SPEAKER NORMALIZED ACOUSTIC MODELING BASED ON 3-D VITERBI DECODING

Toshiaki Fukada *Yoshinori Sagisaka*

ATR Interpreting Telecommunications Research Laboratories
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Tel: +81 774 95 1301, FAX: +81 774 95 1308, E-mail: fukada@itl.atr.co.jp

ABSTRACT

This paper describes a novel method for speaker normalization based on a frequency warping approach to reduce variations due to speaker-induced factors such as the vocal tract length. In our approach, a speaker normalized acoustic model is trained using time-varying (i.e., state, phoneme or word dependent) warping factors, while in the conventional approaches, the frequency warping factor is fixed for each speaker. These time-varying frequency warping factors are determined by a 3-dimensional (i.e., input frames, HMM states and warping factors) Viterbi decoding procedure. Experimental results on Japanese spontaneous speech recognition show that the proposed method yields a 9.7 % improvement in speech recognition accuracy compared to the conventional speaker-independent model.

1. INTRODUCTION

Robust and precise acoustic modeling is an indispensable technique for achieving high recognition performance. In most current speaker-independent speech recognition systems, acoustic models are trained using a large amount of speech uttered by a wide variety of speakers. The spectral distributions often exhibit high variance and hence high overlap among different phonemes. Therefore, recognition performance saturates even if a number of mixtures and states are used or the context is increased. Consequently, research efforts have been conducted to reduce variations due to speaker-induced factors based on speaker normalization [1]~[6], speaker clustering[7] or hybrid methods [8]~[10]. In recent years, many researchers have been working on speaker normalization, since one of the major sources of interspeaker variance is the vocal tract length. Acoustic modeling based on speaker normalization techniques can be roughly divided into two approaches:

1. frequency warping (FWP) [3][4][6]
2. maximum likelihood linear regression (MLLR) [5].

In the conventional FWP-based approaches, the frequency warping factor is fixed for each speaker. That is, these approaches do not have a framework of phoneme or HMM state dependent frequency warping, while in the MLLR-based approach, it is possible to define regression classes and associate a regression matrix with each class. Also, a phoneme or allophone dependent warping procedure would be reasonable, when training speech samples are biased to a certain speaker or gender for some phoneme contexts or allophones.

In this paper, we present FWP-based acoustic modeling in which warping factors are dynamically changed during an utterance. These frequency warping factors are determined by a 3-dimensional (i.e., input frames, HMM states and warping factors) Viterbi decoding procedure. In the proposed method, the recognition procedure can be performed with a one-pass search, while in most current FWP-based approaches, a multiple-pass search is required at the recognition stage.

2. THE FRONT-END

2.1. Mel-cepstral Analysis

We represent the model spectrum $H(e^{j\omega})$ by the M-th order mel-cepstral coefficients $\tilde{c}(m)$ as follows:

$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m)\, \tilde{z}^{-m} \tag{1}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\,\alpha\,| < 1. \tag{2}$$

The phase characteristic of the all-pass transfer function $\tilde{z}^{-1} = e^{-j\tilde{\omega}}$ is given by

$$\tilde{\omega} = \tan^{-1} \frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}. \tag{3}$$

For example, for a sampling frequency of 16kHz, $\tilde{\omega}$ is a good approximation to the mel scale based on subjective pitch evaluations when $\alpha = 0.42$. If we choose $\alpha = 0.46$, the mel scale is quite similar to that used in mel-frequency cepstral coefficient (MFCC) analysis.

To obtain an unbiased estimate, we use the following criterion and minimize it with respect to $\{\tilde{c}(m)\}_{m=0}^{M}$.

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} \{\exp R(\omega) - R(\omega) - 1\}\ d\omega \tag{4}$$

where

$$R(\omega) = \log I_N(\omega) - \log \left|\, H(e^{j\omega}) \,\right|^2 \tag{5}$$

and $I_N(\omega)$ is the modified periodogram of a weakly stationary process $x(n)$ with a time window of length N. Since E is convex with respect to $\tilde{c}(m)$, $\{\tilde{c}(m)\}_{m=0}^{M}$ can be obtained by the Newton-Raphson method[11].

Table 1: Recognition performance. (Phone accuracy %)

feature parameter	without LM	with LM
LPC-cepstrum	44.0	52.2
MFCC	48.2	55.5
Mel-cepstrum	**49.6**	**56.4**

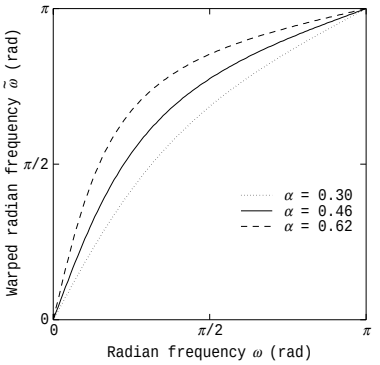

Figure 1: Frequency warping for several α values.

