論文 / 著書情報
Article / Book Information

Title	Home-Environment Adaptation of Phoneme Factorial Hidden Markov Models
Author	Agnieszka Betkowska, Koichi Shinoda, Sadaoki Furui
Journal/Book name	Proc. EUSIPCO 2007, Vol. , No. , pp. 2380-2384
発行日 / Issue date	2007, 9

HOME-ENVIRONMENT ADAPTATION OF PHONEME FACTORIAL HIDDEN MARKOV MODELS

Agnieszka Betkowska, Koichi Shinoda, and Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
email: agabet@furui.cs.titech.ac.jp, {shinoda,furui}@cs.titech.ac.jp

ABSTRACT

We focus on the problem of speech recognition in the presence of nonstationary sudden noise, which is very likely to happen in home environments. To handle this problem, a model compensation method based on a factorial hidden Markov model (FHMM) has been recently introduced. In this architecture, speech and noise processes are modeled in parallel by a phoneme FHMM that is built by combining a clean-speech phoneme hidden Markov model (HMM) and a sudden noise HMM. Here, to increase the robustness of this method further, we apply supervised and unsupervised home-environment adaptation of phoneme FHMMs. A database recorded by a personal robot PaPeRo in home environments was used for the evaluation of the proposed method under noisy conditions. The phoneme home-dependent FHMM achieved better recognition accuracy than the clean-speech home-independent HMM, reducing the overall relative error by 16.2% and 12.3% on average for supervised and unsupervised adaptation, respectively.

1. INTRODUCTION

In recent years, 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. Recent achievements in this area include the android Asimo (Honda), pet robot Aibo (Sony), and family robot PaPeRo (NEC). Speech is the most natural and the easiest way to communicate for humans, so most of these robots are equipped with a speech recognition function because their interfaces 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 real environments due to the presence of nonstationary noise. Therefore, speech recognition devices that are robust against nonstationary noise are in great demand.

Current studies on robust speech recognition can be categorized into the following three groups: speech enhancement, missing data theory, and model compensation. Speech enhancement methods try to suppress noise from the speech signal, but they are only effective for stationary noise. Missing data theory aims at determining the level of reliability of each spectral region in the speech spectrogram. Hence missing data theory only works well for noise that selectively corrupts a small portion of the signal spectrum. Model compensation methods combine noise and speech models during the recognition process, so they are able to deal with nonstationary noise. A promising model compensation method is the factorial hidden Markov model (FHMM), which is an extension of hidden Markov models (HMMs)[1].

The FHMM consists of layers that model loosely coupled processes. They have been used in [4] to increase the robustness of speech recognition systems in the presence of nonstationary sudden noise, which is very likely to occur in home environments. The scheme in [4] has been evaluated with word FHMMs in a word-isolated speech recognition task. An extension to phoneme FHMMs, which can be applied to large-vocabulary continuous speech recognition (LVCSR) systems, has been reported in [5].

In this study, to improve the robustness of the speech recognition system using phoneme FHMMs, we apply adaptation of phoneme FHMMs to a specific home environment. Different people have different voice characteristics and that different places exhibit differences in their noise characteristics. On the basis of that principle, a home-dependent phoneme FHMM is expected to increase the robustness of speech recognition systems compared with that of home-independent phoneme models, which represent common characterstics shared by all homes. For the evaluation of the proposed algorithm, we used a database recorded by a personal robot PaPeRo [6] in home environments. The experiments confirmed that our method improves 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.

Let an FHMM be composed of two HMM layers, Q and R, with N and W states, respectively. The first layer, Q, represents speech, while the second layer, R, models sudden noise. Then, at each time frame, 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 element-wise maximum of the output observations of the two layers is taken [7]. The structure of this FHMM is shown in Figure 1.

2.2 Model formulation

2.2.1 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 [8]. Its transition matrix is defined by the Cartesian product of

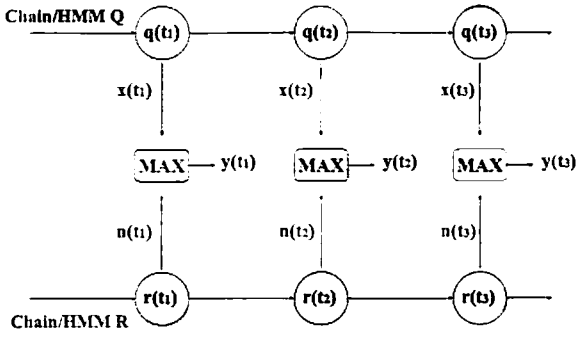

Figure 1: Structure of FHMM composed of two HMMs, Q and R.

