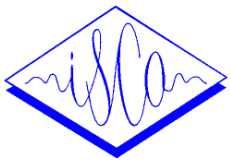

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

HYBRID THRESHOLD APPROACH IN TEXT-INDEPENDENT SPEAKER VERIFICATION

Fangxin Chen, Bruce Millar and Michael Wagner

TRUST Project

Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

In this paper we suggest a hybrid threshold approach for text-independent speaker verification. A maximum distortion (or minimum likelihood) threshold is set for the claimed speaker to perform pre-filtering of the highly dissimilar impostors. Cohort normalisation then is applied to further separate those impostors who are acoustically similar to the claimed speaker. A VQ-distortion based text-independent speaker verification system using this approach achieves better results than the conventional absolute threshold or cohort normalisation methods.

1. INTRODUCTION

The classic method for speaker verification is to measure the difference or similarity between an incoming utterance and a model of the speech of the claimed speaker, then to accept or reject the speaker of the incoming utterance based on comparison with a predetermined absolute threshold. The merit of using such raw measures of difference or similarity in setting the threshold for the decision of acceptance or rejection is that these scores truly reflect the absolute deviation between the client's model and the input speech. However, these raw measures derived from the speech of both impostors and the claimed speaker (client) can vary greatly due to intra-speaker variation, recording environment change, or phonetic variation, and can therefore be characterised by highly overlapping impostor and client distributions. In these situations the Equal-Error Rate (EER), which is the error rate when the absolute threshold is set to produce equal numbers of false rejections and false acceptances, becomes very high and speaker verification performance deteriorates.

An alternative is the dynamic threshold approach which uses cohort normalised likelihood or distortion scores [1,2,3]. In this method a "cohort" of "similar speakers" in the system is chosen. Speakers having characteristics which are very similar to those of each client speaker are individually selected as members of that speaker's cohort. The function of the cohort is to act as a local environment in the space of all speakers against which measures of difference or similarity may be normalised to some degree. The simplest form of such normalisation is to subtract some function of the distance between the incoming speech and this local environment from the distance between the incoming speech and the speech model of the client. This approach is reportedly effective in both text-dependent and text-independent situations [2,3]. The problem, however, is that it overlooks the absolute deviation of the input speech from the client's model, which leads to the possible acceptance of highly dissimilar impostors [4]. The present study applies a hybrid threshold approach which combines both absolute and dynamic threshold approaches to text-independent speaker verification.

2. EXPERIMENTS

The first experiment applied the conventional absolute threshold approach to decision making in a VQ-distortion system. The EER for each client was calculated using raw VQ distortion scores. In the second experiment, cohort normalisation was applied. The cohort normalised scores were calculated with either mean or minimum statistic to determine the contribution of the cohort distance [2]. In this case, the EER was based on the cohort normalised scores. The third experiment was the same as the second experiment except that an absolute threshold based on the raw VQ distortion scores was applied before cohort normalisation. The absolute threshold was determined according to the distribution of VQ distortion scores of each client's training utterances against his/her own trained codebook. Extreme outliers of this distribution were removed by adopting the 99th percentile value as the maximum distortion score. This score was multiplied by a coefficient to provide a robust absolute threshold for this speaker against false rejection. Obviously, the coefficient should be greater than 1 to anticipate the variation of the true speaker's utterance score over time and across changing speech environments. Any test utterance with a raw distortion score greater than this value was immediately rejected as belonging to an impostor without recourse to cohort normalisation. Thus, the cohort normalised scores with mean or minimum statistic take the following form:

if $S_c < KS_{max}$

then $Score_N = S_c - Stat\ [S_i, i=1, n]$

else $Score_N = \infty$

where

S_c is input speech's VQ distortion score for the client's VQ model;

K is the absolute threshold coefficient;

S_{max} is the maximum VQ distortion score of the client's training utterances for his own VQ model;

$Score_N$ is the cohort normalised score;
Stat is either the mean or minimum statistic;
S_i is the VQ distortion score for the ith cohort member's VQ model;
n is the cohort size.

3. METHOD

A set of 24 native English speakers, 12 males and 12 females, with age range 20-50, were selected from our speech database as clients. The speech data for VQ codebook training comprised 30 utterances with five repetitions. The utterances were mainly two-word computer commands (eg. "Save file," "New window"). They were selected so that, as a set, they basically covered the phonemic inventory of Australian English. The test data for each claimed speaker comprised the same list of 30 utterances plus ten additional computer commands with two repetitions (total 170 utterances). The interval between the training and test data recording was greater than one week. The selection of cohort speakers for each client was based on the average distortion scores obtained from other speakers' training utterances to the client's VQ model. The speakers with the lowest distortion scores were selected as the client's cohort. The selection of a cohort for each client was not restricted to the set of 24 clients, but was made from our total speech database of 37 speakers in order to obtain the best possible cohort for each client. The cohort size for this experiment was set to five. For each client, 23 impostors were selected from the database, providing 4080 testing utterances. To avoid bias [2,3], the cohort members were excluded from the impostor population for each client. That is, 23 impostors were selected for each client from the total database minus this client's cohort members.

