A HYBRID SCORE MEASUREMENT FOR HMM-BASED SPEAKER VERIFICATION

Yong Gu and Trevor Thomas

Vocalis Ltd.
Chaston House, Mill Court, Great Shelford
Cambridge CB2 5LD, UK
Email:yong.gu@vocalis.com

ABSTRACT

In speaker verification the world model based approach and the cohort model based approach have been used for better HMM score measurements for verification comparison. From theoretical analysis these two approaches represent two different paradigms for verification decision-making strategy. Two techniques could be combined for a better solution. In the paper we present a hybrid score measurement which combines the world model based technique and the cohort model based technique together. The method is evaluated with the YOHO database. The results show that the combination can lead a better score measurement which improves speaker verification performance. An experimental comparison between the world model based approach and the cohort model based approach with the YOHO database can also be found in the paper.

1. INTRODUCTION

One of challenges for moving HMM from speech recognition to speaker verification (SV) and utterance verification (UV) is to understand the HMM score variation and to define a proper measurement which is comparable across speech sample domains. This is different from recognition as recognition tasks require only score comparison across templates. Some approaches have been proposed for better HMM score measurements using score normalisation techniques in both SV [2][3][4][5] and UV [6][7]. In SV the world model based normalisation and the cohort model based normalisation are two most popular techniques. In [1] two basic verification measurements, qualifier-based measurements and competition-based measurements, are introduced to analyse these two score normalisation techniques from theoretical point of view. From the analysis it can be seen that the cohort model approach and the world model approach are based on two different verification measurement paradigms. Theoretically two methods could be combined for a better solution for SV. In this paper we present a hybrid method which combines the two score normalisation techniques together. The evaluation results demonstrate that such a combination can lead to a better score measurement which improves SV performance. In addition, a comparison between the world model approach and the cohort model approach can also be found in the paper. The comparison results are based on the YOHO database using both the speaker specific (SS) threshold and the speaker independent (SI) threshold method.

2. HMM-BASED VERIFICATION

Verification is a decision-making process that for a given sample and a claimed identity, the verification system gives a value for acceptance or rejection. The system should have knowledge for any claimed identity and for some systems the knowledge of other identities may also be available. Let V be a verification process and I be a claimed identity and K represent knowledge. For an input sample S the verification process can be defined as

$$V:(S,I,K)\rightarrow\{0,1\} \tag{1}$$

The verification process may consist of a measurement M of the input sample with pre-stored templates Ts and a verification decision based on the obtained measurement and a pre-defined threshold θ. Thus

$$V:(S,I,K)=\begin{cases}0 & M(S,I,Ts)<\theta\\ 1 & M(S,I,Ts)\geq\theta\end{cases} \tag{2}$$

There are generally two basic measurements for verification decision-making, qualifier-based measurements and competition-based measurements. With qualifier-based measurements, the system makes a decision based on a calculation using the claimed template only and no other templates are directly involved in the measurement, so the measurement becomes

$$M(S,I,Ts)=P(S,I,Ti) \tag{3}$$

where $P(S,I,Ti)$ is a measurement between sample S and claimed template Ti. With this method, the robustness of the measurement over samples as well as across speaker templates is important for the success of the verification system.

With competition-based measurements, the system makes its decision based on calculations using the claimed template and some other templates. The system takes a relative value of scores from the claimed template and some other templates as a measurement for decision. With this method, the measurement reflects how well the claimed template matched with the sample compared to other templates either using the ratio

$$M(S,I,Ts)=\frac{P(S,I,Ti)}{F(\{P(S,I,Tj)\mid j\neq \mathrm{i}\})} \tag{4}$$

where F is a function over a set of scores, or the difference

$$M(S,I,Ts)=P(S,I,Ti)-F(\{P(S,I,Tj)\mid j\neq i\}) \tag{5}$$

A typical example is to measure the scores from the template of the claimed speaker and most competitive template(s) e.g. cohort model approach for SV [2][4], and second best in UV [6][7]. As the measurement depends on other templates to measure competitiveness, this method requires available selected templates that are somehow representatives of possible testing samples so that the measurement becomes reliable.

