SPEAKER-NORMALIZED HMM-LIKELIHOOD FOR SELECTING A REFERENCE SPEAKER IN SPEAKER-ADAPTIVE SPEECH RECOGNITION

Yasunaga Miyazawa and Shigeki Sagayama

ATR Interpreting Telephony Research Laboratories

ABSTRACT - This paper proposes a principle of speaker-normalized HMM-likelihood for preselecting a reference speaker to be used in speaker-adaptive HMM-based speech recognition. The experimental evaluation of this principle indicates that speaker selection using the speaker-normalized HMM-likelihood is superior to the simple likelihood-based method.

INTRODUCTION

The pre-selection of the reference speaker has been known as an effective method to improve the performance of speaker adaptive speech recognition, because the adaptation performance depends on the relationship between the characteristics of the reference speaker and of the input speaker. For example, adaptation is not easy between male and female speakers.

The difference in static characteristic patterns between speakers, i.e. the distortion of the vector quantization, has been efficiently used as the distance measure for the pre-selection (Sugiyama and Shikano,1986). A speaker selection method based on HMM output likelihoods has also been reported (Rosenberg,1991) as an effective method. In these conventional methods, a reference speaker is selected as the most suitable speaker for the input speaker before speaker adaptation. However, the characteristics of the reference speaker model before and after the speaker adaptation are different. We think that comparing the reference speaker model after the adaptation with the input speech data of the input speaker is more reasonable for the pre-selection of the reference speaker. In order to do this, we use the **speaker-normalized HMM-likelihood.**

In the following section, details of a speaker pre-selection method using the **speaker-normalized HMM-likelihood** are given, and this method using the **vector field smoothing method** (Ohkura and Sagayama, 1992) for speaker adaptation is evaluated by phoneme recognition experiments using 12 male reference speakers and 10 other male input speakers.

SPEAKER PRE-SELECTION USING SPEAKER-NORMALIZED HMM LIKELIHOOD

Difference in speaker characteristics

Figure 1 shows the average LPC-cepstral parameter vectors of five Japanese vowels spoken by three speakers, plotted on the first and second principle component subspace. The pentagons represent the acoustic structures of these three speakers. In conventional speaker selection based on the absolute distances between speakers, reference speaker B would be chosen for input speaker X, although speaker A's structure seems to better fit speaker X's structure through speaker adaptation.

In our proposed speaker selection based on the **speaker-normalized HMM-likelihood**, reference speaker A was actually chosen for input speaker X, and in the conventional speaker selection based on the HMM output likelihood of the reference speaker before adaptation, reference speaker B was chosen. The next subsection explains the principle of speaker selection based on the **speaker-normalized HMM-likelihood.**

Speaker-normalized HMM-likelihood

In this work, Hidden Markov Models (HMMs) are used as the acoustic models of the reference speaker. If the acoustic characteristics of the reference speaker fit well the input speaker's through a speaker adaptation procedure, the resulting likelihood is expected to be large.

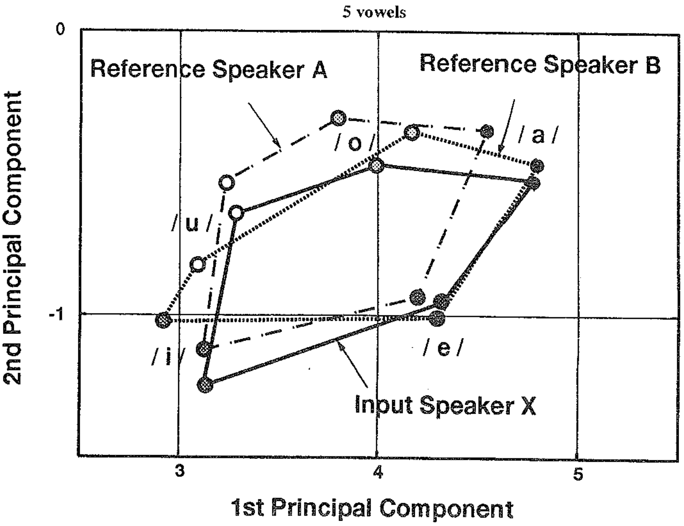

Figure 1: Five Japanese vowels uttered by three speakers, represented by average LPC-cepstral parameter vectors, and plotted on the first and second principle component subspace

Therefore, we define a speaker selection criterion with the likelihood after speaker adaptation. We call this criterion the **speaker-normalized HMM-likelihood**. Speaker selection is done so that the reference speaker gives the highest **speaker-normalized HMM-likelihood**. This reference speaker selection is applicable to any kind of speaker adaptation requiring speaker pre-selection.

Algorithm for pre-selection of reference speaker

In this section, the algorithm of the reference speaker pre-selection method using the **speaker-normalized HMM-likelihood** is described as follows and illustrated in Fig. 2.

