Noise Suppression Using Search Strategy with Multi-Model Compositions

Takatoshi Jitsuhiro[1], *Tomoji Toriyama*[1], *Kiyoshi Kogure*[1]

[1]ATR Knowledge Science Labs.
2-2-2 Hikaridai, "Keihanna Science City," Kyoto, 619-0288, Japan
{takatoshi.jitsuhiro, toriyama, kogure}@atr.jp

10.21437/Interspeech.2007-108

Abstract

We introduce a new noise suppression method by using a search strategy with multi-model compositions that includes the following models: speech, noise, and their composites. Before noise suppression, a beam search is performed to find the best sequences of these models using noise acoustic models, noise-label n-gram models, and a noise-label lexicon. Noise suppression is frame-synchronously performed by the multiple models selected by the search. We evaluated this method using the E-Nightingale task, which contains voice memoranda spoken by nurses during actual work at hospitals. For this difficult task, the proposed method obtained a 21.6% error reduction rate.
Index Terms: speech recognition, noise suppression, model composition, multi-pass search, E-Nightingale project

1. Introduction

We have been working on the E-Nightingale Project to establish the fundamental technology for a knowledge sharing system based on understanding everyday activities and situations[1][2]. It focuses on the medical care domain. We have been collecting voice memoranda recorded by nurses about their actions while working and analyzing activities[3]. Recently, we started to evaluate the performance of speech recognition for these voice memoranda. However, recognizing them is difficult because they are very noisy spontaneous speech that includes many kinds of noise signals and other voices. These data also include general problems of speech recognition in real environments. In this paper, we attempted to suppress such noise signals to obtain better recognized results.

Many noise suppression methods have been proposed to improve the performance of speech recognition for noisy speech. For stationary noise signals, Spectral Subtraction[4] and Parallel Model Combination[5] have been proposed. The Gaussian Mixture Model (GMM) based Minimum Mean-Squared Error (MMSE) method[6] assumes that input noise is stationary but fluctuating. Recently, noise suppression research focuses on non-stationary noise signals, e.g., [7]. Since these methods usually assume that only one kind of noise signal exists, applying them to noisy speech including many kinds of noise signals is difficult. In general, not only stationary noise signals but also accidental noise signals occur in real environments Furthermore, another important problem still remains: obtaining the actual noise signals from input signals.

We propose a noise suppression method by searching for the best multi-label paths using multi-model compositions. First, we consider that a noise suppression process resembles a search process because this process needs to find speech and noise intervals. To obtain time alignments of utterances and noise signals from noisy speech data, we apply a beam search by using acoustic models for speech and noise signals, noise-label n-gram models, and a noise-label lexicon. The most important problem is estimating intervals overlapped by many kinds of sources and suppressing their noise signals. To solve this problem, we make models combining many kinds of sources and use them both for the search and for noise suppression as acoustic models. There are many ideal combinations, but actual existing combinations in real environments are usually limited. If the amount of training data is large enough, and situations for using speech recognition are limited, the coverage of obtained composite models can be small even for open data. Using obtained label sequences for the search, one model for each frame is allocated, and an extension of the GMM-based MMSE method[6] for multi-model compositions can reduce noise signals even if utterances are contaminated by several noise signals.

First, in Section 2, we briefly explain our motivation, the E-Nightingale project, and its recognition task. Next, our proposed method is described in Section 3. In Section 4, we perform experiments and report results, and we conclude this paper in Section 5.

2. E-Nightingale project

Recently, medical malpractice has become a serious social problem. One aim of the E-Nightingale project is to establish a technology using wearable computers and sensor networks to support nursing services[1][2]. To analyze daily nursing activities, we recorded and collected their voice memoranda in real environments while they were working[3]. We asked them to record short sentences about each nursing event using IC recorders with small microphones attached to their chests. Headset microphones could not be used because they disturbed nurses' services. Therefore, the SNRs of recorded speech were usually less than 10 dB. Fig. 1 shows a sample of recorded speech where a nurse said, "The service adjustment meeting is finished." This sample includes a beep, a target utterance needed for our analysis, conversations with a coworker, and other persons' speech as background noise. Recognizing them is very difficult because many kinds of non-stationary noise signals are included, and the utterances are not so long, but they include many kinds of spontaneous speech, e.g., small and ambiguous voices with local accents. These data include so many general and essential speech recognition problems that we cannot solve them all at once. In this paper, we focus on noise suppression techniques.