In the HMM approach, the speech utterance is considered as a sequence of observations O generated by a production model $M(S,W)$ associated with a speaker S and a word W. For a given sample O, a measurement between sample and model is defined as the *a posteriori* probability for model $M(S,W)$ to generate O, $P(M(S,W)|O)$. Using Bayes' Rule the following equation can be derived

$$P(M(S,W)|O)=\frac{P(O|M(S,W))P(M(S,W))}{P(O)} \qquad (6)$$

In speech recognition, speaker S becomes irrelevant and $P(M(W))$ is also reasonably assumed as a constant. Thus the recognition task becomes the solution to this equation

$$w=\arg\max_i\{P(M(W_i)|O)\}=\arg\max_i\{\frac{P(O|M(W_i))}{P(O)}\} \qquad (7)$$

Since $P(O)$ is the same in comparison across the models $M(W_i)$, the measurement can be simplified to $P(O|M(W_i))$. The HMM approach provides a framework of estimating a model $M(W_i)$ and measure $P(O|M(W_i))$.

In SV, the measurements are required to compare on the sample domain O. In such cases, $P(O)$ can not be removed from calculating measurement $P(M(S,W)|O)$ from equation 6, as O is variable in the verification comparison. Given the testing speaker in the verification task is often an open set therefore the probability $P(O)$

$$P(O)=\sum_i^{\infty}P(O|M(S_i))P(S_i) \qquad (8)$$

is not possible to be calculated fully. The measurement $P(O|M(S,W))$ has been proved not robust for verification from experimental evaluation [3]. Thus finding robust measurements have been one of most challenge tasks in HMM-based speaker verification. A number of approaches have been proposed to normalise the score $P(O|M(S,W))$ for better measurement.

3. SCORE NORMALISATION

In SV two most popular score normalisation techniques are the world model based approach and the cohort model based approach. 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, i.e.

$$R_{cohort}=\frac{P(O|M(S_i,W))}{\max_{j\neq i}\{P(O|M(S_j,W))\}} \qquad (9)$$

This leads an HMM score measurement without estimating $P(O)$ and the measurement fits well in theory for verification in a close set. However, for an open set it needs a collection of speakers from which the cohort speaker can be selected. The selection of the competitive speaker does not depend on the claimed speaker but depends on test sample (imposter). Therefore this collection of speakers is required, in some way, representing the testing population. Consequently the performance may depend on the collection of speakers for cohort selection.

The world model approach uses a set of text-dependent speaker independent word models as world models. The score of test utterance from the world models is used to normalise the score from speaker template $P(O|M(S,W))$. Assume that $M(S_{world},W)$ is a world model for word W the normalisation leads a measurement

$$R_{world}=\frac{P(O|M(S_i,W))}{P(O|M(S_{world},W))} \qquad (10)$$

The set of world models is often generated from a large number of speech samples from a large number of speakers. As the SV system is often combined with speech recognition the acoustic models, used for speaker-independent speech recognition, can be used as world models for SV. In [3], the score from the world model was explained as an approximation of $P(O)$ in equation 8. Thus with this approach the verification process takes the qualifier-based measurement of an approximation of $P(M(S,W)|O)$ for the verification decision-making comparison.

From theoretical analysis the cohort model method and the world model method are based two different basic verification measurement paradigms. Two measurements could be combined and a balance of two measurements could lead a better solution for speaker verification measurement. Here a hybrid approach is proposed which combines these two measurements together. A simple way of the combination can be defined as

$$R_{comb}=R_{cohort}^{\alpha}\times R_{world}^{1-\alpha} \qquad (11)$$

where $0\leq\alpha\leq1$. When α is equal to zero the combination becomes the world model measurement and when α is equal to one the measurement becomes same as cohort based. The key issue is to find out if there is a balance point which gives better measurement for SV.

4. EXPERIMENTAL RESULTS

4.1 Experimental Conditions

The experiments were carried out on the YOHO corpus [10]. The YOHO is an American English speaker verification database

which consists of speech data from 138 speakers 106 males and 32 females. The YOHO vocabulary consists of two digits number spoken in sets of three (e.g. thirty-six forty-five eighty-nine). For each speaker there are 4 enrolment sessions of 24 utterances each, and 10 verification sessions of 4 utterances each. In our experiments only 106 male speakers are used as this gives more speakers to evaluate the variation over number of speakers in cohort based normalisation. A single enrolment session of 24 utterances is used for creating a speaker template. For each enrolment session we select 80 speakers to calculate normalisation factor for the cohort approach and the rest of 26 speakers are used for testing. Within the database each speaker provides 40 independent test utterances. In order to estimate the false acceptance rate for each speaker another 40 test utterances are randomly drawn from the test utterances of other 25 speakers. This gives an equal ratio of impostor testing and claimed testing. For each of four sessions a different selection of testing speakers is applied so that the total number of testing speakers for each experiment is equivalent to (26×4). The total number of testing utterances for each experiment is equal to $(26 \times 4 \times 80)$.

