

論文 / 著書情報
Article / Book Information

Title(English)	Comparative Study on Robust Speech Recognition against Nonstationary Noise in the Home Environment
Authors(English)	Agnieszka Betkowska, Koichi Shinoda, Sadaoki Furui
Citation(English)	Proc. Symposium on Large-Scale Knowledge Resources(LKR2007), Vol. , No. , pp. 175-178
発行日 / Pub. date	2007, 3

Comparative Study on Robust Speech Recognition against Nonstationary Noise in the Home Environment

Agnieszka Betkowska, Koichi Shinoda, and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology

{agabet, furui}@furui.cs.titech.ac.jp, shinoda@cs.titech.ac.jp

Abstract

We focus on the problem of robust speech recognition in the presence of nonstationary sudden noise, which is very likely to happen in the home environment. As a model compensation method for this problem, we investigated the use of Parallel Model Combination (PMC) architecture developed from a clean-speech Hidden Markov Model (HMM), and a sudden-noise HMM. We analyzed three problems (1) the time structure of sudden noise, (2) the position of noise in the corrupted speech, and (3) a possible choice for the gain matching term based on the relation between the noise power and the SNR. The study was carried out based on the database recorded in home environments by a personal robot PaPeRo of NEC Corporation.

1. Introduction

A great deal of effort has been devoted to developing personal robots, such as household robots, educational robots, or personal assistants, which interact with human beings in the home environment. Most of those robots are equipped with a speech recognition function because their interface should be sufficiently easy for children and elderly people to control.

While current speech recognition systems give acceptable performance in laboratory conditions, their performance decreases significantly when they are used in actual environments. The major reason for this degradation is that many different kinds of noise exist in actual environments. Developing speech recognition devices that are robust against nonstationary noise is important. There have been many studies on this topic, and they are categorized as follows: speech enhancement, missing data theory, and model compensation.

Speech enhancement aims at suppressing noise in the speech signal with the risk of degrading the original clean signal. Spectral subtraction, filtering techniques, and mapping transformation [1] belong to this category. They are known to be effective when the noise is stationary, but their performance degrades significantly for nonstationary noise.

Missing data theory tries to determine the level of reliability of each spectral region in the speech spectrogram [2], assuming that some portions of the speech spectrum are not contaminated by noise. However, this approach is effective only for noise that selectively corrupts a small portion of the signal spectrum.

Model compensation methods use noise models and combine them with speech models during the recognition process. One example is the well-known HMM composition and decomposition method [3], which can deal with nonstationary noise, but it is computationally expensive. A simplified version of HMM composi-

tion and decomposition is the parallel model combination (PMC) approach [4]. Although computationally less expensive, the gain matching term, which determines the signal-to-noise ratio (SNR), must be manually chosen.

We present a comparative study on robust speech recognition against nonstationary sudden noise, which is very likely to happen in home environments. This noise appears suddenly and lasts for a short time. As a model compensation method we have chosen PMC. We studied three problems: (1) if time structure of sudden noise is informative during recognition process, (2) how the recognition process is influenced by the position of noise in the corrupted speech, and (3) we also investigated a possible choice for the gain matching term based on the relation between the noise power and the SNR. These problems were analyzed by several experiments on the database recorded in home environment by personal robot PaPeRo (NEC Corp).

2. Robust speech recognition using PMC

2.1. Parallel model combination

The objective of PMC is to recognize noisy speech (speech contaminated by noise) using the combination of clean speech and noise models. Clean speech and noise models trained in cepstral domain can be combined in the linear or log-spectral domain by using the mismatch function [4]:

$$O_i^l(t) = \log \left(e^{(S_i^l(t))} + g e^{(N_i^l(t))} \right) \quad (1)$$

where $O_i^l(t)$, $S_i^l(t)$, $N_i^l(t)$ represents the i th element of the observation vector in the log-spectral domain at time t for noisy speech, clean speech, and noise, respectively. The parameter g is the gain matching term. The mismatch function for the i th delta element of the noisy speech observation, $\Delta O_i^l(t)$, is defined as:

$$\begin{aligned} \Delta O_i^l(t) = & \log \left(e^{\Delta S_i^l(t) + S_i^l(t-\tau)} + e^{\Delta N_i^l(t) + N_i^l(t-\tau) + \log(g)} \right) \\ & - \log \left(e^{S_i^l(t-\tau)} + e^{N_i^l(t-\tau) + \log(g)} \right), \end{aligned} \quad (2)$$

where the parameter τ defines the time shift.