The speech data were recorded in a quiet office environment and digitised at 20,000 samples/s after 60-9000Hz bandpass filtering. Seventeen cepstral coefficients were calculated based on mel-frequency analysis (with frame size of 25.6 ms and 10 ms frame shift). In training and testing, a variance-weighted VQ method was used. The variance of each cepstral coefficient for all training data was calculated and VQ codebooks comprising 128 codewords were developed using an inverse-variance-weighted cepstral distance. In subsequent testing, the raw VQ distortion score was calculated as the average inverse-variance-weighted distance between all frames of the test data and their closest codeword in the VQ codebook of the target speaker. The coefficient for the robust absolute threshold was empirically set to 1.1.

4. EXPERIMENTAL RESULTS

Table 1 shows the results of the three experiments. The averaged EER using raw VQ distortion scores FOR all clients is 5.5%. Cohort normalisation using either mean or minimum statistic improves the overall performance, reducing the EER to 4.1% and 3.6%. However, these improvements are not statistically significant. Only the hybrid method, which employs cohort normalisation after application of a robust absolute threshold, produces significant improvement over the absolute threshold approach. (mean statistic [F=4.868, p<0.05], minimum statistic [F=4.135, p<0.05]).

As for the two different statistical methods for calculating cohort scores, our present experiment shows no significant difference at the cohort size of 5 for either the dynamic threshold or the hybrid threshold approaches.

Raw	Cohort		Hybrid	
5.5	mean	min	mean	min
	4.1	3.6	1.3	1.3

Table 1. The averaged verification EER(%) for 24 clients. Raw is the absolute threshold approach which uses the raw VQ distortion scores. Cohort uses normalised distortion scores with either mean or minimum statistic. Hybrid uses cohort normalised scores after application of robust absolute threshold.

5. DISCUSSION

By examining individual EER for all the 24 clients, we find that the cohort normalisation method works particularly well for clients whose raw VQ distortion scores highly overlap with those of the impostors. For example, the EER based on the raw VQ distortion scores for Client 16 is 15.39% (see Fig.1).

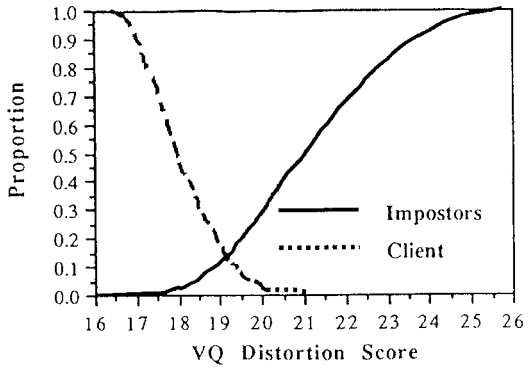

Fig.1 Client 16 and Impostors' Raw VQ Distortion Score Distributions. Proportion indicates the proportion of impostors/client scores that are below/above the abscissa score.

With cohort normalisation using mean statistic, her EER dropped to 0.31% (see Fig.2).

However, there is one problem with this method. Although it generally improves the EERs, there are 6 speakers, accounting for 25% of the testing population, whose EERs are greater than those using the absolute threshold method. For example, Client 11's EER based on raw VQ distortion scores is 5.72% (see Fig.3). Cohort normalisation using mean statistic causes EER to increase by about 14% (see Fig.4).

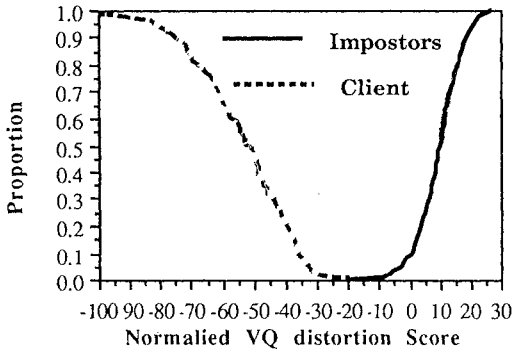

Fig.2 Client 16 and Impostors' Mean Statistic Cohort Normalised VQ Distortion Score Distributions. Proportion indicates the proportion of impostors/client scores that are below/above the abscissa score.

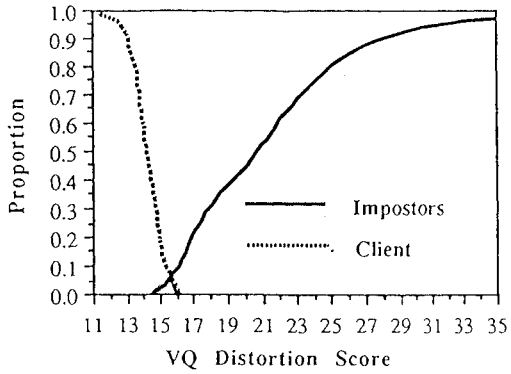

Fig.3 Client 11 and Impostors' Raw VQ Distortion Score Distributions. Proportion indicates the proportion of impostors/client scores that are below/above the abscissa score.

Cohort normalisation using mean statistic causes EER to increase by about 14% (see Fig.4).