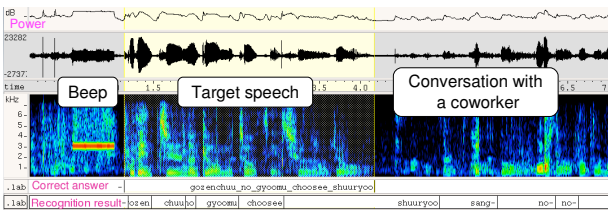

Figure 1: Wave sample including a target speech. Beep prompts speech input. Speaker talked with her coworker after recording her voice memorandum.

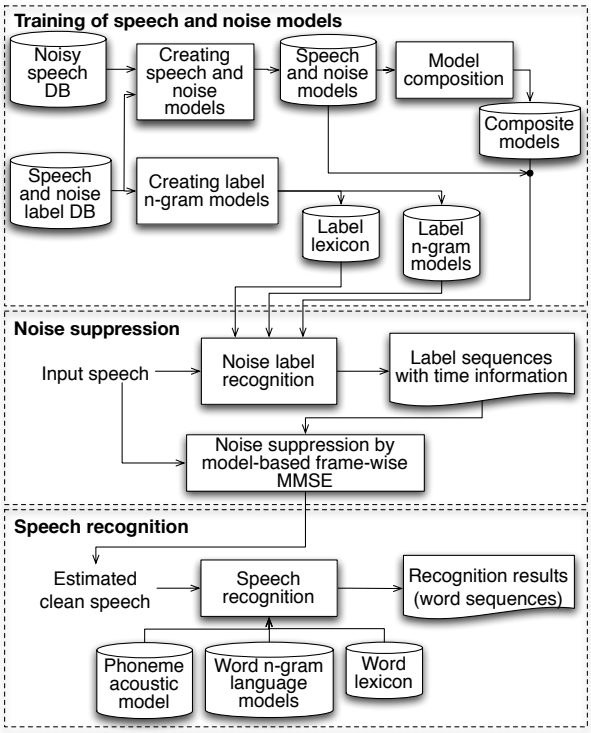

Figure 2: Overview of Multi-Model Noise Suppression

3. Multi-Model Noise Suppression

3.1. Overview

Figure 2 shows an overview of the proposed method. The process is divided into three parts: training of speech and noise models, noise suppression, and usual speech recognition. In training the speech and noise models, first, speech and noise models are trained using multi-layered noise labels that include many kinds of speech and noise labels, for example, target utterances, beep sounds, machine noises, and so on. In this paper, we used GMMs to represent them. Next, to represent overlapped noise signals, some models are combined from these trained models whose details we'll describe later. Second, a lexicon and n-gram models of these labels are generated from noise labels. Furthermore, speaker-adapted models as clean speech models for the noise suppression are trained by using these data.

In the noise suppression process, the above models, that is, the speech and noise models including clean speech models, the label lexicon, and the label n-gram models, are used in a speech recognizer to recognize speech and noise labels, identical to usual speech recognition. Instead of word sequences, sequences of speech or noise labels with time information are obtained by this search. Therefore, this process can be considered a multiple pass search. Using the recognized labels with time information, model-based frame-wise noise suppression is performed. For this approach, time alignments are needed to find which labels are allocated to frames. We extend the GMM-based MMSE method[6] to obtain estimated clean speech by multiple noise models.

Finally, standard speech recognition is performed with phoneme acoustic models, word n-gram models, and a word lexicon for the estimated clean speech, and word sequences are obtained as recognition results.

3.2. Multi-layered labels and composite models

Figure 3 shows an example of multi-layered noise labels and composite models. To consider overlapped noise signals, we

