Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 20593, 9 pages
doi:10.1155/2007/20593

Research Article

Robust Speech Recognition Using Factorial HMMs for Home Environments

Agnieszka Betkowska, Koichi Shinoda, and Sadaoki Furui

Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo 152-8552, Japan

Received 1 February 2006; Revised 19 August 2006; Accepted 17 December 2006

Recommended by Mark Gales

We focus on the problem of speech recognition in the presence of nonstationary sudden noise, which is very likely to happen in home environments. As a model compensation method for this problem, we investigated the use of factorial hidden Markov model (FHMM) architecture developed from a clean-speech hidden Markov model (HMM) and a sudden-noise HMM. While in conventional studies this architecture is defined only for static features of the observation vector, we extended it to dynamic features. In addition, we performed *home-environment adaptation* of FHMMs to the characteristics of a given house. A database recorded by a personal robot called PaPeRo in home environments was used for the evaluation of the proposed method. Isolated word recognition experiments demonstrated the effectiveness of the proposed method under noisy conditions. Home-dependent word FHMMs (HD-FHMMs) reduced the word error rate by 20.5% compared to that of the clean-speech word HMMs.

1. INTRODUCTION

A great deal of effort has been devoted to developing personal robots, such as household robots, educational robots, or personal assistants, that 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 under laboratory conditions, their performance decreases significantly when they are used in actual environments. This is mainly because many different kinds of nonstationary noise exist in actual environments. Developing speech recognition devices that are robust against that 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 composition 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. Therefore, the PMC approach works well only for noise with a relatively stable SNR.

We focus on the problem of speech recognition in the presence of nonstationary sudden noise, which is very likely to happen in home environments. This noise appears suddenly and lasts for a short time, and there is no information about its SNR. We apply a model compensation method based on factorial hidden Markov models (FHMMs) that have been introduced as a possible extension of HMMs in [5] because speech enhancement methods and missing data theory approaches are not suitable for this problem. By using the log-max approximation, FHMM can calculate the output probability of the combined model of speech and noise without any gain matching term even when the SNR varies significantly.

In our proposed method, an HMM for each word in the dictionary and an HMM for sudden noise are created. Then, these models are combined to create an FHMM for each word. We propose an extension to employ dynamic features as well because the FHMM architecture proposed in [6] is applicable only to static features of speech signals. We also investigate other possibilities to increase the FHMM model accuracy by applying home-environment adaptation. A database recorded by a personal robot called PaPeRo [7] in home environments was used for the evaluation of the proposed methods. These experiments confirmed that this method improved the recognition accuracy under noisy conditions.

2. ROBUST SPEECH RECOGNITION USING FHMMs

2.1. FHMM

An FHMM is formed as a dynamic belief network composed of more than one layer. Each layer can be seen as a hidden Markov chain that evolves independently from the other layers. The output observation of the FHMM depends on the current states of all layers at each time t.

Let two HMMs, Q and R, with N and W states, respectively, define an FHMM with two layers. The first layer, Q, represents speech, while the second layer, R, models sudden noise. Then, at each time, the speech and noise processes are described by the FHMM *metastate* (q, r), which is defined as a pair of states, q and r, of HMM Q and HMM R, respectively. Furthermore, we assumed that the elementwise maximum of the output observations of the two layers is taken [8]. The structure of this FHMM is shown in Figure 1.

2.2. Log-max approximation

Log-max approximation is based on the observation that, unless two signals are synchronized, the spectrogram of their mixture is almost the same as the elementwise maximum of the spectrograms of these two signals. The spectrogram of the *noisy* speech, $y(t)$, which is the combination of clean speech, $x(t)$, and sudden noise, $n(t)$, can be easily calculated by using the following approximation [9]:

$$\log |Y(j\omega)| = \max\left(\log |X(j\omega)|, \log |N(j\omega)|\right), \quad (1)$$

where $Y(j\omega)$, $X(j\omega)$, and $N(j\omega)$ are the Fourier transforms of $y(t)$, $x(t)$, and $n(t)$, respectively. This log-max approximation was also shown to hold for Mel frequency spectral coefficients (MFSC) [10], which are defined as the log-energy outputs of the speech signal after they are filtered by a bank of triangular bandpass filters on a Mel frequency scale.

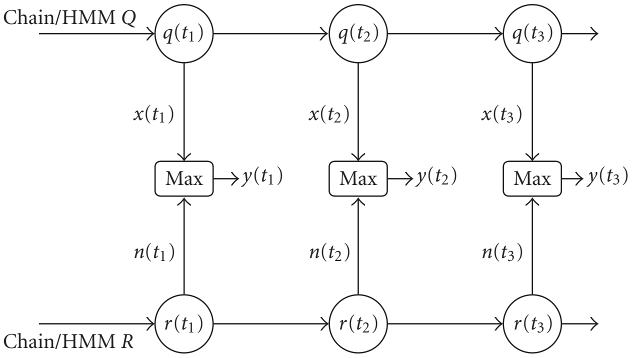


Figure 1: Structure of FHMM composed of two HMMs, Q and R.

