IMPROVEMENT OF SPEAKER RECOGNITION SYSTEM BY INDIVIDUAL INFORMATION WEIGHTING

Se-Hyun Kim, Gil-Jin Jang, and Yung-Hwan Oh
Division of Computer Science, Dept. of EECS, KAIST
373-1, Kusong-dong, Yusong-gu, Taejon 305-701, Korea
shkim@bulsai.kaist.ac.kr
http://bulsai.kaist.ac.kr

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ABSTRACT

In speaker recognition, it is very important to use individual information extracted from speech waves. Most of the speaker recognition methods assume that each part of speech has equal amount of information to represent a speaker, although it differently contribute to speaker recognition. The aim of this paper is to suggest a new scoring method of the HMM, which applies different importance to all the basic portions of a sampled speech waveform. we first define the quantity of the importance of speech frames, propose how to measure it and apply to speaker recognition. The performance of the proposed method was compared to non-weighting HMM based speaker recognition system. In speaker verification experiments, the proposed method reduced equal error rates considerably as compared to a conventional method which treats all speech segments to have the same importance. In speaker identification experiments, the proposed method marked relatively 28% higher recognition rate than the baseline system, and was more robust in long-term variation. These results demonstrate that the proposed method is efficient in measuring speaker information and more appropriate for speaker recognition.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique will make it possible to verify the identity of persons accessing systems, that is, access control by voice in various services like door lock system or the authentication in Electronic Commerce. Speaker recognition task is conducted by measuring and comparing speaker information. So, It is very important that using individual information extracted from speech waves. Speaker recognition can be divided into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. speaker recognition methods can also be divided into text-dependent and text-independent methods [4]. The problem in this paper is restricted to text-dependent domain, where text of speech utterance is known and fixed. Generally, most of the text-dependent speaker recognition methods assume that each part of speech has an equal amount of information to represent a speaker, although it contributes differently to speaker recognition. In this paper we suggest a new scoring method for speaker recognition system. We first define individual information quantity and measure it each frame of speech waves. Using measured quantity, we select frames as the reflection of the speaker characteristics. In scoring process, we give higher weighting factors to such frames and calculate new score weighting it the output probability of HMM. Section 2. provides a background to HMM based speaker recognition system. Section 3. describes the new scoring methods, section 4. gives a description of the experiment and the result. Finally we conclude this paper in section 5.

2. SPEAKER RECOGNITION SYSTEM

The basic structure of the speaker recognition system follows. In speaker identification, a speech utterance from an unknown speaker is analyzed and compared with the models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. In speaker verification, an identity claim is made by an unknown speaker, and an utterance verification, an identity claim is made by an unknown speaker, and an utterance of the unknown speaker is compared with the model for the speaker whose identity is claimed [1]. In HMM based speaker recognition system, the comparing is conducted using log-likelihood measurement. HMM system configuration is described in section 2.1.

2.1. HMM Based Speaker Recognition System

The continuous speech waveform is first blocked into frames and a discrete sequence of feature vectors, $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{T(x)}\}$, are extracted where $T(x)$ corresponds to the total number of frames in the speech signal. We will identify the input speech utterance as its feature vector sequence $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{T(x)}\}$ without confusion. In HMM, the observation probability density function of observing vector $\mathbf{x}$ in the j-th state of i-th word

HMM is given by $b_j^i(\mathbf{x}) = Pr(\mathbf{x}|s_t = j)$ and the transition probability is defined by $a_{jk}^i = Pr(s_{t+1} = k|s_t = j)$ [5]. The optimal path under the Vitebi decoding is the one which attains the highest log-likelihood score. we denote $S^i = \{s_1; s_2; \cdots; s_T\}$ as the optimal path of the input utterance X into i-th word HMM λ_i. Then, the log-likelihood score of the input utterance X along its optimal path in i-th model λ_i, $g_i(X, \lambda_i)$, can be written as

$$\begin{aligned} g_i(\mathbf{X}; \Lambda) &= \frac{1}{T} \log f(\mathbf{X}, S^i|\lambda_i) \\ &= \frac{1}{T}\{\sum_{t=1}^{T}[\log a_{s_{t-1}s_t}^{(i)} + \log b_{s_t}^{(i)}(\mathbf{x}_t)]\} \end{aligned} \quad (1)$$

In above equation, we can see that each frame has the same contribution to the total score. But the investigation of the human propensity for speaker recognition shows that concentrate a few important portion of speech waves and then identify a speaker. So, it is necessary to define individual information of each portion and to reflect the importance of frames.