the transition matrices A_Q and A_R of HMMs Q and R, respectively [8]:

$$a_{(i,j)(k,l)} = a^Q_{ik} a^R_{jl}, \quad 1 \le i,k \le N, \quad 1 \le j,l \le W. \quad (1)$$

2.2.2 Output probability density function estimation for static part of observation vector

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

$$\mathbf{y} \approx \max(\mathbf{x}, \mathbf{n}), \quad (2)$$

where "max(.,.)" 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:

$$p_q(\mathbf{x}) = \sum_{m=1}^{M} c_{qm} N(\mathbf{x}|\mu_{qm}, \Sigma_{qm}) \quad \text{and} \quad (3)$$

$$p_r(\mathbf{n}) = \sum_{m=1}^{M} c_{rm} N(\mathbf{n}|\mu_{rm}, \Sigma_{rm}), \quad (4)$$

where M is the number of Gaussians in each state, μ_{qm} and μ_{rm} are the mean vectors of the m-th mixture components of states q and r, and c_{qm} and c_{rm} are the m-th mixture coefficients, respectively. We assume that the covariance matrices Σ_{qm} and Σ_{rm} of the m-th mixture in states q and r, respectively, are diagonal. Hence, a D-variate Gaussian $N(.|.,.)$ is equivalent to the product of D univariate Gaussians. Then, the *pdf* of the observation vector $\mathbf{y}$ for metastate (q,r) of the FHMM is defined by [2]:

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

where

$$F_q(\mathbf{y}) = \sum_{m=1}^{M} c_{qm} \prod_{d=1}^{D} \int_{-\infty}^{y_d} p_q(x_d) dx_d \quad \text{and} \quad (6)$$

$$F_r(\mathbf{y}) = \sum_{m=1}^{M} c_{rm} \prod_{d=1}^{D} \int_{-\infty}^{y_d} p_r(n_d) dn_d. \quad (7)$$

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

2.2.3 Output probability density function estimation for static and dynamic features

The calculation of the output *pdf* defined in (5) is based on the log-max approximation, which is very effective for static features but cannot be applied directly to the dynamic part of the observation vectors. The element-wise maximum selection operation between dynamic features of two different signals is meaningless and does not approximate the Δ features of the mixed signal because dynamic features contain information about changes in the signal over time.

To incorporate dynamic features of the observation vector into the calculation of the *pdf* of metastate (q,r), three steps are performed [4]. First, for each frame, the dominant signal from the mixed signal is detected by applying the log-max approximation to static features of the mixed signal. Next, if speech is dominant, state q is used to calculate the *pdf* of the dynamic features; otherwise, state r is used. Finally, the output *pdf* of FHMM $p'_{q,r}(\mathbf{y}, \Delta\mathbf{y})$ for static and dynamic features of the observation vector is calculated as follows:

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

where $\Delta\mathbf{y}$ represents the dynamic features of $\mathbf{y}$, and $p_r(\Delta\mathbf{y})$ and $p_q(\Delta\mathbf{y})$ are the output *pdfs* for the dynamic part of the observation vector $\mathbf{y}$ given by states q and r, respectively. The *pdf* $p_{(q,r)}(\mathbf{y})$ is defined in (5). The condition in (8) defines whether process Q or process R is *dominant* at a given time, thus defining which state should be used to calculate the output *pdf* for the Δ features. Terms $F_q(\mathbf{y})$ and $F_r(\mathbf{y})$ are defined in (6) and (7), respectively, and they can be regarded as weighting coefficients.

2.3 Phoneme FHMM versus word FHMM

The theory presented in Sect. 2.1 can be applied to create word FHMMs and phoneme FHMMs. In a word FHMM, the speech layer is represented by a word HMM. In a phoneme HMM, the speech layer is represented by a phoneme HMM. In our FHMM formulation, we assume that the speech signal and the sudden noise have the same duration. While this assumption is usually correct in word FHMMs, it is not clear that it is also correct in phoneme HMMs. The duration of some samples of sudden noise might be longer than the duration of a phoneme. In this case, the misalignment of noise signals may deteriorate the recognition performance.

Nevertheless, in certain applications, such as in large-vocabulary systems, phoneme FHMMs are more desirable because they need less training data, reduce the computational time during the recognition process compared with that of word FHMMs, and require significantly less memory.

