Speech Recognition in a Home Environment Using Parallel Decoding with GMM-Based Noise Modeling

Kohei Machida, Takashi Nose and Akinori Ito
Graduate School of Engineering, Tohoku University
E-mail: {machida@spcom.ecei, tnose@m, aito@spcom.ecei}.tohoku.ac.jp

Abstract—In this paper, we propose a method for noise-robust speech recognition in a home environment based on noise modeling and parallel decoding. There are three basic ideas of the proposed method. First, we model the noise signals observed in the environment using a GMM. Second, we generate multiple noise-reduced signals using the mean vectors of the GMM and decode the signals in parallel. Third, we choose the best recognition result from the multiple recognition results based on the confidence score. The proposed method is very simple and straightforward, yet effective compared with simple noise reduction. The experiments proved that the proposed method is effective for not only noise signals in the database but also for those in the real home environment.

Index Terms: Speech recognition in noise, Noise modeling, FBANK, Gaussian Mixture Model, Confidence measure

I. INTRODUCTION

Development of speech recognition technology has expanded the application field of speech recognition to the home automation or smart home, such as controlling home appliances with voice command [1], [2], [3]. Speech command is a very easy way of controlling complicated appliances, as long as the speech recognition rate is high enough. However, there are many noise sources in a home environment, which degrades the speech recognition performance [4]. Therefore, noise-robust speech recognition techniques are indispensable for realizing speech recognition in a home environment.

A large number of noise-robust speech recognition methods have been proposed so far [5]. Those methods include noise reduction using a microphone array [6], [7], front-end noise reduction methods such as the spectral subtraction [8] or the Wiener filter [9], and model-based methods such as parallel model combination [10], [11].

Noise-robust speech recognition methods for home appliances, especially that based on an embedded speech recognizer, should have several features.

1) The methods should be realized at low cost. If we use a microphone array, we need not only multiple microphones but also an expensive multi-channel A/D converter. Therefore, a single-channel method is desirable.
2) The methods should be realized at low computational cost because it should be implemented into an embedded speech recognizer. Therefore, computationally-expensive methods such as noise reduction based on nonnegative matrix factorization (NMF) [12], [13] will not be suitable for this purpose.
3) The noise in a home environment is non-stationary, such as humans’ talking voice, sounds from a TV, sounds from a vacuum cleaner or footstep sounds. Therefore, the methods that assume stationary noise, such as the spectral subtraction or the Wiener filter, will be inadequate for this purpose.

Considering these conditions, we propose a new noise-robust speech recognition method. The proposed method is based on three basic ideas: noise modeling using a Gaussian mixture model (GMM), parallel decoding of the noise-reduced speech signals, and combination of multiple recognition results. In this paper, we first describe the basic ideas of the proposed method, and then we show the results of the recognition experiments.

II. NOISE MODELING AND PARALLEL DECODING

A. Overview

As mentioned above, many kinds of non-stationary noises will be observed in a home environment. However, kinds of noises can be limited when the recognizer is installed in a specific position (inside a remote controller of a TV set, for example). Therefore, we can train a model that expresses noise signals observed in that environment.

The overview of the proposed method is as follows. We first train a GMM of the environmental noise, and then the mean vectors of the mixture components of the GMM are used as the candidates of noise spectra to be reduced. Using the noise spectra, we obtain multiple candidates of noise-reduced speech signals using the spectral subtraction. These signals are decoded individually, and the recognition results are combined to choose the final recognition result. In the later experiment, we carried out isolated word recognition.

This idea modeling noise signals using a GMM is similar to the concept of parallel model combination (PMC) [10], [11]. The PMC creates a noise HMM/GMM and combines it with the clean speech HMM to generate an HMM of noisy speech. If the noise signals are modeled by an \( n \)-mixture GMM and each state of the clean speech HMM has \( m \) distributions, each state of the generated noisy speech HMM has \( nm \) distributions. Using a large number of mixture densities, we can follow the non-stationary noise in the decoding process.
although the computational cost increases. Our method also uses the noise GMM, but our method is feature-based, where the noise GMM is used to reduce noise, and only the HMM for clean speech is used for recognition. The other difference is the criterion by which the best noise candidate is selected. In the PMC, selection of the best noise is tightly coupled with the decoding process, which means that the noise candidate is selected according to the maximum likelihood criterion. However, as the later experiment shows, the ML criterion does not necessarily give the best performance. As our method chooses the best result from the results of the decoder outputs, we can exploit various criteria for the selection.