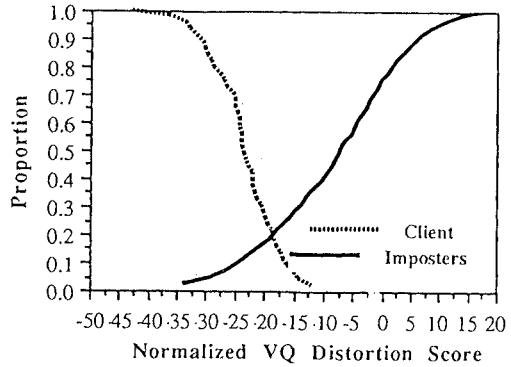

Fig.4 Client 11 and Impostors' Mean Statistic Cohort Normalised Distortion Score Distributions. Proportion indicates the proportion of impostors/client scores that are below/above the abscissa score.

The reason for the increase of EER is that the cohort normalised score of the input speech takes the difference between the raw distortion score from the client's VQ codebook and the mean (or minimum) raw distortion score from the cohort VQ codebooks. If the given utterance's VQ distortion from the client's codebook is lower than from the cohort codebooks, then the utterance will be accepted as the client's no matter whether the absolute VQ distortion from the client's codebook is large or small. This causes possible acceptance of highly dissimilar impostors as clients. To further illustrate this point, we use a hypothetical two-dimensional VQ distance diagram (Fig.5). It should be noted that the real VQ codebook distance has high dimensionality (17 in our case) and that the simplified two-dimensional diagram here is only for purpose of illustration.

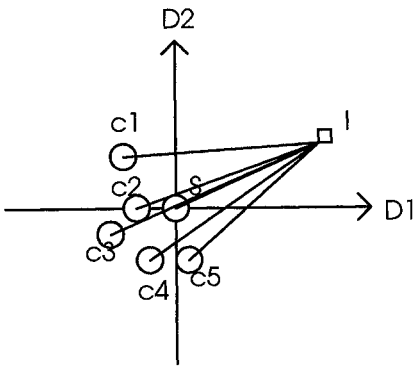

Fig.5 2D Representation of VQ Distance Space. S is a client's VQ codebook. c1-5 are the 5 cohort codebooks selected for that client. I is an impostor's utterance

It can be observed from Fig.5 that among the VQ distortion distances of the impostor's utterance from all the VQ codebooks, the distance from the client's is the shortest. If simple cohort normalisation is applied, the impostor will be falsely accepted as the client no matter what statistic is used or how distant the impostor is from the client in the VQ distortion distance space. To solve this problem, a threshold needs to be applied to reject those impostors whose utterance VQ distortion score is greater than the client's maximum possible distortion score, but which robustly accepts all utterances of the client. Fig.6 shows the greatly improved EER for Client 11 when a robust threshold is applied before cohort normalisation using the mean statistic.

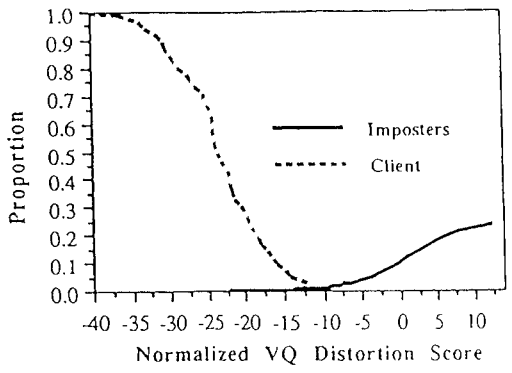

Fig. 6 Client 11 and Impostors' Hybrid Mean Statistic Normalised Distortion Score Distributions. Proportion indicates the proportion of impostors/client scores that are below/above the abscissa score.

6. CONCLUSION

Three different threshold approaches have been investigated for the same text-independent speaker verification task. The dynamic threshold approach based on cohort normalisation is effective in separating impostors from a client when their raw distortion scores are highly overlapping. However, one drawback of this method is that it is vulnerable to distant impostors whose VQ distances from the client's codebook are shorter than those from the cohort members' codebooks. The hybrid threshold approach, which applies a robust threshold before cohort normalisation, can reduce the incidence of this problem and significantly improve the performance of speaker verification.

ACKNOWLEDGEMENT

This research has been carried out on behalf of the Harry Triguboff AM Research syndicate. The helpful discussion of all the TRUST project members and contribution of Dr. Yuqing Gao to the earlier development of the VQ software are gratefully acknowledged.

REFERENCES

[1] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomised phrase prompting", Digital Signal Processing1, pp.89-106, 1991.

[2] A.E. Rosenberg, J.Delong, C.H. Lee, B.H. Huang, and F.K. Soong, "The use of cohort normalised scores for Speaker Verification", Proc. ICSLP, Bandff, pp. I-599-602, 1992.

[3] T. Matsui and S. Furui, "Similarity normalisation method for speaker verification based on a posteriori probability", Proc. ESCA workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, pp.59-62, 1994.

[4] S. Furui, "An overview of speaker recognition technology". Proc. ESCA Workshops on Automatic Speaker Recognition Identification and Verification, Martigny, pp. 1-9, 1994.