2014 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing
NCSP'14, Honolulu, Hawaii, USA. February 28 - March 3, 2014

Speech recognition in noisy conditions based on speech separation using Non-negative Matrix Factorization

Yuxuan Du, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292 Japan
Phone/FAX:+81-0761511391
E-mail: {du_yuxuan, Akagi}@jaist.ac.jp

Abstract

This paper proposes a speech recognition method for applications in adverse noisy environments. Speech recognition in noisy conditions is a challenging problem since speech observed in such conditions is corrupted by noise. To deal with this problem, we integrate *non-negative matrix factorization* (NMF) and *modified restricted temporal decomposition* (MRTD) into a recognition method based on the concept of "auditory scene analysis" (ASA). Experiments were conducted using 100 isolated words in 4 different noise conditions at a *signal to noise ratio* (SNR) of 0 dB. Experimental results showed the proposed method achieved recognition rates of 80%, which is about 50% higher than that of the method based on *dynamic time warp* (DTW).

1. Introduction

Since noise corrupts spoken utterances, computers cannot recognize target sounds correctly. Therefore, to build a reliable speech recognition method for noisy speech signals, it is necessary to reduce the effects of the noise. Many methods were studied to solve this problem. In general, there are two approaches to deal with the effects of noise. The first approach focuses on suppressing the noise in the input signals before recognition. For example, spectral subtraction (SS) [1] was used as a front-end processor of Automatic Speech Recognition (ASR) systems to reduce noise. The second one aims at building ASR models that are able to adapt to the effect of noise. For example, Vector Taylor-series (VTS) model adaptation was used to account for noise [2]. However, there has been lack of ASR method working well in real noise environments, because noise suppression methods in the first approach (such as SS) are not able to deal with non-stationary noise, where adaptation models in the second approach cannot perform in arbitrary real noisy conditions.

On the other hand, human can easily recognize a target sound in various noisy environments, which is commonly referred as the "cocktail party effect" [3]. One factor contributing to the cocktail party effects is the function of an active scene analysis system, which is popularly known as "auditory scene analysis (ASA)" [4]. Recently, Haniu *et al.* [5] proposed a speech recognition method using the idea of ASA. In this method, speech was recognized after verifying the validity of segregation of speech and noise. It is hypothesized that only one target sound exists in input data. The possibility of existence of target sound in the input sound is calculated through verifying segregation process and segregation result. There may be several candidates of target sound existing in the input sound. After calculating the possibility for all candidates of target sound, the target sound which most possible existing in the input sound is selected as recognized word. Therefore, this method can recognize target sounds regardless to the type of noise. The method of Haniu *et al.* [5] achieved relatively high recognition rates in various noise conditions.

However, this method has a high computation cost because the segregation part of the method used a complex selective sound segregation model. Even in the situation where only 10 candidates of target sounds exist in the input sounds, it took over a day to recognize a input sound. This problem makes the method difficult to be used as recognition method in real conditions.

To solve this problem, a new recognition method is proposed based no the basic concept of Haniu *et al* [5]. However, In the proposed method, Non-negative Matrix Factorization (NMF) [6] is used as a separation method to separate speech and noise instead of the selective speech segregation model. Since NMF has computational advantage by avoiding considering all combinations across noise and target sounds [7], computational complexity can be reduced. Modified Restricted Temporal Decomposition (MRTD) [8] is further used to synthesize templates for recognition. MRTD makes it possible to modify the templates for different utterances, which is a basic requirement of speech recognition.Consequently, the proposed method that integrate NMF and MRTD is expected to be a speech recognition method suitable for the concept of ASA.

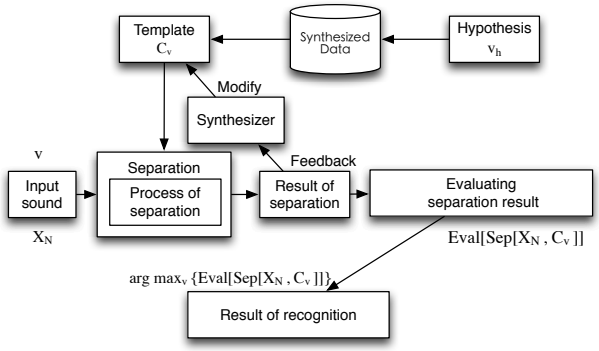

Figure 1: Outline of the proposed method

2. Proposed method

2.1 Outline

The outline of the proposed method is shown in Figure 1. Before speech recognizing, all possible target sounds need to be synthesized. The synthesizer can generate the controllable templates by synthesizing sound data from isolated words to phonemes. Therefore, it is possible for the proposed method to recognize input sounds from isolated words to phonemes.

