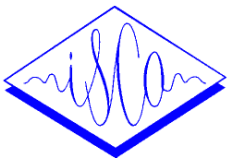

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Successive Cohort Selection (SCS) for Text-Independent Speaker Verification

Eric H. C. Choi and Jianming Song

Motorola Australian Research Centre
Human Interface Lab, Sydney, Australia
Email: {hchoi, jsong}@arc.corp.mot.com

ABSTRACT

A novel cohort selection method, namely, successive cohort selection (SCS) is presented in this paper for text-independent speaker verification. The proposed method computes distance between two models directly and it selects new cohort member based on both the claimed speaker model and the existing cohort members. In addition to this new cohort selection method, we also propose a new score measure to be used in verification stage. This new score measure makes use of the absolute score of an utterance to weigh its normalized score in order to further improve the accuracy of a verification. Experimental results on the YOHO speech corpus have revealed that the proposed cohort selection method together with the new score measure reduce the equal-error-rate (EER) of a text-independent system by four times. With only a cohort size of four, the EER of the system is reduced to 0.72%.

1. INTRODUCTION

A speaker verification system is designed to analyze the voice of a speaker claiming to be a particular individual, and determine whether the claim is true [1]. Recent advance in the technology has provided numerous development opportunities for applications such as authentication for telephone banking, smart card transactions and access control. Compared with other biometrics, for example, fingerprint recognition and retinal scanning for access control, speaker verification through voice has a significant advantage in being much less obtrusive. In addition, it is a low cost technology in high volume applications.

A commonly used approach in speaker verification is to compute a normalization score for an input utterance using the likelihood ratio of the claimed speaker model and a set of cohort models or background speaker models [2][3][4][5]. In these systems, a claimed identity is accepted if the normalized score is larger than a pre-defined threshold. Otherwise, the claimed identity is rejected. Score normalization against a set of cohort models is essential for maintaining stable decision thresholds in different environments since it minimizes the non-speaker related variations in the scores.

Various methods have been proposed for choosing the appropriate speaker models to be used as cohort for a claimant. These methods include choosing the closest speaker models to a claimed speaker model [2], choosing the closest and farthest models [3], and combing those models which are closest into a single model [4]. While these methods are straightforward to implement, there are some drawbacks. The first one is that they require training utterances to find out the appropriate cohort models. If we can compute the distance between two models directly, then we can save a lot of storage due to the training data and in most cases, we can also reduce computation as well. The other drawback is that these methods only consider the distance between a model and a claimed speaker model in choosing the cohort set. However, once a cohort model is selected, the ROC curve of the resultant system is likely to be changed. Therefore the selection of a new cohort member should also be dependent on the existing cohort members as well.

In this paper, we propose a novel method, namely, successive cohort selection (SCS) which does not have the above mentioned drawbacks of the existing methods. Our proposed method computes distance between two models directly and it selects new cohort member based on both the claimed speaker model and the existing cohort members. In addition to a new cohort selection method, we also propose a new score measure used in verification. This new score measure makes use of the absolute score of an utterance to weigh its normalized score in order to further differentiate an impostor whose voice is quite different from those of the claimant and the cohort members.

We will describe our baseline text-independent system in Section 2. Following that will be a detailed description of the SCS method in Section 3 and a definition of relative normalized distance in Section 4. Our experimental results will be reported in Section 5.

2. BASELINE SYSTEM

Our baseline system was developed for text-independent applications with its underlying modeling technology being based on variance weighted vector quantization (VQ) [6]. A speaker's voice is represented by a model

$$\lambda_i = \{\mu_{ij}, \Sigma_i\}; \qquad j \le N_{cb} \tag{1}$$

where μ_{ij} is the j-th codeword of speaker i, Σ_i is the corresponding global diagonal covariance matrix and N_{cb} is the VQ codebook size. Verification of an input utterance $X = [x_1, \ldots, x_T]$ for speaker i is based on a normalized distance $d_n(X|\lambda_i)$, which is defined as:

$$d_n(X \mid \lambda_i) = d(X \mid \lambda_i) - d_c(X \mid \lambda_c); \tag{2}$$

$$d_c(X \mid \lambda_c) = \underset{j \in Cset}{F_{sv}} \; (d(X \mid \lambda_j)); \tag{3}$$

$$d(X \mid \lambda_i) = \frac{1}{T} \sum_{t=1}^{T} \min_j \{d_{wg}(x_t, \mu_{ij}, \Sigma_i)\} \tag{4}$$

where F_{sv} is a statistical function (e.g. minimum), *Cset* represents the members of the cohort set and $d_{wg}(.)$ is the variance weighted Euclidean distance between two vectors [7]. The utterance is accepted if $d_n(.)$ is less than a pre-defined threshold τ_i. Otherwise, the utterance is rejected. A block diagram of the baseline system is shown in Figure 1.

