A study on a speech recognition method based on the selective sound segregation in various noisy environments

Atsushi Haniu, Masahi Unoki, and Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa, 923–1292 Japan
Phone/Fax: +81–761–51–1699/+81–761–51–1149
Email: {a-haniu, unoki, akagi}@jaist.ac.jp

Abstract—This paper shows the effectiveness of our proposed speech recognition method in the noisy environments. In our proposed method, the target sound is recognized based on verifying the possibility of the existence of the target sound by evaluating the selective sound segregation using the hypothesis of the target sound and Bregman's regularities. Since this method does not use any noise model, the proposed method is able to recognize the target sound under various noisy environments. In order to evaluate the proposed method, performance of ASR system using the proposed method in noisy environments was compared with performance of traditional methods under speech recognitions. The results show that the proposed method is more robust than other methods in experimental conditions from SNR = 0 dB to ∞ dB. Therefore, these indicate the proposed method is a useful speech recognition concept in noisy environments.

1. Introduction

In real environments, there are undesirable sounds overlapping unpredictably with the target sound in both the time and frequency domains. They change the values of important features of the target sound unpredictably, which makes it difficult to recognize the target sound in noisy environments. Almost all ASR systems for noisy environments have a front-end processor to reduce noise and/or an acoustic model which is adapted to the noisy environments. For example, the spectral subtraction (SS) method [1] and adaptive filtering are typically used as the front-end processor, while the parallel model combination (PMC) [2]/NOVO [3] method is commonly employed in the adapted acoustic model. The typical SS method is not able to reduce non-stationary noise, and the NOVO method is not able to refine the acoustical model for all kinds of noises in real environments. Therefore, there are no drastic improvements for the system to work in practical environments.

On the other hand, human can easily recognize a target sound in various noisy environments, which is commonly known as "the cocktail party effect" [4]. One factor contributing to the cocktail party effect is regard as a function of an active scene analysis system, called "auditory scene analysis" (ASA) [5]. Bregman reported that the human auditory system uses four psychologically heuristic regularities related to acoustic events to perform the ASA [6]. Unoki et al. [7] proposed a selective sound segregation model using features of the target sound, and Bregman's regularities. We proposed the recognition method [8] using the selective sound segregation model. Our main aim in this paper is to show the effectiveness of our proposed speech recognition concept in the noisy environments.

2. Proposed method

2.1. Outline of the proposed method

The four psychologically heuristic regularities which Bregman argued are as follows [6]: **Regularity 1**: Common onset/offset time, **Regularity 2**: Gradualness of change, **Regularity 3**: Harmonicity of components, and **Regularity 4**: Synchronicity of changes on occurring acoustic event. Unoki et al. [7] proposed a selective sound segregation model to segregate a target sound selectively using the Bregman's four regularities, and features generated from the model of the target sound only, without any priori knowledge of noise. We express this "features generated from the model of the target sound" as a hypothesis. This segregation model has been shown to be robust in complex mixed sounds [7]. When segregating forcibly the target sound from the input sound which does not contain the target sound, this model suspends its own process since this model is not able to segregate the sound which fulfills the Bregman's regularities. Even if the model is able to complete the segregation process in such a situation, the sound obtained as the result is contradictory to the hypothesis of the target sound. Thus, by segregating a sound using a hypothesis, we can verify whether the hypothetical sound exists in the input sound.

We proposed the recognition method [8] using the selective sound segregation model, based on the above-mentioned idea. The concept of the proposed method is

$$v_{\max} = \arg\max_{v} \{\mathrm{Eval}\,[\mathrm{Seg}\,[\boldsymbol{X}_N,\, \boldsymbol{C}_v]]\}, \qquad (1)$$

where X_N is a noisy input sound, $\mathrm{Seg}\,[\boldsymbol{X}_N,\, \boldsymbol{C}_v]$ is the selective sound segregation using the feature vector of $\boldsymbol{C}_v$, and

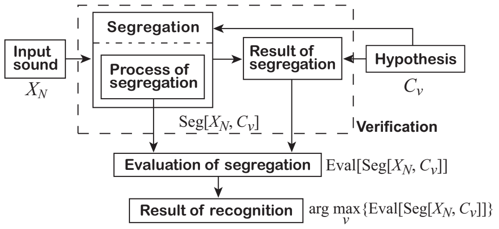


Figure 1: Conceptual block diagram of the proposed method.