We create phoneme FHMMs in two steps. First, clean speech HMMs for each phoneme are trained using clean

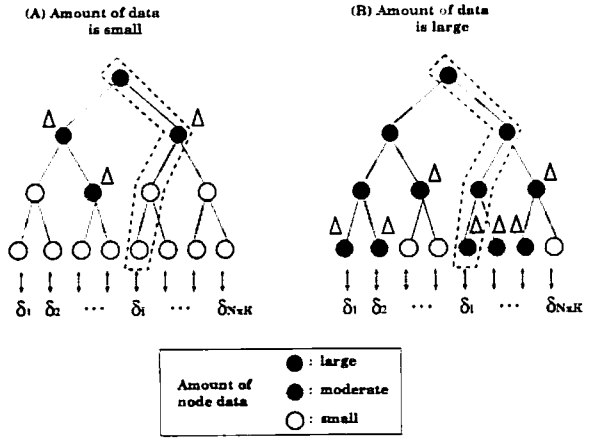

Figure 2: Tree structure for shifts estimation

speech data, and a noise HMM is trained using noise data. Then, the phoneme FHMM that models speech and noise in parallel is created by combining a phoneme HMM for clean speech with a noise HMM for each phoneme, as described in Sect. 2.1. The phoneme HMM and the noise HMM are the mathematical representations of speech and noise layers of the phoneme FHMM, respectively. The resulting transition matrix of phoneme FHMM reflects how the noise signal and the speech signal of the phoneme are mixed.

2.4 Home-environment adaptation

For the FHMM models defined in Section 2.1, the adaptation process is conducted independently for speech layer Q and noise layer R. For the adaptation of each layer, we use a method proposed by Shinoda *et al.* [9]. The effectiveness of this method and that of the (MLLR) method are comparable because both methods are piecewise linear transformations. However, 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 arbitrarily designed.

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 $\hat{\mu}_i$ be the mean of the i-th Gaussian component of the HI-HMM and the corresponding Gaussian component of the HD-HMM, respectively. Then,

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

where δ_i is a shift parameter obtained 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 of Gaussian components (see Figure 2). This tree is constructed by clustering the Gaussian mixtures of all the states 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 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.

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 [6], 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 recorded by PaPeRo were labeled manually and classified into three different types: speech without noise, noisy speech, and noise without speech. Furthermore, each noise sample was labeled with the corresponding noise type: 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 types 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 2,828 samples of speech corrupted by sudden noise. Each sample consists of a small period of silence at the beginning, an uttered word (speech sample) or noise (noise sample) in the middle, and silence at the end. Samples were digitized at a 11,025-Hz sampling rate and analyzed at a 10-ms 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 (commands spoken to the PaPeRo robot). The vocabulary contains 1,492 entries, consisting of words and simple phrases. For simplicity, we treated each phrase as a word. The recognition units in clean-speech HMMs were triphones, and the state output *pdf* for all HMMs was a single Gaussian distribution.

3.2 Effectiveness of supervised and unsupervised adaptation of phoneme FHMMs

To build the phoneme home-dependent FHMM (HD-FHMM), a speech home-independent HMM (HI-HMM) was adapted to the conditions of a given house using the adaptation procedure described in Sect. 2.4. The resulting phoneme home-dependent HMMs (HD-HMM) were combined with the noise HMM as described in Sect. 2.3. We applied supervised and unsupervised home-environment adaptation only for the speech layer in the phoneme FHMM. We did not have a sufficient number of sudden noise samples, so performing adaptation for noise layer R was not possible.

