Speech recognition for mixed speech and music by NMF using various cost functions and noise adaptive training methods

Naoaki Hashimoto, Kazumasa Yamamoto and Seiichi Nakagawa
Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
E-mail: {hashimoto,kyama,nakagawa}@slp.cs.tut.ac.jp

***Abstract*—We investigated speech recognition methods for mixed speech and music that only remove music based on non-negative matrix factorization (NMF). In this paper, we introduced the Euclidean distance of logarithm spectrum D_{LOG} as a distance measure for source separation, which may correspond to the distance measure for speech recognition, and compared it with such traditional distance measures as the Kullback-Leibler divergence and the Itakura-Saito divergence. We improved the speech recognition performance by pooling the estimated speech, the mixed sound, and clean speech to train the acoustic model. For isolated word recognition with NMF using D_{LOG}, we obtained an improvement from the baseline. Using the Itakura-Saito divergence and the "clean, multi-condition and noise-adaptive training model", we reduced the word error rate of 54.7% relative from the case of the "multi-condition training model" on average, from 57.6$\%$ to 80.8$\%$ word recognition rate.**

I. Introduction

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

Several music removal methods have been proposed. Independent component analysis (ICA) based methods [1] have been widely used for sound source separation when multi-channel inputs are available from multiple microphones. Non-negative matrix factorization (NMF) based methods [2] have also been used to separate speech and music from a single microphone. For example, Mesaros et al. divided music into vocal and instrumental sounds for the recognition of sing [3], and Raj et al. divided mixed sound into music and speech for robust automatic recognition of mixed sound [4].

We investigated music removal for input speech with background music from a single microphone using vector quantization [5] and NMF, and applied these methods to the speech recognition of mixed sounds. We improved the speech recognition performance by music removal through two methods [6]. However, since music removal based on NMF requires much computation, it is not practical for real time applications. Therefore, we proposed a fast calculation technique of music removal based on NMF [7]. In previous work [8], for further improvement, we introduced the Itakura-Saito divergence (instead of the Kullback-Leibler divergence) to compare the cost function, the dynamics, and the sparseness constraints of the weight matrix [9] [10].

In this paper, we introduce the Euclidean distance of logarithm spectrum as a distance measure to match the measures of speech recognition and source separation. We also introduced six types of acoustic model training data combinations because the model is robust from training with various data (e.g. different SNRs etc.) and compared them with previous methods [8].

II. Music removal by NMF

A. 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 W and H are estimated by minimizing a cost function under the nonnegativity constraint. In this paper, Y is the amplitude spectrogram of the observed signal, Y_{ij} is an element of the Y, and WH is the amplitude spectrogram of the estimated signal, $(WH)_{ij}$ is an element of the WH. n is the frequency bin, m is the frame size, and r is the codebook size. We compared the following four cost functions.

(a) Kullback-Leibler divergence

Kullback-Leibler divergence, which is usually used as a 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, H is updated until D_{KL} converges [2]:

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

Omitted the update rule W because not update in this paper.

(b) Euclidian distance

We also use the Euclidian distance to compare the cost functions and define it as

$$D_{EU} = \sum_{i,j} \left(Y_{ij} - (WH)_{ij} \right)^2. \tag{4}$$

Using the following updating rules, H is updated until D_{EU} converges [2]:

$$H_{kj} \leftarrow H_{kj} \frac{\sum_i Y_{ij} W_{ik}}{\sum_i W_{ik} (WH)_{ij}}. \quad (5)$$

(c) Itakura-Saito divergence
The cost function based on the Itakura-Saito divergence is suitable for speech recognition [8]. It is defined as

$$D_{IS} = \sum_{i,j} \left(\frac{Y_{ij}}{(WH)_{ij}} - \log \frac{Y_{ij}}{(WH)_{ij}} - 1 \right). \quad (6)$$

