A COMPETITION-BASED MEASUREMENT FOR HMM 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

Score normalisation with cohort speaker models has been widely used in HMM-based speaker verification. Most of the proposed methods are based on the framework of the hypothesis testing. Based on this framework an overall average of all cohort scores is often used for normalisation, which leads a log likelihood ratio (LLR) for verification. In this paper we use a competition-based criterion to define the measurement. Based on this criterion a measurement is proposed, which reflects the competitiveness of claimed model against cohort models for a given testing utterance. The evaluation is carried out on the YOHO database.

1. INTRODUCTION

Speaker verification (SV) has drawn a great deal of interest in commercial environment with continuously increasing verification performance improvement. The HMM approach, which is successfully used for speech recognition, has been also widely used for text-dependent SV. One of challenges for moving HMM from speech recognition to SV is to understand the HMM score variation and to find a proper measurement which is comparable across speech sample domains. This is different from recognition as recognition tasks require only score comparison across models for a given utterance. Some approaches have been proposed for better HMM score measurements with score normalisation techniques. Score normalisation based on some composite models trained with a number of speakers are used for both text-dependent and text-independent speaker verification [1][2]. Score normalisation with cohort speaker models are also widely used [3][4]. In the cohort model approach HMM scores from claimed speaker model and cohort speaker models are used to re-calculate a measurement score for verification decision. The re-calculation is often based on the framework of the hypothesis testing [3]. Thus the cohort speaker models are considered to form a score for an alternative hypothesis and an overall average of cohort scores is used for score normalisation. In this paper we introduce a different approach to re-calculate a score with cohort speaker models. The new approach is based on a criterion of competitiveness[5]. With this criterion the measurement score for verification is to reflect how competitive the claimed model matching with the testing sample comparing with cohort models. Hence a competition-based measurement is defined on the score from claimed speaker model and sorted scores from cohort speaker models. This proposed measurement is evaluated on YOHO database [6] in the paper. It shows that the proposed measurement can improve the SV performance significantly. The evaluation results demonstrate that the SV equal error rate can be reduced by about 30% in comparison with the baseline approach.

2. HMM FOR SPEAKER VERIFICATION

In the HMM pattern matching, a 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. With the HMM framework the model M(S,W) can be estimated from a set of speech samples and a measurement P(O|M(S,W)) between a given sample O and a model M(S,W) can be calculated. However it is a common experience that the measurement P(O|M(S,W)) does very much depend on circumstances of sample O and the measurement score can vary significantly with variation on background noise, channel conditions etc. In recognition the task is to select one with the best score from a list of models based on such measurement for a given sample. The score variation is more consistent across the templates so that the comparison is less affected. In verification, however, the measurements are required to compare across the sample domain O. In such case the measurement P(O|M(S,W)) has been proved not robust for verification.

Using score normalisation techniques has improved SV performance significantly. In [3] score normalisation based on hypothesis testing framework is proposed in which the normalisation is to integrate both the null hypothesis H0: i(O) = c and the alternative hypothesis H1: i(O) ∈ I - c

\[ LR = \frac{L(i(O) = c)}{L(i(O) \in I - c)} \]  

(1)

where \( L(i(O) = c) \) is the likelihood from the claimed speaker i.e. \( P(O|Mc) \) and \( L(i(O) \in I - c) \) is the likelihood of alternative hypothesis. In the cohort model approach a set of models from cohort speakers is available and the scores with these models for a given test utterance can be used together with the score from claimed speaker model to form a measurement for verification decision. The average score from all cohort speaker models is
used as an approximation of the likelihood function for alternative hypothesis. This leads a measurement on log domain, log likelihood ratio (LLR) for verification decision.

In this paper we introduce a different approach to re-calculate the score with cohort models for verification. The new approach uses a competition-based criterion. With this criterion the measurement is defined to reflect how competitive the claimed model matching with the testing sample comparing with cohort models. The concept of this criterion is introduced in [5] and the relative score of claimed model with the best cohort score is used successfully for speaker verification. In this paper we extend this approach further by using a number of top scores from the cohort models and a weighting function is also introduced to reflect the significance of each score on particular sorting index. Thus the normalised score still reflects the competitiveness of the claimed model and the cohort models in general. Moreover by using more cohort scores for normalisation it leads a more stable measurement score for verification.

3. A COMPETITION-BASED MEASUREMENT

Given a set of cohort models \((M_1, M_2, \ldots, M_N)\) for a test utterance \(O\) the HMM likelihood scores \(L = P(O|M)\) from these cohort models are sorted and formed a measurement vector \(V\) together with the score from the claimed speaker model \(L_c = P(O|M_c)\)

\[ V = (L_c, L_1, L_2, \ldots, L_N) \]

where \(L_i > L_j\) if \(i < j\) for \(1 \leq i, j \leq N\).