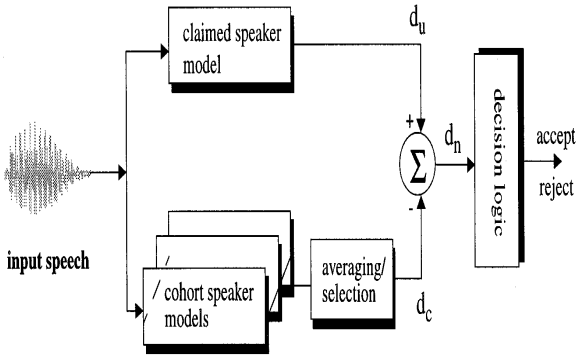

Figure 1: Speaker verification system block diagram

3. SUCCESSIVE COHORT SELECTION

Commonly used cohort selection methods choose speaker models which have the smallest distances from a claimed speaker model as the cohort set, without considering the effect of existing cohort members. However, it is observed that the inclusion of a new model into a cohort set can change the ROC curve of the resultant system. That is, it changes both the distribution of the false acceptance rate for impostors and that of the false rejection rate for the legitimate speaker. Therefore the model thus chosen is optimized respective to the original system only. It is clear that a model chosen by this kind of selection methods is not optimized with respect to the change in characteristics of the system, which occurs due to the inclusion of cohort models into the system. To account for the interactions among cohort members and a claimed speaker in computing the normalized distance for verification, we have proposed a novel successive cohort selection (SCS) method. Moreover, in contrast to the ordinary methods which utilize training utterances to select a cohort set, the SCS method proposed here selects cohort models directly from a set of available speaker models. It thus requires less computation and storage to find out the cohort set. In our SCS method, the distance (D_{ij}) between two speaker models λ_i, λ_j is defined as:

$$D_{ij} = \sum_{k=1}^{N_{cb}} \min_l (dist(\mu_{il}, \mu_{jk})\}; \tag{5}$$

$$dist(x, y) = [\sum_{l=1}^{L} \frac{(x_l - y_l)^2}{\sigma^2_{xl} + \sigma^2_{yl}}]^{1/2} \tag{6}$$

where N_{cb} is the codebook size, μ_{jk} is the k-th codeword of λ_j, x_l and y_l are the l-th components of the respective vectors x and y, σ^2_{xl} and σ^2_{yl} are the l-th diagonal elements of the respective global covariance matrices and L is the vector size. The idea here is to compare two models by considering how far their individual codewords are separated. Note that D_{ij} is asymmetric, that is, D_{ij} is not necessarily equal to D_{ji}.

To select M cohort members for a speaker k, i.e. $coh^k(m)$, $m = 1, \ldots, M$, the SCS can be formulated as follows:

$$coh^k(1) = \arg\min_j \{D_{kj}\}; \qquad k \ne j; \tag{7}$$

$$coh^k(m) = \arg\min_i \{D_{ki} - \underset{j \in Cset(m)}{F_{cs}} (D_{ji})\};$$

$$i \ne j, k; \quad m \ge 2; \tag{8}$$

$$Cset(m) = \{coh^k(1), \ldots\ldots, coh^k(m-1)\} \tag{9}$$

Once the desired number of cohort members have been found out from a pool of speaker models, the selection process is completed. The models in *Cset*(M+1) can then be used as the cohort members for the speaker k.

Two different types of the statistical function $F_{cs}(.)$ are used here. It can be either a minimum function:

$$\underset{j\in Cset(m)}{F_{cs}}(D_{ji}) = \min_{j\in Cset(m)}\{D_{ji}\} \quad (10)$$

or an average function defined as:

$$\underset{j\in Cset(m)}{F_{cs}}(D_{ji}) = \frac{1}{m-1}\sum_{j\in Cset(m)}\{D_{ji}\} \quad (11)$$

In Section 5, we will show some experimental results for using these two different statistical functions in cohort selection.

4. RELATIVE NORMALIZED DISTANCE

One problem with the use of normalized distance in verification is that the difference between two large absolute distances can be the same as that between two small absolute distances. Therefore the use of normalized distance cannot differentiate the two different cases. But if we interpret the absolute distance as a measure of how close an utterance is to a speaker model, utterance with smaller distance should have a higher likelihood of being uttered by the claimed speaker. In other words, an appropriate distance measure should give a smaller distance value for the case where the absolute distance between an utterance and the claimed speaker model is small. One simple but effective solution is to use relative normalized distance (d_{rn}) proposed as below:

$$d_{rn}(X\mid\lambda_i) = \frac{d(X\mid\lambda_i) - d_c(X\mid\lambda_c)}{d(X\mid\lambda_i)} = 1 - \frac{d_c(X\mid\lambda_c)}{d(X\mid\lambda_i)} \quad (12)$$

Straightly speaking, $d_{rn}(.)$ itself is not really a distance measure since it can be a negative value. Nevertheless, we choose to use this terminology for consistency as we basically treat this value as a score to decide the verification outcome. As before, if the relative normalized distance of an utterance is less than a threshold, the utterance is accepted. Otherwise, it is rejected.