2.2. Effectiveness of Mel-cepstral Analysis

To show the effectiveness of mel-cepstral analysis, we performed simple recognition experiments using the TIMIT database. 3 state and 5 mixture context-independent HMMs (61 phone sets) were trained for LPC-cepstrum, MFCC and mel-cepstrum. A phoneme bigram was used for the language model (LM). Table 1 shows the differences in recognition performance for these three types of feature parameters. We can see from these results that mel-cepstral analysis is one of the most useful pre-processing methods for speech recognition.

2.3. Frequency Warping

Frequency warping can be done by changing α in Eq. (2). Figure 1 shows examples of frequency warping for several α values.

3. NORMALIZATION PROCEDURES

3.1. 3-D Viterbi Decoding

The key point of the proposed normalization procedures is to perform a Viterbi search on a 3-D trellis space composed of input frames, HMM states and warping factors (see Fig 2). Note that the conventional frequency warping based normalization is done by finding a warping factor for each speaker which yields the highest likelihood among all possible warping factors by a 2-D (i.e., input frames and HMM states) trellis search (see Fig 3).

3.2. Training Procedure

In the training stage, as transcriptions of speech are known, a transition of warping factors can be obtained by aligning the HMM states with the maximum likelihood criteria. The following procedure is used for acoustic model training:

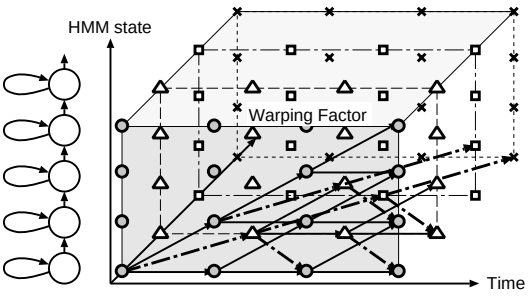

Figure 2: The proposed FWP scheme.

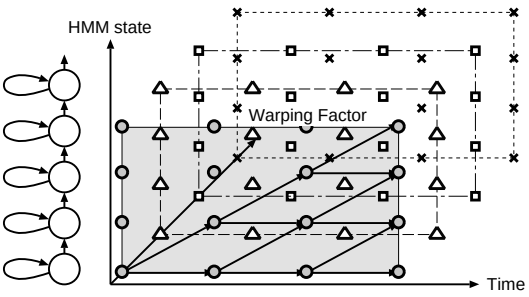

Figure 3: The conventional FWP scheme.

1. Set the initial warping factor to $\alpha = 0.46$, for all speakers and generate the initial HMM.
2. Align the training utterances based on 3-D Viterbi decoding using the current HMM and find the optimal warping factor for each HMM state.
3. Train an HMM using the feature vector sequence of the optimal warping factors
4. Go to step 2 until there is no significant change between consecutive training iterations.

In this procedure, we apply constraints to the 3-D Viterbi decoding procedure so as not to change the warping factor too rapidly (see **3.3**).

3.3. Recognition Procedure

In the recognition stage, a transition matrix of warping factors and a phoneme (or word) sequence of speech are obtained by finding an optimal path with the highest likelihood. Figure 4 shows the recognition algorithm. In this figure, $\mathbf{S}$, Q, D and N are the initial state sets, number of states, number of warping factors and number of frames, respectively. π, P, $a(q',q)$, $f(d',d)$, b and $\mathbf{x}$ are the initial state probability, accumulated probability, transition probability from state q' to q, transition probability from warping factor d' to d, output probability and feature vector, respectively.

In this paper, the transition probability of warping factor $f(d',d)$ is given as:

$$f(d',d) = \begin{cases} 1.0, & |d'-d| \leq w \\ 0.0, & |d'-d| > w. \end{cases} \quad (6)$$

Initialization:

for $q = 1$ **to** Q

  for $d = 1$ **to** D

    if $(q, d) \in \mathbf{S}$ **then**

$$P(q, d, 0) = \log \pi(q, d) \quad , \text{where} \sum_{(q,d)\in\mathbf{S}} \pi(q, d) = 1$$

    else

$$P(q, d, 0) = -\infty$$

Recognition:

for $n = 1$ **to** N

  for $q = 1$ **to** Q

    for $d = 1$ **to** D

$$P(q, d, n) = \max_{q', d'} \{P(q', d', n-1) + \log a(q', q) + \log f(d', d)\} + \log b(q, \mathbf{x}(d, n))$$

Figure 4: Recognition algorithm.

Here, we set to $w = 1$ for inter-phoneme state transitions and $w = 0$ for intra-phoneme state transitions (here denoted as FWP1). These constraints can be considered reasonable because the warping factor is not expected to change too rapidly. Note that the proposed speaker normalization procedure is equivalent to the conventional method (e.g. [3][4][6]) if $w = 0$ for any state transition (here denoted as FWP0).

