AN IMPLEMENTATION AND EVALUATION OF AN ON-LINE SPEAKER VERIFICATION SYSTEM FOR FIELD TRIALS

Yong Gu and Trevor Thomas

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

ABSTRACT

This paper presents a HMM-based speaker verification system which was implemented for a field trial. One of the challenges for moving HMM from speech recognition to speaker verification is to understand the HMM score variation and to define a proper measurement which is comparable across speech samples. In this paper we define two basic verification measurements, a qualifier-based measurement and a competition-based measurement, and examine score normalisation approaches using these two measurements. This leads to some useful theoretical differentiation between cohort model and world model approaches used for HMM score normalisation. We adopted a world model method for score normalisation in the system. The adaptive variance flooring technique is also implemented in the system. The paper presents evaluation results of the implementation.

1. INTRODUCTION

The CAVE project (CAller VErification in Banking and Telecommunications) was a two-year EU-funded project to assess speaker verification (SV) technology in banking and telecommunication applications. The project involved nine European partners and was completed in late 1997. This paper presents a SV system that was implemented for field trials in the CAVE project. The system adopts an HMM-based approach and world model method for score normalisation. The system is text-dependent and has been evaluated using the SESP database collected with a consideration of simulation for field trial in a Dutch calling card application.

One of challenges for moving HMM from speech recognition to speaker verification 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 score measurement in both SV [2][3][4] and utterance verification (UV) [6][7]. In Section 2 we define two basic verification measurements, qualifier-based measurements and competition-based measurements. With these measurements we examine the two main score normalisation methods used in HMM-based SV, the cohort model approach and the world model approach, which leads to some useful theoretical differentiation between these two methods. Section 3 presents the implementation of the SV system. Section 4 gives some evaluation results. The adaptive variance technique (AVF) [1][5], which is used in the system, is evaluated. Evaluations on different amount of enrolment data are also given in the section. The performance is very encourage. Section 5 summarises conclusion.

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

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

where $P(S,I,T_i)$ is a measurement between sample S and claimed template T_i. 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 verification decision making. 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

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

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*, $P(M(S,W) \mid O)$. Using Bayes' Rule the following equation can be derived

$$P(M(S,W) \mid O) = \frac{P(O \mid M(S,W))P(M(S,W))}{P(O)} \quad (6)$$

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

$$w = \arg\max_i \{P(M(Wi) \mid O)\} = \arg\max_i \{\frac{P(O \mid M(Wi))}{P(O)}\} \quad (7)$$

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

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 $P(M(S,W) \mid O)$ 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 $P(O)$

$$P(O) = \sum_i^{\infty} P(O \mid M(Si))P(Si) \quad (8)$$

is not possible to be calculated fully. The measurement $P(O \mid M(S,W))$ 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.

Two main approaches, the cohort model method [2][4] and the world model method [3][10], have been proposed to normalise the score $P(O \mid M(S,W))$ for better measurement. In the cohort model approach, a competition-based measurement is adopted. 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.

$$R_{cohort} = \frac{P(O|M(Si,W))/P(O)}{\max_{j \neq i}\{P(O|M(Sj,W))/P(O)\}} \quad (9)$$

Apparently the cohort model approach leads to a competitive-based measurement without estimating $P(O)$. Theoretically it fits well for a close set verification. However the selection of the competitive speaker does not depend on the claimed speaker but depends on test sample. The test sample from different speaker (imposter) may lead different selection of cohort speaker. This makes it difficult to specify the cohort speaker beforehand [3]. Online selection of the cohort speaker leads to more computation in the online system. As this approach is based on competition-based measurements, the verification theoretically depends on the existing speaker templates to represent unknown imposters. This could also lead some instabilities of SV system for open set verification tasks.

With the world model approach a set of text-dependent speaker independent word models, is used as world model to normalise the score $P(O \mid M(S,W))$. This set of world models is generated from a large number of speech samples from a large number of speakers. In this approach, assuming that $M(S_{world},W)$ is a world model for word *W* the score $P(O \mid M(S_{world},W))$ can be calculated. Therefore the normalisation score becomes

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

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 $P(M(S,W) \mid O)$ for the verification decision making comparison. As the SV system is often combined with speech recognition the world models are available in the system and no other speaker templates are required for this measurement. So this approach becomes easy to implement. Theoretically, the world model approach seems to be more stable for open set applications since the score normalisation does not involve any other speaker templates.