Using the following updating rules, H is updated until D_{IS} converges [11]:

$$H_{kj} \leftarrow H_{kj} \sqrt{\frac{\sum_i \frac{Y_{ij}}{(WH)_{ij}} \frac{W_{ik}}{(WH)_{ij}}}{\sum_i \frac{W_{ik}}{(WH)_{ij}}}}. \quad (7)$$

(d) Euclidean distance of the logarithm spectrum
MFCC is typically used as a feature for speech recognition because the cepstral distance is effective for speech recognition. The cepstrum's Euclidean distance equals to the logarithm spectrum's Euclidean distance. On the other hand, since the amplitude spectrum is typically used for NMF, we considered the evaluation gap between speech recognition and sound source separation. In this paper, we introduce the Euclidean distance of the logarithm spectrum as a distance measure for source separation that is suitable for speech recognition. The cost function is defined as

$$D_{LOG} = \sum_{ij} \left(\log \frac{Y_{ij}}{(WH)_{ij}} \right)^2. \quad (8)$$

Using the following updating rules, W and H are updated until D_{LOG} converges:

$$H_{kj} \leftarrow H_{kj} \sqrt{\frac{\sum_i \frac{Y_{ij}}{(WH)_{ij}} \frac{W_{ik}}{(WH)_{ij}}}{\sum_i p(\xi_{i,j}) \frac{W_{ik}}{Y_{ij}}}}, \quad (9)$$

$$W_{ik} \leftarrow W_{ik} \sqrt{\frac{\sum_j \frac{Y_{ij}}{(WH)_{ij}} \frac{H_{kj}}{(WH)_{ij}}}{\sum_j p(\xi_{i,j}) \frac{H_{kj}}{Y_{ij}}}}, \quad (10)$$

$$p(\xi_{i,j}) = \frac{2 \log \xi_{i,j}}{\xi_{i,j}} + \frac{1}{\xi_{i,j}^2}, \quad \xi_{i,j} = \frac{\sum_k W_{ik} H_{kj}}{Y_{ij}}.$$

These rules were derived by referring to [13].

B. Sound source separation by NMF

We separate speech and music in the same way as [8] referring to [12]. Fig. 1 shows an overview of our NMF method.

Our procedure can be summarized by the following steps:

1) Obtain the basis matrices for speech and music by VQ and combine them to form W.

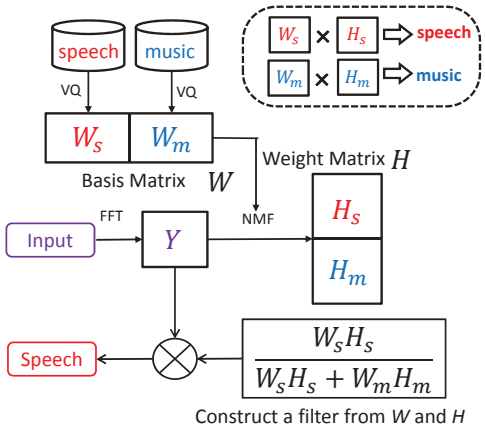

Fig. 1. *Overview of music removal by NMF test*

2) Create matrix Y from the amplitude spectrogram of the input sound.
3) Obtain weight matrix H by the iterative updating rule in Eq.(3), (5), (7) and (9) (W is fixed).
4) Construct a filter from W and H which is obtained by NMF.
5) Separate speech and music by multiplying the filter to the amplitude spectrogram of the input signal.

III. Comparison of Cost Functions

Figures 2 compares of the cost functions. The cost of D_{IS} exceeds that of D_{EU} and D_{KL} when $(WH)_{ij}$ is larger than Y_{ij}. D_{EU} the same even if estimated signal $(WH)_{ij}$ is smaller or larger than Y_{ij}. In contrast, D_{KL} and D_{IS} and D_{LOG} impose a more excessive cost if $(WH)_{ij}$ is less than Y_{ij}. In addition, D_{IS} and D_{LOG} give the same cost (either large or small) of the amplitude because they only depend on the ratio of Y_{ij} to $(WH)_{ij}$. D_{LOG} costs harder than D_{IS}. We discuss the effect of each cost function for speech recognition.

