A Comparative Study of Speaker Adaptation Methods for HMM-Based Speech Recognition

(HMM 음성인식 시스템을 위한 화자적응 방법들의 성능비교)

(Myoung Wan Koo, Chong Kwan Un, Hwang Soo Lee)

구 명 완* 은 종 관* 이 황 수*

ABSTRACT

In this paper, we compare the performances of speaker adaptation algorithms which consist of two stages of processing for an HMM-based speech recognition system. We compare three kinds of VQ adaptation methods which may be used in the first stage to reduce the total distortion error for a new speaker : label prototype adaptation, adaptation with a codebook from adaptation speech itself, and adaptation with a mapped codebook. We then compare the performance of four kinds of HMM parameter adaptation methods which may be used in the second stage to transform HMM parameters for a new speaker : adaptation by the Viterbi algorithm, that by the DTW algorithm, that by the iterative alignment algorithm, and that by the fuzzy histogram algorithm. The results show that adaptation based on the fuzzy histogram algorithm yields the highest accuracy in an HMM-based speech recognition system.

요 약

본 논문에서는 HMM을 이용한 음성인식 시스템에서 2단계로 이루어지는 화자적응 알고리즘의 성능비교를 수행하였다. 첫단계는 새로운 화자와의 거리차이를 줄여주는 VQ 적응방식들로 구성되는 이 방식들 중에서 lable prototype 적응, 적응음성으로부터 구성된 VQ 코우드 북을 사용한 적응 및 사상 코우드 북을 사용한 적응등의 알고리즘 성능비교를 하였다. 두번째 단계는 새로운 화자를 위해 HMM 파라미터를 변환시켜주는 HMM 파라미터 적응방식들로 이루어지는데 이 방법들 중에서 Viterbi 알고리즘, DTW 알고리즘, iterative alignment 알고리즘 및 fuzzy histogram 알고리즘의 성능을 비교하였다. 성능비교 결과 fuzzy histogram 알고리즘에 의한 화자적응 방식이 최고의 인식율을 나타내었다.

1. Introduction

A Hidden Markov model(HMM) typically requires a large amount of input speech to obtain reliable probability estimates for training the recognition system. In practice, however, it is not usually possible to get such large speech data for training. To remedy this problem a speaker adaptation procedure is used. The purpose of speaker adaptation is to yield acceptable recognition performance even for speakers who have not provided

*Center for Speech Information Research
Korea Advanced Institute of Science and Technology

enough speech to train the HMMs.

There are many kinds of speaker adaptation procedures. These can be divided into two categories : supervised and unsupervised adaptation according to whether the short size of training speech (adaptation speech) is already known or not. Here, we focus on the supervised adaptation procedure because of its good performance. One of the supervised adaptation procedures is the spectral mapping technique in which codebook adaptation and HMM parameter adaptation are done. In the stage of codebook adaptation, a codebook is generated for a new speaker, and in the stage of HMM parameter adaptation, welltrained HMMs are transformed from a prototype speaker to a new speaker. Recently, the label prototype adaptation for codebook adaptation and the use of the Viterbi algorithm for HMM parameter adaptation were studied.[1] Also, the codebook mapping algorithm for codebook adaptation and the use of dynamic time warping (DTW) for HMM parameter adaptation were investigated.[2] In addition, the codebook generation from adaptation speech itself for codebook adaptation and the iterative alignment improvement in the DTW algorithm for HMM parameter adaptation were considered[3][4]

In this paper, we perform a comparative study of speaker adaptation methods in which two-stage adaptation is done for HMM-based speech recognition. We first describe the two-stage speaker adaptation system, and then consider our HMM-based speech recognition system. Then, we provide experimental results of various speaker adaptation methods.

2. Two-stage speaker adaptation

2.1 Codebook adaptation

So far, several codebook adaptation methods for a new speaker have been proposed. These may be categorized as follows. One method is the label prototype adaptation in which the original codewords are modified by the K-means algorithm. The adaptation data for a new speaker is encoded by using the original codebook, and then all the spectral feature vectors with the same codeword are averaged. The original codewords are replaced by the newly averaged feature vectors. This procedure is continued until the feature vectors converge to some prespecified values.[1] Another is the codebook generation from adaptation speech itself.[4] The last one is the codebook mapping method in which a mapped codebook $k_i^{(A\to B)}$ is obtained from the adaptation speech as

$$k_i^{(A)} \to k_i^{(A\to B)} = \sum_{j=1}^{c} h_{ij} \cdot k_j^{(B)} / \sum_{j=1}^{c} h_{ij}, \qquad (1)$$

where $k_i^{(A)}$ and $k_j^{(B)}$ are code vectors of the new speaker and the reference speaker, respectively, h_{ij} is a histogram of code vector correspondences obtained by DTW, and c is the size of the codebook. [2]

2.2 HMM parameter adaptation

In HMM parameter adaptation methods based on the spectral mapping, the trained HMM parameters for the old speaker are transformed for a new speaker by using two kinds of mapping functions. One is based on the Viterbi algorithm and the other is based on the DTW algorithm.