In the proposed method, it is assumed that a target sound v exists in the input noisy sound X_N. At first we assume that $v = v_h$, where v_h is a candidate target sound. Then corresponding template C_v of v_h is used to separate the target sound v and noise. Result of separation can be used to give feedbacks to synthesizer. Then synthesizer can modify the template C_v to approach (such as time warping) the input sound X_N. After the modifcation, the template is used to separate the target sound and noise again. Possibility of existence of the candidate target sound v_h in the input sound X_N is verified by evaluating the result of separation, specified as $Eval\{Sep[X_N, C_v]\}$. This verification is repeated for all the candidates of the target word v_h. The recognition result is $v = arg\ max_v\{Eval\{Sep[X_N, C_v]\}\}$.

2.2 Non-negative Matrix Factorization (NMF)

NMF [6] decomposes $n \times m$ matrix X into $n \times k$ basic matrix B and $k \times m$ activation matrix G as follows:

$$X \approx B \times G \tag{1}$$

All elements of the matrices X, B, G are under the constraint of non-negativity. X is estimated by minimizing a cost function between X and $B \times G$. For speech analysis, we use the Itakura-Saito divergence as follows:

$$D_{IS}(x_{ij}, b_i gj) = \frac{x_{ij}}{t_i^T v_j} - log\frac{x_{ij}}{t_i^T v_j} - 1 \tag{2}$$

Following rule is used to update B and G until D_{IS} is converged.

$$b_{ik} \leftarrow b_{ik}\sqrt{\frac{\sum_j \frac{x_{ij}}{\hat{x}_{ij}}\frac{g_{kj}}{\hat{x}_{ij}}}{\sum_j \frac{g_{kj}}{\hat{x}_{ij}}}} \tag{3}$$

$$g_{kj} \leftarrow g_{kj}\sqrt{\frac{\sum_i \frac{x_{ij}}{\hat{x}_{ij}}\frac{b_{ik}}{\hat{x}_{ij}}}{\sum \frac{b_{ik}}{\hat{x}_{ij}}}} \tag{4}$$

2.3 Modified Restricted Temporal Decomposition (MRTD)

MRTD is proposed by Nguyen *et al.* [8] to synthesize sounds in low rate speech coding. This is an analysis procedure based on a linear model of the effects of co-articulation. The output of this method is a linear approximation of a time sequence of spectral parameters in terms of a series of time-overlapping event functions and an associated series of event vectors ,as follows:

$$\hat{y}(n) = \sum_{k=1}^{K} \alpha_k \phi_k(n),\ \ 1 \leqslant n \leqslant N \tag{5}$$

where α_k and ϕ_k are respectively the kth event vector and kth event function. $\hat{y}(n)$ is the approximation of $y(n)$ produced by MRTD model. In the proposed method, Mel-Frequency Cepstrum Coefficient (MFCC) is used as a spectral parameter. MFCCs of each target sound are synthesized by MRTD as templates.

The event targets α_k ($k = 1, \dots, K$) are initialized with the samples of MFCC vector trajectory $y(n_k)$. α_k are the n_kth vector of $y(n)$ which get minimal values of spectral feature transition rate (SFTR), which is calculated as:

$$SFTR:\ \ s(n) = \sum_{i=1}^{P} c_i(n)^2,\ \ 1 \leqslant n \leqslant N \tag{6}$$

where

$$c_i(n) = \frac{\sum_{m=-M}^{M} m y_i(n+m)}{\sum_{m=-M}^{M} m^2},\ \ 1 \leqslant i \leqslant P \tag{7}$$

In Eq. (6) and (7), P is the dimension of spectral parameter, and the window size of SFTR calculation is $2M$. After α_k are initialized, the corresponding event targets $\phi_k(n)$ can be calculated by MRTD algorithm. Therefore, it is possible to warp the time of sampling by modifying the event functions and it is possible to deal with different utterances of the same word by modifying the event targets.