3. SPEAKER VERIFICATION SYSTEM

Figure 1. shows an overall diagram of the SV system. The system is implemented in a real-time fashion. The incoming speech is initially converted into a sequence of acoustic feature vectors in the pre-processing of the feature analysis. In the enrolment, a set of speech utterances is collected from a speaker and each sentence is checked and segmented by speech recogniser. These segmental units are then classified according to their recognition result and a collection of segmental units for each vocabulary item is used to produce a HMM model. A set of speaker-dependent word models is produced and used as a speaker template for verification. In verification, the speech feature vectors are matched with a speaker template according to the given speaker identify and word sequence. A recognition process is used to check whether the input utterance is an expected sentence before the verification. The incoming utterance is also matched with a sequence of world models to obtain scores for normalisation. The scores from both speaker dependent and world models are then used for final verification binary decision making for rejection or acceptance.

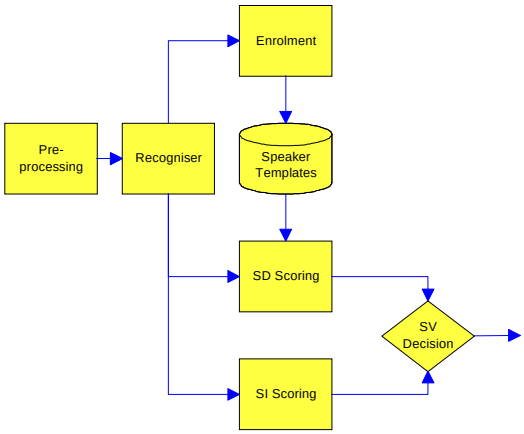

Figure 1. Block diagram of the speaker verification system

3.1 Acoustic Analysis

Speech signals were recorded at 8 kHz from public telephone network. 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 same front-end processing is used for both speaker verification and speech recognition.

3.2 Speaker Template Building

A word-based HMM is used to represent the speaker template. In the system only digits are used for verification. A speaker template consists of a set of speaker dependent digit models. A digit model compromises 12 states with a single mixture per state. Diagonal covariance is used for each mixture and 3 learnable transitional probabilities, itself; next state and the one after are applied. A speech recogniser is used to check the incoming utterances and segment digit string into word level by matching enrolment utterance with speaker independent digit models. Training tokens are collected from enrolment utterances for each digit to build the speaker dependent digit model.

3.3 Verification

A Viterbi algorithm is used for recognition and HMM score measurement for verification. Each speech utterance is checked by the recogniser prior verification. This has been crucial to ensure the stability of the HMM score. The world model score normalisation method is adopted in the system. Two sets of speaker independent digit models, male and female, which are used for speech recognition, are also used as world models in the verification. In the process two scores are calculated for which one is from matching with speaker dependent models and another is from matching with world models male or female (the better one). Same silence models are equally applied on matching in the beginning and end of the sentences and between two words. With world model based score normalisation the log likelihood ratio

$$R_{\log} = \{\log(P(O \mid M(S_{claimed}))) - \log(P(O \mid M(S_{world})))\} / N$$

is computed from two scores for verification decision. N is the number of frames in the test utterance.

3.4 Adaptive Variance Flooring Technique

The adaptive variance flooring (AVF) technique, which is used in [1][5], is adopted in our system. With this technique a global variance of feature vectors is computed from a significant amount of speech data. A flooring variance is defined as a proportion of this global variance i.e. global variance multiplied by constant factor. This flooring variance is then applied in the process of speaker template building. During modelling if any element in the estimated variance vector is less than the corresponding element in the flooring variance the estimated variance element is replaced by the flooring variance element. In [1][5], the AVF technique has been applied in the process of world model modelling process. In our system the AVF for world model is slightly different. As we use existing speaker independent speech recognition word models as world models, we adopt an approach to apply to the AVF technique to the resultant speaker independent models rather than in the modelling process.