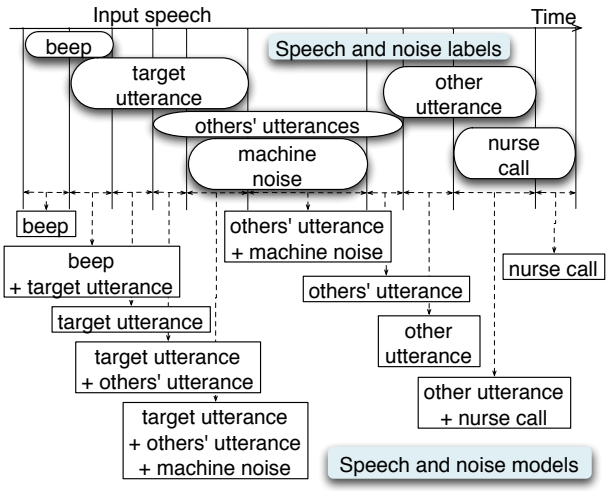

Figure 3: Example of multi-layered labels and composite models

first made each speech or noise model and then combinations among them. As shown at the bottom of Fig. 3, a multi-layered label sequence can be represented by a sequence of composite models that are generated by the combinations of mixture components of a few models in the same manner as [5]. When the best multi-label sequence is obtained by noise recognition, different kinds of noise signals can be identified for each frame, and clean speech can be estimated by GMM-based MMSE extended to plural noise models.

3.3. GMM-based noise suppression

As frame-synchronous noise suppression, GMM-based noise suppression[6][8] can be used. We extend it for multi-model compositions. Assuming that speech and many kinds of noise signals are uncorrelated, the output of the Mel-filter bank of input noisy speech is

$$X(i) = S(i) + \sum_{n=1}^{N} N_n(i), \tag{1}$$

where i is the frame index, $S(i)$ is the clean speech, $N_n(i)$ is the n-th kind of noises, and N is the number of noises. In the log Mel-spectral domain, Eq. (1) can be written as

$$\begin{aligned}\mathbf{x}(i) &= \mathbf{s}(i) \\ &\quad + \log\left[1 + \exp\left\{\log\left(\sum_{n=1}^{N}\exp\left(\mathbf{n}_n(i)\right)\right) - \mathbf{s}(i)\right\}\right] \\ &= \mathbf{s}(i) + g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)),\end{aligned} \tag{2}$$

where $g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i))$ is the mismatch factor between clean speech $\mathbf{s}(i)$ and noisy observation $\mathbf{x}(i)$. For noise compensation, we start from a MMSE estimator for log Mel-spectral energy coefficients of clean speech as follows:

$$\begin{aligned}\hat{\mathbf{s}}(i) &= E\left[\mathbf{s}(i)|\mathbf{x}(i)\right] \\ &= \mathbf{x}(i) - E\left[g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i))|\mathbf{x}(i)\right].\end{aligned} \tag{3}$$

We model the clean speech signal by GMM with K distributions as follows:

$$p(\mathbf{s}) = \sum_{k=1}^{K} w_{\mathbf{s},k}\mathcal{N}(\mathbf{s}; \mu_{\mathbf{s},k}, \mathbf{\Sigma}_{\mathbf{s},k}), \tag{4}$$

where $w_{\mathbf{s},k}$, $\mu_{\mathbf{s},k}$, and $\mathbf{\Sigma}_{\mathbf{s},k}$ are the mixture weight, the mean vector, and the covariance matrix of the k-th mixture component, respectively. In the same manner, we assume that the n-th noise signal can be modeled as

$$p(\mathbf{n}_n) = \sum_{l=1}^{L} w_{\mathbf{n}_n,l}\mathcal{N}(\mathbf{n}_n; \mu_{\mathbf{n}_n,l}, \mathbf{\Sigma}_{\mathbf{n}_n,l}), \tag{5}$$

where $w_{\mathbf{n}_n,l}$, $\mu_{\mathbf{n}_n,l}$, and $\mathbf{\Sigma}_{\mathbf{n}_n,l}$ are the mixture weight, the mean vector, and the covariance matrix of the l-th mixture component, respectively.

Using the above assumptions, Eq. (3) can be written as

$$\hat{\mathbf{s}}(i) \simeq \mathbf{x}(i) - \sum_{m=1}^{M} P(m|\mathbf{x}(i)) g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)),$$

where M is the number of mixture components generated by the model combination. Probability $P(m|\mathbf{x}(i))$ is estimated using the composite model:

$$P(m|\mathbf{x}(i)) = \frac{w_{\mathbf{x},m}\mathcal{N}(\mathbf{x}(i); \mu_{\mathbf{x},m}, \mathbf{\Sigma}_{\mathbf{x},m})}{\sum_{m'=1}^{M'} w_{\mathbf{x},m'}\mathcal{N}(\mathbf{x}(i); \mu_{\mathbf{x},m'}, \mathbf{\Sigma}_{\mathbf{x},m'})},$$