2.2. Score Normalization

In speaker verification the world model based approach and the cohort model based approach have been used for score normalization [2]. The cohort model approach adopts a competition-based measurement. For a simple form of this method, a measurement is defined as a ratio of the score from the claimed speaker template with the score from most competitive speaker template. The world model approach uses a set of text-dependent speaker independent word models as world models. The scores of test utterance from the world models are used to normalize the score from speaker template. This normalization techniques are not directly applied to speaker verification system. Because all models have the same world model for same test utterance in world model based normalization method, it is useless. In cohort model based method, the cohort model has no meaning for speaker model.

3. INDIVIDUAL INFORMATION WEIGHTING

To improve performance of speaker recognition system, we make efficient use of all the clues, that is, speaker identity information. as describing in section 2., humans give different contribution to each of speech segment [3]. This paper proposes a scoring method for speaker recognition, which applies different importance to all basic portions of a sampled speech waveform. First we define the quantity of the importance of speech frames and propose a new scoring method applied to speaker recognition.

3.1. Measuring Individual Information

Speaker identity is related with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope and in the supra-segmental features of speech. Speaker recognition

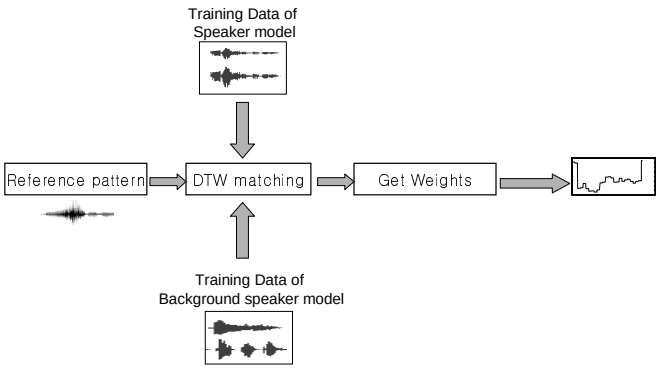

Figure 1: Block diagram of the weighting process of measuring information weighting

system identify speaker using these speaker identities. In this paper, we suggest a method to measure the quantity of speaker identify in each speech frames. To define the quantity of the individual information contained in each frame of speech signal, we use F-ratio, which is the technique to select proper feature parameters in speaker recognition.

$$\text{F-ratio} = \frac{\text{inter-speaker variance}}{\text{intra-speaker variance}} \quad (2)$$

As shown in equation 2, the portion which make larger inter speaker variance and smaller intra speaker variance is well reflected speaker identity. for measuring intra-speaker variance, we use training data of specific speaker i. Between each training data of speaker model, we first select reference pattern of that speaker. Then, we calculate distance of each training data of speaker model i. Because the distance is the frame by frame distance, we must align boundary of reference pattern and each training data pattern before using time alignment algorithm such as DTW (dynamic time warping). In this process, we define inter speaker variance using speaker i's reference pattern and other speaker's with same text data. These process is shown in figure 1. In frame by frame distance measurement, there are 3 case of frame matching. Namely 1 frame to 1 frame, 1 frame to n frames and n frames to 1 frame. Considering these frame matching condition, we define frame distance as equation 3.

$$\delta(\mathbf{x}_t^{ref},\ \mathbf{x}_{t_1;\cdots;t_K}^n) = \frac{1}{K}\sum_{i=1}^{K} ||\mathbf{x}_t^{ref} - \mathbf{x}_{t_i}^n||_e \quad (3)$$

In this equation, distance measure can be used such as Euclidean distance measure, Mahalanobis distance etc. In the training phase, we align all the training data and calculate each frame distance. By using inter speaker distance and intra speaker distance, we calculate the quantity of individual information of each frames. this value can be written as

$$w_t = \frac{\frac{1}{M}\sum_{all\ m}\delta(\mathbf{x}_t^{ref},\ \mathbf{x}_{t_1;\cdots;t_J}^m)}{\frac{1}{N-1}\sum_{n\neq ref}\ \delta(\mathbf{x}_t^{ref},\ \mathbf{x}_{t_1;\cdots;t_K}^n)} \quad (4)$$

