SPEECH RECOGNITION USING FHMMS ROBUST AGAINST NONSTATIONARY NOISE

Agnieszka Betkowska, Koichi Shinoda, and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology, Tokyo, Japan
{agabet,furui}@furui.cs.titech.ac.jp,shinoda@ks.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. 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. A database recorded by a personal robot called PaPeRo in home environments was used for the evaluation of the proposed method under noisy conditions. While we presented a recognition system using isolated-word FHMMs in our previous work, here we evaluated the effectiveness of the phoneme FHMMs.

Index Terms— speech recognition, robustness, speech enhancement, FHMM

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 *a priori* information about its SNR. The SNR changes from one sentence to another; it also changes within one sentence. At each moment, SNR depends on speaker, noise source, and robot position. Hence, preparing an appropriate PMC model for each utterance is practically impossible. We applied a model compensation method based on factorial hidden Markov models (FHMMs) that have been introduced as a possible extension of HMMs in [5] to solve 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. We also proposed an extension to employ dynamic features as well because the FHMM architecture proposed in [6] is applicable only to static features of speech signals. In our previous work [7], the proposed method was evaluated with the use of word FHMMs. Here, we discuss phoneme FHMMs to extend our work to large vocabulary continuous speech recognition (LVCSR) system. First, an HMM for each phoneme in the dictionary and an HMM for sudden noise are created. Then, these models are combined to create an FHMM for each phoneme. A database recorded by a personal robot called PaPeRo [8] in home environments was used for the evaluation of the proposed method. The experiments confirmed that our method improved the recognition accuracy under noisy conditions.

2. ROBUST SPEECH RECOGNITION USING FHMMS

2.1 FHMM

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 element-wise maximum of the output observations of the two layers is taken [9]. 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 spectrum of their mixture is almost the same as the element-wise maximum of the spectrums of these

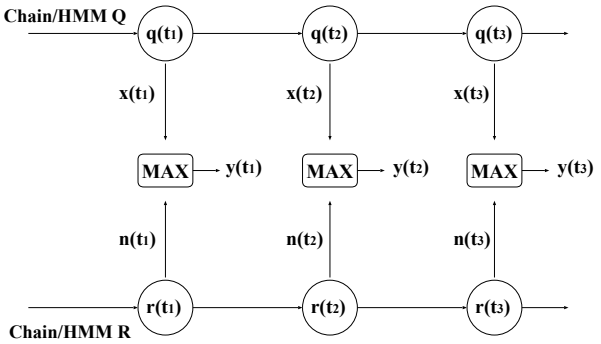

Fig. 1. Structure of FHMM composed of two HMMs, Q and R.

two signals. The spectrum 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 [6] :

$$\log|Y(j\omega)| = \max(\log|X(j\omega)|, \log|N(j\omega)|), \tag{1}$$

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) [6], 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.

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 [10]. Its transition matrix is defined by the Cartesian product between the transition matrices A_Q and A_R of HMMs Q and R, respectively [10]:

$$a_{(i,j)(k,l)} = a^Q_{ik} a^R_{jl}, \quad 1 \le i,k \le N, \quad 1 \le j,l \le W. \tag{2}$$

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:

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

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:

$$p_q(\boldsymbol{x}) = \sum_{m=1}^{M} c_{qm} N(\boldsymbol{x}|\boldsymbol{\mu}_{qm}, \boldsymbol{\Sigma}_{qm}) \quad \text{and} \tag{4}$$

$$p_r(\boldsymbol{n}) = \sum_{m=1}^{M} c_{rm} N(\boldsymbol{n}|\boldsymbol{\mu}_{rm}, \boldsymbol{\Sigma}_{rm}), \tag{5}$$

where M is the number of Gaussians in each state, $\boldsymbol{\mu}_{qm}$ and $\boldsymbol{\mu}_{rm}$ 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 $\boldsymbol{\Sigma}_{qm}$ and $\boldsymbol{\Sigma}_{rm}$ 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 $\boldsymbol{y}$ for metastate (q,r) of the FHMM is defined by [6]:

$$p_{(q,r)}(\boldsymbol{y}) = p_q(\boldsymbol{y}) F_r(\boldsymbol{y}) + p_r(\boldsymbol{y}) F_q(\boldsymbol{y}), \tag{6}$$

where

$$F_q(\boldsymbol{y}) = \sum_{m=1}^{M} c_{qm} \prod_{d=1}^{D} \int_{-\infty}^{y_d} p_q(x_d) dx_d \quad \text{and} \tag{7}$$

$$F_r(\boldsymbol{y}) = \sum_{m=1}^{M} c_{rm} \prod_{d=1}^{D} \int_{-\infty}^{y_d} p_r(n_d) dn_d. \tag{8}$$

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.

