ROBUST SPEECH RECOGNITION USING NOISE SUPPRESSION BASED ON MULTIPLE COMPOSITE MODELS AND MULTI-PASS SEARCH

Takatoshi Jitsuhiro, Tomoji Toriyama, Kiyoshi Kogure

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

ABSTRACT

This paper presents robust speech recognition using a noise suppression method based on multi-model compositions and multi-pass search. In real environments, many kinds of noise signals exists, and input speech for speech recognition systems include them. Our task in the E-Nightingale project is speech recognition of voice memoranda spoken by nurses during actual work at hospitals. To obtain good recognized candidates, suppressing many kinds of noise signals at once to find target speech is important. First, before noise suppression, to find speech and noise label sequences, we introduce multi-pass search with acoustic models including many kinds of noise models and their compositions, their n-gram models, and their lexicon. Second, noise suppression based on models is performed using the multiple composite models selected by recognized label sequences with time alignments. We evaluated this approach using the E-Nightingale task, and the proposed method outperformed the conventional method.

Index Terms— speech recognition, noise suppression, model composition, multi-pass search, E-Nightingale project

1. INTRODUCTION

Our laboratories have been working on the E-Nightingale Project to establish fundamental technology for a knowledge sharing system based on understanding everyday activities and situations[1]. We focus on the prevention and reduction of medical malpractice in medical care domains. As one of our research activities, we have been collecting voice memoranda recorded by nurses about their services while working to analyze their actual daily activities[2]. Recently, we started to evaluate the performance of speech recognition for these voice memoranda. However, recognizing them is difficult because they involve 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.

Many noise suppression methods have been proposed to improve the performance of speech recognition for noisy speech. For stationary noise signals, Spectral Subtraction[3] and Parallel Model Combination[4] have been proposed. The Gaussian Mixture Model (GMM) based Minimum Mean-Squared Error (MMSE) method[5] assumes that input noise is stationary but fluctuating. Recently, noise suppression research has focused on non-stationary noise, including a sequential EM approach[6] and particle filter[7]. Since these methods usually assume that only one kind of noise signal exists, applying them to noisy speech that includes many kinds of noise signals is difficult. In general, many kinds of accidental noise signals occur in real environments. Furthermore, obtaining the actual noise signals from input signals is very difficult. We must consider how to detect noise signals, suppress them, and find the target speech.

We propose a new noise suppression method based on multi-pass search using multi-model compositions[8]. This method is called Multi-Model Noise Suppression (MM-NS). For speech recognition in real environments, it is necessary to find speech and noise intervals. Therefore, we consider that a noise suppression process should include a kind of search process. To obtain time alignments of utterances and noise signals from noisy speech data, we apply a multi-pass search 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. In [9], they use Multi-class AdaBoost to detect noise signals that suddenly occur and contaminate speech. To solve this problem, we make models for many kinds of sources, combine some of them, and use these models both for the search and noise suppression as acoustic models. Composite models can find overlapped intervals. Many ideal combinations may be considered but actual existing combinations in training data 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 by multi-pass search, one model for each frame can be allocated. We define an extension of the GMM-based MMSE method[5] for multi-model compositions that can reduce noise signals even if utterances are contaminated by several noise signals.

The rest of our paper is organized as follows. 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 in the world. One aim of the E-Nightingale project is to establish technology using wearable computers and sensor networks to support nursing services[1]. To analyze daily nursing activities, we collected nurses' voice memoranda in real environments while nurses were working[2]. We asked them to record short sentences about each nursing event using IC recorders with small microphones attached to their chests as in Fig. 1. It was difficult to use headset microphones because they were cumbersome and disrupted nursing services. Therefore, we used small microphones on their chests, and the SNRs of recorded speech were usually small, i.e., less than 10 dB. Figure 2 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 analysis, conversations with a coworker, and other persons' speech as background noise. Recog-