In the feature extraction a filter bank process is used to produce 32 filter bank coefficients every 15 ms and these filter bank coefficients are then transformed to 12 cepstral coefficients by cosine transformation. The 12 delta cepstral coefficients are derived from cepstral coefficients every 5 frames. The dynamic cepstral normalisation technique (also referred as cepstral mean subtraction), which was developed by Vocalis (former Logica) in EU SUNDIAL project [8], is applied to cepstral coefficients to remove long time shift on individual cepstral coefficient. Thus, the overall feature vector consists of 12 normalised cepstral coefficients and 12 delta cepstral coefficients.

The world models are produced using a large number of speech tokens from a large number of speakers, separately from YOHO. The modelling process was optimised to produce speaker-independent word models for speech recognition. Each model compromises 8 states with 10 mixtures per state. For each digit two models are produced to represent male and female.

A set of speaker-dependent word models is used to represent the speaker template. In enrolment this set of word models is generated from 24 enrolment utterances. Each digit model compromises 12 states with a single mixture per state. In verification a silence model is applied on the matching in the beginning and end of the sentences and between two words. For the world model method as both male and female speaker-independent digit models are used the one with better score is selected for score normalisation.

4.2 Baseline Results

Figure 1. shows the verification EER of the world model approach and the cohort model approach. This gives a comparison between two approaches and also defines a baseline performance for the evaluation of the hybrid approach. As described in Section 4.1 the experiment is based on an open set scenario in which all the test speakers are not used in the enrolment and normalisation factor calculation. In the figure SV results over cohort size are presented with both the speaker specific (SS) threshold and the speaker independent (SI) threshold method. The SS threshold method has been widely used for SV performance measurement as suggested in [9]. However lack of proper way to set the threshold for individual speaker leads a significant gap between laboratory result and real application system. Therefore the SI threshold method remains attractive for real SV applications. In this paper the evaluation results are presented using both SI threshold and SS threshold method. For cohort size 80 the results are derived from 8320 testing utterances of 104 speakers as described in Section 4.1. For cohort size 20 and 40 the set of 80 cohort speakers is divided into four and two groups respectively. For each group an experiment is conducted. The results in the figure represent an average EER of four sets of tests for the size 20 and an average of two sets of tests for the size 40 to cover all 80 of cohort speakers being used. With the world model approach a set of speaker-independent speech recognition models is used so the results from this approach does not change over cohort size.

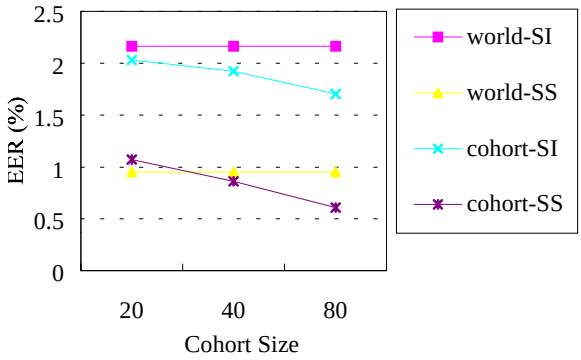

Figure 1. SV performance using the cohort model approach and the world model approach

From the figure it can be seen that the cohort model approach depends on the size of population from which the cohort speaker is selected. With SI threshold the cohort model approach gives better performances over all of three cohort sizes. With SS threshold the cohort model approach gives better performances for cohort size more than 40, but not for cohort size 20. In general, the cohort model approach gives close performance as the world model approach at about cohort size 20, and gets better as the cohort size increases. However the increase of cohort size also implies the increase of computation for the cohort model based normalisation. The results also suggest that the comparison between two approaches with different threshold methods could lead to different conclusions. This gives another reason for using both threshold methods for our evaluations.

4.3 Evaluation of Hybrid Approach

Figure 2. shows SV EER using the hybrid approach with variation of combination parameter α. The experiment is to locate an optimal value of parameter α. The results in the figure are the average EER of cohort size 20, 40 and 80, and they are calculated by using SI threshold method. In the figure the optimal point for parameter α is equal to 0.6 where the hybrid

approach show about 15% error reduction from the cohort model method and about 20% reduction from the world model method.

