Fast NMF Based Approach and VQ Based Approach Using MFCC Distance Measure for Speech Recognition From Mixed Sound

Shoichi Nakano*, Kazumasa Yamamoto*, and Seiichi Nakagawa*
*Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Japan.
E-mail: {snakano,kyama,nakagawa}@slp.cs.tut.ac.jp Tel: +81-532-44-6777

***Abstract*—We have considered a speech recognition method for mixed sound, consisting of speech and music, that removes only the music based on vector quantization (VQ) and non-negative matrix factorization (NMF). Instead of conventional amplitude spectrum distance measure, MFCC distance measure which is not affected by the pitch is introduced. For isolated word recognition using the clean speech model, an improvement of 53% word error reduction rate was obtained compared with the case of not removing music. Furthermore, a high recognition rate, close to clean speech recognition was obtained at 10dB. For the case of the multi-conditions, our proposed method reduced the error rate of 67% compared with the multi-conditions model.**

I. Introduction

Speech recognition performance is significantly reduced in noisy environments. Therefore, for speech recognition in the presence of noise, it is necessary to reduce the effect of the noise. The spectral subtraction and Wiener filter based methods are general techniques for noise removal. Although these methods are valid for stationary noise, they are not effective for non-stationary noise. In this paper, we consider speech recognition in speech with background music that constitutes non-stationary signals. Several music removal methods have been proposed for separating speech and music using a single microphone, such as the binary masking [1] and non-negative matrix factorization (NMF) [2] methods. Methods for sound source separation when multi-channel inputs are available from multiple microphones based on independent component analysis (ICA) have been widely used [3].

For mixed speech into a single channel, there was a monaural speech separation and recognition challenge, where keywords in sentences spoken by a target talker was identified with a background talker saying similar sentences [4]. Main approaches for this task were based on missing feature theory, speaker dependent/independent models CASA (Computational Auditory Scene Analysis) and NMF [5]. Although *Grais and Erdogan* proposed a regularized NMF using Gaussian mixture priors, the SNR was not improved for 20dB [6].

We considered music removal for input speech with background music from a single microphone using vector quantization [7] and non-negative matrix factorization, and applied these methods to speech recognition in mixed sounds consisting of speech and music [8] [9]. In [8], we obtained the improvement of speech recognition rate by the music removal through the two methods. However, music removal based on NMF requires much computation, so it is not practical. Therefore, in [9], we proposed a fast calculation technique of music removal based on NMF and improvement of VQ method. In this paper, we propose as a further improvement, instead of conventional amplitude spectrum distance measure, the introduction of MFCC distance measure which is not affected by the pitch

II. Music removal by NMF

In recent years, the use of NMF has been studied to solve the sound source separation problems of separating music into vocal sound and instrumental sound [13] and separating mixed sound into music and speech [14].

A. Nonnegative Matrix Factorization

NMF decomposes $n\times m$ matrix V into $n\times r$ matrix W and $r\times m$ matrix H.

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

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

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

B. Applied to the sound source separation of NMF

In this paper, we refer to the idea of phoneme recognition using NMF in [15] to separate speech and music in mixed sound. Matrix V is composed of an amplitude spectrogram; that is, a sequentially arranged amplitude spectrum for each frame of input sound as a column vector. Matrix V 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, that is, $W = [W_s W_m]$. In the experiment, we fixed W, because the VQ code vectors are considered to be representative basis vectors. We used VQ code vectors for speech and music sound as basis vectors for W_s and W_m, respectively. After this processing,

$$V \approx W_s H_s + W_m H_m \tag{3}$$

can be separated into $W_s H_s$ and $W_m H_m$ corresponding to speech and music, respectively. In this paper, to obtain estimated spectrum of speech and music, we construct a filter from

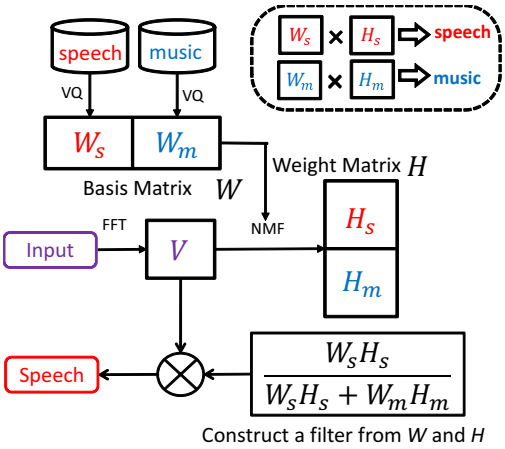

Fig. 1. Overview of music removal by NMF.

