Fast NMF based approach and improved VQ based approach for speech recognition from mixed sound

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

***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). This paper describe fast calculation technique of music removal based on NMF and improvement using a VQ method. For isolated word recognition using the clean speech model, an improvement of 46% 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 10 dB. For the case of the multi-conditions, our proposed method reduced the error rate of 50% 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 and CASA (Computational Auditory Scene Analysis).

We considered music removal for input speech with background music from a single microphone using vector quantization [5] and non-negative matrix factorization, and applied these methods to speech recognition in mixed sounds consisting of speech and music [6]. In [6], 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 this paper, we propose a fast calculation technique of music removal based on NMF and improvement of VQ method. NMF as methods for suppressing music was proposed to construct a Wiener filter from the amplitude spectra obtained by complex non-negative matrix factor deconvolution (NMFD) considering multi-frames [7], and using samples of large amounts of data as basis vectors of music [11]. In this paper, however, we use it to restore the speech spectrum after constructing the filter from the amplitude spectrum obtained by NMF using the VQ code vectors as basis vectors.

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 [10] and separating mixed sound into music and speech [11].

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

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

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

$$H_{ij} \leftarrow H_{ij}\frac{\sum_k W_{ki}V_{kj}/(WH)_{kj}}{\sum_k W_{ki}} \tag{3}$$

$$W_{ij} \leftarrow W_{ij}\frac{\sum_k H_{jk}V_{ik}/(WH)_{ik}}{\sum_k H_{jk}} \tag{4}$$

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

B. Applied to the sound source separation of NMF

In this paper, we refer to the idea of phoneme recognition using NMF in [12] 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, $W = [W_s W_m]$. H is obtained

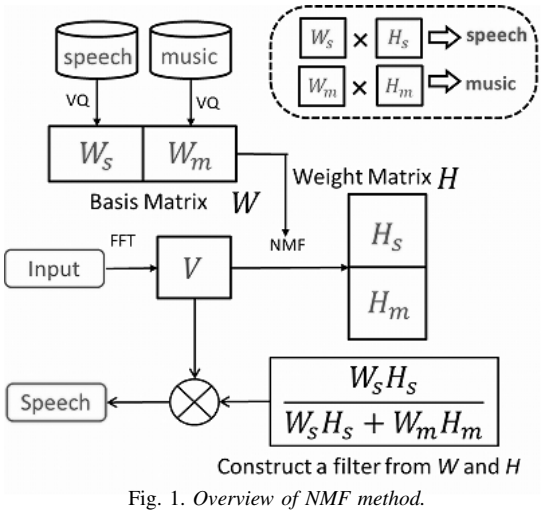

Fig. 1. *Overview of NMF method.*

from W using the update rule in (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 speech and music sound as basis vectors for W_s and W_m, respectively. After this processing,

$$V \approx W_sH_s + W_mH_m \tag{5}$$

can be separated into W_sH_s and W_mH_m corresponding to speech and music, respectively. In this paper, to obtain estimated spectrum of speech and music, we construct a filter from the decomposed results, which multiplies the input signal, as follows:

$$\hat{S} = V \otimes \frac{W_sH_s + C_1}{W_sH_s + W_mH_m + C_2} \tag{6}$$

$$\hat{M} = V \otimes \frac{W_mH_m + C_1}{W_sH_s + W_mH_m + C_2} \tag{7}$$

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 (we used $C_1 = C_2 = 1$), the operator $\otimes$ and all division are element wise multiplication and division, respectively. Figure 1 shows an overview of our NMF method.

The 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 V 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 that obtained by NMF.
5) Separate speech and music by multiplying the filter to amplitude spectrogram of input signal.

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

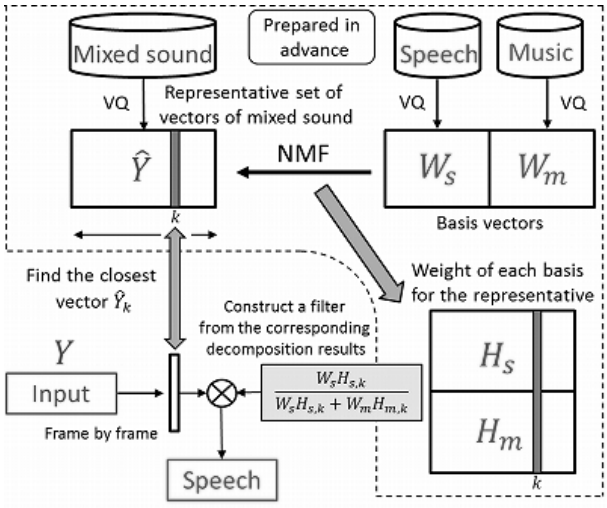

Fig. 2. *Overview of fast calculation technique.*

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) Obtains 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 $3 \sim 5$ 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.