where the m-th component of the noisy signal is the model combining the k-th component of the speech and the l_{nm}-th components of several noise signals $\mathbf{N}_m$ selected from $\{\mathbf{n}_1, \ldots, \mathbf{n}_N\}$. We define its weight as $w_{\mathbf{x},m} \equiv w_{\mathbf{s},k} \cdot \prod_{n=1,\mathbf{n}_n \in \mathbf{N}_m}^{N} w_{\mathbf{n}_n,l_{nm}}$. Its mean vector and covariance matrix are estimated by applying the first order Taylor series expansion[9] as follows:

$$\mu_{\mathbf{x},m} \approx \mu_{\mathbf{s},k} + g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)), \tag{6}$$

$$\mathbf{\Sigma}_{\mathbf{x},m} \approx (\mathbf{I} + \mathbf{H}_\mathbf{s}) \mathbf{\Sigma}_{\mathbf{s},k} (\mathbf{I} + \mathbf{H}_\mathbf{s})^T + \sum_{n=1,\mathbf{n}_n \in \mathbf{N}_m}^{N} \left(\mathbf{H}_{\mathbf{n}_n,l_{nm}} \cdot \mathbf{\Sigma}_{\mathbf{n}_n,l_{nm}} \cdot \mathbf{H}_{\mathbf{n}_n,l_{nm}}^T \right), \tag{7}$$

where $\mathbf{H}_\mathbf{s}$, and $\mathbf{H}_{\mathbf{n}_n,l_{nm}}$ are diagonal matrices whose diagonal elements are $\partial g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i))/\partial \mathbf{s}$, and $\partial g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i))/\partial \mathbf{n}_{n,l_{nm}}$, respectively. However, in this paper, for simplicity, when more than two models were combined, first, two models were selected and combined. Next, another model was added to this composite model, and then, this composition was continued until all models were combined into one model. Therefore, each process of model combination is the same as that of single noise model. However, the approximation errors by Eq. (6) and (7) are accumulated to combined models.

3.4. Mismatch factor

Since this approach can detect voice activity intervals, noise suppression can be done separately for each intervals. We define the mismatch factor $g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i))$ as follows:

$$g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)) = \begin{cases} \mu_{\mathbf{x},m} - \mu_{\mathbf{s},k} & \text{for target utterances,} \\ \mu_{\mathbf{x},m} - \varepsilon & \text{for the others,} \end{cases} \tag{8}$$

where $\mu_{\mathbf{x},m}$ is composed from $\mu_{\mathbf{s},k}$ for a speech interval and ε is a small positive number that can control the power of residual signals after noise suppression.

4. Experiments

4.1. Experimental setup

The E-Nightingale data were recorded in some hospital in Osaka, Japan. The data collected the first day were used for evaluation. The length of each file was 10 sec including one target utterance. The data of the second day were used as training data to adapt the acoustic models to speakers and create noise GMMs for noise suppression. In this paper, diagonal covariance matrices were used for all distributions. These sets were recorded by IC recorders, iAUDIO G3 made by COWON Japan Inc., with a 32 kHz sampling rate and 16 bits. After recording, their sampling rates were converted to 16 kHz. The bandwidth of the microphone ranged from 100 Hz to 10 kHz. The test data included 208 utterances with 1,051 words spoken by eight speakers who were selected as the common speakers included in the test data and the adaptation data.

As a speech recognizer and training tools, we used the ATRASR large-vocabulary speech recognition system version 3.6 developed by ATR Spoken Language Communication Labs. A feature vector consists of 12 MFCCs, 12ΔMFCCs, and Δlog power extracted from frames of 20 ms with 10 ms frame shift of data recorded with 16 kHz sampling rate. Cepstral mean subtraction (CMS) was applied. Clean speech Japanese acoustic models were trained using 37-hour speech including dialogues and read speech from the ATR travel arrangement task corpus. Phoneme HMMs with 2,086 states were generated by the MDL-SSS algorithm. Since all test speakers were females in this paper, we only used a female acoustic model. To obtain the female acoustic model with five mixture components, 21-hour female speech data from the above data were used. As language models, word bigram and trigram models were used and trained using 9,936 utterances, i.e., nine-day data extracted from the E-Nightingale text database. The vocabulary contained 2,636 words and included all words of the evaluation set. For the perplexity of all of the evaluation data, 39.4 for word bigram models, and 39.3 for word trigram models were obtained, which is a small performance difference. Also, for some speakers, higher perplexity, over 100, was obtained. Therefore, this task poses many difficulties for language modeling too.