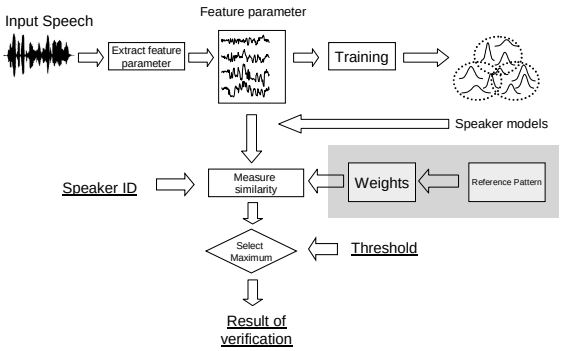

Figure 2: Block diagram of proposed Speaker verification system

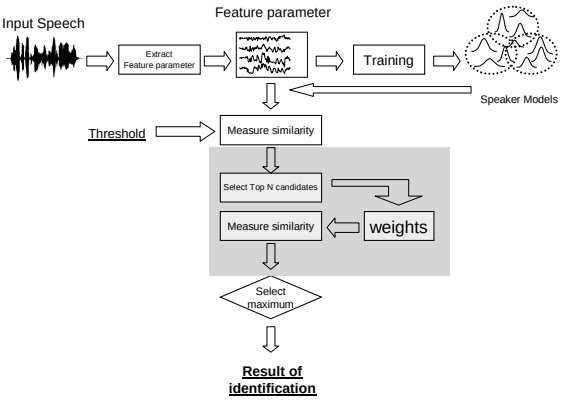

Figure 3: Block diagram of proposed Speaker identification system

where M is the number of ith speaker's background model data and N is the number of training data of speaker i. w_t is normalized individual information quantity of t frame in speech signal. A denominator represents intra speaker variance and a numerator represents intra speaker variance. As the greater value it has, the better individual information is reflected.

3.2. Speaker Verification System

The measured quantity is used as a weighting factor in speaker verification system and incorporated into a scoring process. as shown in section 2.1., log-likelihood score is calculated and defined as the weight of the frames. We weigh it to the output probability of HMM. The weighted log-likelihood score can be written as

$$\begin{aligned} g_i^{new}(\mathbf{X};\Lambda) &= \tfrac{1}{T}\log f^{new}(\mathbf{X}, S^{(i)}|\lambda_i) \\ &= \tfrac{1}{T}\left\{\sum_{t=1}^{T}\left[\log a^{(i)}_{s_{t-1}s_t} + \log\left\{b^{(i)}_{s_t}(\mathbf{x}_t)\times w_t^{(i)}\right\}\right]\right\} \\ &= \tfrac{1}{T}\left\{\sum_{t=1}^{T}\left[\log a^{(i)}_{s_{t-1}s_t} + \log b^{(i)}_{s_t}(\mathbf{x}_t) + \log w_t^{(i)}\right]\right\} \end{aligned} \tag{5}$$

Figure 2 shows the block diagram of proposed speaker verification system. In training phase, we select reference pattern between training data and calculate individual information quantity of each frame. In testing phase, we time-align the test speech with reference pattern and calculate each frame weights. Next we compare this score with threshold, and decide whether accept or reject. these algorithm have some more processes than original system. So time complexity is higher than baseline system. But most of process is conducted in training phase, this is not serious concerned.

3.3. Speaker Identification System

In speaker identification system, we use measured individual information for normalization. There are two main normalization technique in speaker verification, that is world model based normalization and cohort model based normalization. But these techniques can not be applied directly to speaker identification. Figure 3 shows the block diagram of proposed speaker identification system. Proposed system is divided two part. First, like general HMM based system, we calculate likelihood ratio between each model and test utterance. Second phase, by comparing the above likelihood ratio with the likelihood of top 1 model, we select top n candidates within specific threshold. We assume that each candidates are the claimed speaker and time-align test utterance and candidate model i using DTW matching. Then we calculate the weighting factor of each test utterance frames on the assumption that test utterance comes from speaker i. This process is conducted for all n candidates. Finally we calculate new score for each candidates and decide the model id that have maximum likelihood score.