Let $\mathbf{P}(s)$ represent the discrete probability density function(pdf) defined over a fixed set N of spectral templates. It is given by $\mathbf{P}(s)=[p(k_1|s), p(k_2|s), \cdots, p(k_N|s)]$, where $p(k_i|s)$ is the probability of spectral template k_i at state s of the HMM. If we denote the quantized spectrum from the original speaker as k_i, $1 \leqq i \leqq N$, and that

from the new speaker as $\hat{k}_j$ $1 \leqq j \leqq N$, where i and j are indices denoting the quantized spectra, the probability that the new speaker produces a new quantized spectrum $\hat{k}_j$ is given by

$$p(\hat{k}_j/s) = \sum_{i=1}^{N} p(k_i/s)\, p(\hat{k}_j/k_i), \tag{2}$$

where the probability for spectrum $\hat{k}_j$, given k_i, is assumed to be independent on s. The mapping probability $p(\hat{k}_j/k_i)$ $(= T_{ij})$ for all i and j forms an $N \times N$ matrix T, which can be interpreted as a probabilistic transformation matrix from one speaker's spectral space to another in each state. Then the discrete pdf $\mathbf{P}(s)$ at state s for a new speaker can be computed as $\mathbf{P}(s)=\mathbf{P}(s)\,\mathbf{T}$. Since the transformation between speakers cannot be modeled by a single transformation, we transform phoneme-dependently the speech of one speaker to that of another by using a different matrix $\mathbf{T}(ph)$ for all the HMM models involving a phoneme ph. In practice, we interpolate the phoneme-dependent transformation matrix $\mathbf{T}(ph)$ with the phoneme-independent transformation matrix $\mathbf{T}$ to get

$$\mathbf{T}_{ip} = \lambda(N_{ph})\mathbf{T}(ph) + [1 - \lambda(N_{ph})]\mathbf{T}, \tag{3}$$

where $\mathbf{T}_{ip}$ indicates the interpolated transformation matrix, and $\lambda(N_{ph})$ depends on the number of observed frames of phoneme ph, N_{ph}. The mapping probability $p(\hat{k}_j/k_i)$ is estimated by counting the correspondence between the VQ codeword indices of adaptation speech and those of reference speech. The correspondence between the adaptation speech and the reference speech can be obtained by Viterbi alignment or DTW.

Let the Viterbi alignment between a state s and a VQ codeword index k for adaptation speech data w, $1 \leqq w \leqq W$, at frame t, $1 \leqq t \leqq T$, be $s=V[k(w,t)]$. Probabilistic count $C(\hat{k}_j, k_i)$ of the correspondences between $\hat{k}_j$ and k_i are computed for all w and t as

$$C(\hat{k}_j, k_i) = \sum_{k(w,t)=\hat{k}_j} P(k_i/V[k(w,t)]), \quad 1 \leqq w \leqq W,\ 1 \leqq t \leqq T. \tag{4}$$

When the DTW algorithm is used, the counts $C(\hat{k}_j, k_i)$ of the correspondences are obtained from the accumulation of the code vector correspondences along the DTW optimal path. Then, the probability of $\hat{k}_j$, given k_i, is obtained as

$$P(\hat{k}_j/k_i) = C(\hat{k}_j, k_i) / \sum_{\hat{k}_j=1}^{N} C(\hat{k}_j, k_i). \tag{5}$$

When the spectral space of the new speaker is significantly different from that of the old speaker, the alignment produced by the DTW algorithm may not correspond to the phonetically correct alignment. In that case, an iterative alignment algorithm can be used in order to improve the accuracy of the alignment. In each iteration of the algorithm, the adaptation speech is aligned to the reference speech, and each spectral frame of the adaptation speech is shifted by an amount that is dependent on the index of the corresponding VQ value of its aligned reference frame. Using a shifted spectral frame of adaptation, the adaptation speech is realigned to the reference speech. This procedure is continued until the mean-squared error(mse) of the alignment is similar to that of the alignment resulting from the

previous iteration.[3]