the decomposed results, which multiplies the input signal, as follows:

$$\hat{S} = V \otimes \frac{W_s H_s + C_1}{W_s H_s + W_m H_m + C_2} \tag{4}$$

$$\hat{M} = V \otimes \frac{W_m H_m + C_1}{W_s H_s + W_m H_m + C_2} \tag{5}$$

where $\hat{S}$ is estimated amplitude spectrogram of speech, $\hat{M}$ is estimated amplitude spectrogram of music, C_1 and C_2 are constant values for smoothing, the operator $\otimes$ and all division are element wise multiplication and division, respectively. Figure 1 shows an overview of our NMF method.

C. Fast calculation technique of NMF based approach

The normal NMF method described in Section II-B requires to perform the matrix decomposition for each input speech, so it is not practical due to the large amount of calculation. In this paper, we propose a fast calculation technique of NMF based approach. The technique is to achieve in advance an approximate separation based on NMF by creating a VQ codebook from mixed sound of the training data, decompose the matrix of VQ code vectors, then use the results of decomposition corresponding to the input speech. Figure 2 shows an overview of the proposed method.

The method consists of the following steps.

1) Obtain the representative spectrum $\hat{Y}$, W_s and W_m for mixed sound, speech and music through VQ clustering.
2) NMF decomposes a representative spectrum $\hat{Y}$ of mixed sound, and obtain the weight matrix H.
3) Calculate the distance between input sound Y and each column of $\hat{Y}$, and find the index of the column has the closest distance.
4) Construct a filter from H corresponding to the obtained index and the basis W.
5) Separate speech and music by multiplying the filter to amplitude spectrogram of input sound.

Steps 1 and 2 are performed in advance. Steps 3–5 are performed frame by frame for each input speech. Since the matrix decomposition by NMF is conducted only once in advance, the amount of computation is greatly reduced.

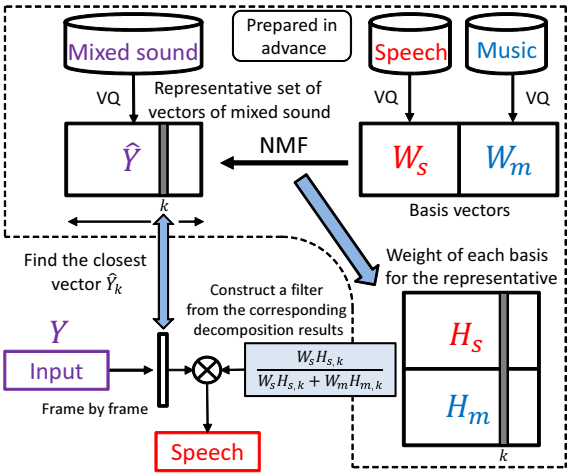

Fig. 2. Overview of FastNMF.

III. Music Removal by VQ Method

This method is a simple novel method for separating noisy speech using an example based method, which simplifies a statistical method [11][12].

Figure 3 shows an overview of music removal by our VQ method [7]. The method consists of the following steps performed in the amplitude spectrum domain.

1) Clean speech data and music data are prepared as training data. The music data ($M(i)$) are added to the clean speech data ($S(i)$) to create noisy speech data ($Y(i) = S(i) + M(i)$) with variations in the SNRs, where i represents the frame number.
2) A set of pairs of noisy speech data and the corresponding speech data are prepared in a spectral domain, $\{Y(i) = S(i) + M(i), S(i)\}$, where $i = 1, 2, \cdots, I$. I denotes the number of frames in the training sample.
3) A VQ codebook is generated from the feature vectors using the Linde-Buzo-Gray (LBG) algorithm. In this process, only the noisy speech amplitude spectrum is used for VQ clustering, $\{\hat{Y}(k), \hat{S}(k)\}$, where $k = 1, 2, \cdots, K$. K denotes the codebook size.
4) Using the input sound signal (noisy speech) $Y(j)$ as the key, the codebook index is searched for the closest matching codebook to the noisy speech input by comparing with the noisy speech spectrum in the codebook. $D(j,k) = ||Y(j) - \hat{Y}(k)||$
5) Construct a filter from the found code vector and it applies to the input sound signal. C_3 and C_4 are constant values for smoothing.