2.3. Model formulation

2.3.1. Transition matrix

The FHMM with layers Q and R defined in Section 2.1 can be represented by a traditional HMM with $N \times W$ states [11]. Its transition matrix is defined by the Cartesian product between the transition matrices A_Q and A_R of HMMs Q and R, respectively [11]:

$$a_{(i,j)(k,l)} = a_{ik}^{Q} a_{jl}^{R}, \quad 1 \le i,\ k \le N,\ 1 \le j,\ l \le W. \quad (2)$$

2.3.2. Output probability density function estimation

For each frame, let $\boldsymbol{y} = (y_1, y_2, \ldots, y_D)^T$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^T$, and $\boldsymbol{n} = (n_1, n_2, \ldots, n_D)^T$ be the D-dimensional MFSC vector for noisy speech, clean speech, and noise, respectively. Then, output $\boldsymbol{y}$ of the FHMM for each frame is given by the log-max approximation

$$\boldsymbol{y} \approx \max(\boldsymbol{x}, \boldsymbol{n}), \quad (3)$$

where "max$(\cdot,\cdot)$" stands for the operation selecting the element-wise maximum. This approximation is based on the assumption that, at each time and at each frequency band, one of the mixed signals is much stronger than the other. Hence, the contribution to the output probability density function (*pdf*) from the weaker signal can be neglected.

Let the output *pdfs* for state q in HMM Q and state r in HMM R be represented by the mixture of Gaussians

$$\begin{aligned} p_q(\boldsymbol{x}) &= \sum_{m=1}^{M} c_{qm} N(\boldsymbol{x} \mid \boldsymbol{\mu}_{qm}, \boldsymbol{\Sigma}_{qm}), \\ p_r(\boldsymbol{n}) &= \sum_{m=1}^{M} c_{rm} N(\boldsymbol{n} \mid \boldsymbol{\mu}_{rm}, \boldsymbol{\Sigma}_{rm}), \end{aligned} \quad (4)$$

where M is the number of Gaussians in each state, $\boldsymbol{\mu}_{qm}$ and $\boldsymbol{\mu}_{rm}$ are the mean vectors of the mth mixture components of states q and r, and c_{qm} and c_{rm} are the mth mixture coefficients, respectively. We assume that the covariance matrices $\boldsymbol{\Sigma}_{qm}$ and $\boldsymbol{\Sigma}_{rm}$ of the mth mixture in states q and r, respectively, are diagonal. Hence, a D-variate Gaussian $N(\cdot \mid \cdot,\cdot)$

is equivalent to the product of D univariate Gaussians. Then, the *pdf* of the observation vector $\boldsymbol{y}$ for metastate (q, r) of the FHMM is defined by [6]

$$p_{(q,r)}(\boldsymbol{y}) = p_q(\boldsymbol{y})F_r(\boldsymbol{y}) + p_r(\boldsymbol{y})F_q(\boldsymbol{y}), \tag{5}$$

where

$$\begin{aligned} F_q(\boldsymbol{y}) &= \sum_{m=1}^{M} c_{qm} \prod_{d=1}^{D} \int_{-\infty}^{y_d} p_q(x_d)\,dx_d, \\ F_r(\boldsymbol{y}) &= \sum_{m=1}^{M} c_{rm} \prod_{d=1}^{D} \int_{-\infty}^{y_d} p_r(n_d)\,dn_d. \end{aligned} \tag{6}$$

Symbols $p_q(x_d)$ and $p_r(n_d)$ represent the dth univariate Gaussians in states q and r of HMM Q and HMM R, respectively.

2.4. Extension of FHMM to dynamic features

Temporal changes in the speech spectrum provide important clues about human speech perception and are helpful in describing speech trajectory. The most popular approach to represent this information is to use $\boldsymbol{\Delta}$ coefficients, which are calculated as follows:

$$\boldsymbol{\Delta y}_t = \frac{\sum_{\tau=1}^{G} \tau(\boldsymbol{y}_{t+\tau} - \boldsymbol{y}_{t-\tau})}{\sum_{\tau=1}^{G} \tau^2}, \tag{7}$$

where $\boldsymbol{y}_t$ and $\boldsymbol{\Delta y}_t$ stand for static coefficients and dynamic coefficients, respectively, of the observation vector $\boldsymbol{y}$ in frame t. Parameter τ defines the time shift. It is known that representation containing both static and dynamic features has better performance in speech recognition than a representation with only static features [12].

The calculation of the output *pdf* defined in (5) is based on the log-max approximation. Although this approximation is very effective for static features, it cannot be applied directly to the dynamic part of observation vectors. The elementwise maximum operation between dynamic features of two different signals is meaningless and does not approximate the $\boldsymbol{\Delta}$ features of the mixed signal because dynamic features contain information about changes in the signal over time.