For noise suppression, speaker-independent GMM with 512 mixture components and 24-order outputs of log Mel-filter bank were trained using eight-hour dialogue speech from the above training data. All GMMs were trained by HTK version 3.3. "FBANK" was used as feature vectors. Speaker-adapted GMMs were obtained by the Maximum A Posteriori probability (MAP) estimation. For each speaker, about 200 mixture components remained after adaptation. These SI-GMM and SD-GMMs were considered models of target utterances, and their estimated intervals were utterances needed by the recognition system. The other speech and noise models were generated as GMMs with four mixture components. The number of basic noises was 32, and the number of noises in the label lexicon including multi-labels was 194. The Out of Label Vocabulary (OOLV) rate for the test set was 3%. Multi-label bigram and trigram models were used. Test set perplexity by label bigram and trigram models were 8.08 and 6.47, respectively. To obtain multi-label sequences from input noisy speech, multi-label GMMs, lexicon, and n-gram models were used by the ATRASR speech recognizer as usual speech recognition.

MAP-VFS[11] was used as the speaker-adaptation method. For the Multi-Model Noise Suppression, noise intervals become almost clean if noise recognition works well, but signals remain noisy when noise intervals cannot be estimated. Therefore, using the labels obtained by noise recognition, phoneme models and a silence model were separately trained to obtain speaker-dependent and noisy silence models.

We evaluated recognition performance for (1) baseline, i.e., without noise suppression, "Baseline," (2) Single-Model Noise Suppression (SM-NS) with speaker-independent GMM, "SM-NS (SI)," (3) SM-NS with speaker-adapted GMM, "SM-NS (SD)," (4) method (3) + speaker adaptation of acoustic model by MAP-VFS, "(3) + MAP-VFS," (5) Multi-Model Noise Suppression (MN-NS) with result labels by noise recognition and speaker-dependent GMM, "MM-NS (RLAB)," (6) MM-NS with speaker-dependent GMM and MAP-VFS, "MM-NS (RLAB) + MAP-VFS," and (7) MM-NS with manual labels and speaker-dependent GMM, and MAP-VFS, "MM-NS (MLAB) + MAP-VFS." (7) is the ideal case for (6). SM-NS means that one distribution is used to represent input noises. This distribution is estimated from 10 frames at the beginning of each input file. In MM-NS, background noise is estimated in the same manner as the noise distribution of the SM-NS.

Table 1: Average SNR

Method	SNR [dB]
(1) Baseline	8.25
(2) SM-NS (SI)	13.43
(3) SM-NS (SD)	10.24
(4) (3) + MAP-VFS	
(5) MM-NS (RLAB)	14.19
(6) MM-NS (RLAB) + MAP-VFS	
(7) MM-NS (MLAB) + MAP-VFS	57.39

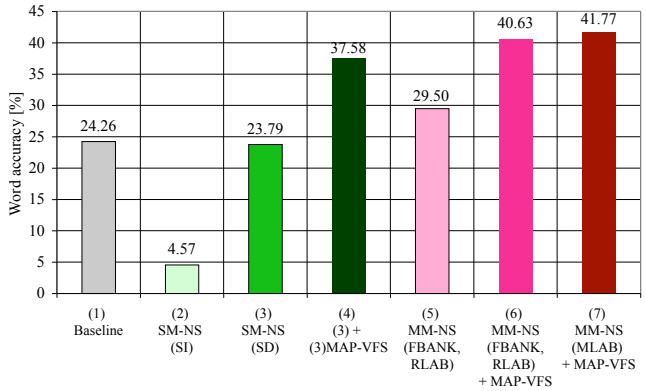

Figure 4: Word accuracy for each method

4.2. Experimental results

Table 1 shows the average SNR for each noise suppression method. To calculate these SNRs, target utterance intervals were extracted, and noise power was calculated from 500-ms intervals at the sides of speech intervals.

Since noise periods are almost clean using very small positive number for ε in Eq. (8), the obtained SNR in (7) MM-NS (MLAB) is very high. MM-NS (RLAB), (5) and (6), obtained much more noisy signals than (7) because result labels include many mistakes. However, the SNR obtained by (5)(6) MM-NS (RLAB) was higher than that obtained by conventional methods (2), (3), and (4).

