91.8	51.2
$fastNMF\_D_{KL}$	-	5.2	17.6	71.1	90.4	46.1
$fastNMF\_D_{IS}$	-	5.7	17.2	69.4	88.1	45.1
clean speech	-	98.8 ($SNR=\infty$)				

Table 2: Recognition rate for matched condition model[%]

method	λ	SNR				
		-5dB	0dB	10dB	20dB	ave
mixed sound	-	25.0	59.3	94.4	98.5	69.3
NMF_D_{KL}	-	40.9	70.4	91.1	95.1	74.4
NMF_D_{IS}	-	53.2	80.6	94.8	97.7	81.6
$NMF\_D_{KL}+C$	-	48.4	78.7	95.1	97.5	79.9
$NMF\_D_{KL}+C+\Delta$	0.5	52.2	80.4	95.0	97.3	81.2
$NMF\_D_{KL}+C+sp$	1.0	52.1	80.3	94.9	97.0	81.1
$NMF\_D_{IS}+C$	-	59.7	85.2	97.2	98.9	85.3
$NMF\_D_{IS}+C+\Delta$	0.1	53.5	82.1	96.3	98.1	82.5
$NMF\_D_{IS}+C+sp$	0.2	48.1	81.7	96.4	98.3	81.1
$fastNMF\_D_{KL}$	-	23.7	60.5	94.1	98.7	69.3
$fastNMF\_D_{IS}$	-	27.0	64.1	94.5	98.8	71.1

4.3. Evaluation by SDR and spectrogram

As an evaluation metric, we used SDR (Source to Distortion Ratio) in the spectral domain as follows:

$$SDR = 10\log_{10}\frac{\sum_{n,f} S_{n,f}^2}{\sum_{n,f}(S_{n,f}-\hat{S}_{n,f})^2} \qquad (21)$$

where $S_{n,f}$ is the target signal spectrum, $\hat{S}_{n,f}$ is the estimated signal spectrum, $n=\{1,\cdots,N\}$ is the time frame index, and $f=\{1,\cdots,F\}$ is the frequency bin[20]. Table 3 shows the obtained SDR for speech after music removal and for music after speech removal, respectively. The NMF using D_{IS} cost function is worse SDR for speech than D_{KL} cost function, but better SDR for music. It is considered that the speech as well as music is cut when music was removed. We guess that this fact leads to better recognition in the matched condition models for D_{IS}. Notice that NMF improved SDR with speech in low SNRs (-5 - 10dB), but it did not improve in high SNRs, such as 20dB.

Figure 4 shows the spectrogram of the mixed sound of speech and music and the spectrograms of speech estimated by NMF. NMF using the D_{IS} cost function can remove the music better than the D_{KL} cost function, especially in the period of only music.

5. Conclusions

In this paper, as a music removal method for speech recognition in mixed sound, we introduced the Itakura-Saito divergence and the constraints for improving NMF and evaluated them by an isolated word recognition experiment with 200 words .

Table 3: SDR in the spectral domain

method	$S_{n,f}$ in (21)	SNR				
		-5dB	0dB	10dB	20dB	ave
NMF_D_{KL}	speech	4.96	7.89	12.96	16.45	10.56
NMF_D_{IS}	speech	3.94	5.67	9.85	13.29	8.19
NMF_D_{KL}	music	7.62	4.57	1.57	0.89	3.66
NMF_D_{IS}	music	11.63	7.46	2.95	1.67	5.93

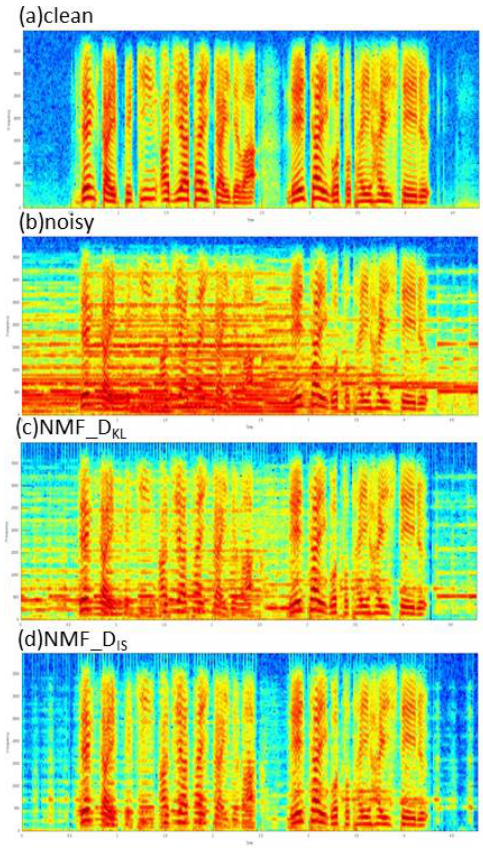

Figure 4: *Comparison of spectrogram for NMF ($SNR=0dB$). "ZeNkai PekiN Aziataikaidewa..."*

In the clean model, by the NMF using the D_{KL} cost function and introducing smoothing and sparseness constraints, we obtained 29.5% word error rate reduction compared with the NMF using the D_{KL} cost function (at 20dB, from 92.2% to 94.5%). In the matched condition model, by the NMF using the D_{IS} cost function, we obtained 53.1% word error rate reduction compared with the D_{KL} cost function (at 20dB, from 95.1% to 97.7%). By introducing smoothing, we obtained 77.6% word error rate reduction compared with the D_{KL} cost function (at 20dB, from 95.1% to 98.9%). These results show that the D_{IS} cost function was more effective in the matched condition model case, and we improved the recognition rate by our introduced constraint only with D_{KL} cost function, not with D_{IS}.