4. EXPERIMENTS

To investigate the relative effectiveness of the proposed method, we conducted continuous speech recognition experiments on a Japanese spontaneous speech database[12].

4.1. Conditions

230 speakers were used for training and 42 speakers for evaluation. A 26-dimensional feature vector (12-dimensional mel-cepstrum + power and their derivatives) computed with a 25.6 msec window duration and a 10 msec frame period were used for acoustic modeling. First, shared-state HMMs (800 states in total) with 5 Gaussian mixture components per state[13] were trained by using an initial warping factor of $\alpha = 0.46$ for all speakers (gender-independent HMM; GI-HMM). Then, we generated two kinds of speaker normalized models (i.e., FWP0 and FWP1) described in **3**. As for the FWP0 training, the best warping factor was determined for each speaker. The normalization session described in **3.2** was repeated four times. The GI-HMM topology was consistently used for every iteration. For feature parameter sets, 9 kinds of warping factors ($D = 9$) were considered in steps of 0.04 from $\alpha = 0.30$ to 0.62. Gender-dependent HMMs (GD-HMM) were also used for comparison. We used spontaneous speech recognizer using cross-word context constrained word graphs[14]. The test vocabulary consists of about 7,000 words, and the variable-length N-gram[15] was used for the language model.

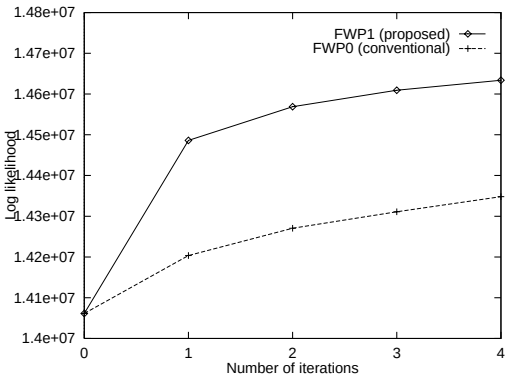

Figure 5: Log likelihood.

4.2. Comparison of Speaker Normalized Models

The increase of the total log-likelihood during the iterative acoustic model training can be seen in Fig. 5. The solid line shows the case for FWP1 and the dotted line for FWP0. The likelihood of iteration 0 is the likelihood of the GI-HMM. FWP1 yielded a consistently higher acoustic likelihood than FWP0 for each iteration. From these results, we can expect that the proposed speaker normalized model based on 3-D Viterbi decoding reduces interspeaker variability more than the conventional normalization method and results in a certain improvement in speech recognition.

The mean and standard deviation of the warping factors for each iteration are shown in Fig. 6 and Fig. 7, respectively. These statistics are calculated from the distribution (histogram) of the warping factors obtained from the 3-D Viterbi based alignment. We can see from these figures that the standard deviations of the proposed method (FWP1) are greater than those of the conventional method (FWP0). This is because the proposed method has a chance to vary the warping factor for each phoneme during the utterance, while the warping factor is fixed for each speaker in the conventional method.

4.3. Recognition Results

The recognition results are shown in Table 2. From these results, it can be seen that the proposed speaker normalized model (FWP1) yielded a better performance than GI-HMM, GD-HMM and the conventional speaker normalized model (FWP0). Recognition performances of multiple iterations are shown in Table 3. In this table, FWP0 and FWP1 with no iteration mean that 9 kinds of feature parameters were used as inputs and recognition was performed using the GI-HMM with $w = 0$ for FWP0 and $w = 1$ for FWP1. From this table, multiple training iterations improve the recognition performances for both FWP0 and FWP1 cases. It is also interesting that using the 3-D Viterbi based decoding procedure with the unnormalized model (i.e., GI-HMM) still gives a 6.1 % improvement over the GI-HMM (from 74.6 % to 76.2 %).

In our experiment the recognition result with the highest likelihood among the several frequency warping factors, is determined in a time-synchronous one-pass beam search.

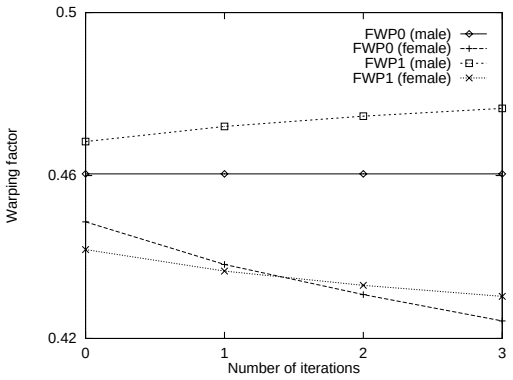

Figure 6: Mean of warping factors.