$$\hat{S}(j) = Y(j) \times \frac{\hat{S}(\hat{k}) + C_3}{\hat{Y}(\hat{k}) + C_4}, \quad \hat{k} = \arg\min_k D(j,k) \tag{6}$$

6) Restore the speech signal from the spectrum leaving only the speech component.

Steps 1–3 constitute the training phase. The spectrum obtained in step 5 is converted to MFCC as feature parameters for speech recognition. In this paper, we divide the spectral vector into four sub-vectors to enlarge the size of the actual codebook.

In [8], after creating a VQ codebook of the spectrum pair of mixed and music, the estimated speech was represented by

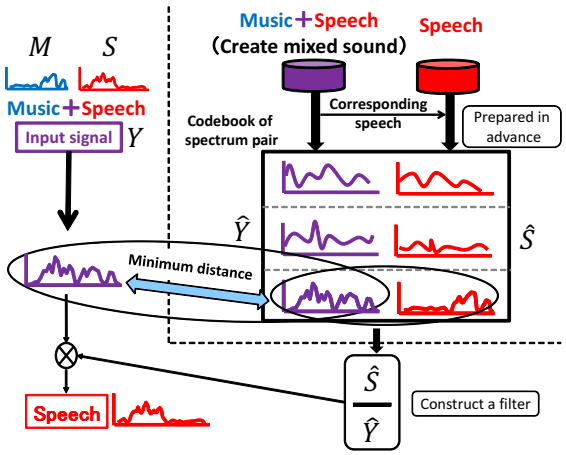

Fig. 3. Overview of music removal by VQ method.

$\hat{S}(j) = Y(j) - \hat{M}(\hat{k})$. However, this approach was worse than an filtering approach based on Eq.(6) [9].

IV. Vector Quantization Based on MFCC distance measure

In the conventional method described in Sections 2 and 3, only the amplitude spectrum is used as a feature vector. However, the amplitude spectrum variation due to the influence of the pitch is large. So, it is considered a framework to reduce the affect. In this section, we propose to introduce an MFCC distance based on a feature vector, the affect of the pitch is removed at the time of creation of the VQ codebook and search of a code vector. Cepstrum distance is defied by the low-order 20-dimensional coefficients.

A. Fast NMF method combined with MFCC distance

When we create a vector representation of the mixed sound by VQ, the combined vector of the low-order cepstrum and the amplitude spectrum is used. A VQ codebook is created by clustering the low-order cepstrum part. In music removal phase, an optimum code vector is found by the distance between the low-order cepstrum of the input speech and the low-order cepstrum portion of the VQ codebook. Then, a filter from the decomposition results of the amplitude spectrum corresponding to the found one is constructed. In addition, it is also conceivable, instead of the Eq.(2), to change with the cepstrum Euclidean distance measure. However, it is difficult to derive the update rules, rather it became worse because of the use at the approximate update rule.

B. VQ method combined with MFCC distance

In the conventional VQ method, the code vector consists of a pair of mixed sound and corresponding clean speech amplitude spectrum. A VQ codebook is created clustering by the amplitude spectrum of mixed sound.

In the proposed method, in addition to the amplitude spectrum of mixed sound and clean speech, a code vector includes the low-order cepstrum of mixed sound. A new VQ codebook is created by clustering in the low-order cepstrum part. In music removal phase, an optimum code vector is found by the distance between the low-order cepstrum of the input speech and the low-order cepstrum portion of the VQ codebook. Then, a filter from the decomposition results of the amplitude spectrum corresponding to the found one is constructed.