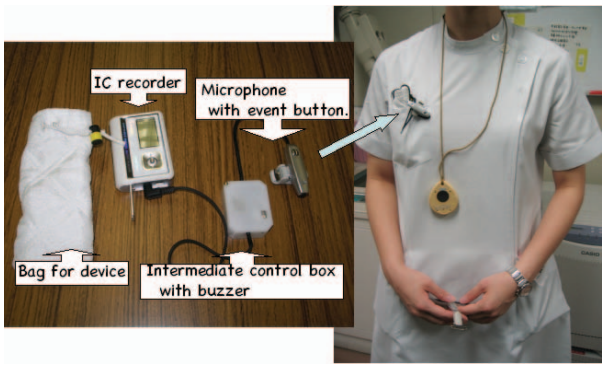


Fig. 1. Recording device

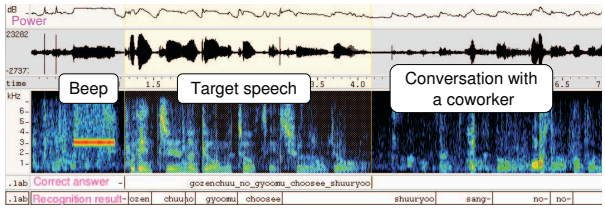


Fig. 2. Wave sample including target speech. Beep prompts speech input. Speaker talked with her coworker after recording her voice memorandum.

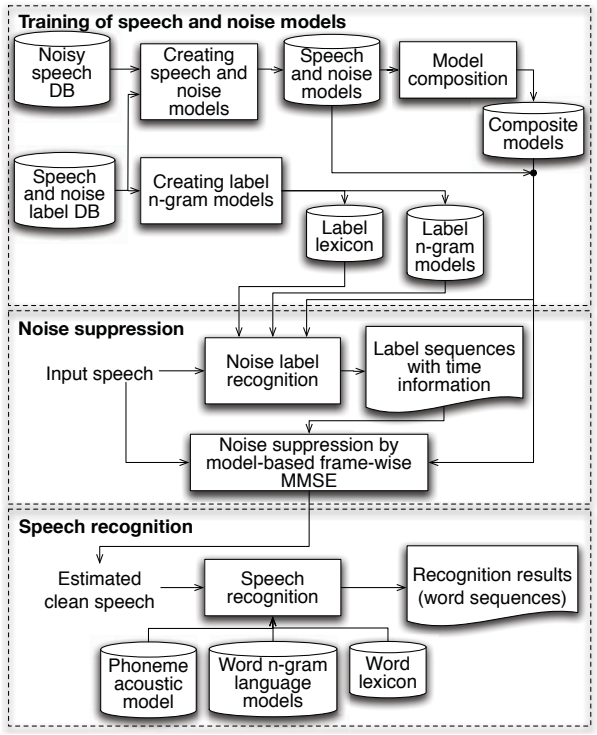


Fig. 3. Overview of Multi-Model Noise Suppression

nizing such voice memoranda 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 many general and essential speech recognition problems. First, we focus on noise suppression in this paper.

3. MULTI-MODEL NOISE SUPPRESSION

3.1. Overview

Figure 3 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. 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. The details of multi-layered labels and composite models will be given in Section 3.2. In this paper, we used GMMs to represent them. Next, to represent overlapped noise signals, models are combined from these trained models. Second, a lexicon and the n-gram models of these labels are generated from noise labels. Furthermore, speaker-adapted models as clean speech models for noise suppression are trained 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; that is identical to usual speech recognition. Instead of word sequences, sequences of speech and noise labels with time information are obtained by this search. Therefore, this process can be considered a multi-pass search. Using 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[5] to obtain estimated clean speech by multiple noise models. Its details will be described in Section 3.3.

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

3.2. Multi-layered labels and composite models

Figure 4 shows an example of multi-layered noise labels and composite models. To consider overlapped noise signals, we first made each speech or noise model and then combinations among them. As shown at the bottom of Fig. 4, a multi-layered label sequence can be represented by a sequence of composite models generated by combinations of the mixture components of a few models in the same manner as [4]. We call a label of composite model a multi-label. Each multi-label is added to a lexicon as one entry.

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.

This approach needs manual labels for noise signals at first. It may be expensive, but it is unavoidable to obtain a model for each noise signal if you want to suppress noise signals elaborately. Furthermore, if the performance of noise label recognition is good enough, unsupervised training is also available.