2.4 Extension of FHMM to dynamic features [7]

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:

$$\boldsymbol{\Delta y}_t = \frac{\sum_{\tau=1}^{G} \tau(\boldsymbol{y}_{t+\tau} - \boldsymbol{y}_{t-\tau})}{\sum_{\tau=1}^{G} \tau^2}, \tag{9}$$

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 [11].

The calculation of the output *pdf* defined in (6) 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 element-wise 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:

$$p'_{(q,r)}(\boldsymbol{y}, \boldsymbol{\Delta y}) = \begin{cases} p_{(q,r)}(\boldsymbol{y}) p_q(\boldsymbol{\Delta y}), \\ \qquad \text{if } p_r(\boldsymbol{y}) F_q(\boldsymbol{y}) < p_q(\boldsymbol{y}) F_r(\boldsymbol{y}), \\ p_{(q,r)}(\boldsymbol{y}) p_r(\boldsymbol{\Delta y}), \quad \text{otherwise}, \end{cases} \tag{10}$$

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 (6). The condition in (10) 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.

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 [8], which was used in the houses of 12 Japanese families (H01-H12). The whole database contains 74640 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. 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, which we call *recorded noisy speech*. Each sample is defined as the recording that consists of silence, uttered word (speech sample) or noise (noise sample), and silence. Samples were digitized at the 11,025 Hz sampling rate, and analyzed at a 10 msec 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. 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. A word HMM was designed for each entry in the vocabulary by concatenating the states of the silence HMM and triphone HMMs according to their corresponding sequence in the given entry. A noise 'word' 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. The word FHMM for a given word was created by combining the word HMM for clean speech and the noise 'word' HMM. Differently from word FHMMs, an phoneme FHMM that models speech and noise in parallel for a given phoneme was created by combining triphone HMMs for clean speech with the noise HMM, as described in 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. An evaluation test with 548 utterances at each SNR was prepared.

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 2. The FHMMs performed better than their corresponding clean-speech HMMs. The FHMMs defined only for static features

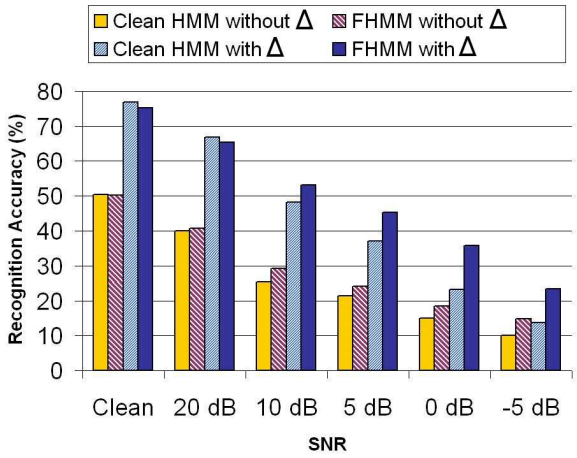

Fig. 2. Recognition rates of speech artificially corrupted by sudden noise. HMMs (baseline) and the proposed method (phoneme FHMMs) with and without Δ features.

improved the recognition accuracy by 4.7% absolute at -5 dB, by 3.6% absolute at 0 dB, by 2.7% absolute at 5dB, and by 4.0% 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 4.8% absolute at 10 dB, by 8.1% absolute at 5 dB, by 12.7% absolute at 0 dB, and by 9.7% absolute at -5 dB. As the SNR increased, however, the difference between the baseline clean-speech HMMs and the proposed FHMMs decreased, giving a slight 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 (10)], the wrong HMM model is used to calculate the *pdf* of Δ features of the clean speech signal. Hence, the initial error is amplified and is more difficult to correct.

Next, we evaluated the effectiveness of FHMMs in real conditions. 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. In our previous work [7], we created word FHMMs and showed their effectiveness in real conditions. Here, we compared the word FHMMs, clean speech HMMs and phoneme FHMMs with Δ features. The results are shown in Figure 3. The results given by clean-speech HMMs and word HMMs were the same. The phoneme FHMM exhibited better performance than that of clean speech HMMs, giving improvement ranging from 3.4% to 11.7% absolute (H10 to H05), respectively, for almost all houses except H04.