Algorithm for Pre-selection of Reference Speaker

1. Adapt each of all reference speaker HMMs to the input speaker using the training speech data
2. Evaluate speaker-normalized HMM-likelihood by using the training speech data
3. Select the reference speaker whose speaker-normalized HMM likelihood is the largest

Vector field smoothing method

In this work, we use our **Vector Field Smoothing method** for the speaker adaptation method to evaluate the **speaker-normalized HMM-likelihood** principle. In the following, the principle of the **Vector Field Smoothing method** is explained and shown in Fig. 3 .

A speaker adaptation method can be regarded as a retraining problem, where a limited amount of training data is available. To retrain the HMM, the following two problems must be addressed:

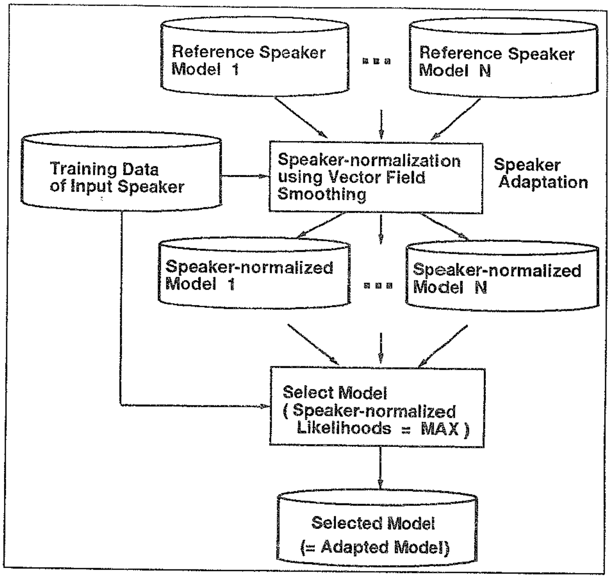
Reference Speaker Model 1
Reference Speaker Model N
Training Data of Input Speaker
Speaker-normalization using Vector Field Smoothing
Speaker Adaptation
Speaker-normalized Model 1
Speaker-normalized Model N
Select Model (Speaker-normalized Likelihoods = MAX)
Selected Model (= Adapted Model)

Figure 2: Speaker selection using speaker-normalized likelihood and simple likelihood

Problem 1 : Some phonemes may be missing in the typically small amount of training data available. Therefore, some phoneme CDHMMs may not be trained.

Problem 2 : Errors in estimating CDHMM parameters can result from insufficient training data.

The vector field smoothing method is applied in order to deal with these general retraining problems related to insufficient training data. The vector field smoothing method is carried out in three steps, i.e. **Concatenation Training**, **Interpolation** and **Smoothing**. First, concatenation training trains the mean vectors of the Gaussian distributions in phoneme CDHMMs. Second, interpolation is carried out by transferring the untrained mean vectors to the unknown speaker's voice space which employs an interpolated transfer vector. This transfer vector is interpolated by looking at the differences between mean vectors in CDHMMs before and after training. **Problem 1** is solved in this step. However, since the training data may not represent the true distribution of each phoneme, the transfer vectors calculated by these estimated distributions may be disorderly. To solve **Problem 2**, the third step of **smoothing** is applied. The vector field is smoothed in accordance with the restored directions of the transfer vectors. Here, the smoothness of the vector field is controlled by the value of **fuzziness**. Increasing the value of **fuzziness** makes the vector field smoother and the transfer vectors more parallel.

RECOGNITION OF 23 PHONEMES

Experimental conditions

The speaker pre-selection method based on the **speaker-normalized HMM-likelihood** is evaluated using the **Vector Field Smoothing method** and the ATR speech database. Each of the HMMs of the 12 male reference speakers was adapted to each of the 10 male input speakers'. The feature was a 34-dimensional vector consisting of 16 cepstral coefficients, 16 Δcepstral coefficients, logarithmic power and Δlogarithmic power. The analysis conditions are listed in Table 1. The HMM of the reference speaker consisted of four states, three loops and three mixtures per state. Each of the mixture components