3.3. GMM-based Multi-Model Noise Suppression

We extend GMM-based noise suppression[5][10] for multi-model compositions. We extend it for multi-model compositions. Assum-

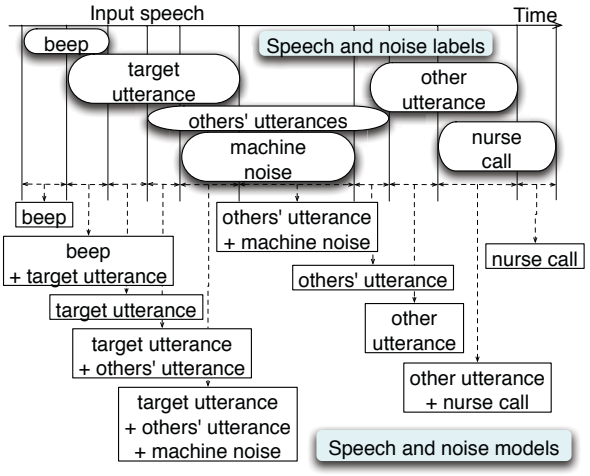


Fig. 4. Example of multi-layered labels and composite models

ing 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 amount of noises. In the log Mel-spectral domain, when $\mathbf{s}(i) = \log S(i)$, $\mathbf{n}_n(i) = \log N_n(i)$, and $\mathbf{x}(i) = \log X(i)$, Eq. (1) can be written as

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

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 $\mathbf{x}(i)$.

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

$$p(\mathbf{s}) = \sum_{k=1}^{K} w_{\mathbf{s},k}\mathcal{N}(\mathbf{s}; \mu_{\mathbf{s},k}, \mathbf{\Sigma}_{\mathbf{s},k}), \tag{3}$$

where $\mathcal{N}()$ means a Gaussian distribution. $w_{\mathbf{s},k}$, $\mu_{\mathbf{s},k}$, 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 by a GMM with L distributions as

$$p(\mathbf{n}_n) = \sum_{l=1}^{L} w_{\mathbf{n}_n,l}\mathcal{N}(\mathbf{n}_n; \mu_{\mathbf{n}_n,l}, \mathbf{\Sigma}_{\mathbf{n}_n,l}), \tag{4}$$

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, clean speech, $\hat{\mathbf{s}}(i)$, can be approximated in the same manner as [5].

$$\hat{\mathbf{s}}(i) \simeq \mathbf{x}(i) - \sum_{m=1}^{M} P(m|\mathbf{x}(i))g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)), \tag{5}$$

where M is the number of mixture components dependent on the combined noise GMMs. Probability $P(m|\mathbf{x}(i))$ is estimated using the composite model:

$$P(m|\mathbf{x}(i)) = \frac{w_{\mathbf{x},m}\mathcal{N}(\mathbf{x}(i); \mu_{\mathbf{x},m}, \mathbf{\Sigma}_{\mathbf{x},m})}{\sum_{m'=1}^{M} w_{\mathbf{x},m'}\mathcal{N}(\mathbf{x}(i); \mu_{\mathbf{x},m'}, \mathbf{\Sigma}_{\mathbf{x},m'})}, \tag{6}$$

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 $\{\mathbf{n}_1, \ldots, \mathbf{n}_N\}$. The l_{nm}-th component means the l_{nm}-th original component of the n-th model included in the m-th composite component. 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[11] as follows:

$$\mu_{\mathbf{x},m} \approx \mu_{\mathbf{s},k} + g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)), \tag{7}$$

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{x},m} &\approx (\mathbf{I} + \mathbf{H}_{\mathbf{s}})\,\mathbf{\Sigma}_{\mathbf{s},k}\,(\mathbf{I} + \mathbf{H}_{\mathbf{s}})^T \\ &\quad + \sum_{n=1,\mathbf{n}_n\in\mathbf{N}_m}^{N}\left(\mathbf{H}_{\mathbf{n}_n,l_{nm}} \cdot \mathbf{\Sigma}_{\mathbf{n}_n,l_{nm}} \cdot \mathbf{H}_{\mathbf{n}_n,l_{nm}}^T\right),\end{aligned} \tag{8}$$