2.2. Parameter estimation

2.2.1. Output probability density function estimation

In order to calculate the parameters of the noisy speech HMMs, an expectation of the mismatch function should be calculated. As

there is no closed-form solution for this problem, a log-max approximation can be applied [6].

Let $\hat{\mu}^{\Delta l}$, $\mu^{\Delta l}$, and $\tilde{\mu}^{\Delta l}$ be the 2P-dimensional mean of mixture component in corrupted speech HMM, clean speech HMM and noise HMM, respectively, in the log-spectral domain.

$$\begin{aligned}\hat{\mu}^{\Delta l} &= [\hat{\mu}_1^l, \hat{\mu}_2^l, \dots, \hat{\mu}_P^l, \Delta \hat{\mu}_1^l, \dots, \Delta \hat{\mu}_P^l] \\ \mu^{\Delta l} &= [\mu_1^l, \mu_2^l, \dots, \mu_P^l, \Delta \mu_1^l, \dots, \Delta \mu_P^l] \\ \tilde{\mu}^{\Delta l} &= [\tilde{\mu}_1^l, \tilde{\mu}_2^l, \dots, \tilde{\mu}_P^l, \Delta \tilde{\mu}_1^l, \dots, \Delta \tilde{\mu}_P^l]\end{aligned}$$

Then, the static mean of the corrupted speech HMM is given by [6]

$$\hat{\mu}_i^l = \log(e^{\mu_i^l} + g e^{\tilde{\mu}_i^l}). \quad (3)$$

Assuming stationarity, the delta mean is given by

$$\begin{aligned}\Delta \hat{\mu}_i^l &= \log \left(e^{\Delta \mu_i^l + \mu_i^l} + e^{\Delta \tilde{\mu}_i^l + \tilde{\mu}_i^l + \log(g)} \right) \\ &\quad - \log \left(e^{\mu_i^l} + e^{\tilde{\mu}_i^l + \log(g)} \right).\end{aligned} \quad (4)$$

2.2.2. Transition matrix estimation

A PMC built from clean speech HMM S (with K states) and noise HMM N (with Z states) can be represented by a traditional HMM with $K^2 \times Z^2$ states. Its transition matrix is defined by the Cartesian product between the transition matrices A_S and A_N of HMMs S and N , respectively [7]:

$$a_{(i,j)(k,l)} = a_{ik}^S a_{jl}^N, \quad 1 \leq i, k \leq K, \quad 1 \leq j, l \leq Z. \quad (5)$$

3. Problems in the home environment

There are several problems we encounter in home environments. Firstly, the starting point and the duration of nonstationary noise is difficult to be predicted. Even the same type of noise can last for a different amount of time. As we will show later, the starting point of nonstationary noise can have significant impact on the recognition accuracy. Secondly, more than one type of noise can appear at a time. Therefore robust speech recognition must be able to deal with different combinations of noises. Thirdly, the SNR of corrupted speech samples is not constant, as it is a function of speaker, robot and noise source positions. Therefore a question is how to alleviate the difference between the SNR of the training and testing data, and how to define SNRs for these sets if each noise and speech sample pair has different SNR. In this paper we mainly address the following issues: (1) the time structure of sudden noise, (2) the position of noise in the corrupted speech, and (3) a possible choice for the gain matching term based on the relation between the noise power and the SNR.

In order to verify if the time structure is informative during recognition, we tested different model structures. We fixed all the parameters of HMMs except for those responsible for time modeling. If the time structure is relevant, the PMC based on the model that properly represents the time structure should yield better performance than the PMC based on HMM with no ability of time structure modeling.

In order to define a possible choice for the gain matching term, we observed in Eq. (1) that the gain matching term should reflect the relationship between noise power in the training set and noise power in the testing set. The most natural way is to follow the

same process as during the synthesis of the clean speech and noise samples into the noisy speech samples at the desired SNR. During this process the noise sample is multiplied by the amplification factor. We calculate the gain matching term in similar manner to the amplification factor. However, the calculation is done over the averaged power for speech and noise samples in the training set [$P(\text{speech})$ and $P(\text{noise})$]. Therefore, the gain matching parameter g for the given testing set is defined as:

$$g = \left(10^{-\frac{u}{10}} \right) \frac{P(\text{speech})}{P(\text{noise})}, \quad (6)$$

where u is the expected SNR for the testing set.