As future works, we will combine the constraints and address other kinds of music than a piano trio. Futhermore, we plan to do some experiments with acoustically mixed sound of music and speech instead of the electronic mixing.

6. References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction ,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27,no. 2, pp.113-120, 1979.

[2] J. S. Lim and A. V. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-26, no.3, pp.197-210, 1978.

[3] H. Itou, Y. Ohishi, C. Miyajima, N. Kitaoka and K. Takeda, ”Source separation based on binary masking using Bayesian network,” IPSJ, vol.2008, no.72, pp.51–56, 2008. (in Japanese)

[4] D. D. Lee and H. S. Seung, ”Algorithms for Non-negative Matrix Factorization,” Proc. NIPS, pp.556–562, 2000.

[5] M.A. Casey and A. Westner, ”Separation of mixed audio sources by independent subspace analysis,” Proc. International Computer Music Conference, Berlin, Germany, pp.154-161, Aug. 2000.

[6] M. Cooke, J. R. Hershey and S. T. Rennie, ”Monaural speech separation and recognition challenge,” Computer Speech and Language, Vol.24, no.1, pp.1-15, 2010.

[7] K. Yamamoto and S. Nakagawa, ”Evaluation of privacy protection techniques for speech signals,” Proc. IPMU, pp.653–662, 2010.

[8] S. Nakano, K. Yamamoto and S. Nakagawa, “Speech recognition in mixed sound of speech and music base on vector quantization and non-negative matrix factorization,” Proc. INTERSPEECH , pp.1781–1784, 2011.

[9] S. Nakano, K. Yamamoto and S. Nakagawa, “Fast NMF based approach and improved VQ based approach for speech recognition from mixed sound,” Proc. APSIPA, OS.15-SLA.7, 2012.

[10] A. Mesaros and T. Virtanen, “Recognition of phonemes and words in singing,” Proc. ICASSP, pp.2146–2149, 2010.

[11] B. Raj, T. Virtanen, S. Chaudhuri and R. Singh, “Non-Negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition,” Proc. INTERSPEECH, pp.717–720, 2010.

[12] C. Févotte, N. Bertin, and J-L. Durrieu, ”Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis, ”Neural Comput., vol. 21, no. 3, pp. 793-830, 2009.

[13] H. Kameoka, “Non-negative matrix factorization”, Journal of the Society of Instrument and Control Engineers,vol. 51,no. 9, pp. 835-844, 2012.(in Japanese)

[14] H. Sawada, “Nonnegative Matrix Factorization and Its Applications to Data/Signal Analysis”, The Institute of Electronics, Information and Communication Enginee, vol.95, no.9, pp.829-833, 2012.(in Japanese)

[15] B. Schuller and F. Weninger, “Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization,” Proc. ICASSP, pp.5054–5057, 2010.

[16] T. Virtanen, “Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria,” IEEE Trans. Audio, speech and Language Processing, vol.15, no.3, pp.1066-1074, 2007.

[17] B. Cauchi, S. Goetze, S.Doclo, “Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization,” Proc. of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments, pp. 28-33,2012.

[18] S.Makino, K.Niyada, Y.Mafune, K.Kido, “Tohoku University and Panasonic isolated spoken word database,”Acoustical Science and Technology, vol.48,no.12,pp.899-905, 1992(in Japanese)

[19] L. Deng, A. Acero, M. Plumpe, and X. Huang,“Large-vocabularyspeech recognition under adverse acoustic environments,” Proc ICSLP, Beijing, China, 2000, pp. 806809.

[20] E.Vincent, R.Gribonval and C.Fevotte “Performance measurement in blind audio source separation”, IEEE Trans. Audio, Speech, Language Process., Vol. 14, No.4, pp.1462–1469 2006.

[21] C. Joder, F. Weninger, D. Virette, and B. Schuller, “A comparative study on sparsity penalties for nmf-based speech separation:Beyond Lp-norms,” Proc. ICASSP, pp.858–862, 2013. ICASSP 2013

[22] Z.Duan, G. J. Mysore and P. Smaradis, “Speech enhancement by online non non-negative spectrogram decomposition in non-stationary noise environments,” Proc. INTERSPEECH, 03b. 06, 2012.

[23] Demir, C, M. Saraclar and A. T. Cemgil, “Single-channel Speech-music separation for robust ASR with mixture models,” IEEE Trans. Audio, Speech and Lang. Process., Vol.21, No.4, pp.725–736, 2013.

[24] L.Nikolay, K.Mikhail, “Non-negative Matrix Factorization with Linear Constraints for Single-Channel Speech Enhancement,” Proc. INTERSPEECH, pp.446–449, 2013.