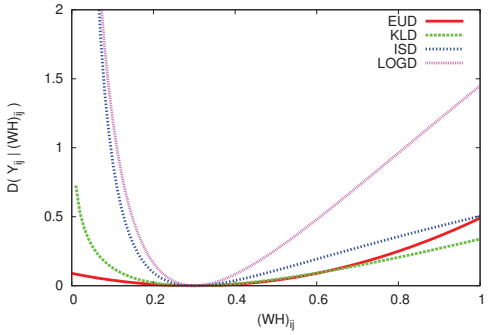

Fig. 2. Comparison of cost functions ($Y_{ij} = 0.3$)

IV. Training Acoustic Models

By adding mixed sound and/or estimated speech to clean speech to train an acoustic model, the remained music not removed by NMF or the distortion caused by NMF is compensated. We trained acoustic models of speech recognition with the following training data sets in six ways; (a) Trained by clean speech : *clean model*, (b) Trained by mixed sound : *multi-condition training model*, (c) Trained by clean speech + mixed sound : *clean and multi-condition training model*, (d) Training by estimated speech : *noise adaptive training*

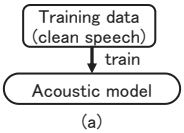

Fig. 3. Clean model

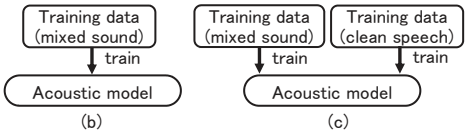

Fig. 4. Multi-condition training model

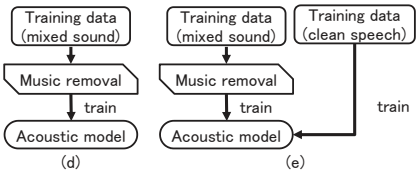

Fig. 5. Noise adaptive training model

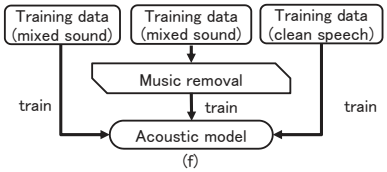

Fig. 6. Clean, multi-condition and noise-adaptive training model

model[14], (e) Training by clean speech + estimated speech : *clean and noise adaptive training model*, (f) Training by clean speech + mixed sound + estimated speech : *clean, multi-condition and noise adaptive training model*. Figs. 3, 4, 5, and 6 show the conceptual diagrams of the acoustic model training. Models (c) and (e) are conventionally used [8], (f) is our new proposed model, and (a) and (b) are the baseline models.

V. Experiments

A. Experimental setup

We experimentally conducted recognition evaluation using 200 isolated words from 20 speakers in the Tohoku University and Matsushita word speech database [15]. We used 15 speakers for the training and 5 speakers for the test. We used "Piano trio in G minor Op.8" (piano, violin, cello) as music data and divided it into training (6min, 30sec) and test parts(3min, 20sec). We also experimented with two jazz pieces: "I'll close my eyes" for training (4min, 43sec) and "Bye Bye Blackbird" for a test(4min, 11sec)) (piano, bass, drums). The audio data were sampled at 12 kHz in mono-mode. Speech analysis in the NMF methods was done with 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.

We constructed acoustic models for speech recognition as entire word based HMMs, with 14 states and 8 (clean model) or 16 (multi-condition and/or noise-adaptive training 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 and training data at 20, 10, 0, and -5 dB SNRs. We conducted recognition experiments using the four cost functions and the six acoustic models and compared them.

B. Speech recognition result

Table I shows the speech recognition rate by six acoustic models and the four cost functions.

1) Comparison of acoustic models: The recognition results of *clean, multi-condition and/or noise-adaptive training models* (c), (e), and (f) outperformed *clean model* (a) for all of the distance measures. Similar results were obtained in the vector quantization method [6] and the FastNMF method [7].