4. Experiments

4.1. Experimental conditions

For our studies, we used a database recorded by a personal robot called PaPeRo developed by NEC Corporation [5], which was used in the houses of 12 Japanese families (F01-F12). The database contains 74,640 sounds, each of which was detected by the speech detection algorithm equipped in PaPeRo. These sounds were classified manually into three categories: *clean speech* (speech without noise), *speech corrupted by noise*, and *noise* (noise without speech).

In this study, we used 16,000 samples of clean speech, and 480 recordings of sudden noise such as doors slamming, knocking, and falling objects. The statistics for each family are shown in Figure 1 and Figure 2. Samples were digitized at the 11,025 Hz sampling rate, and analyzed at a 10 msec frame period. Mel frequency cepstral coefficient parameters consisting of 12 static features and 12 Δ features were used as the input features in each frame. We developed a system for recognizing isolated Japanese words. The vocabulary contains 1492 entries, consisting of words and simple phrases (for simplicity we treated each phrase as a word).

The samples from eight families (F02-F06, F08, F09, and F11) were used for training the HMMs of clean speech and sudden noise. The test set was prepared as follows. From each of the remaining 4 families, all samples of sudden noise and 137 samples of clean speech were taken. Then, each clean speech sample was paired with a sudden noise sample that was selected randomly from the noise samples in the remaining 4 families. Next, the paired speech and noise samples were mixed at different SNRs: $-5, 0, 5, 10$, and 20 dB.

To achieve the desired SNR for each pair of speech and noise samples, the power of speech and noise was calculated as follows. Let $w(i)$ be the power in the i th frame of the signal s . In addition, let $C := \{i | w(i) \geq \lambda\}$, i.e., C is the set of indices in which the i th frame has power greater than or equal to threshold λ . The power of signal s is defined by

$$P(s) = \frac{\sum_{i \in C} w(i)}{W}, \quad (7)$$

where W is the number of frames in set C . For our experiments, we set the threshold $\lambda = 400$ for speech, and $\lambda = 50$ for noise. These values were optimized in our preliminary experiments. A clean speech sample and a sudden noise sample were synthesized in such a way that the center point of these two samples was located at the same point (*middle position*). Additionally two more

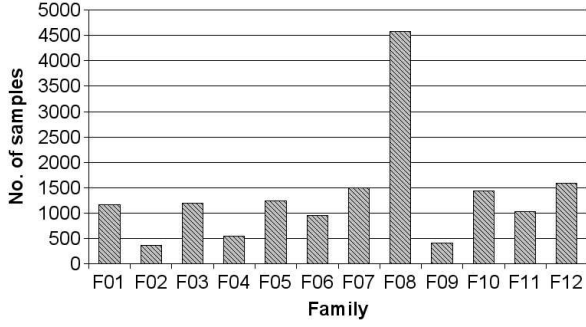


Figure 1: Number of clean speech samples used in our experiments

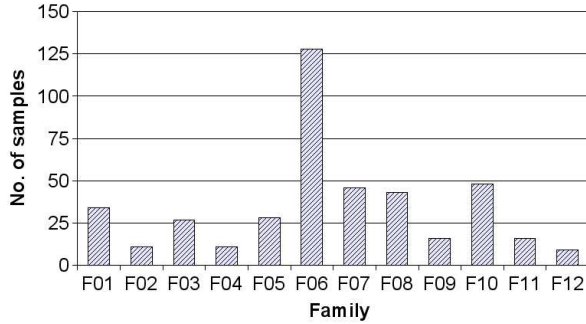


Figure 2: Number of sudden noise samples used in our experiments

sets were created where the center point of the noise sample was shifted in relation to the center point of the speech sample by -0.36 sec (*frontal position*) and by $+0.36$ sec (*backward position*). An evaluation test with 548 utterances at each SNR was prepared for each noise position.