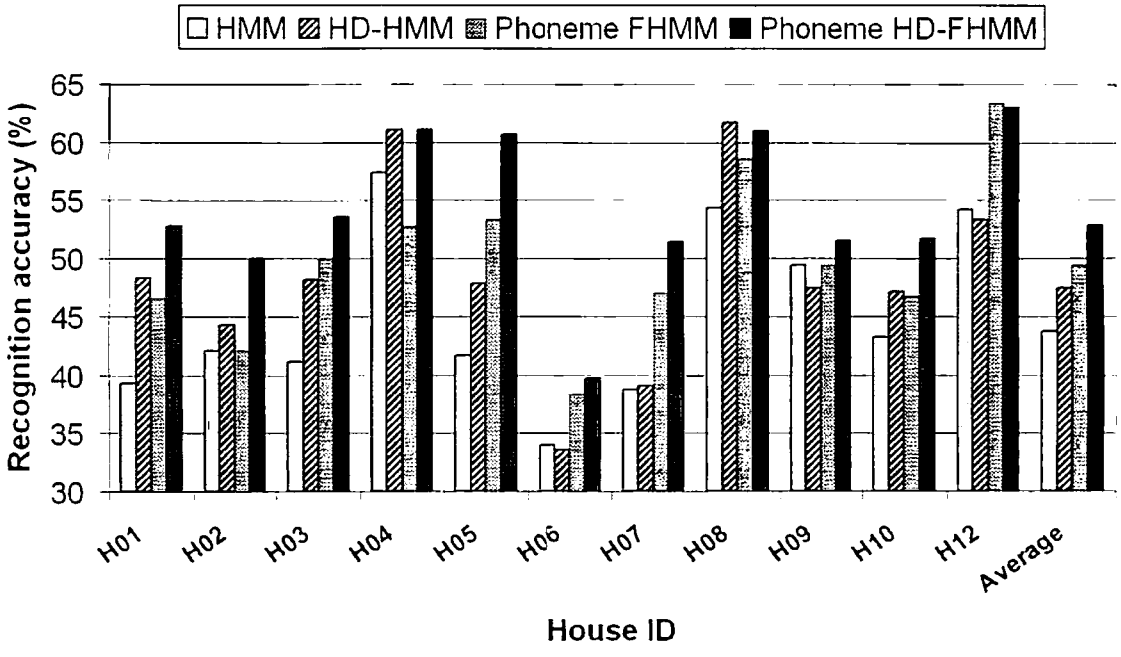

Figure 3: Results for supervised adaptation

For assessment, we used the *"leave-one-out"* method, where the training and testing processes were repeated for each house, except for H11, which had too few noisy speech samples for testing. For each house, the training data consisted of samples of clean speech from all other houses. The noisy speech samples of the given house were used as a testing set. For supervised and unsupervised adaptation of speech layer Q, we used 183 clean speech samples from each house, which were not included in the training or testing sets. In supervised adaptation, we used the true transcription of the adapted data, which was manually prepared. On the other hand, in unsupervised adaptation, the transcription of the adaptation test was obtained via the speech-recognition process.

The results of supervised and unsupervised phoneme FHMM adaptation are given in Figures 3 and 4, respectively. On average, supervised adaptation of the phoneme FHMM reduced the relative error by 6.7% compared to that of the phoneme HI-FHMM and 16.2% compared to that of HI-HMMs. The best absolote improvement of 19.0% was obtained for house H05. The unsupervised adaptation gave worse results that those of supervised adaptation. This is expected because the system might use incorrect labels for given samples during the adaptation process. Nevertheless, applying unsupervised speech adaptation to phoneme HI-FHMM decreased the relative error by 2.6% compared to that of HI-FHMM.

We also compared the adaptation performance of phoneme and word FHMMs (see Figure 5). The Word FI-FHMM, the word HD-FHMM (unsupervised adaptation), and the word HD-FHMM (supervised adaptation) gave 17.9%, 20.5% and 25.2% relative error reduction, respectively, compared to those of HI-HMMs.

The word FHMM performs better than the phoneme FHMM because the construction of the transition matrix of the phoneme FHMM is based on the assumption that the noise duration and the phoneme duration are similar. However, in practice, this assumption does not necessarily hold. In our experiments, the average duration of Japanese phonemes was 65 ms and that of sudden noise samples was 143 ms. A careful design of the transition matrix taking into account variations on the duration of phonemes and noise is expected to reduce the performance difference between word and phoneme FHMMS.

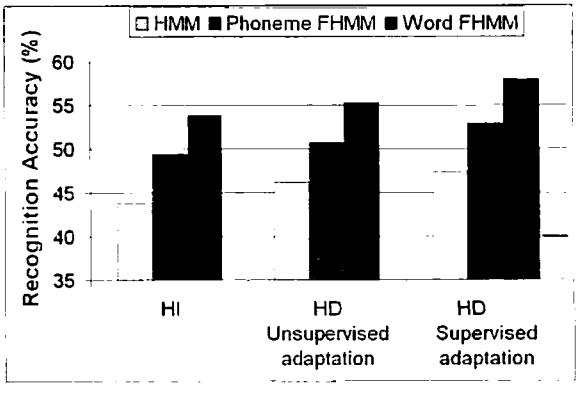

Figure 5: Averaged results of home environment adaptation

4. CONCLUSION AND FUTURE WORK

We investigated the impact of phoneme FHMM adaptation for speech recognition in the presence of nonstationary sud-