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 [8][9].

Figure 3 shows an overview of music removal by our VQ method [5]. 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

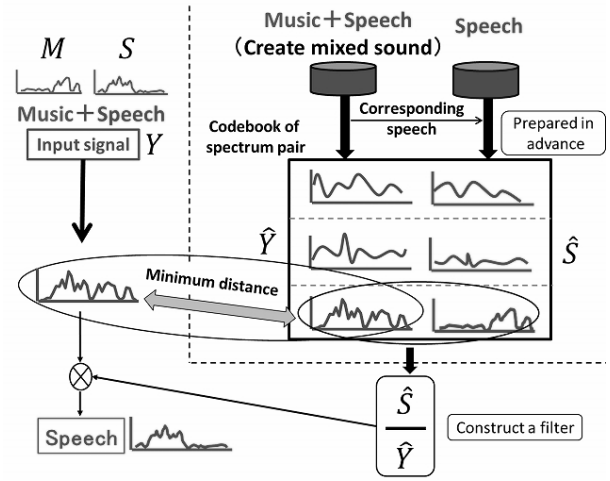

Fig. 3. *Overview of music removal by VQ method.*

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 (we used $C_3 = C_4 = 1$).

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

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

Steps $1 \sim 3$ 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 [6], after creating a VQ codebook of the spectrum pair of mixed and music, the estimated speech was represented by $\hat{S}(j) = Y(j) - \hat{M}(\hat{k})$. We call this *original VQ method*. In this paper, we proposed to construct a filter as well as NMF method and smoothing.

IV. 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.

The conditions for speech analysis in the VQ method were a 512 pts Hanning window and a 256 pts frame shift. Music was added to the training data at 20, 10, 0, and -5 dB SNRs for training the VQ codebook. The dimensions of the code vector were 256 (for noisy speech) + 256 (for clean speech) (frequency bins) with a codebook size of 8192. In this experiment, 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.

The conditions for speech analysis in the NMF method were a 256 pts Hanning window and a 128 pts frame shift. Matrix W, base vectors, was composed by both speech and music code vectors of size 512 constructed using the VQ technique.

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, 39 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 -5 dB 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:

$$P = (1 - \alpha)P_{VQ} + \alpha P_{NMF} \quad (8)$$

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

B. Experimental results

Table I gives the recognition results for HMMs trained with the clean speech data. Although a speech recognition rate of 98.8% was obtained for the clean speech data, when the music sound was added, the rate decreased markedly. Application of the original VQ method resulted in no improvement at 20 dB or 10 dB, and in fact, the results deteriorated. On the other hand, an improvement was observed for all SNRs when applying the proposed VQ method. In addition, an improvement of several percent was obtained with smoothing ($C = 1$). By applying the NMF method, significant improvement was obtained over and above that for the VQ method. With the application of the fast technique, although the results were worse than those for the normal NMF method, similar improvement to the VQ method was obtained. By applying a combination of the two methods, further improvement was obtained. For example, combining the proposed VQ method and the fast technique of the NMF method resulted in a reduction in the error rate of 41% at 20 dB and 46% at 10 dB, compared with "no processing".

Next, Table II 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. With the mixed sound or multi-condition model, a recognition rate of 98.5% was obtained at 20 dB. This rate is similar to that of clean speech recognition. For all other SNRs, the performance was significantly improved compared with that of the clean speech model. Although an improvement was obtained using the proposed methods with the "mixed

TABLE I
Word recognition rate for clean speech model [%].

input/method	SNR			
	-5 dB	0 dB	10 dB	20 dB
no processing	2.2	7.8	53.4	86.1
VQ (*original* [6])	4.6	12.3	45.0	66.0
VQ (proposed, $C = 0$)	7.8	19.4	71.0	89.0
(a) VQ (proposed, $C = 1$)	8.0	20.0	74.1	90.9
(b) NMF	21.1	43.4	83.2	93.2
(c) fast technique of NMF	5.2	17.6	71.4	90.4
combination (a+b)	21.1	43.4	83.3	93.6
combination (a+c)	8.0	21.9	74.7	91.8
clean speech	98.8			

TABLE II
Word recognition rate for matched condition [%].