The idea of combining different features for robust speech recognition was also proposed by Zhang et al. [14]. They combined several acoustic features such as MFCC, PLP and RASTA using the dynamic Bayesian network. Compared with their work, our framework is much simpler and we can use a standard decoder for recognition.

### B. Noise modeling and reduction

In this work, we used FBANK coefficients [15] as a feature of the noise model. We can use either FBANK or MFCC as the feature vector for modeling the noise signals. We chose FBANK rather than MFCC because we wanted to model the noise signal including its magnitude.

When the recognition system is installed, we assume that the system first observes the environmental noise and trains the noise GMM. When the number of mixture component is \( M \), we obtain \( M \) mean vectors as representative spectra of the noise observed at that environment.

When we recognize the input speech, we first calculate the noise-reduced speech signals using the mean vectors of the GMM and spectral subtraction. Let the \( m \)-th mean vector of the GMM be \( \hat{N}_m(l) \) and the FBANK coefficients of the \( t \)-th frame be \( F(l,t) \), where \( l \) denotes the channel of FBANK. Then the \( m \)-th candidate of noise-reduced speech is calculated as

\[
\hat{S}_m(l,t) = \log(SS(\exp(F(l,t)), \exp(\hat{N}_m(l)))) 
\]

\[
SS(X,N) = \begin{cases} 
X - \alpha N & \text{if } X > \alpha N \\
\beta N & \text{otherwise} 
\end{cases}
\]

Here, \( SS(X,N) \) is a function of spectral subtraction for the observed spectrum \( S \) and the noise spectrum \( N \), where \( \alpha > 0 \) and \( 0 < \beta \ll 1 \) are the parameters. This calculation is based on subtraction between FBANK coefficients, which is known to be as effective as that between the amplitude spectra [16].

### C. Decoding the signal and combining the results

After calculating the noise-reduced FBANK coefficients for all mean vectors of the GMM, we calculate the MFCC from the FBANK coefficients, and the speech signals are decoded individually. After calculating the results, we determine the final recognition result from the results of the speech recognizers. In the later experiment, we tested three criteria: the result with the highest score (i.e. ML criterion), the result with the highest confidence score [17], and the majority vote. When using the majority vote and more than one result obtained the largest number of votes, the final result was determined based on the recognition score among the candidates with the largest number of votes.

### III. Experiment I

#### A. Experimental conditions

The task of the recognition was isolated word recognition. The acoustic model was phonetic-tied-mixture triphone trained from JNAS database [18]. The evaluation data were 212-word set spoken by 60 speakers from the Tohoku University-Matsushita word speech database [19]. We used four kinds of environmental noise data from JEITA noise database [20]: Exhibition hall, Station, Factory and Computer room. Length of each of those noise signal was 30 minutes. The first 20 minutes were used for training, and the signals extracted from the last 10 minutes were added to the clean speech. The average signal-to-noise ratio (SNR) values at the four environments are shown in Table I. Other conditions are shown in Table II.

#### B. Recognition result

Figure 1 shows the average recognition results under the four noise environments. In this figure, “Score” is the result when the final results were chosen based on the recognition score, “CMScore” is that based on the confidence score, “Majority” is that based on the majority vote, “Just before” is the result when the noise part just before the speech was used as the noise signal for the spectral subtraction, and “Max” is the result when one of the mean vectors of the noise GMM that gave the best result was chosen \textit{a posteriori}. Note that, in “Max” condition, one best mean vector for all utterances was chosen, while the mean vector was chosen utterance by utterance in “Score”, “CMScore” and “Majority” conditions. The recognition rate without noise reduction was 64.6%.

These results show that the method based on the confidence measure (“CMScore” condition) gave the best recognition rate.
Not only “CMScore” condition, but also other two methods (“Score” and “Majority”) surpassed the conventional spectral subtraction method (“Just before” condition), which suggests the importance of estimating the noise actually superimposed on the speech, rather than using the silent part just before the speech signal expecting that the same noise signal continues until the end of the utterance.