The best performance was obtained by NMF based on D_{IS} integrated with the *clean, multi-condition and noise-adaptive training model*(f). Significant improvement was obtained from the *no processing* (mixed sound(c)) (80.8% vs. 57.6% on average for the piano trio piece). Furthermore, improvement was obtained by the *model* (f) from the *no processing* (mixed sound(c)) at 20dB (97.4% vs 96.6%), although no improvement was obtained of *model* (e), which is conventionally used [8]. Therefore, these results show the effect of adding mixed sound to the estimated speech to train the acoustic model.

For NMF based on D_{KL}, the recognition result of the *model* (c) without estimated speech outperformed the *model* (f). Perhaps this result was caused by the leftover music and sound distortion. It is different from the distance measure shown in the next subsection.

A similar tendency was also observed for the jazz background music.

2) Comparison of cost functions: The NMF based on D_{LOG} improved the recognition rate by the *clean model* (a) for all the SNRs in comparison with *no processing* (mixed sound(a)) (45.9% vs 28.8% on average for the piano trio piece). It also outperformed D_{KL} for high SNRs and in the all cases using the training style (f).

However, D_{LOG} did not outperform D_{IS} for all cases. WH's estimation might be less than Y_{ij} because D_{LOG} imposes a cost faster than D_{IS} when $Y_{ij} < (WH)_{ij}$ and is nearly equal when WH is around Y_{ij}, and this might not be good for speech recognition. Therefore, the best cost function is both a large cost when $Y_{ij} > (WH)_{ij}$ and a small cost when $Y_{ij} < (WH)_{ij}$. Furthermore, NMF using D_{LOG} can not estimate signals well because the base vectors are composed using the VQ technique based on linear spectral. We must verify the behavior when the base vectors are composed by NMF. As expected, D_{EU} showed the worst performance.

A similar tendency was also observed for the jazz background music.

C. Objective evaluation

As an evaluation metric for NMF itself, we used the Source to Distortion Ratio (SDR) in the spectral domain and Cepstral Distortion (CD) as follows:

$$SDR = 10\log_{10}\frac{\sum_{n,f} S_{n,f}^2}{\sum_{n,f}(S_{n,f} - \hat{S}_{n,f})^2} \qquad (11)$$

TABLE I
RECOGNITION RATE - PIANO TRIO - [%]

Method/input	Training style	SNR				
		−5dB	0dB	10dB	20dB	ave
NMF_D_{EU}	(a)	4.0	11.7	53.6	88.0	39.3
	(c)	24.2	50.0	88.6	95.7	64.6
	(d)	15.0	33.9	73.0	83.9	51.5
	(e)	14.0	36.2	76.2	90.3	54.2
	(f)	23.5	49.0	85.0	94.3	63.0
NMF_D_{KL}	(a)	7.1	20.5	64.4	89.2	45.3
	(c)	30.4	59.4	91.5	96.3	69.4
	(d)	17.0	38.8	72.9	82.2	52.7
	(e)	19.1	42.9	80.0	91.7	58.4
	(f)	34.5	58.6	88.5	94.9	69.1
NMF_D_{IS}	(a)	8.2	20.7	73.0	92.6	48.6
	(c)	30.2	53.4	90.7	95.7	67.5
	(d)	28.4	59.7	87.0	92.4	66.9
	(e)	32.8	65.7	92.5	96.4	71.9
	(f)	51.4	78.4	95.8	97.4	80.8
NMF_D_{LOG}	(a)	6.9	18.4	66.4	91.7	45.9
	(c)	26.3	48.3	85.9	95.5	64.0
	(d)	27.4	55.1	83.4	89.6	63.9
	(e)	28.2	57.3	88.6	94.8	67.2
	(f)	42.2	67.6	92.0	96.1	74.5
Mixed sound	(a)	1.1	2.9	31.4	79.8	28.8
	(b)	11.8	30.4	83.4	92.6	54.6
	(c)	13.1	33.8	86.9	96.6	57.6
Clean speech	(a)	98.8				