We created a PMC as follows. First, we constructed clean-speech HMMs and an HMM for sudden noise. The recognition units in clean-speech HMMs were triphones, which were trained using clean-speech data. An HMM for sudden noise was trained using sudden noise samples. Then, for each entry in the vocabulary, a word HMM was designed by concatenating the states of the silence HMM and triphone HMMs according to their corresponding sequence in the given entry. A noise HMM [which consists of three states of silence, different number of states of sudden noise (ranging from one to three), followed by additional three states of silence] was built in similar manner. The state output *pdf* for all states of speech HMMs was a single Gaussian distribution. Finally, a PMC that models speech and noise in parallel for a given word was created by combining the word HMM for clean speech and the noise HMM.

4.2. Time structure of sudden noise

We investigated three different sudden noise model structures. An HMM with one state with a single Gaussian distribution is the simplest possible model, therefore a PMC built from this model should give the worst accuracy. Experimentally, we found out that the optimal number of states for sudden noise HMM is three. Therefore, the second HMM had three states and one Gaussian per state. As

Table 1: Gain matching term for different SNRs.

SNR (dB)	g
-5	15.84
0	5.01
5	1.58
10	0.46
20	0.05

we were interested in time structure of noise and its importance in the PMC architecture, the third model had the same total number of Gaussians as the second one, but differed in the time structure modeling. It had one state and three Gaussians, and, unlike the second model, it ignored the time structure of the noise.

From these noise HMMs and clean speech HMMs, three different PMC models were created and used for recognition of the *middle position* test set. The results are shown in Figure 3. For almost all SNR, the model with three mixtures and one state performed better than model with three states and with one mixture per state. The latter model gave slightly higher accuracy only for SNR 0 dB. The same experiment was conducted for *frontal position* and *backward position* test sets and similar results were obtained. These results did not follow our expectation that the time structure of noise is important. This might be due to the fact that the sudden noise model parameters might not be estimated correctly, since the number of sudden noise samples was limited. Therefore experiments with more reliable database are needed.

4.3. The noise position in corrupted speech

Next, we investigated how the recognition results are influenced by the position of noise in the corrupted speech. Based on the results from the last section, we decided to use a PMC built from a noise HMM with one state and three Gaussians per state. We performed the recognition on three test sets: *frontal position*, *middle position* and *backward position* test set. The results are shown in Figure 4. The worst accuracy was achieved when the noise sample was in the middle of the speech sample. The best accuracy was achieved when noise appeared at the end of the speech. On average, the difference in recognition performance between these two sets was more than 10%. In low SNR this dissimilarity was even bigger, reaching about 20% at -5 dB. However at 20 dB, PMC gave higher recognition score for *middle position* test set. It can be explained by the fact that the noise is masked by speech and therefore cannot be observed in high SNR. However, when noise appears at the beginning or at the ending of the speech signal, it is partially combined with silence and therefore it can still cause distortions in the recognition process. Interestingly, there is a noticeable difference in the recognition results for *frontal position* and *backward position* test sets. The results indicated that noise that occurred at the beginning of speech causes more confusion than the noise that occurred at the end. It might be due to the fact that a Japanese word mostly starts with a consonant and ends with a vowel, and consonant recognition performance might be more influenced by the sudden noise than that of vowels.

4.4. Gain matching term

We investigated the value for the gain matching term that is used to alleviate the mismatch between the SNR of the training set and the SNR of the testing set. We calculated the gain matching term for different SNR according to Eq. (6). The averaged SNR for the

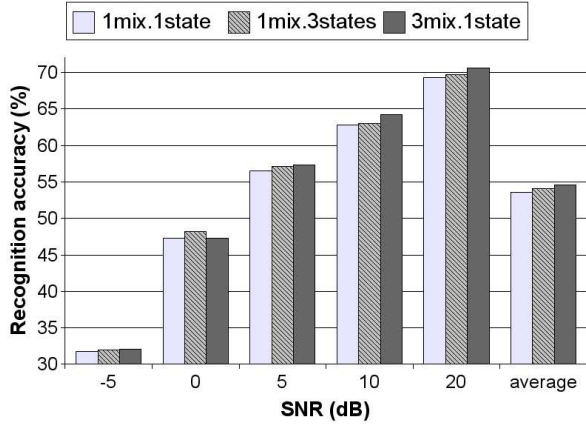


Figure 3: Recognition results given by three different PMC that model differently time structure of sudden noise