4. EXPERIMENTS

A number of experiments were conducted in order to evaluate the usefulness of the weighted scoring method for both speaker verification and speaker identification.

4.1. Database

In each experiment, a total of 60 speaker models were trained with 5 fixed 4-digit utterances for each model. Speech utterance set is consisted of korean 4-digit utterances. We collect speech utterance in laboratory environment. Speaker is consisted of 10 male speaker and 2 female speaker. First we collect 5 utterances for each model training. Then, we collect 8 utterances during 2-3 days interval. So total 480 trials are tested in speaker identification experiment. And 28800 trials are tested in speaker verification experiment. Long-term variation is another critical problem in speaker recognition. In this paper, we collect test data for long-term variation experiment after 1 year later first training data are collected.

4.2. Experimental Results

12-order mel-frequency cepstrum, delta mel-frequency cepstrum, energy and delta energy were used as input pa-

Table 1: Equal Error Rate of Speaker verification system

	6 states	8 states	10 states
baseline system	11.0%	9.3%	8.2%
weighted system	6.9%	6.2%	5.6%
error reduction rate	37.4%	33.9%	31.7%

rameters. Recognizer is consisted of HMVQM [6] which have a different number of states and codebook sizes. Proposed speaker verification system is compared with baseline system which treats all speech segments to have the same importance. The performances are compared with the various numbers of HMM states. Speaker verification experiment result is shown in table 1. As shown in table, regardless of HMM states number, proposed speaker verification system outperformed the baseline system. We can get 37.4% EER (equal error rate) reduction in 6 states.

Table 2: Recognition rate of Speaker identification system

	6 states	8 states	10 states
baseline system	89.2%	91.0%	92.5%
weighted system	92.5%	93.5%	94.4%
error reduction rate	30.6%	27.8%	25.3%

In speaker identification experiment, the proposed method marked relatively 28% higher recognition rate than the baseline system. in 6 states environment, total 140 trials entered top-N test. 88 tests changed their result. 86 tests are correctly changed and 2 tests give wrong answer. So the performance improvement is relatively 30.6%.

Table 3: long-term variation

	6 states	8 states	10 states
baseline system	15.9%	13.6%	13.9%
weighted system	6.1%	5.7%	1.0%

Long-term variation is one of the critical problem in speaker recognition. In such a case that the time interval between training data collection time and testing data collection time is too far, recognition rate is radically decreasing. Table 3 shows long-term variation experiment result of proposed system and baseline system. the average recognition rate reduction of baseline system is 14.5%. It is similar to recognition reduction of general long term variation system. But the average recognition reduction of proposed system is 4.3%. this is due to the individual information is not easily variable as time is going.

5. CONCLUSION

This paper proposes a speaker recognition method which applies different importance to all basic portions of a sampled speech waveform. To define the quantity of the speaker information contained in each frame of speech signal, we use F-ratio measure, which is the technique to select proper feature parameters in speaker recognition. In speaker verification experiments, the proposed method reduced equal error rates considerably compared to a conventional method, which treats all speech segments to have the same importance. In speaker identification experiments, the proposed method marked relatively 28% higher recognition rate than the baseline system, and was more robust in long-term variation. These results demonstrate that the proposed method is efficient in measuring individual information and more appropriate for speaker recognition.

6. REFERENCES

[1] Kuldip K. Paliwal Chin-Hui Lee, Frank k. Soong. *Automatic Speech and Speaker Recognition - Advanced Topics*. Kluwer Academic Publishers, 1992.

[2] Yong Gu and Trevor Thomas. A Hybrid Score Measurement for HMM-Based Speaker Verification. In *Proceedings of ICASSP*, pages 317–320, 1999.

[3] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Communications*, 22:1–15, 1997.

[4] Douglas O'Shaughnessy. Speaker Recognition. *IEEE ASSP Magazine*, pages 4–17, 10 1986.

[5] Lawrence R. Rabiner and Biing-Hwang Juang. An introduction to Hidden Markov model. *IEEE ASSP Magazine*, pages 4–16, 1 1986.

[6] Yun Seong-Jin. Performance improvement of speaker recognition system for small training data. Master thesis, KAIST, in Korean, 1994.