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. Their d-th diagonal elements are

$$\begin{aligned}&\frac{\partial g_d}{\partial s_d} \\ &= -\left[1 + \exp\left\{\mu_{s,k,d} - \log\left(\sum_{n=1}^{N}\exp(\mu_{n_n,l_{nm},d})\right)\right\}\right]^{-1}, \\ &\frac{\partial g_d}{\partial n_{n,d}} \\ &= \left[1 + \exp\left\{\mu_{s,k,d} - \log\left(\sum_{n=1}^{N}\exp(\mu_{n_n,l_{nm},d})\right)\right\}\right]^{-1} \\ &\cdot \frac{\exp(\mu_{n_n,l_{nm},d})}{\sum_{n'=1}^{N}\exp(\mu_{n_{n'},l_{n'm},d})},\end{aligned}$$

respectively. If covariance matrices are diagonal, composite models can be obtained by combining them incrementally. Incremental combined models are identical to models combined all at once.

After the mismatch factor is estimated, in the same manner as [10], the impulse response is calculated for the negative value of mismatch factor, and clean speech is estimated by filtering in the time domain by convoluting the impulse response with the input noisy speech.

3.4. Mismatch factor

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

$$\begin{aligned}&g(\mathbf{s}(i), \mathbf{n}_1(i), \ldots, \mathbf{n}_N(i)) \\ &= \begin{cases} \mu_{\mathbf{x},m} - \mu_{\mathbf{s},k} & \text{for target utterances,} \\ \mu_{\mathbf{x},m} - \varepsilon & \text{for the others,} \end{cases}\end{aligned} \tag{9}$$

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 a hospital in Japan. Data collected the first day were used for evaluation. The length of each file was 10 sec including one target utterance. Data from the second day were used as training data to adapt the acoustic models to speakers and to create noise GMMs for noise suppression. In this paper, diagonal covariance matrices were used for all distributions. Table 1 shows the details of the experimental conditions. Test data included 208 utterances with 1,051 words spoken by eight speakers who were selected as ordinary speakers included both in the test and adaptation data.

For noise suppression, the HTK was used for extracting feature parameters and training GMMs. 24-order outputs of log Mel-filter bank "FBANK" were used as feature parameters. We compared MFCCs with FBANKs only for noise label recognition before noise suppression. We evaluated a speaker-independent (SI) GMM with 512 mixture components and a speaker-dependent (SD) GMM adapted to each speaker by Maximum A Posteriori probability (MAP) estimation[12]. These SI- 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. In this training data, 32 kinds of noise models including a target speech model were obtained. The total number of models with the composite models was 194. In the obtained composite models, the maximum number of models combined into one model was three. Multi-label bigram and trigram models were used for noise label recognition.

As a speech recognizer and training tools, we used the ATRASR large-vocabulary speech recognition system developed by ATR Spoken Language Communication Labs. Its decoder was used both for noise label and word recognition. In this decoder, the bigram model was used with the acoustic model at the first pass, and the trigram model was used to rescore candidates at the second pass. As acoustic models for word recognition, phoneme HMMs with 2,086 states generated by the MDL-SSS algorithm[13] were used. Since all test speakers were females, we only used a female acoustic model. MAP-VFS[14] was used as the speaker-adaptation method. For Multi-Model Noise Suppression, noise intervals became almost clean if noise recognition worked well, but they remained noisy when noise intervals couldn't be estimated. Therefore, the result's labels were obtained by noise recognition, and then phoneme models and a silence model were separately trained to obtain speaker-dependent and noisy silence models.