V. Experiments

A. Experimental setup

A recognition evaluation was carried out through an experiment using 200 isolated words from 20 speakers in the Tohoku University and Matsushita word speech database. For training data, we used 15 speakers, and for test data, we used the rest 5 speakers. We used the piano trio (mixed instruments of piano, violin, cello. First movement of Piano Trio in G minor Op.8) as the music data. The audio data were sampled at a frequency of 12 kHz in mono-mode. The word section was extracted by visual inspection.

In a representative vector set of mixed sound in FastNMF, the code vector is combined by the low-order cepstrum and the amplitude spectrum ($20 + 256 = 276$), and the codebook size is 4096. The conditions for speech analysis in the NMF method were a 512pts Hanning window and a 256pts frame shift. Matrix W, base vectors, was composed by both speech and music code vectors of size 512 constructed using the VQ technique.

The conditions for speech analysis in the VQ method were a 512pts Hanning window and a 256pts frame shift. Music was added to the training data at 20, 10, 0, and -5dB SNRs for training the VQ codebook. In the conventional method, the dimensions of the code vector were 256 (for noisy speech) + 256 (for clean speech) (frequency bins) with a codebook size of 8192. Here, we divided the spectral vector into four sub-vectors (64 dimensions each) to enlarge the size of the actual codebook; that is, the VQ represent 8192^4 distinct spectra. On the other hand, in the proposed method, a combined vector with the corresponding cepstrum was not split. A code vector is combined vector of the low-order cepstrum and the amplitude spectrum ($20+256+256 = 532$), and the codebook size is 8192.

In addition, constant values for smoothing were set to $C_1 = C_2 = C_3 = C_4 = 1$.

Acoustic models for speech recognition were constructed by whole word based HMMs, with 14 states and 8 mixtures of Gaussians (diagonal covariance matrix). As features we used 12 dimensions of the MFCCs, their deltas, double-deltas, delta power, and double-delta power (in total, 38 dimensions) obtained with a 25 ms window size and 10 ms frame shift.

Music was added to the 1000 words in the test data at 20, 10, 0, and -5dB SNRs. In addition, as a combination method, we combined VQ and NMF approaches. The likelihoods after removing the music by the VQ method and by the NMF method were linearly integrated as follows:

$$P = (1 - \alpha)P_{VQ} + \alpha P_{NMF} \tag{7}$$

where α is the an interpolation coefficient that is varied in increments of 0.1 from 0.0 to 1.0.

We conducted the recognition experiments by using two models; clean speech model and matched condition model.

All experiments were run on an Intel Xeon X5365 CPU of 3.0 GHz with 32 GB RAM.

TABLE I
WORD RECOGNITION RATE FOR CLEAN SPEECH MODEL [%].

input/method	SNR			
	−5dB	0dB	10dB	20dB
no processing	2.2	7.8	53.4	86.1
VQ (*original* [9])	8.0	20.0	74.1	90.9
VQ (proposed)	9.4	27.3	74.8	93.4
NMF	21.1	43.4	83.2	93.2
FastNMF (*original* [9])	5.2	17.6	71.4	90.4
FastNMF (proposed)	8.2	17.5	66.1	91.5
combination (*original* [9])	8.0	21.9	74.7	91.8
combination (proposed)	10.4	28.4	75.1	93.4
clean speech	98.8			

TABLE II
REAL-TIME-FACTOR FOR PROPOSED METHODS.

method	VQ method	NMF	FastNMF
RTF	0.26	10.83	0.21

B. Experimental results

Table I gives the recognition results for HMMs trained with the clean speech data. For the proposed VQ method, the improvement was obtained from our previous original method as 93.4% from 90.9% at 20dB, 74.8% from 74.1% at 10dB, 27.3% from 20.0% at 0dB and 9.4% from 8.0% at −5dB, respectively. For the proposed FastNMF, the improvement was obtained from our previous original method as 91.5% from 90.4% at 20dB and 8.2% from 5.2% at −5dB. However, there was no improvement for some cases. In combination of both methods, the improvement was obtained more than no-combination at the all SNRs except 20dB. From the case "no processing", the absolute improvement was 7.3% at 20dB (error reduction rate of about 53%) and 21.7% at 10dB (error reduction rate of about 47%), respectively. Table II shows the processing time for proposed methods in real time factor (RTF).