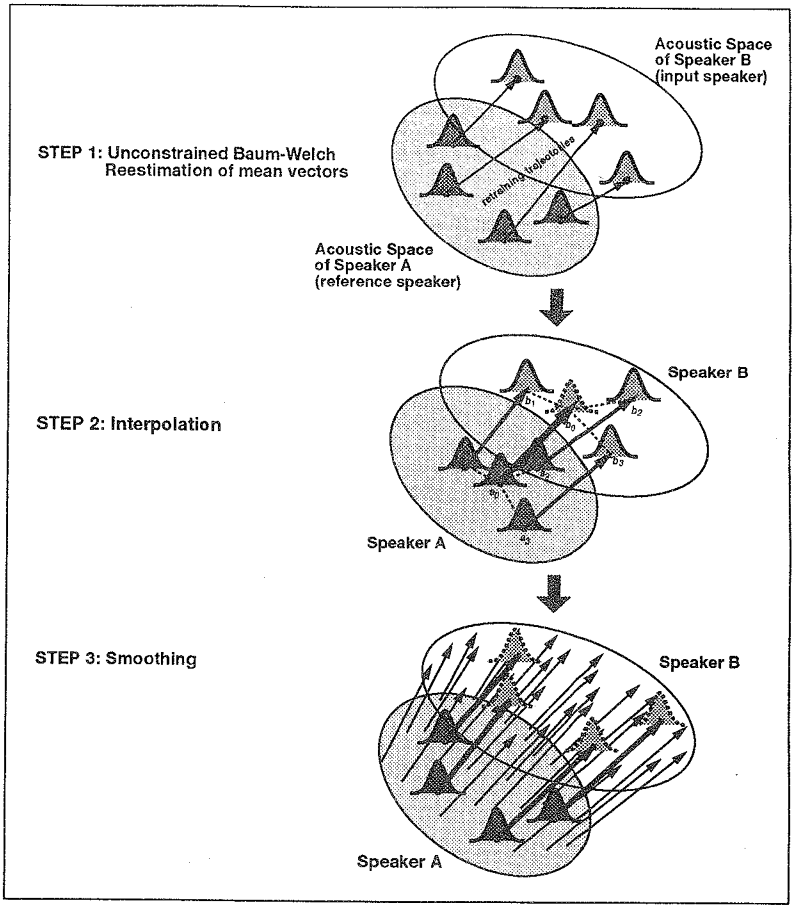

Figure 3: The vector field smoothing method (VFS)

Table 1: Analysis conditions

pre-emphasis	$1 - 0.98z^{-1}$
window length	20 ms (Hamming window)
window shift	5 ms
LPC analysis order	16
LPC cepstrum order	16
Δ window length	50 ms

had a diagonal covariance matrix. The phoneme HMMs of the reference speaker were trained using 720 isolated words and the speaker independent phoneme HMMs were trained using all isolated words (8640 words) of the 12 male speakers.

Experiments

This speaker selection principle was experimentally tested in comparison with the conventional selecting method, a speaker-independent model and a case not involving speaker selection. First, the phone models of the reference speakers were represented by continuous mixture density HMMs and were trained with about 720 isolated words. Second, through the speaker adaptation technique of the **Vector Field Smoothing** (Ohkura and Sagayama, 1992), they were adapted to each input speaker using 20, 50, 100 and 200 isolated words as speaker adaptation data. Finally, the speaker-adapted phone models were tested on the speech recognition of 23 Japanese phonemes extracted from about 2600 isolated words of the input speaker.

Table 2 shows the recognition rates of 23 Japanese phonemes. These results show that speaker selection based on the **speaker-normalized HMM-likelihood** consistently performs better than simple likelihood-based methods and methods without speaker pre-selection. It also performs better than speaker-independent phone models if more than 50 words are used for adaptation.

In Fig. 1, reference speaker A was chosen for input speaker X using this speaker selection based on the **speaker-normalized HMM-likelihood,** and reference speaker B was chosen using the conventional speaker selection based on the HMM output likelihood before adaptation. Intuitively, speaker A's structure seems to better fit speaker X's structure. When 50 isolated words of speaker X are used for the adaptive training, the phoneme recognition rate of speaker A was 81.9% and that of speaker B was 78.9%.

Table 2: Recognition rates (%) of 23 Japanese phonemes for different speaker pre-selection methods

Method of Speaker Selection	Number of Training Words for Speaker Adaptation				
	0	20	50	100	200
none	62.8	76.3	80.7	82.9	84.5
simple likelihood	76.4	79.9	83.3	85.5	86.8
speaker-normalized likelihood	76.4	80.3	83.9	86.3	87.4
speaker-independent HMM	81.6				

CONCLUSION

This paper described a principle of speaker pre-selection based on the **speaker-normalized HMM-likelihood.** This method was evaluated on the recognition of 23 Japanese phonemes by using the **vector field smoothing method** for the speaker adaptation. This evaluation indicated that speaker selection using the **speaker-normalized HMM-likelihood** is superior to the simple likelihood-based methods.

In the future, this method will be evaluated for phrase recognition and for a large amount of reference speakers.

ACKNOWLEDGMENTS

The authors wish to thank Dr. A. Kuramatsu for his support of this work. We would also like to acknowledge Mr. K. Ohkura for his useful advice.

REFERENCES

Ohkura, K. and Sagayama, S. (1992) *Speaker Adaptation based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs*, submitted to ICSLP92,

Rosenberg, A.E. et al. (1991) *Connected Word Talker Verification Using Whole Word Hidden Markov Models*, Proc. ICASSP91, pp. 381-384, 57.S6.4.

Sugiyama, M. and Shikano, K. (1986) *Unsupervised Learning Algorithm for Vowel Templates Based on Minimum Quantization Distortion*, Review of Electri. Commun. Lab., Vol.34, No.3 , pp.357-362.