Eval[·] is a function for evaluating validity of the selective sound segregation. In this paper, a feature vector of $\boldsymbol{C}_v$ is used as a hypothesis, and we assumed additive noise environments. A conceptual block diagram of the proposed method is shown in Fig. 1. The validity of the selective segregation process is to evaluate whether the segregated sound fulfills the Bregman's regularity. The validity of the selective segregation results is to evaluate whether the segregated sound is contradictory to the hypothesis. The selective sound segregation is evaluated by the validity of both the segregation process and the segregation results. These evaluations would be repeated through all the possible categories and the category with the maximum value $v_{\max}$ of the validity would be identified. In this way, the proposed method can recognize the target sound using the hypothesis without being influenced by noisy environments.

In ASR systems using the method (i) noise reduction methods as a pre-processor, recognition becomes remarkably difficult with non-stationary noise, because it is assumed that noise is stationary in almost all noise reduction method. In ASR systems using the method (ii) an acoustical model which is adapted to the noisy environments, a priori knowledge of the noise environment is necessary. The recognition accuracy of such systems would decrease drastically in unknown noisy environments. On the contrary, the proposed method does not use any noise model, and the priori knowledge of the noise environment. In addition, the proposed method verifies the possibility of the existence of the target sound by evaluating the selective sound segregation using the hypothesis of the target sound, as well as computes the distance between the input sound and the hypothesis. Consequently, the proposed method would be able to recognize the target sound under various noisy environments.

2.2. ASR system based on the proposed method

The system based on the concept of the proposed method was implemented with four functional blocks: ① signal analyzer, ② hypothesis manager, ③ segregation block, and ④ recognition block. The functional block diagram of the proposed system is shown in Fig. 2. The details of the system are described in [9]. The brief description of the ASR system based on the proposed method is described in the following document.

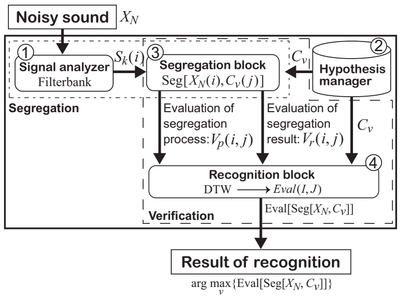


Figure 2: Functional block diagram of proposed system.

In order to process the input sound in both time and frequency domains, the signal analyzer (①) analyzes the input sound in the time-frequency domain into the instantaneous amplitude $S_k(i)$ using a K-channels constant bandwidth gammertone filterbank and Hilbert transform technique [10]. $k(= 1, 2, \ldots, K)$ is an index that denotes a channel number of the filterbank. $i(= 1, 2, \ldots, I)$ denotes process time index of the input sound.

The hypothesis manager (②) generates a running spectrum of a reference pattern as hypothesis and provides the hypothesis to the segregation block and the recognition block. The hypothesis $\boldsymbol{C}_v(j)$ is made from the clean speech using the same filterbank of the signal analyzer. $j(= 1, 2, \ldots, J)$ denotes process time index of the hypothesis, v $(= 1, 2, \ldots, V)$ is an index of categories, V is number of categories, and $\boldsymbol{C}_v$ is the v-th category. The hypothesis manager (②) holds the hypothesis of all the categories

The segregation block (③): Seg$[\boldsymbol{X_N}(i), \boldsymbol{C}_v(j)]$, which is based on the selective sound segregation method [7], segregates the target sound using the hypothesis and Bregman's regularities. The target sound is predicted with fulfilling the **Regularity 2** using Kalman filter described in [10]. In order to fulfill the **Regularity 3**, the system estimates F0 of the target sound, and determines ℓ-th harmonic components $(\ell = 1, 2, \ldots, L)$ using the estimated F0. After that, the system segregates the most similar sound to the hypothesis using the hypothesis of the target sound.