C. Effect of number of mixture

Figure 2 shows relationships between the number of mixture of the noise GMM and recognition rate. In this experiment, we changed the number of mixture of the noise GMM from 1 to 64. We used the confidence score for choosing the final result. When the number of mixture was one, we approximated all of the noise signal under that environment by only one Gaussian distribution. We can see that only one noise distribution improved the recognition rate compared with the result without noise reduction (64.6%), and more mixture gave more improvement. Using two mixtures gave a relatively large improvement (2.2 points), and the recognition rate using more than 16 mixtures did not improve anymore. Therefore, we can say that we should use at least two mixtures for the noise GMM, and more number of mixtures up to 16 will gradually improve the recognition rate. Because increasing the number of mixture also increases the computational cost, we need to determine the appropriate number of mixture according to the requirement of recognition speed.

IV. EXPERIMENT 2: RECOGNITION IN A REAL HOME ENVIRONMENT

Next, we recorded the environmental noise sound in a real home environment, and carried out an experiment to model the noise sound observed in the home. The sound was recorded in a real home where a family (parents and a junior-high-school kid) lives. The sounds were recorded in a living room. Figure 3 shows the overview of the room. We installed two microphones at the east and west side of the living room (Microphone (E) and (W) in Fig. 3, respectively), and only the sounds recorded by Microphone (W) were used in the experiment. The sounds were recorded using Marantz PMD660 in 16 bit linear format, mono, 48kHz sampling, on 30th December, 2013. The length of the recorded sounds was 12 hours. 10 environmental noise signals were extracted from the beginning of every one hour of the recorded signal. All the other part of the signal was used for training of the GMM. Then we mixed the speech signal of the word speech used in the previous section with the 10 environmental noise signals. The average SNR of the mixed signal was 9.12 dB.

Using the mixed word speech signals and the noise GMM, we carried out an experiment to recognize those speech signals. The other conditions of the experiment were the same as that shown in Table II. Figure 4 shows the recognition result. This result shows that the simple spectral subtraction (“Just before” in the figure) does not improve the recognition rate. There could be two reasons for this result: first, the noise level in this experiment was lower than the previous experiment (thus the SNR was higher), and then the baseline accuracy was much higher than that in the previous section. The second reason is that almost all noises in this environment were non-stationary, which made the simple spectral subtraction inefficient. The proposed method gave an improvement of
almost three points, showing the effectiveness of the proposed method even in a real home environment.

V. DISCUSSION

As shown, the proposed method outperformed the simple SS with very simple framework. One problem with the proposed method is that the method is applied to only the isolated word recognition task, which does not seem to be sufficient for a real task. Here, we believe that it is not difficult for the proposed method to apply to the continuous speech recognition task using a method of multiple candidate combination, such as ROVER [21] or the consensus network [22]. In the proposed framework, one candidate input stream copes with only one kind of noise, and therefore the method cannot switch the noise within one stream even when the noise changes within a sentence. However, we can expect that noise candidates of some of the other input streams are similar to the noise superimposed to the current speech signal, as long as the noise signal observed at the environment are appropriately modeled by the noise GMM. Therefore, we expect that one of the recognition candidates gives correct recognition results with high confidence score as long as the noise does not change much within a word, which makes possible to improve the recognition result via the recognition result combination such as ROVER.

VI. CONCLUSION

In this paper, we propose a method for noise-robust speech recognition in a home environment based on noise modeling and parallel decoding. The basic ideas of the proposed method are (1) to model the noise signals observed in the environment using a GMM, (2) to generate multiple noise-reduced signals using the mean vectors of the GMM and decode the signals in parallel, (3) to obtain multiple recognition results and combine them to choose the best recognition result. The proposed method is very simple and straightforward, yet effective compared with simple noise reduction. The experiments proved that the proposed method is effective for not only noise signals in the database but also that in the real home environment.

VII. ACKNOWLEDGMENT

Part of this work is supported by NEC Communication Systems, Ltd.

REFERENCES