Table III shows the recognition results for HMMs trained with a matched condition or multi-condition model. The "matched condition" refers to speech recognition by an acoustic model trained under the same conditions as the test speech. For the proposed VQ method, the improvement was obtained from our previous original method at all SNRs. In particular, significant improvement was obtained in the low SNRs as 72.6% from 66.6% at 0dB and 42.3% from 35.4% at −5dB. On the other hand, for the proposed FastNMF, in some cases, the improvement was not obtained as in the case of clean speech models. In combination of both methods, the improvement was 45.9% from 37.7% at −5dB and 76.2% from 72.1% at 0dB, respectively. In compared with "mixed sound training model", the error reduction rates were 29.6% at −5dB, 47.4% at 10dB, 53.6% at 10dB and 66.7% at 20dB for the case of combination of "mixed sound" and two proposed methods, respectively.

VI. CONCLUSIONS

In this paper, we proposed to introduce MFCC distance measure instead of conventional amplitude spectrum distance measure, which was not affected by the pitch. By applying these methods to speaker independent isolated word recognition of 200 words, we obtained the significant improvement. In the model of the clean speech, the word error reduction rate of 53% was obtained in comparing the conventional method (in 20dB). In the matched condition, in the combination of three methods, the high recognition rates of about 98% at 10dB and about 80% at 0dB were obtained.

In future works, we incorporate the framework that takes into account of the dynamic feature and apply to more complex tasks.

TABLE III
WORD RECOGNITION RATE FOR MATCHED CONDITION [%].

method/model	SNR			
	−5dB	0dB	10dB	20dB
(a) mixed sound	25.0	59.3	94.4	98.5
(b) VQ (*original*)	35.4	66.6	95.7	98.5
(c) VQ (proposed)	42.3	72.6	96.0	99.0
NMF	48.3	76.1	94.0	97.1
(d) FastNMF (*original*)	22.1	61.1	94.1	98.6
(e) FastNMF (proposed)	37.3	66.1	93.3	98.6
combination (b+d)	37.7	72.1	97.2	99.4
combination (c+e)	45.9	76.2	96.8	99.0
combination (a+b)	37.5	73.6	96.7	98.7
combination (a+c)	43.4	76.3	97.4	99.5
combination (a+b+d)	39.6	77.3	97.9	99.6
combination (a+c+e)	47.2	78.7	97.4	99.5

REFERENCES

[1] H. Itou, Y. Ohishi, C. Miyajima, N. Kitaoka, and K. Takeda, "Source separation based on binary masking using Bayesian network," IPSJ SIG Tech. Rep. SLP, Vol.2008, No.68, pp.51–56, 2008. (in Japanese)

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," Proc. NIPS 2000, pp.556–562, 2000.

[3] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," Proc. Int. Comp. Music Conf., pp.154–161, 2000.

[4] M. Cooke, J. R. Hershey, and S. T. Rennie, "Monaural speech separation and recognition challenge," Comp. Speech & Lang., Vol.24, No.1, pp.1–15, 2010.

[5] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," Comp. Speech & Lang., Vol.27, No.3, pp.621–633, 2013.

[6] E. M. Grais and H. Erdogan, "Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation," Comp. Speech & Lang., Vol.27, No.3, pp.746-762, 2013.

[7] K. Yamamoto and S. Nakagawa, "Evaluation of privacy protection techniques for speech signals," Proc. IPMU 2010, 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 2011, 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 ASC 2012, OS.15-SLA.7, 2012.

[10] Y. Kitano, H. Kameoka, K. Kashino, N. Ono, and S. Sagayama, "Wiener filtering steered by complex NMFD with application to background music suppression," Proc. ASJ 2009 Spring Meeting, 3-9-6, pp.719–720, 2009. (in Japanese)

[11] R. Blouet, G. Rapaport, I. Cohen, and C. Fevotte, "Evaluation of several strategies for single sensor speech/music separation," Proc. ICASSP 2008, pp.37–40, 2008.

[12] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," IEEE Trans. Audio, Speech, & Lang. Proc., Vol.14, No.1, pp.191-199, 2006.

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

[14] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," Proc. INTERSPEECH 2010, pp.717–720, 2010.

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