Therefore, we assume that the HMM for the dominant signal, which was selected based on static features of mixed signals, can be used to calculate the *pdf* for the dynamic features as well. We incorporated $\boldsymbol{\Delta}$ features by defining the output *pdf* of FHMM $p'_{q,r}(\boldsymbol{y}, \boldsymbol{\Delta y})$ as

$$p'_{(q,r)}(\boldsymbol{y}, \boldsymbol{\Delta y}) = \begin{cases} p_{(q,r)}(\boldsymbol{y})p_q(\boldsymbol{\Delta y}) & \text{if } p_r(\boldsymbol{y})F_q(\boldsymbol{y}) < p_q(\boldsymbol{y})F_r(\boldsymbol{y}), \\ p_{(q,r)}(\boldsymbol{y})p_r(\boldsymbol{\Delta y}) & \text{otherwise}, \end{cases} \tag{8}$$

where $\boldsymbol{\Delta y}$ represents the dynamic features of $\boldsymbol{y}$, and $p_r(\boldsymbol{\Delta y})$ and $p_q(\boldsymbol{\Delta y})$ are the output *pdfs* for the dynamic part of the observation vector $\boldsymbol{y}$ given by HMM Q and HMM R, respectively. The *pdf* $p_{(q,r)}(\boldsymbol{y})$ was defined in (5). The condition in (8) defines whether process Q or process R is *dominant* at a given time, thus defining which HMM should be used to calculate the output *pdf* for the $\boldsymbol{\Delta}$ features. Terms $F_q(\boldsymbol{y})$ and $F_r(\boldsymbol{y})$ can be regarded as weighting coefficients.

2.5. Home-environment adaptation

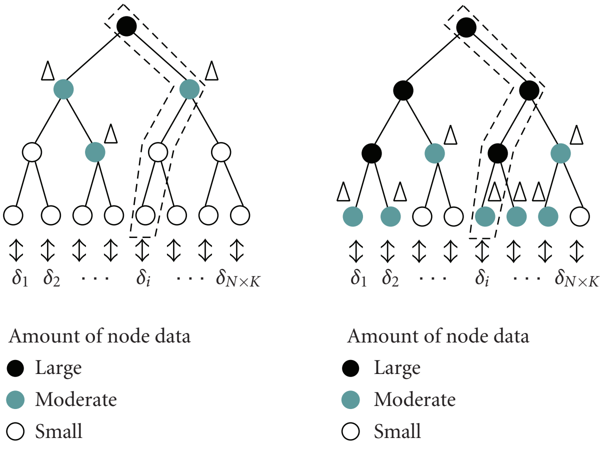
It is generally observed that different groups of people exhibits differences in their voice characteristics and different places exhibits differences in their noise characteristics. Therefore, a home-dependent FHMM (HD-FHMM) adapted to a specific house is expected to yield better performance than that of a home-independent FHMM (HI-FHMM), which represents common characteristics shared by all houses. For the FHMM models defined in Section 2, the adaptation process is conducted independently for speech layer Q and noise layer R.

We use a method proposed by Shinoda and Watanabe [13] for the adaptation of the HMM of each layer. The effectiveness of this method and that of the (MLLR) method [14] are comparable because both methods are piecewise linear transformations. In Shinoda's algorithm, the tree structure is more flexible because the number of branches in each level and the depth of the tree can be chosen manually. In this method, the mean of each Gaussian component in the home-independent HMM (HI-HMM) is mapped to the unknown mean of the corresponding Gaussian component in the home-dependent HMM (HD-HMM). Let μ_i and $\widehat{\mu}_i$ be the mean of the ith Gaussian component of the HI-HMM and the corresponding Gaussian component of the HD-HMM, respectively. Then,

$$\widehat{\mu}_i = \mu_i + \delta_i, \quad i = 1, \ldots, N \times M, \tag{9}$$

where δ_i is a shift parameter from the mean of the HI-HMM, N is the number of states in the model, and M is the number of Gaussian components in each state. Shift δ_i is estimated using a training algorithm such as the forward-backward algorithm or the Viterbi algorithm. The number of δ_i is so large $(N \times M)$ that the correct estimation of these shifts with a limited amount of adaptation data is often very difficult.

To overcome this problem, the proposed method controls the number of shifts to be estimated by using a tree structure (see Figure 2). This tree is constructed by clustering the Gaussian mixtures of the HI-HMM with a top-down clustering method that employs the k-means algorithm. The Kullback-Leibler divergence is used as a measure of distance between two Gaussians. In such a tree, each leaf node i corresponds to Gaussian mixture i, and a tied-shift Δ_j is defined for each nonleaf node j. Using this tree structure, we can control the number of free parameters according to the amount of data available. When we do not have a sufficient amount of data, a tied-shift Δ_j in the upper part of the tree is applied to all the Gaussian components below node j. As the amount of data increases, tied-shifts in the lower levels are chosen for adaptation. To control this process, we use a threshold that defines the minimum amount of data needed to estimate Δ_j. This threshold represents the number of data frames needed for the precise estimation of the shifts attached to each node and is chosen experimentally.