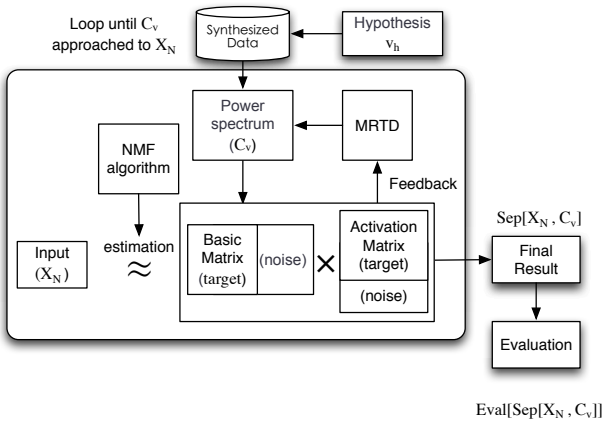

Figure 2: Separation based on NMF and MRTD

2.4 Separation and recognition based on NMF and MRTD

To calculate the possibility of existence of the target sound v in the input noisy sound X_N as $Eval[Sep[X_N, C_v]]$, NMF and MRTD are used as shown in Figure 2. When estimating the input sound X_N using NMF algorithm, the template C_v generated by MRTD is used as a part of basic matrix. This part is used to represent the target sound v in X_N. The rest part of basic matrix is generated randomly to represent the noise in X_N. By using the updating rules in Eq. (3), (4), the target sound part will not be changed, but the noise part will be updated. Therefore target sound v can be separated with noise regardless to the types of noise. Since the target sound part of basic matrix is set, if the target sound v is included in the input sound X_N, corresponding sound part of activation matrix will become a unit matrix. However, the differences between utterances of the same word may make the sound part of activation matrix different to a unit matrix. Thus, feedbacks are sent to MRTD and the MRTD can modify the template to make the sound part of activation matrix to be more like a unit matrix. On the other hand, if the target sound v is not included in the input sound X_N, the sound part of activation matrix will much differ form a unit matrix even the template is modified using the feedbacks.

By using NMF and MRTD, the presence of the target sound v in X_N can be judged by whether the sound part of activation matrix is similar to a unit matrix. Therefore, for values of sound part of activation matrix, distribution rate of the values near diagonal is a standard to evaluate similarity of C_v and X_N. As $v_{max} = arg\ max_v\{Eval\{Sep[X_N, C_v]\}\}$, the target sound v, which has the highest distribution rate of the values near diagonal in the sound part of activation matrix, is the recognition result.

3. Evaluation of the proposed method

To verify validity of the proposed method in first step, experiments for recognizing isolated words in various noise conditions based on proposed method were carried out. As a well-known template based recognition method, recognition experiments based Dynamic Time Warp (DTW) [9] method were carried out as a reference method to compare with the proposed method.

3.1 Experimental conditions

Clean speech data used in the experiments are FW03 Japanese data corpus [10]. As the first step to evaluate the proposed method in various noise conditions, a simplest condition that the same data were used to generate templates and input sounds. The first 100 isolated 4 mora Japanese words that uttered by speaker "fto" in FW03 data base were chosen as speech data. For the proposed method and method based on DTW, the templates were generated by MRTD using the 100 isolated clean words. The speech data were downsampled from 48 kHz to 16 kHz. MFCC was chosen as spectral parameter to synthesize the templates. The number of coefficients of MFCC was 13 and filter bank was set to 20. Window length and frame shift were 20 ms and 10 ms respectively. Since NMF has the constraint of non-negativity for X, B, and G in Eq. (1), MFCC was transformed to power spectrum before using algorithm of NMF. While using the DTW algorithm, MFCCs of templates and inputs were transformed to power spectrum, and the divergences between the templates and inputs were calculated based on power spectrum.

Noisy input sounds were generated by adding noise to the clean speech data. White noise, pink noise, speech-like babble noise and unsteady factory noise were used to mix with the target sounds. For each kind of noise, the SNRs of input sounds were 0, 10, 20, and ∞. Therefore, in addition to clean condition, there were 12 noisy conditions. For each experiment condition, 100 words with randomly generated noise were inputed once respectively.

3.2 Results and discussion

Correct recognition rates in noisy conditions of DTW-based method and the proposed method are shown in Figure 3 and 4, respectively.

In Figure 3, recognition rates of DTW-based method are significantly decreased in noisy conditions, especially at 0 dB of SNR. Results in Figure 4 showed the effectiveness of the proposed method in comparision with the recognition method based on DTW.