Table 2 shows evaluation patterns. As a conventional method, we evaluated Single-Model Noise Suppression (SM-NS) that uses one distribution for noise modeling. This distribution is estimated from 100 ms at the beginning of each input file. Pattern (1) used the speaker-independent acoustic model (SI-AM), i.e., the female SI-AM, without noise suppression. As a clean speech GMM used in noise suppression, (2) speaker-independent (SI), and (3) speaker-dependent (SD) GMMs were used, respectively. The SD-GMM was obtained by noise suppression with SI-GMM and speaker adaptation because we did not have any clean speech data for speaker adaptation. In (4), no-processed data were recognized by the SD-AMs adapted using pattern (3). On the other hand, the data processed by (3) were recognized by the SD-AMs in (5). As for our proposed methods, Multi-Model Noise Suppression (MM-NS), we evaluated two types of noise label recognition with (6) & (8) FBANK and (7) & (9) MFCC. Both (6) & (7) used SI-AM, and (8) & (9) used the SD-AMs. (6), (7), (8), and (9) used recognized labels (RLAB) for noise

Table 1. Experimental conditions

Common conditions	
Recording device	IC recorder: iAUDIO G3, COWON Japan Microphone: RASTA BANANA RBENS02 (Handsfree microphone for cellular phones) Band width: 100Hz–10kHz
Analysis conditions	16kHz sampling rate, 16 bit Frame shift: 10 ms, frame length: 20 ms
Test data	8 females (208 utterances, 1,051 words)
Recognition for noise labels	
Tools	HTK Ver.3.3 (GMMs' feature parameters and training) ATRASR Ver.3.6 (decoding and training N-gram models)
Feature parameters	24 Mel-filter bank (FBANK) (for search and noise suppression)
	12 MFCC and 0th MFCC (only for search)
Acoustic models	32 basic GMMs (speech and noise) Training data: about 1 hour 162 composite models
	Clean speech GMM: Speaker-Independent (SI) GMM: 512 mixture components Speaker-Dependent (SD) GMM: about 200 mixture components
Language models	Multi-label bigram, multi-label trigram Training data: 354 utterances
Lexicon	194 multi-labels
Speech recognition	
Tools	ATRASR Ver.3.6
Feature parameters	12 MFCC, 12 ΔMFCC, Δlog power Cepstral Mean Subtraction (CMS)
Acoustic models (AMs)	Phoneme HMM: 2,086 states with 5 mixture components Silence HMM: 3 states with 10 mixture components
AM Training DB	Topology training data: 37 hours ATR Travel conversation DB (TRA), phoneme-balanced sentences
	Re-training DB: 21 hours (female only)
Language models (LMs)	Word bigram, word trigram (Classes for given and family name)
	Test set perplexity: bigram: 39.4, trigram: 39.3 Out of Vocabulary (OOV) rate: 2.36%
LM Training DB	E-Nightingale data 9 days, 9,936 utterances
Lexicon	2,636 words

Table 2. Evaluation patterns

(1) Baseline, SI-AM	without noise suppression (NS) and with a speaker-independent AM
(2) SM-NS (SI) + SI-AM	Single Model NS with speaker-independent GMMs
(3) SM-NS (SD) + SI-AM	Single Model NS with speaker-dependent GMMs
(4) (1) + (3)SD-AM	(1) with speaker-dependent AMs made by (3)
(5) (3) + (3)SD-AM	(3) with speaker-dependent AMs made by (3)
(6) MM-NS (FBANK, RLAB) + SI-AM	Multi-Model NS with result labels by FBANK and SI-AMs
(7) MM-NS (MFCC, RLAB) + SI-AM	Multi-Model NS with result labels by MFCC and SI-AMs
(8) MM-NS (FBANK, RLAB) + SD-AM	Multi-Model NS with result labels by FBANK and SD-AMs
(9) MM-NS (MFCC, RLAB) + SD-AM	Multi-Model NS with result labels by MFCC and SD-AMs
(10) MM-NS (MLAB) + SD-AM	MM-NS with manual labels and SD-AMs

Table 3. Average SNR

Method	SNR [dB]
(1)&(4) Baseline	8.25
(2) SM-NS (SI)	13.43
(3)&(5) SM-NS (SD)	10.24
(6)&(8) MM-NS (FBANK, RLAB)	14.19
(7)&(9) MM-NS (MFCC, RLAB)	16.18
(10) MM-NS (MLAB)	57.39

Table 4. Label accuracy, Voice Activity Detection (VAD) correct, and VAD accuracy

Method	Multi-label LM	Label accuracy [%]	VAD correct [%]	VAD accuracy [%]
(6)&(8) MM-NS (FBANK, RLAB)	bigram	32.27	90	-11
	trigram	33.96	88	40
(7)&(9) MM-NS (MFCC, RLAB)	bigram	35.02	85	6
	trigram	38.23	83	30

suppression, but (10) used manual labels (MLAB). These proposed methods from patterns (6) to (10) used SD-GMMs, and background noise models were estimated in the same manner as the noise distribution of SM-NS.