Δ
δ_1 δ_2 · · · δ_i · · · $\delta_{N\times K}$
Amount of node data
● Large
● Moderate
○ Small

(a) Amount of data is small.

(b) Amount of data is large.

Figure 2: Tree structure for shifts estimation.

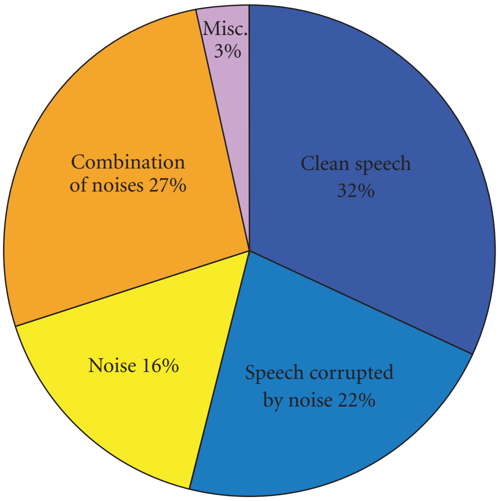


Figure 3: Database statistics.

3. EXPERIMENTS

3.1. Experimental conditions

For the evaluation of the proposed method, we used a database recorded by a personal robot called PaPeRo developed by NEC Corporation [7], which was used in the houses of 12 Japanese families (H01–H12). 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). The database statistics are shown in Figure 3. Each sample in the categories of *clean speech* and *speech corrupted by noise* was transcribed manually. Furthermore, each sample in *speech corrupted by noise* and *noise* categories was labeled with the corresponding noise types. We defined the following noise types: TV, human distant speech, sudden noise, motor, kitchen sounds, electrical sounds, footsteps, robot speech, and miscellaneous (undefined noise). There is a large variety of noise in the home environment, so each sample can contain more than one noise type.

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. We also used 2828 samples of speech corrupted by sudden noise, which we call *recorded noisy speech*. The statistics for each house are shown in Figure 4. Samples were digitized at the 11 025 Hz sampling rate, and analyzed at a 10- millisecond frame period. Log filter-bank parameters consisting of 24 static features, 24 Δ features, and Δ energy 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).

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 nine states (three states of silence, three states of sudden noise, and the remaining states also of silence), was built in a similar manner. The state output *pdf* for all HMMs was a single Gaussian distribution. Finally, an FHMM 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, as described in Section 2.1.

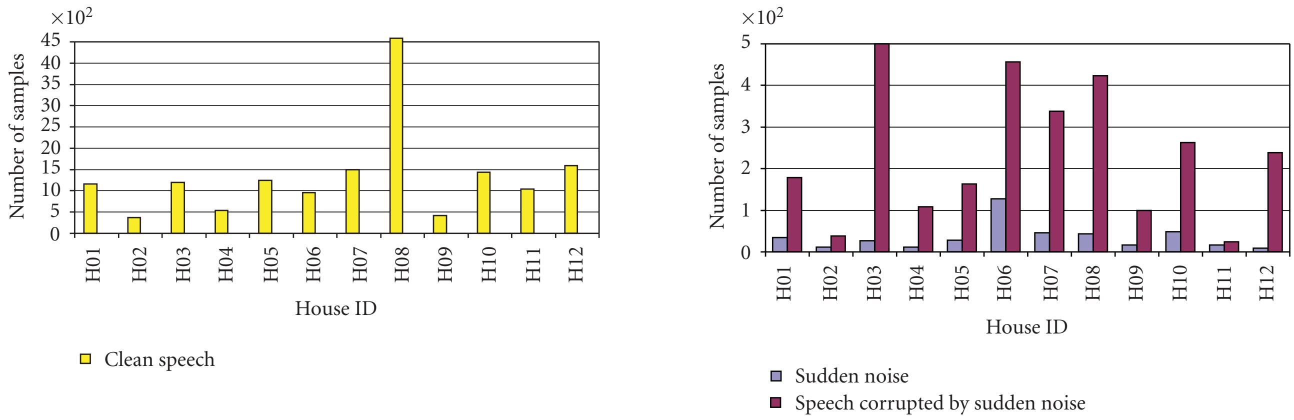
3.2. Effectiveness of FHMMs

First, we evaluated the effectiveness of the proposed FHMMs. In this experiment, the samples from eight houses (H02–H06, H08, H09, and H11) 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 houses, 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 houses. 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 ith frame of the signal s. In addition, let $C := \{i \mid \mathbf{w}(\boldsymbol{i}) \geq \lambda\}$, that is, C is the set of indices in which the ith frame has power greater than or equal to threshold λ. The power of signal s is defined by

$$P(s) = \frac{\sum_{i\in C} \mathbf{w}(i)}{W}, \qquad (10)$$

where W is the number of frames in set C. For our experiments, we set threshold $\lambda = 400$ for speech, and $\lambda = 50$