4. EVALUATIONS

4.1 Database

The evaluation has been carried out on the SESP database collected by KPN research for SV evaluation. SESP contains telephone utterances of 24 male and 24 female speakers calling with different handsets and from wide range of places. Some of the calls were made from another country. In our experiments the 21 male and 20 female speakers for whom there is sufficient speech material to use. The evaluation has been carried out to simulate speaker verification in a telephone calling card application for a field trial in Dutch PTT. The utterances used for experiments are calling card number, a sequence of 14 digits in Dutch.

4.2 AVF Technique Effects

The first evaluation result in Figure 2 is to demonstrate the effectiveness of AVF technique in SV. The result in the figure are based on Gender-Balanced Sex-Independent Equal Error Rate (EER), following the EAGLES recommendation [9]. These results are based on 8 utterances for enrolment. About 3000 utterances are used for testing. The **Baseline** result is obtained from the system without AVF. The equal error rate is 1.4%. **Spkr-only** represents the result by applying the AVF in speaker template training for which the EER is 0.7%. By further flooring the world model the EER is reduced to 0.5% as indicated by

Both. The experiment shows that applying AVF technique in building speaker template can reduce over 40% of errors. With AVF applying to world models the error rate can be reduced further about 30%.

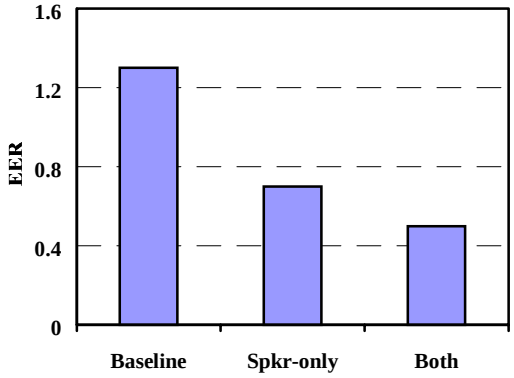

Figure 2. Comparison of verification error rate using AVF technique

4.3 Enrolment Data Effects

In the second experiment we evaluate the verification performance with different number of utterances for enrolment. In this experiment we compared the verification results by using 4, 6 and 8 enrolment utterances respectively. Their equal error rates are shown in Figure 3.

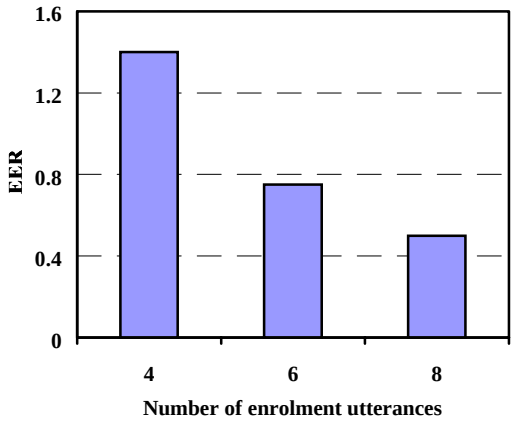

Figure 3. Comparison of error rates using different number of utterances for enrolment

It is shown in this figure that the error rate can be reduced dramatically with increasing enrolment utterances. The error rate nearly be cut by half with increasing enrolment utterances from 4 to 6 and further reduced from 0.75% to 0.5% by adding two more utterances. This experiment is limited by the database size the result is unable to indicate when the error rate may converge. However the result suggests that increasing enrolment data can improve the system performance dramatically, particularly when the amount of enrolment data is limited.

5. CONCLUSIONS

The paper presents an implementation and evaluation of online SV system based on HMM approach and world model method for score normalisation. The best EER is 0.5% with 8 utterances for enrolment. The AVF technique is applied in the implementation and the evaluation result shows that such technique gives over 60% error reduction. In the paper we also defines two basic verification measurements, qualifier-based measurement and competition-based measurement, and examine cohort model approach and world model approach using two measurements which leads some useful theoretical differentiation of two methods.

6. ACKNOWLEDGEMENT

This work is partially supported by EU Telematics Programme through CAVE project (LE-1930).

7. REFERENCES

1. Bimbot F., Hutter H,-P., Jaboulet C., Koolwaaij J., Lindberg J., Pierrot J.-B., "Speaker Verification the Telephone Network: Research activities in the CAVE Project", *Proc. Eurospeech-97, pp. 971-974*

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. Melin H., "Optimizing Variance Flooring in HMM-based Speaker Verification", *COST-250, Ankara, April 1998*

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