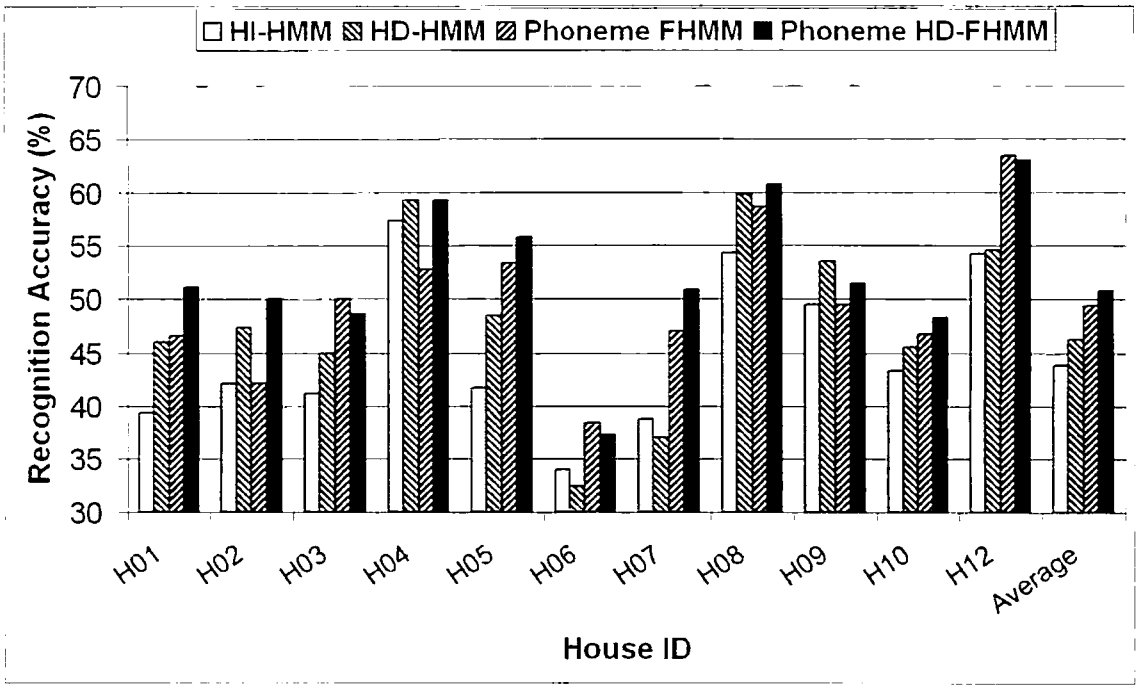

Figure 4: Results for unsupervised adaptation

den noise, which is very likely to be present in home environments. The proposed phoneme HD-FHMMs achieved better recognition accuracy than clean-speech HI-HMMs, reducing the overall relative error by 16.2% and 12.3% on average for supervised and unsupervised adaptation, respectively. Although phoneme HD-FHMMs did not outperform word HD-FHMMs in our experiments, they require significantly less training data and reduce the computational time of speech recognition compared with that of word FHMMs. Hence, phoneme FHMMs are more desirable candidates to be used in more complex tasks, especially in LVCSR systems.

We created a noisy phoneme FHMM by combining an HMM for clean speech and an HMM for noise, both of which have simple structures in this study. In addition, HMMs created with more complex structures (more Gaussians per state, different HMM topologies, and number of states) need to be investigated. 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.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Z. Ghahramani and M. I. Jordan, "Factorial Hidden Markov Models," Machine Learning, 29, pp. 245-275, 1997.

[2] N. A. Deoras and M. Hasegawa-Johnson, "A Factorial HMM Approach to Simultaneous Recognition of Isolated Digits Spoken by Multiple Talkers on One Audio Chanel." in Proc. ICASSP, pp. 861-864, 2004.

[3] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Transactions on Acoustic, Speech, and Signal Processing, vol. 34, pp. 52-59, 1986.

[4] A. Betkowska, K. Shinoda, and S. Furui, "Robust speech recognition using factorial HMMs for home environments." Eurasip Journal on Applied Signal Processing, in press, 2007.

[5] A. Betkowska, K. Shinoda, and S. Furui, "Speech recognition using FHMMs robust against nonstationary noise." in Proc. ICASSP 2007, to be presented.

[6] T. Iwasawa, S. Ohnaka, and Y. Fujita, "A Speech Recognition Interface for Robots using Notification of Ill-Suited Conditions," in Proc. of the 16th Meeting of Special Interest Group on AI Challenges, pp. 33-38, 2002.

[7] T. S. Roweis, "One Microphone Source Separation," Neural Information Processing Systems, vol. 13, pp. 793-799, 2000.

[8] B. Logan and P. Moreno, "Factorial HMMs for Acoustic Modeling," in Proc. ICASSP, pp. 813-816, 1998.

[9] K. Shinoda and T. Watanabe, "Speaker Adaptation with autonomous control using tree structure," in Proc. EuroSpeech95, pp. 1143-1146, 1995.