(a)

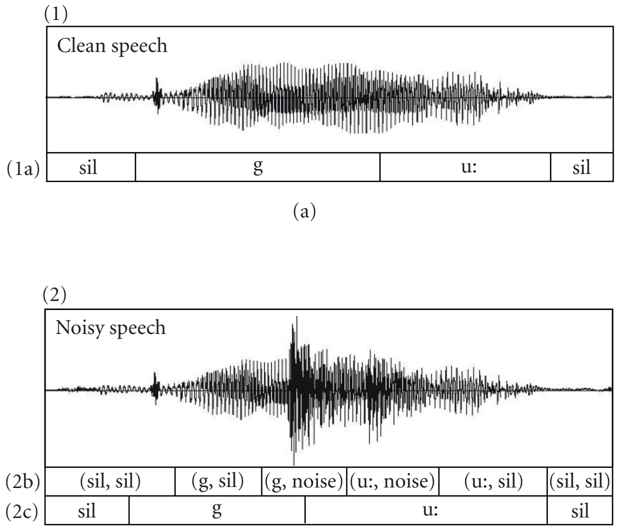
(b)

Figure 4: Number of samples in our experiments.

for noise. These values were optimized in preliminary experiments.[1] Sudden noise has a much shorter duration than speech, so we investigated three different ways of synthesizing the two signals: adding a noise sample at the beginning, in the middle, and at the end of the speech sample. In preliminary experiments, we found that a synthesized noisy speech signal with noise at the center was the most difficult task for the speech recognizer. Therefore, a clean speech sample and a sudden noise sample were synthesized such that the midpoint of these two samples was located at the same time. An evaluation test with 548 utterances at each SNR was prepared.

An example of the segmentation given by an FHMM is shown in Figure 5, where artificial noisy speech (in Figure 5(2)) is made by synthesizing the clean speech sample (Figure 5(1)) and a sudden noise sample at an SNR of 0 dB. The FHMM correctly detected the noisy part of the signal as well as the speech sequence, as shown in Figure 5(2b). Its alignment was only slightly different from that of Figure 5(1a) for clean speech, while the segmentation of Figure 5(2c) given by the clean-speech HMM was different from that shown in Figure 5(1a).

We compared the recognition accuracies of clean-speech HMMs without Δ features, clean-speech HMMs with Δ features, FHMMs with Δ features, and FHMMs without Δ features for the five different SNRs. The results averaged over the four houses are shown in Figure 6. The FHMMs performed better than their corresponding clean-speech HMMs. The FHMMs defined only for static features improved the recognition accuracy by 6.2% absolute at −5 dB, by 6.4% absolute at 0 dB, by 1.8% absolute at 5 dB, and by 4.8% absolute at 10 dB. When Δ features were included, further improvement was obtained. The proposed FHMM improved the recognition accuracy obtained from clean-speech HMMs by 2.0% absolute at 10 dB, by 3.5% absolute at 5 dB, by 7.7% absolute at 0 dB, and by 6.1% absolute at −5 dB. As the SNR increased, however, the difference between the baseline clean-speech HMMs and the proposed FHMMs decreased, giving an advantage to the conventional HMM at 20 dB SNR and under clean conditions. This may be because slight mismatches between the training data and the test data in the clean part of the noisy speech were misrecognized as noise when the SNR is high. When the recognizer chooses the noise as the stronger signal [see (8)], the wrong HMM model is used to calculate the *pdf* of Δ features of the clean speech



(a)

(b)

Figure 5: Segmentation for sound "gu:". Segmentations (1a) and (2c) are given by clean-speech HMMs. Segmentation (2b) is given by FHMM, where each unit is represented by a pair of speech and noise units.

[1] We decided not to use the standard SNR measure methods, such as NIST standard, because they calculate the SNR from the noisy speech only, not from the clean speech and the noise separated signals (http://www.nist.gov/smartspace/snr.html).