5. EXPERIMENTS

5.1 Experimental Setup

The SCS method proposed here was tested on the YOHO speech corpus [8]. The YOHO corpus contains combination lock phrases of three two-digit numbers (e.g. thirty-six, forty-five, eighty-nine) spoken by 138 speakers (106 males and 32 females). For each speaker, there are four enrollment sessions with 24 phrases per session and ten verification sessions with four phrases per session. The average length of the verification utterances is only 1.7 seconds.

In the experiments, 12 mel-frequency cepstral coefficients (MFCC) generated every 10ms with a frame size of 32ms were used as feature vectors. For each speaker, training data from all four enrollment sessions were used to train the speaker model. Each speaker model was represented as a 128-codeword VQ codebook trained by the LBG algorithm [9], together with a global diagonal covariance matrix. For testing, we only performed one-phrase verification and all 40 verification utterances per speaker were tested. For each of the 138 speakers, all the remaining 137 speakers were used as impostors to test that speaker model. The results reported below were obtained by using speaker-dependent verification threshold and the average equal-error-rate (EER) across all speakers is reported for each experiment.

5.2 Results

Without using any cohort models, a baseline EER of 2.92% with standard deviation of 3.12% was obtained. Using the SCS method to choose four cohort models and relative normalized distance in verification, we achieved the results as shown in Table 1. This table summarizes the verification results obtained by using minimum or average function for combing cohort scores in both cohort selection and verification.

Table 1: EER(%) for a cohort size of 4, SCS and d_{rn}

F_{cs}	F_{sv}	Avg EER (%)	Std dev. (%)
min	min	1.75	2.91
min	avg	1.00	1.77
avg	avg	0.72	1.37

As observed from the above table, the use of average as a combining function is more appropriate in both cohort selection and verification. Furthermore, it seems that applying average function in verification can provide

more improvement in accuracy than applying it in cohort selection.

To better compare our novel SCS method with other methods which only choose the nearest speakers (NRS) as the cohort set, we had performed another set of experiments and the results are shown in Table 2. Also shown in the table are the different results for using d_n instead of d_{rn} in verification.

Table 2: EER(%) for a cohort size of 4, F_{cs}= F_{sv} = average

Cohort selection	Distance measure	Avg EER (%)	Std dev. (%)
NRS	d_n	2.55	4.03
	d_{rn}	1.35	2.50
SCS	d_n	1.01	1.60
	d_{rn}	0.72	1.37

The above results readily demonstrate the effectiveness of using SCS and relative normalized distance. By just replacing d_n with d_{rn}, we have achieved a relative error reduction of about 50% for the NRS cohort selection and about 30% reduction for the SCS method. Meanwhile, by using SCS instead of NRS, we have obtained a relative error reduction of about 60% for using d_n and about 50% reduction for using d_{rn}. Compared with an average EER of 2.92% for using no cohorts, the best result represents a more than four times reduction in error rate.

6. CONCLUSION

A novel successive cohort selection method has been proposed and developed. In contrast to ordinary methods which utilize training utterances in the process of cohort selection, the SCS method selects cohort models directly from a set of VQ codebooks. It thus achieves large reduction in computation and storage for the cohort selection process. Moreover, this method improves verification accuracy by taking into account the interactions among a claimed speaker model and the existing cohort models. To further improve accuracy and robustness, a distance measure which combines normalized distance and absolute distance has been proposed. A series of experiments on the YOHO corpus has revealed the effectiveness of the proposed method.

In view of the effectiveness of using relative normalized distance in verification stage, we believe that further improvement in verification accuracy should be possible by also using relative normalized distance when applying SCS in cohort selection. Further work following this line will be pursed.

7. REFERENCES

[1] O'Shaughnessy D., "Speaker Recognition", IEEE ASSP Magazine, pp 4-17, Oct. 1986.

[2] Rosenberg A. E. et al, "The Use of Cohort Normalized Scores for Speaker Verification", Proc. ICSLP'92, vol.2, pp.599-602, 1992.

[3] Reynolds D. A., "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", Speech Communication, vol. 17, nos. 1-2, pp. 91-108, Aug. 1995.

[4] Liu C. S. et al, "Speaker Verification Using Normalized Log-Likelihood Score", IEEE Trans. Speech and Audio Processing, vol. 4, no.1, pp. 57-60, Jan. 1996.

[5] Liu W. et al, "On Optimum Normalization Method Used for Speaker Verification", Proc. ICSLP'98, vol. 2, pp. 165-168, Dec. 1998.

[6] Zhu X. et al, "Text-Independent Speaker Recognition Using VQ, Mixture Gaussian VQ and Ergodic HMMs", Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 55-58, Apr. 1994.

[7] Deller J. R. et al, Discrete-Time Processing of Speech Signals, Macmillian Publishing, New York, Chap. 1, pp. 62, 1993.

[8] Campbell J. P., Jr., "Testing with the YOHO CD-ROM voice verification corpus", Proc. ICASSP'95, vol. 1, pp. 341-344, 1995.

[9] Linde Y. A. et al, "An Algorithm for Vector Quantizer Design", IEEE Trans. Communications, vol. 28, pp. 84-95, Jan. 1980.