On average, phoneme FHMMs achieved 12.8% relative error reduction compared to that of clean speech HMMs. The word FHMMs performed better than phoneme FHMMs giving 17.9% of relative error reduction. It might be due to the fact that in phoneme FHMMs the noise duration is that of the phoneme, while in reality it is longer, hence word FHMM better models the phenonema. On the other hand, using phoneme FHMMs is more desirable, especially in large vocabulary recognition systems. Additionally, phoneme FHMMs re-

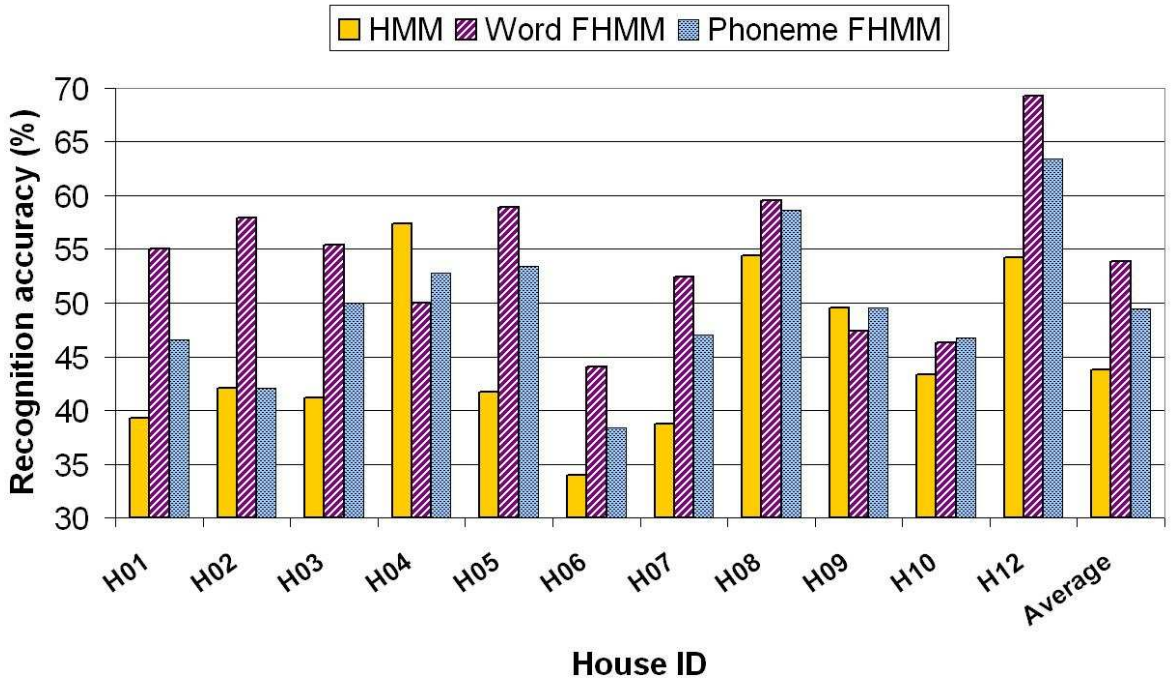

Fig. 3. Results for real recorded noisy speech data

duces the computational time which is quite high in the case of word FHMMs.

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. The usability of FHMMs was further investigated by using a recorded noisy speech test set. The overall relative error reduction given by phoneme FHMMs with Δ features was 12.8% compared to that given by the clean-speech HMMs.

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. HMMs created with more complex structures (more Gaussians per state, different HMMs 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. FHMMs for the combination of different noises should also be investigated.

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 permittion 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," Prince-Hall, 2001.

[2] M. Cook, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Communication, vol. 34, pp. 267-285, 2001.

[3] A. P. Varga and R. E. Moore, "Hidden Markov model decomposition of speech and noise," in Proc. ICASSP, pp.845-848, 1990.

[4] M. J. F Gales and S. J. Young, "HMM Recognition in Noise Using Parallel Model Combination," in Proc. EuroSpeech, pp.837-840, Berlin, 1993.

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

[6] 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.

[7] 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.

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

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

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

[11] 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.