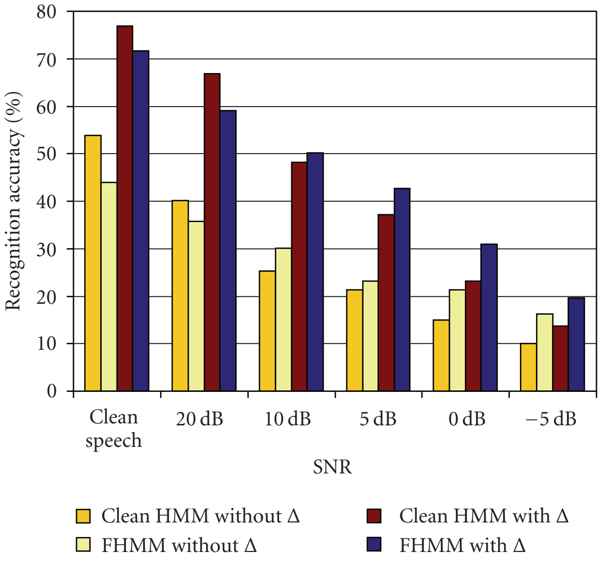


Figure 6: Recognition rates of speech artificially corrupted by sudden noise. HMMs (baseline) and the proposed method (noisy FHMMs) with and without Δ features.

signal. Hence, the initial error is amplified and is more difficult to correct. A second possible explanation for the reduced performance of our proposed algorithm compared to the baseline is the way that word FHMM is constructed (see Section 3.1). Although the noise signal is not detected, the output *pdf* for a given sequence of observation vectors is practically calculated by the combination of all clean-speech HMM states and silence states of noise HMM. The silence states do not perfectly model the noise absence and this may cause the degradation of the output *pdf* given by FHMMs.

3.3. Comparison with matched HMM

Knowing whether our FHMM is more effective than *matched* HMMs that were trained by using noisy speech samples (HMMs trained in matched condition) is important. However, we could not construct such HMMs due to the insufficient number of recorded noisy speech samples. To overcome this problem, we adapted the clean speech HMMs to noisy speech using the available recorded noisy speech samples. For adaptation, we used 1811 recorded noisy speech samples from eight houses (H02–H06, H08, H09, and H11). The recorded noisy speech samples from the other 4 houses, 1017 samples, were used for evaluation. The results are given in Figure 7. While the matched HMMs improved the recognition accuracy by 1.1% absolute, FHMMs improved it by 13% absolute. This result confirmed that FHMMs are more effective.

This result can be explained by the characteristics of sudden noise. This sudden noise appears for only a limited time, so it corrupts only a small portion of each speech sample; the remaining part of the sample is not distorted. Hence, obtaining an HMM that represents the characteristics of sudden noise and clean speech well is very difficult.

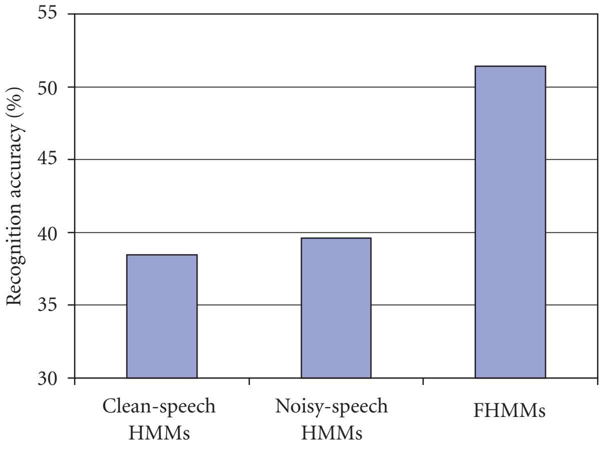


Figure 7: Recognition rates of clean HMMs (baseline), noisy HMMs, and noisy FHMMs for recorded noisy speech.

On the contrary, our FHMMs represent such characteristics efficiently. The speech and noise layers in FHMM compete in each frame to represent each observation vector of noisy speech. Depending on whether the speech or noise signal is stronger, the FHMM switches the layers to calculate the output probability.

3.4. Home-environment adaptation

Finally, we evaluated the effect of home-environment adaptation of FHMMs. We applied supervised and unsupervised home-environment adaptation only for speech layer Q in the FHMM. We did not adapt noise layer R because for most of the houses there was not a sufficient number of sudden noise samples to perform the adaptation. In supervised adaptation, we assumed that the transcription of clean speech samples for adaptation is known. On the contrary, in unsupervised adaptation, samples are unlabeled and their transcription is performed via a speech recognition process.

For the evaluation, we used a *"leave-one-out"* method, where the training and testing process was repeated for each house, except for H11, which had a very small number of noisy speech samples. For each house, the training data consisted of samples of clean speech from all other houses. Recorded noisy speech samples of the given house were taken for a testing set. The sizes of the test sets were different for each house, ranging from 24 to 500 samples (see Figure 4). For supervised and unsupervised adaptation, we used a randomly chosen set of 183 clean speech samples from each house, which were not included in the training or testing set. The adaptation procedure was as follows. First, a clean-speech home-independent HMM (HI-HMM) and noise HMM were trained using clean speech samples and noise samples as the training data, respectively. Then, the speech HI-HMM was transformed to HD-HMM for the given house using the adaptation procedure described in Section 2.5. We constructed the tree structure for shifts as a binary tree with four levels. The threshold, which defines

Table 1: Threshold for home-environment adaptation—minimum amount of data needed to estimate Δ_j.