The recognition block (④) evaluates the segregation block using the hypothesis of the target sound and Bregman's regularities, and then recognizes the target sound. Validity of the segregation process is computed based on **Regularity 4** in time domain. Validity of the segregation results is computed using similarity between the segregated results and the hypothesis in frequency domain. In order to evaluate how much the segregated sound fulfills **Regularity 4**, the system computes an average cross correlation $V_p(i, j)$ between each amplitude envelope $\boldsymbol{X}_\ell(i, j)$ of the segregated sound's harmonics from the time index $i - s$ to i. This system considers that $V_p(i, j)$ is validity of the segregation process. In order to evaluate how much the segregated sound is similar to the hypothesis of the target sound, the system calculates the similarity $V_r(i, j)$ between the segregated results and the hypothesis. To calculate the similarity with the distribution $\sigma_k^2(i)$ in the estimation of the segregation block (③) by the Kalman filter, the similarity $V_r(i, j)$ between the segregated results and the hypothesis is calculated using the normalized normal distribution. This system considers that $V_r(i, j)$ is validity of the segregation result. Validity of the segregation block $V(i, j)$ is computed by unifying $V_p(i, j)$ and $V_r(i, j)$ using Dempster–Shafer theory [11]. In order to absorb the time distortion of the input sound and the hypothesis, an appropriate selective sound segregation is decided based on DTW technique [12]:

$$\begin{aligned} E(i, j) =& V(i, j)+ \\ & \max\{E(i-1, j), E(i-1, j-1), E(i, j-1)\} , \end{aligned} \tag{2}$$

where $E(i, j)$ is total validity of the selective sound segregation process and result, and $E(1, 1) = V(1, 1)$, and $E(I, J) = \mathrm{Eval}[\mathrm{Seg}[X_N, C_v]]$. Therefore, the implemented system recognizes the target sound using Eq. (1).

3. Evaluation of the proposed method

3.1. Experimental conditions

In order to evaluate the performance of the proposed method, speech recognition experiments for recognizing five Japanese vowels (/a/, /e/, /i/, /o/, and /u/) and words (/hai/ and /iie/) in noisy environments were carried out. Clean speech data used in the experiments are the JEIDA Japanese common speech data corpus [13]. Each speaker uttered each word and phoneme four times in this database. In these experiments, one of the speech data uttered four times by Japanese male speaker (Speaker number 1) was used as a reference pattern, and the remainder of the data were used as clean speech data.

Noisy input sound was made by adding noise data to the clean speech data on a computer. The "machinegun," "babble," and "pink" were selected as non-stationary, speech-like, and stationary noise data respectively from the NOISEX-92 [14] noise database. The SNRs of input sound were at 0, 10, 20, and ∞ dB. The sampling frequency of the speech and noise is 16 kHz, and the channel number of the filterbank is 400.

Comparing the performance of the proposed method, the recognition experiments were carried out with four DTW-based ASR systems: [**System A**]: an ASR system not using a front-end processor and an adaptation process, [**System B**]: an ASR system with (i) a pre-processor, [**System C**]: an ASR system with (ii) an adaptation model, and [**System D**]: an ASR system employing the proposed method. The **System A**, **B**, and **C** use the same DTW method. The **System A**, **B**, and **C** have the same signal analyzer as the **System D**. Similarly, all templates and hypothesis were computed from the clean speech data using the filterbank. The **System A** computes the local distance $d(i, j)$ between an input sound vector $\mathbf{S}(i) = \{S_1(i), S_2(i), \ldots, S_K(i)\}$ and a template vector $\boldsymbol{C}_v(j)$ as follows:

$$d(i, j) = \{1 - \langle \boldsymbol{S}(i), \boldsymbol{C}_v(j)\rangle / (\|\boldsymbol{S}(i)\| \, \|\boldsymbol{C}_v(j)\|)\} . \tag{3}$$