method/model	SNR			
	-5 dB	0 dB	10 dB	20 dB
(d) mixed sound (matched)	25.0	59.3	94.4	98.5
VQ (*original* [6])	25.4	57.6	91.3	95.7
(e) VQ (proposed, $C = 1$)	35.4	66.6	95.7	98.5
(f) NMF	48.9	76.2	94.0	97.4
(g) fast technique of NMF	22.1	61.1	94.1	98.6
combination (e+f)	53.8	79.6	97.0	99.2
combination (e+g)	37.7	72.1	97.2	99.4
combination (d+e)	37.5	73.6	96.7	98.7

sound" or "matched model" at 0 dB and -5 dB, the results were almost the same at 20 dB and 10 dB. Combining the two methods resulted in a significant improvement. The best results were obtained using a combination of the proposed VQ method and the fast technique of the NMF method, with reductions in error rates of 60% (SNR= 20 dB), 50% (10 dB), 31% (0 dB) and 17% (-5 dB) compared with those of the "matched model".

Figure 4 shows a number of spectrograms (SNR=10 dB), visually confirming the improvement in each method.

Finally, we compared the computation times for music removal by the NMF methods. The normal NMF method requires a computation time about 20 times greater than real-time. In contrast, the proposed fast calculation technique can be processed in less than real-time. All experiments were run on an Intel Xeon X5365 CPU of 3.0 GHz with 32 GB RAM.

V. Conclusions

In this paper, we proposed a fast calculation technique of music removal based on NMF and the improvement of VQ method. 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 46% was obtained by the music removal (in 10dB). In the matched condition, the combination of the fast NMF and improved VQ reduced the error rate of 50% $\sim$ 60% for 10dB and 20dB. We achieved real-time processing in the fast calculation technique of NMF method.

In future works, we intend applying this proposed method to various mixed sound including composite music sounds, and so on. For this purpose, we need to revise the VQ procedure in the spectral domain, for example, from a FFT based spectrum to a cepstrum based smoothed spectrum.

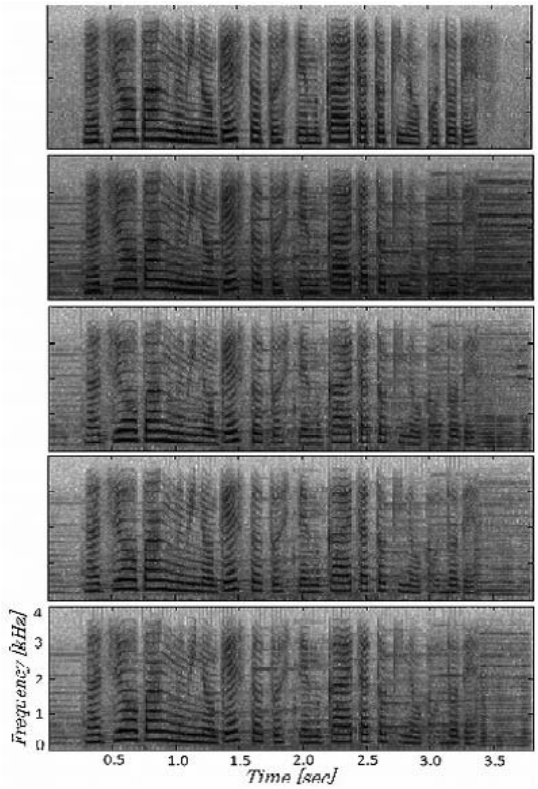

Fig. 4. *Examples of spectrograms (SNR= 10 dB). From top to bottom: clean speech, noisy speech, and speech restored using VQ method, NMF method, and fast technique of the NMF method.*

References

[1] H. Itou, Y. Ohishi, C. Miyajima, N. Kitaoka and K. Takeda, "Source separation based on binary masking using Bayesian network," IPJS, vol.2008, no.72, 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. International Computer Music Conference, Berlin, Germany, Aug. 2000.

[4] 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.

[5] K. Yamamoto and S. Nakagawa, "Evaluation of privacy protection techniques for speech signals," Proc. IPMU 2010, 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 2011, pp.1781–1784, 2011.

[7] Y. Kitano, H. Kameoka, K. Kashino, N. Ono and S. Sagayama, "Wiener Filtering Steered by Complex NMFD with Application to Background Music Suppression," ASJ spring 2009, 3-9-6, pp.719–720, 2009. (in Japanese)

[8] 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.

[9] L. Benaroya, F. Bimbot and R. Gribonval, "Audio Source Separation With a Single Sensor," IEEE Trans. Audio, speech and Language Processing, vol.14, no.1, pp.191-199, 2006.

[10] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," Proc. ICASSP 2010, 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 2010, pp.717–720, 2010.

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