House ID	Number of frames
01	10
02	25
03	10
04	18
05	7
06	17
07	26
08	17
09	18
10	40
12	55

the minimum amount of data needed to estimate Δ_j, was chosen experimentally for each house (see Table 1). A noisy HD-FHMM was created from the HD-HMM and the noise HMM, as described in Section 3.1.

We compared the home-independent FHMMs (HIFHMMs) and home-dependent FHMMs (HD-FHMMs). The results are shown in Figures 8 and 9. When supervised adaptation was applied, the HD-FHMM exhibited better performance than that of HI-FHMMs, giving improvement ranging from 1.1 to 16.7% absolute (H06 to H04), respectively, for almost all houses except H12.

On average, HD-FHMMs achieved 8.9% relative error reduction compared to that of HI-FHMMs. Using the HI-FHMMs resulted in a 17.9% relative error reduction compared to that of HI-HMMs. Overall, the proposed method reduced the relative error rate by 25.2%, compared to that of HI-HMMs.

Unsupervised adaptation gave slightly worse results than those of supervised adaptation, as shown in Figure 10. This was expected because the system might use incorrect labels for given samples during the adaptation process. Nevertheless, applying the speech-adaptation process to HI-FHMMs resulted in a 20.5% relative error rate reduction, taking clean-speech HI-HMMs as a baseline. The HD-FHMMs outperformed HI-FHMMs by 1.5% absolute.

In addition, this experiment demonstrated the effectiveness of HI-FHMM over HI-HMMs, as in Section 3.2. This time, instead of using synthesized data, we used actual recorded noisy data. In 7 houses out of 11, the recognition accuracy was improved by more than 10% absolute.

4. CONCLUSION AND FUTURE WORK

We investigated the use of FHMMs for speech recognition in the presence of nonstationary sudden noise, which is very likely to be present in home environments. The proposed FHMMs achieved better recognition accuracy than clean-speech HMMs for different SNRs. In the best case, at 0 dB, an

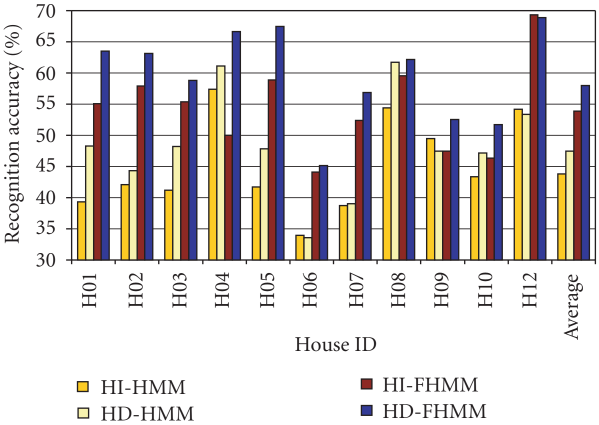


Figure 8: Results for supervised home-environment adaptation.

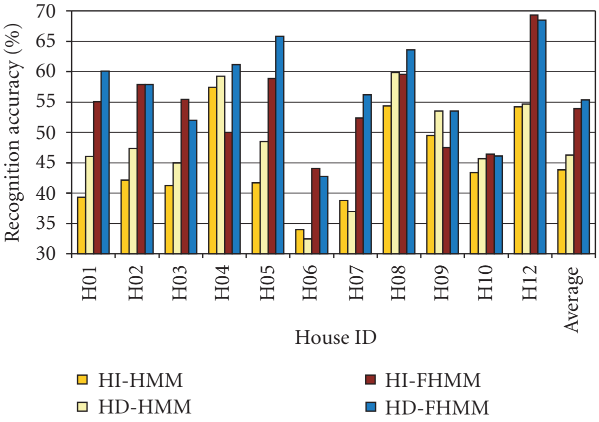


Figure 9: Results for unsupervised adaptation.

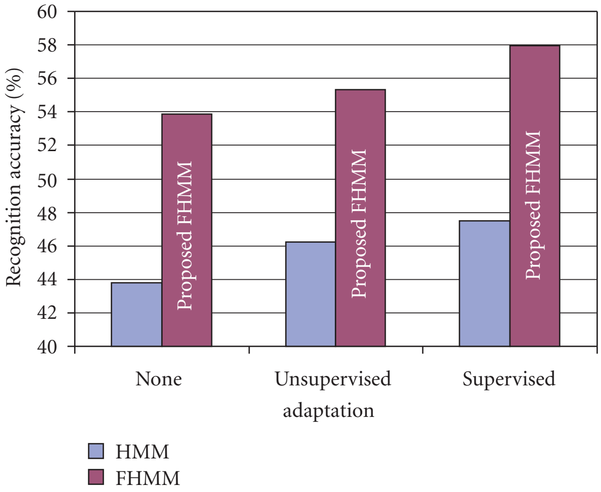


Figure 10: Averaged results of home-environment adaptation.

improvement of 7.1% absolute was obtained. The usability of FHMMs was further investigated by using a recorded noisy speech test set. The overall relative error reduction given by FHMMs with Δ features was 17.9% compared to that given by the clean-speech HMMs. Our experiments also demonstrated the effectiveness of home-environment adaptation. We achieved relative error-reduction rates of 20.5% and 25.2% for unsupervised and supervised adaptation, respectively.