4.2. Experimental results

Table 3 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 both sides of speech intervals. Since noise intervals were almost clean in the ideal case of our proposed method, (10) MM-NS (MLAB), the obtained SNR was very high. (6) & (8), and (7) & (9) MM-NS (RLAB) obtained many more noisy signals than (10) because result labels included many mistakes. However, the SNRs obtained by our proposed methods from patterns (6) to (9) were higher than those obtained by conventional methods (2) and (3).

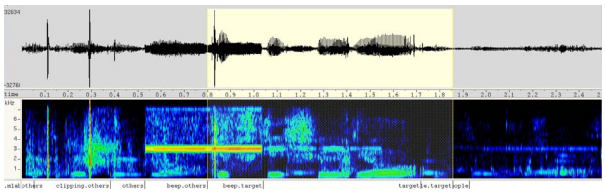
The Out of Label Vocabulary (OOLV) rate can be defined in the same manner as the Out of Vocabulary (OOV) rate. In this test set, the OOLV rates for single and multi-labels were 1.12% and 3.77%, respectively. The current training data for noise models only included 354 utterances. If training data increase, OOLV rates will be reduced. Furthermore, it is enough that noise models just cover frequent noise for noise suppression.

Test set perplexity for the multi-label bigram and multi-label trigram models was 8.08 and 6.47, respectively. This shows that it is meaningful to use multi-label n-gram models.

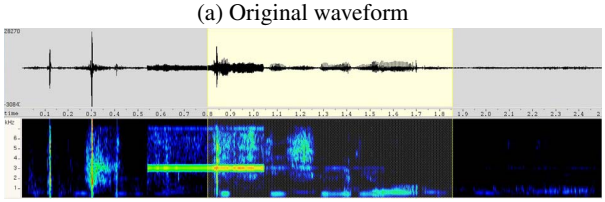
We also evaluated the performance of the Label Accuracy (LA) and Voice Activity Detection (VAD) by MM-NS search. Table 4 shows LA, VAD correct, and VAD accuracy. The LA is defined in the same manner as word accuracy. For noise label recognition, MFCCs obtained better LA rates than FBANK outputs because MFCCs can smooth spectral envelopes and emphasize spectral characteristic. LA rates can show the correctness of label sequences without time information. Our proposed method needs time alignments to allocate noise models for noise suppression. However, exact label sequences with time information are not so important. Detecting target speech intervals is more important. Therefore, we evaluated the correctness of VAD. To evaluate VAD, according to [15], the VAD correct and accuracy rates are defined as VAD correct $= N_c/N_u$, VAD accuracy $= (N_c - N_f)/N_u$,, where N_u 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. Including 500-ms margins, (6) & (8) MM-NS (FBANK) with trigram models was better than (7) & (9) MM-NS (MFCC) with trigram models. The false alarms both for them are large, but, for speech recognition, it is more preferable than deletion errors.

Figure 5 shows sample waveforms: (a) original waveform, (b) conventional method, i.e., pattern (3) & (5) SM-NS (SD) and (c) proposed method using manual labels, i.e., (10) MM-NS (MLAB). The conventional method, (b), only reduced background noise but could not suppress the other noise signals, for example, beep sounds, others' speech, and so on. Our proposed method, (c), suppressed not only background noise but also the other noise signals overlapping the target speech.