The fuzzy VQ can also be adopted to calculate a correspondence histogram in order to improve the dependence on training vocabularies.[2][6] Fuzzy VQ represents an input vector X by using a weighted linear combination of VQ code vectors, $\hat{X}$. The fuzzy membership function u_i is obtained as

$$u_i = 1/[\sum_{j=1}^{k} (d_i/d_j)^{1/(m-1)}], \quad (6)$$

and an input vector can be calculated with the fuzzy membership function as

$$\hat{X} = \sum_{i=1}^{k} [(u_i)^m \cdot k_i] / \sum_{i=1}^{k} (u_i)^m, \quad (7)$$

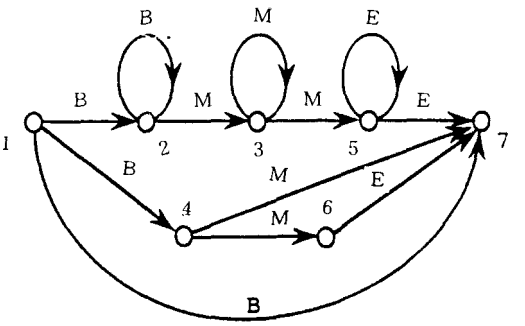
where $d_i = ||X - k_i||$, k_i is a code vector in codebook K and the values of k and fuzziness m are fixed as 6 and 1.6, respectively. Here a fuzzy histogram algorithm is utilized in order to have the fuzzy membership functions. This algorithm uses the mapped codebook obtained by (1), but the code vector correspondence, k_j, k_i, along the DTW optimal path is calculated as

$$C(\hat{k}_j, k_i) = u_j^{(A)} \cdot u_i^{(B)}, \quad (8)$$

where $u_j^{(A)}$ and $u_i^{(B)}$ are the fuzzy membership functions for the new speaker A and the original speaker B, respectively, and $C(k_j, k_i)$ is the probabilistic count given by(4).

3. Description of the baseline system

We established a speaker-dependent baseline system based on phoneme-like subword units (47 units). We used Korean speech which consisted of phonetically balanced 100 word vocabulary. Average number of syllables in the vocabulary was 1.6. Ten repetitions were pronounced by a male speaker who was designated as a reference speaker. The speech was sampled at 10 kHz and represented every 10 msec by three sets of parameters : (1) mel-scaled cepstral coefficients ; (2) their differential coefficients ; and (3) log power and differential log power. These parameters are vector quantized separately into three codebooks, each with 256 entries. The HMM subword model has 7 states and 12 transitions with three output probabilisty density functions(pdf's) as shown in Fig. 1. This subword model is similar to that of Lee's.[5] Three iterations of the forward-backward algorithms on the four repetitions of the manually segmented 100 word vocabulary were run and smoothed by the co-occurrence method in order to initialize our phoneme-like units.[5] We modeled each phone with a context-independent HMM after runing another five iterations of the forward-backward algorithm on those vocabulary and another three repeitions of 100 word by words, recognition of 96.3% was 0.0 braned from the recognition lest with other wets of 100 words.



B,M,E : output pdf's

Fig. 1. HMM subword model

4. Evaluation experiments

Speaker adaptation experiments were conducted for two male and one female speakers excluding the reference speaker. Each speaker pronounced 100 words four times. We used one repetition of 100 words as adaptation speech and three repetitions of 100 words as test words. We compared the performances of the speaker adaptation algorithms with two-stage adaptation methods.

4.1 Comparison of codebook adaptation methods

Three kinds of codebook adaptation methods were compared. In order to compare the performance of codebook adaptation methods, HMM parameter adaptation based on the DTW algorithm was done together. The results are shown in Table 1. The label prototype adaptation method yielded the recognition rate of 89.0% for the first candidate (called the top 1), 95.8% for the first and second candidates (called the top 2). The adaptation with a codebook generated from the adaptation speech itself gave the recognition rate of 89.7% for the top 1, 96.3% for the top 2, and that with a mapped codebook yielded 89.9% for the top 1, and 96.4% for the top 2. These results show that it is desirable to make a codebook from adaptation speech itself for a new speaker, and that codebook adaptation with a mapped codebook yields better recognition rate.

4.2. Comparison of HMM parameter adaptation methods

In order to find a proper scheme for HMM parameter adaptation, we first compared the performance of adaptation by the Viterbi algorithm with that by the DTW algorithm. The codebook for a new speaker was generated from adaptation speech itself. The results are shown is Table 2. The adaptation by the Viterbi algorithm gave the recognition rate of 88.1% for the top 1, and 95.8% for the top 2, and that by the DTW algorithm yielded the recognition rate of 89.7% for the top 1, and 96.3% for the top. 2 From these results, it can be concluded that the DTW algorithm is better than the Viterbi algorithm for HMM parameter adaptation. Next, we evaluated adaptation by the iterative alignment algorithm and that by the fuzzy histogram algorithm. These results are