We created a noisy FHMM by combining an HMM for clean speech and an HMM for noise, both of which have simple structures in this study. HMMs created with more complex structures (more Gaussians per state, different HMMs topologies, and number of states) need to be investigated. In addition, we designed an FHMM for each word, while a phone is a more natural unit for our problem. Our next plan is to create phone FHMMs and combine them with phone HMMs in various ways. In our experiments, we used MFSC features because they follow the log-max approximation. In the future, we would like to apply more robust features to FHMM architecture. Moreover, we only considered one kind of noise at a time; however, in home environments there are many other kinds of noise such as footsteps, TV sounds, and distant speech. FHMMs for the combination of different noises should also be investigated. Finally, we performed the adaptation using only clean speech; however, the same can be done in a similar manner for noise. In the future, we would like to perform home-environment adaptation of FHMMs using noise and noisy speech samples and test online adaptation as well.

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 the PaPeRo database.

REFERENCES

[1] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory Algorithm and System Development*, Prentice-Hall, Upper Saddle River, NJ, USA, 2001.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[3] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '90)*, vol. 2, pp. 845–848, Albuquerque, NM, USA, April 1990.

[4] M. J. F. Gales and S. J. Young, "HMM recognition in noise using parallel model combination," in *Proceedings of the 3rd European Conference on Speech Communication and Technology (EuroSpeech '93)*, vol. 2, pp. 837–840, Berlin, Germany, September 1993.

[5] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[6] A. N. Deoras and M. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 1, pp. 861–864, Montreal, Quebec, Canada, May 2004.

[7] T. Iwasawa, S. Ohnaka, and Y. Fujita, "A speech recognition interface for robots using notification of III-suited conditions," in *Proceedings of the 16th Meeting of Special Interest Group on AI Challenges*, pp. 33–38, 2002.

[8] T. S. Roweis, "One microphone source separation," in *Proceedings of Neural Information Processing Systems (NIPS '00)*, vol. 13, pp. 793–799, Denver, Colo, USA, 2000.

[9] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[10] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] B. Logan and P. Moreno, "Factorial HMMs for acoustic modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 2, pp. 813–816, Seattle, Wash, USA, May 1998.

[12] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[13] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous control using tree structure," in *Proceedings of the 4th European Conference on Speech Communication and Technology (EuroSpeech '95)*, pp. 1143–1146, Madrid, Spain, September 1995.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

Agnieszka Betkowska received the M.S. degree in computer science from Adam Mickiewicz University, Poland, in May 2002. From September 2002 to March 2003, she had been a Marie-Curie Fellowship Holder in the Research Center of Information Technologies, Karlsruhe, Germany. Since April 2003, she has been awarded the Monbukagakusho (Japanese Government) Scholarship. Currently, she is pursuing the Ph.D. degree at the Tokyo Institute of Technology. Her current interest is in robust speech recognition technology.

Koichi Shinoda received the B.S. degree in 1987, M.S. degree in 1989, both in physics from the University of Tokyo, and Dr. Eng. degree in computer science from Tokyo Institute of Technology in 2001. In 1989, he joined NEC Corporation, Japan, and was involved in research on automatic speech recognition. From 1997 to 1998, he was a Visiting Scholar with Bell Labs, Lucent Technologies, Murray Hill, NJ.

From October 2001 to March 2002, he was an Associate Professor with the University of Tokyo. He is currently an Associate Professor with Tokyo Institute of Technology. His research interests include speech recognition, audio and video information retrieval, statistical pattern recognition, and human interface. He received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from IEICE in 1998. He is a Member of IEEE, ISCA, ASJ, IPSJ, JSAI.

Sadaoki Furui is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction, and he has authored or coauthored over 700 published articles. He is a Fellow of the IEEE, the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), and the Acoustical Society of America. He has served as President of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA). He has served as a Member of the Board of Governors of the IEEE Signal Processing (SP) Society and Editor-in-Chief of both the Transaction of the IEICE and the Journal of Speech Communication. He has received the Yonezawa Prize, the Paper Award, and the Achievement Award from the IEICE (1975, 1988, 1993, 2003, 2003), and the Sato Paper Award from the ASJ (1985, 1987). He has received the Senior Award and Society Award from the IEEE SP Society (1989, 2006), the Achievement Award from the Minister of Science and Technology and the Minister of Education, Japan (1989, 2006), and the Purple Ribbon Medal from Japanese Emperor (2006). In 1993, he served as an IEEE SPS Distinguished Lecturer.