These systems recognize the target sound using DTW technique:

$$\begin{aligned} D(i, j) =& d(i, j)+ \\ & \min\{D(i-1, j), D(i-1, j-1), D(i, j-1)\} , \end{aligned} \tag{4}$$

$$v_{\max} = \arg\min_v D(I, J) , \tag{5}$$

where $D(i, j)$ is the distance between the input sound vector and the template vector at (i, j), and $D(1, 1) = d(1, 1)$. The **System B** has a noise reduction pre-processor using the SS method [1] as follows:

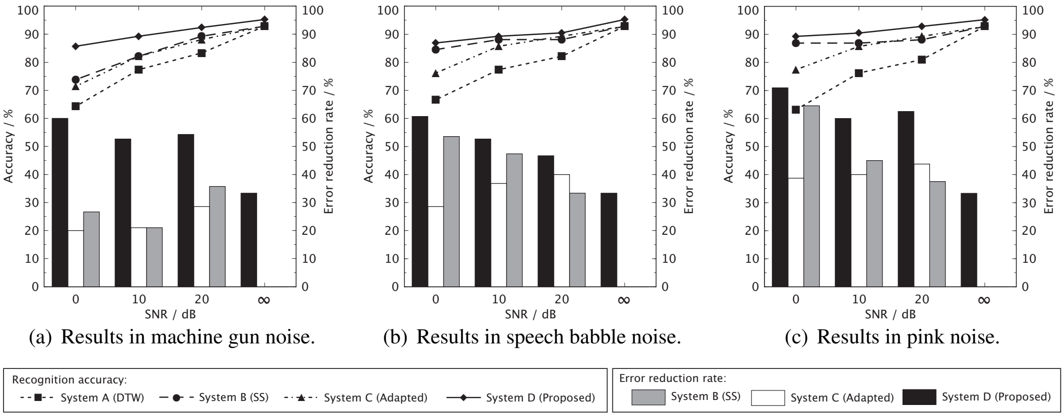
$$\widehat{X}_k(i) = \begin{cases} S_k(i) - N_k, & S_k(i) - N_k > 0.01\, S_k(i) \\ 0.01\, S_k(i), & \text{otherwise} , \end{cases} \tag{6}$$

where $\widehat{\boldsymbol{X}}(i) = \{\widehat{X}_1(i), \widehat{X}_2(i), \ldots, \widehat{X}_K(i)\}$ is the estimated speech signal using SS. In the DTW of the System B, the input sound vector is $\widehat{\boldsymbol{X}}(i)$. The **System C** has templates $\boldsymbol{C}_v(j)$ to which the noise spectrum was added, instead of acoustic model adaptation. In the DTW of the System C, the template vector is $\boldsymbol{C}'_v(j)$. In these experiments, N_k is a time average spectrum of the noise which was added to the clean speech data.

3.2. Results and discussion

The recognition accuracy and error reduction rate of the ASR systems in the "machinegun," "babble," and "pink" noise environments are displayed in Figs. 3(a), 3(b), and 3(c), respectively. The left vertical axis is recognition accuracy. The right vertical axis is error reduction rate to System A. The horizontal axis is SNR of the input sounds.

The recognition accuracy of the **System D** at 0 dB SNR is decreased less than 10 % comparing that of SNR = ∞ dB in all noisy environments. These results indicate that the proposed method was able to recognize the target sound regardless of the feature of noise. The reason for obtaining these results is that the proposed method does not have any noise model. Therefore, the results prove that the proposed method is able to recognize the target sound in various noises. The results show that error reduction rate of the **System D** is between around 50 % and 70 % in all noisy environments. The **System D** is more robust than other ASR systems from SNR = ∞ dB to 0 dB. The **System D** recognized the target sound in the situation where the other ASR systems were not able to recognize the target sound well. Since the distance between the input sound and the template increases with decreasing SNR, the recognition accuracy of other ASR systems only using the distance as a measure of recognition decreases with decreasing SNR. On the other hand, the proposed method verifies the possibility of the existence of the target sound by evaluating the selective sound segregation using the hypothesis of the target sound as well as the distance. Therefore, the robustness of the **System D** lies on the fact that the proposed method evaluates the selective sound segregation using validity of



(a) Results in machine gun noise.