$$CD = \frac{1}{N}\sum_{n}\sqrt{\left(C^{x}_{n,l} - C^{y}_{n,l}\right)^{2}} \qquad (12)$$

where $S_{n,f}$ is the target signal spectrum, $\hat{S}_{n,f}$ is the estimated signal spectrum, $C^{x}_{n,l}$ is MFCC of the target signal, $C^{y}_{n,l}$ is MFCC of the estimated signal, $n = \{1, \cdots, N\}$ is the time frame index, $f = \{1, \cdots, F\}$ is the frequency bin, and $l = \{1, \cdots, L\}$ is the filter index.

Tables II and III show the SDR and the CD for speech after music removal, respectively. The SDR of the NMF based on the D_{IS} or D_{LOG} cost function was worse than the D_{KL} cost function. Thus, even though SDR is often used as a measure of speech enhancement, it does not seem to correspond directly with speech recognition accuracy. On the other hand, the CD of the NMF based on the D_{IS} cost function was best performance, followed by D_{LOG}.

TABLE II
SOURCE TO DISTORTION RATIO - PIANO TRIO -

Method	SNR				
	−5dB	0dB	10dB	20dB	ave
NMF_D_{EU}	3.03	6.23	10.84	12.96	8.26
NMF_D_{KL}	3.97	7.14	12.56	16.07	9.94
NMF_D_{IS}	3.50	5.41	9.77	13.00	7.92
NMF_D_{LOG}	3.23	4.91	8.77	11.85	7.19

TABLE III
CEPSTRAL DITORTION - PIANO TRIO -

Method	SNR				
	−5dB	0dB	10dB	20dB	ave
NMF_D_{EU}	20.93	19.03	15.19	11.42	16.64
NMF_D_{KL}	20.16	18.21	14.37	10.91	15.91
NMF_D_{IS}	19.80	17.81	13.93	10.66	15.55
NMF_D_{LOG}	20.07	18.07	14.07	10.70	15.73

VI. CONCLUSIONS

In this paper, as a music removal method for speech recognition in mixed sound, we introduced the Euclidean distance of the logarithm spectrum, compared it with other cost functions using six types of acoustic models, and evaluated them by an isolated word recognition experiment with 200 words .

Although NMF using the Euclidean distance of the logarithm spectrum obtained improvement from the *no processing* of all the SNRs and outperformed the Kullback-Leibler divergence based NMF, it did not outperform the Itakura-Saito divergence based NMF.

The recognition rate was improved by a training model that was trained by adding mixed sound and/or estimated speech.

As future works, we plan to do online learning of basis vectors by semi-supervised NMF so that it can deal with various music sounds, and voice activity detection under music and extensions to large vocabulary continuous speech recognition.

REFERENCES

[1] 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.

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

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

[4] 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.

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

[6] 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.

[7] 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.

[8] N. Hashimoto, S. Nakano, K. Yamamoto and S. Nakagawa, "Speech recognition based on Itakura-Saito divergence and dynamics/sparseness constraints from mixed sound of speech and music by non-negative matrix factorization," Proc. INTERSPEECH , pp.2749–2753, 2014.

[9] 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.

[10] J. Eggert, E. Korner. "sparse coding and NMF", Neural Networks, vol. 4, pp. 2529-2533, 2004.

[11] 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.

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

[13] T. Higuchi, H. Kameoka, "Multipitch analysis based on cepstrum distance regularization", Proc. IPSJ SIG-MUS, 2014-MUS-104(10), 2014. (in Japanese)

[14] L. Deng, A. Acero, M. Plumpe, and X. Huang,"Large-vocabularyspeech recognition under adverse acoustic environments," Proc ICSLP, Beijing, China, pp. 806-809, 2000.

[15] 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)