The proposed method got high recognition rates in various noise conditions regardless to noise model. Furthermore, to recognize a input sound with 100 candidate target sounds, it

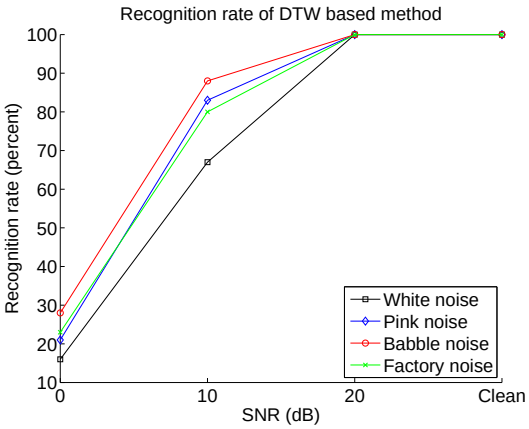

Figure 3: Recognition rate of DTW based method

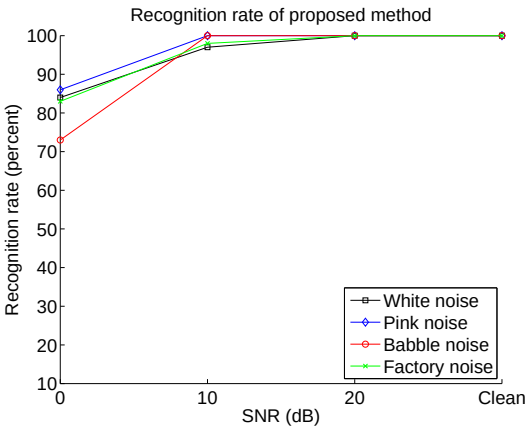

Figure 4: Recognition rate of proposed method

took only about 20 minutes in comparison with 1 day of the method of Haniu *et al* [5]. By reducing the amount of candidates of target sounds using such as speech forecast, it is possible to further reduce recognition time. This indicates that the proposed method is validity in various noise condition and it is potentially used in real conditions.

4. Conclusion

This paper proposed a speech recognition method in various noise conditions based on the concept of ASA in which speech data is recognized through verifying validity of segregation. The proposed method was implemented by NMF and MRTD. To evaluate the validity of the proposed method, experiments of recognizing isolated Japanese words in various noise conditions were carried out. Experimental results showed our proposed method achieved higher recognition rates in comparison with that of DTW based method, while processing time was much lower comparing with the method of Haniu *et al* [5]. This suggests that the proposed method is applicable and possible to be a ubiquitous recognition method in real conditions.

Acknowledgements

A part of this work was supported by the Strategic Information and Communications R & D Promotion Programme (SCOPE; 131205001) of the Ministry of Internal Affairs and Communications (MIC), Japan. This study was also supported by the Fostering ICT Global Leader Program.

References

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. Acoust. Speech Signal Process. ASSP-27, pp. 113-120, 1979.

[2] A. Agarwal and Y. M. Cheng, “Two-stage mel-warped Wiener filter for robust speech recognition”, Proc. ASRU, 1999.

[3] E. C. Cherry, “Some experiments on the recognition of speech with one and with two ears”, J. Acoust. Soc. Am., vol.25, no.5, pp. 975-979, Sept. 1953.

[4] A. S. Bregman, “Auditory Scene Analysis: The perceptual Organization of Sound”, MIT Press, Cambridge, MA., 1990.

[5] A. Haniu, M. Unoki and M. Akagi, “A speech recognition method based on the selective sound segregation in various noisy environments”, NCSP, 2008.

[6] D. D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization”, 2000.

[7] S. J. Rennie, J. R. Hershey and P. A. Olsen, “Single-Channel Multitalker Speech Recognition”, IEEE Signal Processing Magazine, pp. 66-80, 2010.

[8] P. C. Nguyen, T. Ochi and M. Akagi, “Modified Restricted Temporal Decomposition and Its Application to Low Rate Speech Coding”, ” IEICE TRANSACTIONS on Information and Systems, E86-D(3) : pp. 397-405, 2003.

[9] D. Ellis , “Dynamic Time Warp (DTW) in Matlab”, http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/, 2003.

[10] A. Shigeaki, K. Kondo, S. Sakamoto and Y. Suzuki, “Speech Data Set for Word Intelligibility Test based on Word Familiarity (FW03)”, NII Speech Resources Consortium, 2006.