We also evaluated the performance of the label accuracy (LA) rates and voice activity detection (VAD) by MM-NS search. LA is calculated in the same manner as word accuracy. LA rates for the all test sets were 32.27% for the bigram model and 33.96% for the trigram model. To evaluate VAD, according to [12], the correct VAD rate is defined as $Corr = N_c/N$, and the accuracy rate is defined as $Acc = (N_c - N_f)/N$, where N is the number of utterance intervals, N_c is the number of correct utterance intervals, and N_f is the number of false utterance intervals. Using trigram model rescoring and 200-ms margins, the correct rate was 85%, and the accuracy rate was 33%. The false alarm is large, but for speech recognition it is more preferable than deletion errors. In this paper, we did not extract detected speech intervals: we only used the decision for Eq. (8).

Figure 4 shows the word accuracy rate for each method. (1) Baseline performance was 24.26%, and the proposed method, (6) MM-NS (RLAB) + MAP-VFS, obtained 40.82%. The error reduction rate was 21.6%. The performance of method (6) is very close to that of method (7), which is the ideal case of method (6). The proposed method outperformed (4) SM-NS (SD) + MAP-VFS, whose performance was 37.58%. The error reduction rate was 4.89%. Therefore, the proposed method is more effective than the conventional method.

5. Conclusion

We proposed multi-model frame-synchronous noise suppression with a search strategy to reduce many kinds of noise signals. Applying conventional noise suppression methods to noisy speech contaminated by several kinds of noise signals is difficult. To reduce the noise signals of intervals overlapped by several kinds of sources, models combined with several models were used. To estimate the intervals of several sources included in input data, beam search was performed using speech and noise models including composite models, noise-label n-gram models, and a noise-label lexicon. Using noise-label sequences with time information obtained by the search process, the GMM-based MMSE method extended to multi-model compositions was performed noise suppression. To evaluate this method, we used the E-Nightingale task recorded in real situations and environments. Experimental results show that our proposed method is more effective than the conventional method.

6. Acknowledgements

This research was supported by the National Institute of Information and Communications Technology of Japan. We'd like to thank the nurses for their cooperation, ATR-SLC's members for their tools and advice, and ATR-KSL's members for their advice and for making our database.

7. References

[1] K. Kogure, "Toward a knowledge sharing system based on understanding everyday activities and situations – Introduction to the E-Nightingale Project –," Proc. of the Workshop on Knowledge Sharing for Everyday Life 2006 (KSEL2006), pp. 1–8, 2006.

[2] http://www.e-nightingale.org/

[3] H. Ozaku, A. Abe, K. Sagara, N. Kuwahara, K. Kogure, "A task analysis of nursing activities using spoken corpora," Advances in Natural Language Processing (Ed. A. Gelbukh), Research in Computing Science 18, pp. 125–136, Instituto Politecnico Nacional, Mexico, 2006.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol. 27, no. 27, pp. 113–120, 1979.

[5] M. F. J. Gales, "Model-based techniques for noise robust speech recognition," PhD thesis, University of Cambridge, 1995.

[6] J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," Proc. of EUROSPEECH2001, vol. 1, pp. 221–224, 2001.

[7] M. Fujimoto, S. Nakamura, "A non-stationary noise suppression method based on particle filtering and Polyak averaging," IEICE Trans. Inf. & Syst., vol. E89-D, no. 3, 2006.

[8] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, S. Nakamura, "Hands-free speech recognition and communication on PDAs using microphone array technology," Proc. of ASRU2005, pp. 302–307, 2005.

[9] P. J. Moreno, "Speech recognition in noisy environments," PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1996.

[10] T. Jitsuhiro, T. Matsui, S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," IEICE Trans. on Information and Systems, vol. E87-D, no. 8, pp. 2121–2129, 2004.

[11] M. Tonomura, T. Kosaka, S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," Computer Speech and Language, vol. 10, pp. 117–132, 1996.

[12] N. Kitaoka, et al., "Progress report of SLP noisy speech recognition evaluation WG: Individual evaluation framework for each factor affecting recognition performance," IPSJ SIG Technical Reports, vol. 2006, No. 136, 2006-SLP-64, pp. 1–6, 2006 (in Japanese).