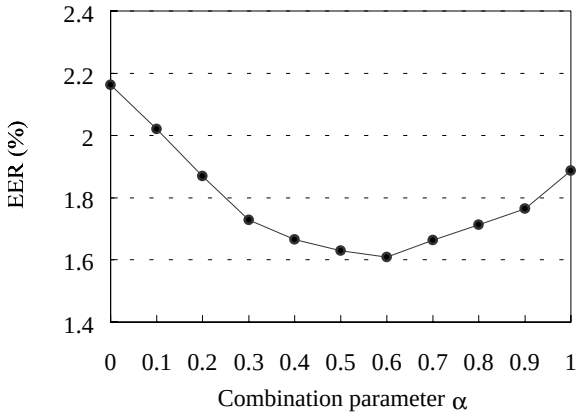

Figure 2. Average SV EER using the hybrid approach versus combination parameter α

Figure 3. illustrates a comparison of the hybrid approach with the cohort model approach and the world model approach with both SI and SS threshold methods. The parameter α value 0.6 is used in the experiment. In the figure the hybrid method shows overall improvements with both SI and SS threshold. For cohort size 80 it reduces EER from 2.16% (world) and 1.70% (cohort) to 1.56% using SI threshold, and from 0.95% (world) and 0.61% (cohort) to 0.52% using SS threshold. For cohort size 20 the hybrid approach reduces EER from 2.16% (world) and 2.03% (cohort) to 1.67% using SI threshold, and from 0.95% (world) and 1.07% (cohort) to 0.69% using SS threshold.

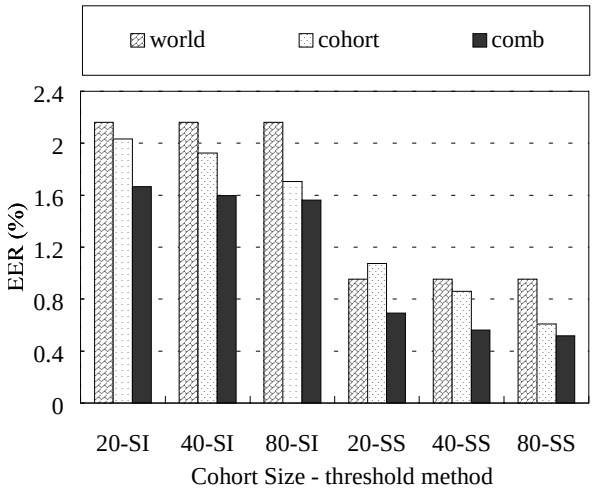

Figure 3. Comparison of the hybrid approach with the cohort model approach and the world model approach

5. CONCLUSIONS

From the theoretical analysis the cohort model approach and the world model approach are based on two different verification measurement paradigms. Combination of these two could lead a better balance of two measurements. In the paper a hybrid method has been described that combines two score normalisation techniques together. The results demonstrate that the hybrid combination can lead to a better HMM score measurement for verification, which improves SV performance. As the SV system is often combined with speech recognition the world models are available in the system such approach can be easily applied. A comparison between the cohort model approach and the world model approach is also given in the paper. The results indicate that SV performance by the cohort model based approach significantly depends on the cohort size. It gives close performance as the world model approach for cohort size 20, better performance as the cohort size increases.

7. ACKNOWLEDGEMENT

This work is partially funded by EU Telematics Programme through PICASSO project (LE4-8369).

6. REFERENCES

[1] Gu Y. and Thomas T., "An Implementation and Evaluation of an On-line Speaker Verification System for Field Trial", *to be appear in ICSLP-98*

[2] Rosenberg A.E., Delong J., Huang C. H. and Soong F. K., "The use of Cohort Normalized Scores for Speaker Verification", *Proc. ICLSP, pp. 99-106, 1996*

[3] Matsui T. and Furui S., "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model", *Speech Communication, Vol 17. pp. 109-116, 1996*

[4] Higgins A., Bahler L., and Porter J., "Speaker Verification using randomized phrase prompting", *Digital Signal Processing, Vol. 1. pp. 89-106, 1991*

[5] Carey M. J. and Parris E.S., "Speaker Verification using connected words", *Proc. Institute of Acoustics, Vol. 14, No.6, pp. 96-100, 1992*

[6] Caminero-Gil, F.J., Torre de la, C., Hernandez-Gomez, L.A., and Martin-del Alamo, C. "New N-Best Based Rejection Techniques for Improving a Real-time Telephonic Connected Word Recognition System", *Eurospeech-95, pp. 2099-2102*

[7] Tan B. T., Gu Y. and Thomas T., "Evaluation and Implementation of A Voice-Activated Dialing System with Utterance Verification", *to be appear in ICSLP-98*

[8] Gu Y. and Thomas T., Reported *in SUNDIAL Project (P2218) Report D6, EU ESPRIT Programme, 1993*

[9] Bimbot F. and Chollet G., "Assessment of Speaker Verification System", *In: Spoken Language Resources and Assessment, EAGLES Handbook, 1995*

[10] Campbell J., "Testing with the YOHO CD-ROM Voice Verification Corpus", *Proc. ICASSP-95, Vol. 1., pp. 341-344*