Table 1. Performance comparison of codebook adaptation methods : label prototype adaptation, adaptation with a codebook generated from the adaptation speech, and adaptation with a mapped codebook
(a) reference →male 1 b) reference →male 2
(c) rdference →female (b) average

adaptation method	recognition rate(%)							
	(a)		(b)		(c)		(d)	
	top 1	top 2	top 1	top 2	top 1	top 2	top 1	top 2
without adaptation	58.3	77.3	49.7	64.3	41.0	52.0	49.7	64.6
label prototype adaptation	92.7	98.3	87.7	95.0	87.0	94.0	89.0	95.8
codebook from adaptation speech	93.0	98.3	89.7	95.7	86.3	95.0	89.7	96.3
mapped codebook	93.7	98.0	88.3	95.7	87.7	95.7	89.9	96.4

Table 2. Performance comparison of HMM parameter adaptation methods : adaptation by the Viterbi algorithm, adaptation by the DTW algorithm, adaptation by the iterative alignment algorithm, and adaptation by the fuzzy histogram algorithm
(a) reference →male 1 b) reference →male 2
(c) rdference →female (b) average

adaptation method	recognition rate(%)							
	(a)		(b)		(c)		(d)	
	top 1	top 2	top 1	top 2	top 1	top 2	top 1	top 2
by the Viterbi	93.3	99.3	82.3	91.7	88.7	96.3	88.1	95.8
by the DTW	93.0	98.3	89.7	95.7	86.3	95.0	89.7	96.3
by the iterative alignment	94.0	98.3	89.3	96.0	87.0	95.0	90.1	96.4
by the fuzzy histogram	94.3	99.0	89.3	95.0	87.7	96.3	90.4	96.4

also in Table 2. The adaptation by the iterative alignment algorithm produced the recognition rate of 90.1% for the top 1 and 96.4% for the top 2 , and that by the fuzzy histogram algorithm yielded the recognition rate of 90.4% for the top 1 and 96.4% for the top 2. From these results, it can be concluded that the DTW algorithm is again better than the Viterbi algorithm for HMM parameter adaptation, and that adaptation algorithm by the fuzzy histogram results in the best recognition tate.

5. Conclusion

We have presented comparative performance results of speaker adaptation algorithms for an HMM-based speech recognition system. A baseline system with the recognition tate of 96.3% was first established. With this, we investigated two-stage speaker adaptation methods. First, we studied three kinds of VQ adaptation methods that may be used for the reduction of the total distortion error for a speaker in the first stage ; label prototype adaptation, adaptation with a codebook from adaptation speech itself, adaptation with a mapped codebook. Then we compared four kinds of HMM parameter adaptation methods that may be used in the second to transform HMM parameters for a new speaker : adaptation by the Viterbi algorithm, that by the DTW algorithm, that by the iterative alignment algorithm, and that by the fuz zy histogram algorithm. The experiments showed that adaptation based on the fuzzy histogram algorithm yielded the best result for speaker adaptation in an HMM based speech recognition system.

References

1. M. Nishimura and K. Sugawara, "Speaker adaptation method for HMM-based speech recognition," *Proc. ICASSP 88*, Paper S5.7, 1988.
2. S. Nakamura and K. Shikano, "Speaker adaptation applied to HMM and neural networks," *Proc. ICASSP 89*, Paper S3.3, 1989.
3. M. Feng, "Iterative normalization for speaker-adaptive training in continuous speech recognition," *Proc. ICASSP 89*, Paper S12.4, 1989.
4. M. Feng, "Improved speaker adaptation using text dependent spectral mappings," *Proc. ICASSP 88*, Paper S3.3, 1988.

5. K.F Lee and H.W. Hon, "Speaker-independent phone recognition using hidden Markovmodels," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-37, pp. 1641-1648, 1989.

6. T. Hanazawa et al., "ATR HMM-LR continuous speech recognition system," *Proc. ICASSP 90*, Paper S2.4, 1990.

Koo Myoung WAN

Education : 82.2 B.S, Electronic Engineering, Yonsei University
85.2 M.S, Electrical and Electronics Engineering, Korea Advanced Institute of Science and Technology
91.8 Ph. D, Electrical and Electronics Engineering, Korea Advanced Institute of Science and Technology

Experience : 85.4 ~ Research engineer, Korea Telecom Research Center

Fields of Interest : Speech recognition system, Neural network, Speech coding and synthesis, Switching system software

▲Chong Kwan Un

Professor, Electrical and Electronics Engineering, Korea Advanced Institute of Science and Technology (Vol. 9 No. 4)

▲Hwang Soo Lee

Associate Professor, Electrical and Electronics Engineering Korea Advanced Institute of Science and Technology (Vol. 6 No. 3)