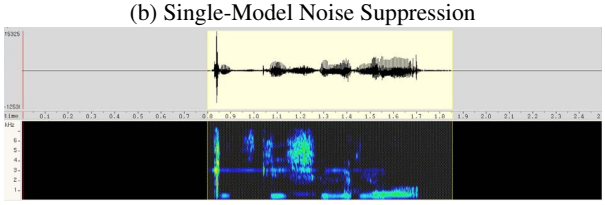
Figure 6 shows the word accuracy rate for each method. The conventional method using SI-GMM, i.e., (2) SM-NS (SI), obtained much lower performance than the baseline, (1). Since the data included a lot of speech like others' speech and the target person's speech, insertion errors were increased after background noise suppression. Therefore, in this task, speaker-adapted acoustic models are needed to extract target utterances. Pattern (3) used SD-GMM in noise suppression, but its performance was still slightly lower than (1) baseline's performance. However, our proposed method, (6) and (7), with SI-AMs outperformed the conventional method. Next, we describe the performance of the methods using SD-AMs. Compared to (4), the conventional method, (5), obtained small error reduction rate, 1.64%. On the other hand, our proposed methods obtained higher improvements. Compared to (4), (8) MM-NS (FBANK, RLAB) and (9) MM-NS (MFCC, RLAB) obtained 6.45% and 7.64% error reduction rates, respectively. Therefore, the proposed method is more effective than the conventional method. Although pattern (10) is the ideal pattern, the performance of pattern (9) was very close to pattern (10). This shows that MFCC is better than FBANK at finding label sequences.



(a) Original waveform



(b) Single-Model Noise Suppression



(c) Multi-Model Noise Suppression with manual labels

Fig. 5. Sample waveforms by noise suppression

5. CONCLUSION

We proposed multi-model noise suppression with a multi-pass search strategy to reduce many kinds of noise signals in realistic environments. It is difficult for conventional noise suppression methods to estimate clean speech from noisy speech contaminated by several kinds of noise signals. To reduce the noise signals of intervals overlapped by several kinds of sources, the multi-model composition was used. To estimate the intervals of several sources included in input data, a multi-pass search was performed using noise acoustic models with 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 for 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.

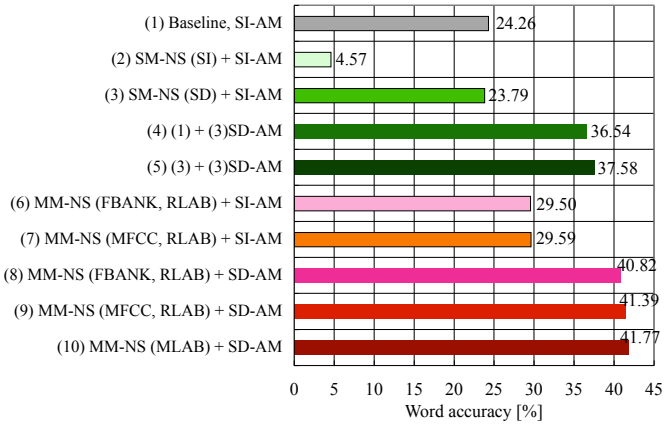


Fig. 6. Word accuracy

6. ACKNOWLEDGEMENTS

This research was supported by the National Institute of Information and Communications Technology of Japan. We'd like to thank all 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 –," in *Proc. the Workshop on Knowledge Sharing for Everyday Life 2006 (KSEL2006)*, 2006, pp. 1–8.

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

[3] 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.

[4] M. F. J. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, University of Cambridge, 1995.

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

[6] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," *Speech Communication*, vol. 42, no. 1, pp. 5–23, 2004.

[7] M. Fujimoto and S. Nakamura, "A non-stationary noise suppression method based on particle filtering and Polyak averaging," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 922–930, 2006.

[8] T. Jitsuhiro, T. Toriyama, and K. Kogure, "Noise suppression using search strategy with multi-model compositions," in *Proc. INTERSPEECH2007*, 2007, pp. 1078–1081.

[9] N. Miyake, T. Takiguchi, and Y. Ariki, "Noise detection with multi-class AdaBoost," in *IEICE Technical Report*, 2006, vol. NLC2006-30, SP2006-86, pp. 7–12, (in Japanese).

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

[11] P. J. Moreno, *Speech recognition in noisy environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[12] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

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

[14] M. Tonomura, T. Kosaka, and 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.

[15] N. Kitaoka et al., "Progress report of SLP noisy speech recognition evaluation WG: Individual evaluation framework for each factor affecting recognition performance," in *IEICE Technical Report*, 2006, vol. NLC2006-29, SP2006-85, pp. 1–6, (in Japanese).