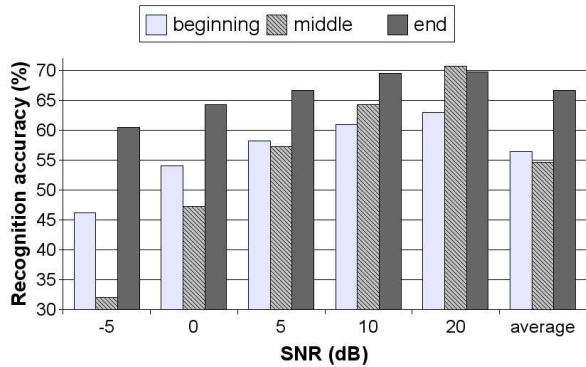


Figure 4: Recognition results for different noise position in the corrupted speech

training set was 7dB. Table 1 shows the value of g for different SNRs.

For each SNR a PMC model was created with the corresponding value of the gain matching term. These models were compared with a PMC with constant gain matching term $g = 1$, which assumes that no mismatch of the test and train set SNR occurs. The results are given in Table 2. The gain matching term improved recognition accuracy in high SNRs. Applying $g < 1$ for higher SNR ($\text{SNR} > 7$ dB) is equivalent to shifting the PMC model parameters towards those of the clean speech HMM. Such model is more likely to recognize almost clean speech than noisy speech PMC (with $g = 1$). On the other hand, PMC models with the estimated g did not improve accuracy in low SNRs. The results of our experiments in low SNR can be influenced by the fact that the SNR varied significantly between each clean speech and noise speech samples in the training set. Therefore the gain matching term g calculated over their averaged powers may not reflect correctly the difference between the training and testing sets.

5. Conclusion and future work

We presented a comparative study on robust speech recognition against sudden noise in the home environment. Firstly, for sudden noise time structure, more studies with larger database are needed. Secondly, we found out that the position of the sudden noise influ-

Table 2: The recognition accuracy (%) under different SNRs

Test set	-5 dB	0 dB	5 dB	10 dB	20 dB
<i>frontal</i>					
g constant	48.3	56.0	60.6	63.9	66.4
g matched	32.1	51.4	59.7	64.0	68.4
<i>middle</i>					
g constant	32.1	47.3	57.3	64.2	70.6
g matched	25.5	48.3	56.6	65.5	73.5
<i>backward</i>					
g constant	60.4	64.2	66.6	69.5	69.7
g matched	58.2	61.9	64.4	67.3	67.3

ences significantly the recognition performance in our experiments using Japanese. In the future, more experiments using different languages should be performed to exam language dependency for this problem. Finally, our suggested adapting gain matching term was effective in high SNRs. However, when it is difficult to define the SNR for the training set, it may be better to keep the gain matching term constant during the calculation of the PMC parameters, as in low SNRs it did not improve recognition performance. This study was conduct only for sudden noise. In the future, different kinds of noise and their combination should be taken into account.

. Acknowledgments

This work is supported by 21th Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources". We thank NEC Corporation for the permission to use PaPeRo database.

. References

- [1] X. Huang, A. Acero, and H. Hon, "Spoken language processing: a guide to theory algorithm and system development," Prince-Hall, 2001.
- [2] M. Cook, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Communication, vol. 34, pp. 267-285, 2001.
- [3] A. P. Varga, and R. E. Moore, "Hidden Markov model decomposition of speech and noise," in Proc. ICASSP, pp.845-848, 1990.
- [4] M.J.F Gales and S.J. Young, "HMM Recognition in Noise Using Parallel Model Combination," in Proc. EuroSpeech, pp.837-840, Berlin, 1993.
- [5] T. Iwasawa, S. Ohnaka, and Y. Fujita, "A Speech Recognition Interface for Robots using Notification of III-Suited Conditions," in Proc. of the 16th Meeting of Special Interest Group on AI Challenges, pp. 33-38, 2002.
- [6] S. G. Pettersen, M. H. Johnsen, and T. A. Myrvoll, "Joint Bayesian Predictive Classification and Parallel Model Combination for Robust Speech Recognition," in Proc Eurospeech, pp. 373-376, 2005.
- [7] B. Logan, and P. Moreno, "Factorial HMMs for Acoustic Modeling," in Proc. ICASSP, pp. 813-816, 1998.