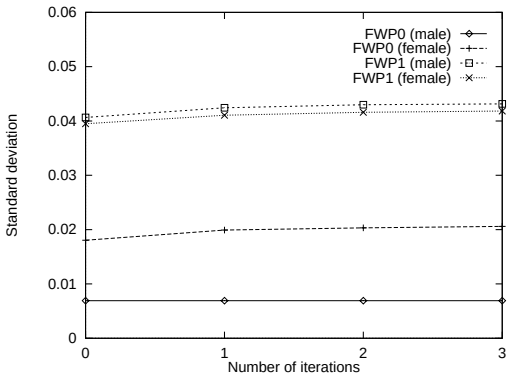

Figure 7: Standard deviation of warping factors.

The conventional speaker normalized model FWP0 (73.3 %) was surprisingly slightly worse than the GI-HMM (74.6 %). This could happen because we kept a constant beam-width, limited by memory requirements of the search engine, across all experiments, which produced search errors in the FWP0 case due to large local fluctuations in likelihood. Nevertheless, FWP1 achieved a 9.7 % improvement compared to the GI-HMM (from 74.6 % to 77.1 %). Note that in most current FWP-based approaches, a multiple-pass search is required at the recognition stage, while in the proposed method, the recognition procedure to find the best hypothesis and simultaneously select the best (time-dependent) warping factor can be performed with a one-pass search.

5. CONCLUSION

In this paper, we have proposed a new acoustic modeling technique based on a 3-D Viterbi decoding procedure which aims at normalizing speaker's variability. This method has a framework which makes it possible to vary the frequency warping factor with arbitrary units (i.e., state, phoneme, word, etc.) during an utterance. The conventional frequency warping based acoustic modeling can be viewed as a special case of the proposed modeling (i.e., $w = 0$ in Eq. (6)). The experimental results on spontaneous speech recognition showed that the proposed models yielded a 9.7 % improvement in word accuracy compared to the standard speaker-independent model.

Table 2: Word accuracy and relative improvement from GI-HMM (%).

acoustic model	accuracy	improvement
GI-HMM	74.6	—
GD-HMM	74.0	-2.3
FWP0	73.3	-5.0
FWP1	**77.1**	**9.7**

Table 3: Recognition performance improvements for FWP0 and FWP1 after four times of training iterations.

acoustic model	no. of iterations	accuracy (%)
FWP0	0	72.7
FWP0	4	73.3
FWP1	0	76.2
FWP1	4	77.1

6. REFERENCES

[1] F. -H. Liu, R. M. Stern, X. Huang and A. Acero: "Efficient cepstral normalization for robust speech recognition," *Proc. of DARPA Speech and Natural Language Workshop*, pp. 69-74 (1993).

[2] E. Eide and H. Gish: "A parametric approach to vocal tract length normalization," *Proc. ICASSP-96*, pp. 346-348 (1996).

[3] L. Lee and R. Rose: "Speaker normalization using efficient frequency warping procedures," *Proc. ICASSP-96*, pp. 353–356 (1996).

[4] S. Wegmann, D. McAllaster, J. Orloff and B. Peskin: "Speaker normalization on conversational telephone speech," *Proc. ICASSP-96*, pp. 339–341 (1996).

[5] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul: "A compact model for speaker-adaptive training," *Proc. ICSLP-96*, pp. 1137-1140 (1996).

[6] P. Zhan and M. Westphal: "Speaker normalization based on frequency warping," *Proc. ICASSP-97*, pp. 1039–1042 (1997).

[7] T. Kosaka, S. Matsunaga and S. Sagayama: "Speaker-independent speech recognition based on tree-structured speaker clustering," *Computer Speech and Language*, Vol. 10, pp. 55-74 (1996).

[8] M. Padmanabhan, L. Bahl, D. Nahamoo and M. Picheny: "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems," *Proc. ICASSP-96*, pp. 701–704 (1996).

[9] D. Pye and P. Woodland: "Experiments in speaker normalisation and adaptation for large vocabulary speech recognition," *Proc. ICASSP-97*, pp. 1047–1050 (1997).

[10] P. Zhan, M. Westphal, M. Finke and A. Waibel: "Speaker normalizaiton and speaker adaptation – A combination for conversational speech recognition," *Proc. EUROSPEECH-97*, pp. 2087–2090 (1997).

[11] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai: "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, pp. I-137–I-140 (1992).

[12] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura and Y. Sagisaka: "Japanese speech databases for robust speech recognition," *Proc. ICSLP-96*, pp. 2199–2202 (1996).

[13] M. Ostendorf and H. Singer: "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, Vol. 11, pp. 17–41 (1997).

[14] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka: "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," *Proc. ICASSP-96*, pp. 145–148 (1996).

[15] H. Masataki and Y. Sagisaka : "Variable-order n-gram generation by word-class splitting and consecutive word grouping," *Proc. ICASSP-96*, pp. 188–191 (1996).