(b) Results in speech babble noise.

(c) Results in pink noise.

Figure 3: Results in various noisy environments.

the segregation process and the segregation results. As a result, the proposed method is useful for a speech recognition system to work in various noisy environments.

The results show that the recognition accuracy of the **System B** was seriously influenced by noisy environments. In the "machinegun" noise environment, the recognition accuracy of the **System B** falls to about 70 %. The reason of the results is that SS barely reduce non-stationary noises. The results show that the recognition accuracy of the **System C** was seriously influenced by noisy environments. In high SNR, the error rate of the **System C** is between 10 % and 20 %, whereas the error rate of the **System C** increases seriously in 10 dB SNR or less. Since the distances among the templates become small, recognizing the target speech is a difficult task for the **System C** in low SNR.

4. Conclusion

This paper described an outline of a novel robust speech recognition method based on the selective sound segregation concept. In order to evaluate the performance of the proposed method, speech recognition experiments of recognizing five Japanese vowels and words in noisy environments were carried out. The experimental results revealed that ASR system based on the proposed concept is more robust than other ASR systems. Therefore, the proposed concept verifying the possibility of the existence of the target sound, without any noise model, is useful for a speech recognition system in noisy environments.

Acknowledgments

This work was supported by a Grant-in-Aid for Science Research (No. 17650047) by the Ministry of Education, Culture, Sports, Science, and Technology in Japan.

References

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, and Signal processing, vol.27, pp.113-120, Apr. 1979.

[2] M.J.F. Gales, and S.J. Young, "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech and Audio Processing, vol.4, no.5, pp.352-359, Sept. 1996.

[3] F. Martin, K. Shikano, Y. Minami, and Y. Okabe, "Recognition of noisy speech by composition of hidden markov models," IEICE Technical Report, vol.SP92-96, pp.9-16, Dec. 1992.

[4] 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.

[5] A.S. Bregman, Auditory Scene Analysis : The Perceptual Organization of Sound, MIT Press, Cambridge, MA., 1990.

[6] A.S. Bregman, "Auditory scene analysis : hearing in complex environments," in Thinking in Sound: The Cognitive Psychology of Human Audition, eds. S. McAdams, and E. Bigand, chapter 2, pp.10-36, Oxford University Press, 1993.

[7] M. Unoki, M. Kubo, A. Haniu, and M. Akagi, "A model-concept of the selective sound segregation: –a prototype model for selective segregation of target instrument sound from the mixed sound of various instrument –," Jornal of Signal Processing, vol.10, no.4, pp.407-417, Apr. 2006.

[8] A. Haniu, M. Unoki, and M. Akagi, "A study on a speech recognition method based on the selective sound segregation in noisy environment," 2005 RISP International Workshop on Nonlinear Circuits and Signal Processing, pp.403-406, Hawaii, USA, March 2005.

[9] A. Haniu, M. Unoki, and M. Akagi, "A study on a speech recognition method based on the selective sound segregation in noisy environment," JCA2007, pp.P-2-10, June 2007.

[10] M. Unoki, and M. Akagi, "A method of signal extraction from noise-added signal," Electronics and communications in Japan, Part 3, vol.80, no.11, pp.1-11, Nov. 1997, Translated into English from IEICE Trans. Fundamentals (Japanese Edition), Vol. J80-A, No.3. March 1997, pp. 444–453.

[11] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, N.J., 1976.

[12] H. Sakoe, and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust., Speech, and Signal processing, vol.26, no.1, pp.43-49, Feb. 1978.

[13] S. Itabashi, "A noise database and japanese common speech data corpus," The Journal of the Acoustical Society of Japan (Japanese Edition), vol.47, no.12, pp.951-953, Dec. 1991.

[14] A. Varga, and H.J.M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, vol.12, no.3, pp.247-251, July 1993.