The index for the cohort score is different from the index of the cohort model \((M_1, M_2, \ldots, M_N)\). For different test utterance the index for the score could map with different cohort model depending their competitiveness with the particular utterance. The model identity associated with cohort scores is less irrelevant in score vector \(V\). The competition-based measurement (CMP) is defined as a function on the measurement vector \(V = (L_c, L_1, L_2, \ldots, L_N)\). Considering exponential nature of HMM score we define this CMP measurement as

\[ S_{CMP} = \log(L_c) - \sum_{i=1}^{K} w_i \times \log(L_i) \quad (2) \]

which gives a relative measurement between the score from the claimed model and a weighted average of the top \(K\) cohort scores on log-likelihood domain. On the likelihood domain this measurement is equivalent to the following equation, a weighted geometric mean (WGM) of the sorted cohort scores.

\[ R_{WGM} = \frac{P(O|M_c)}{\prod_{j=1}^{K} P(O|M_j)^{w_j}} \quad (3) \]

When \(K = N\) and \(w_i = 1/N\) this gives a normalisation approach with a geometric mean (GM) of cohort scores.

In this paper we investigate a number of variations in the measurement equation in order to find a better score measurement and optimise for verification. The variations are listed in Table 1.

1. Choose different \(K\) and use the average top \(K\) cohort scores as a normalisation factor i.e.

\[ w_i = \begin{cases} 1/K & i \leq K \\ 0 & i > K \end{cases} \]

2. With individual weight (IW) for the top \(K\) scores and \(\sum_{i=1}^{K} w_i = 1\)

3. With shared weight (EW) for selected cohort scores and no restriction on the summation of weights.

4. With individual weight for selected cohort scores and no restriction on the summation.

| Table 1. SV performance for combined sorted cohort scores |

The reason to constraint the summation of weights equal to one is to keep the exponential factor same on denominator and numerator in Equation 3, which we call a symmetrical (SYM) normalisation. Thus this measurement approach is equivalent to the geometric mean approach. For variation 3 and 4 in the table the resultant weights may not satisfy the condition that the summation of the weights is equal to one. This leads different exponential factor on denominator and numerator in Equation 3, which we call an asymmetric (ASYM) normalisation. Most of score normalisation methods can be considered as symmetric. Compare with the symmetric normalisation the asymmetric approach gives less restriction on the search space for weights. Mathematically the asymmetric factor may have a significance, which reflects a proportional relation between the claimed score and the factor for competitiveness measurement. It would be interesting to know whether the asymmetric factor can lead better measurement solutions.

In the approach the weights in Equation 2. are determined using linear discriminant analysis (LDA) [7]. The Fisher’s discriminant criterion is used for our evaluation. For variation 2 the weights are trained on \((\log(L_c/L_1), \log(L_c/L_2), \ldots, \log(L_c/L_N))\). This results in that the summation of weights in the equation is equal to one. For variation 4 the weights are trained on the vector \((\log(L_c), \log(L_1), \log(L_2), \ldots, \log(L_N))\). For variation 3 a single weight can be trained on the vector \((\log(L_c), \sum_{i=1}^{K} \log(L_i))\). Thus the obtained weights may lead the asymmetric normalisation.
4. EXPERIMENTAL RESULTS

4.1 Experimental Conditions

The experiments were carried out on the YOHO corpus [6]. 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 used for cohort model purpose 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 \((26 \times 4)\). The total number of testing utterances for each experiment is equal to \((26 \times 4 \times 80)\) for cohort size 80. Two other cohort sizes 20 and 40 are used in the evaluation. For these cohort sizes the set of 80 cohort speakers is divided into four and two groups respectively. For each group an experiment is conducted. The average results of four group for size 20, two group for size 40 is presented respectively to cover all 80 of cohort speakers being used. Speaker independent threshold is used in verification decision.

In the feature extraction the front-end processing produces 12 cepstral coefficients and 12 delta cepstral coefficients every 15 ms. The dynamic cepstral mean normalisation 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 giving a feature vector size of 24. The evaluation is carried out with text-dependent SV. A set of speaker-dependent digit models is trained to represent the speaker template. Each speaker template is built 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. Each HMM score is normalised by the sentence length for score analysis.

4.2 Normalisation with Top \(K\) Cohort Scores

This first experiment is to show the significance of \(K\) most competitive cohort scores for normalisation in comparison with the top 1 score and the overall average cohort scores for normalisation. Figure 1. shows SV equal error rate (EER) using different \(K\) with three cohort size 20, 40 and 80. The horizontal axis indicates that how many cohort scores are selected to form a normalisation factor. To unify over different cohort sizes we use a percentage of total cohort speakers as a scale for the experiment. The top 1 score method is indicated by 1. In general the results clearly indicate that the sorted cohort scores after certain point (30%) is counter-productive for normalisation. In the graph it is shown that the top 1 and the overall average give close performance. The error rate can be reduced over 20% by selective parameter \(K\) at 30% of cohort sizes for all three sizes. All three curves gives relative same trend in the figure.

![Figure 1. SV performance using normalisation with top \(K\) cohort scores](image1)

In the second experiment we are going to compare with all four variations to look at the differences using individual weights (IW) and shared weights (SW), symmetric (SYM) normalisation and asymmetric normalisation (ASYM). Cohort size 40 is used for this experiment. For LDA 50% of test scores are used for training the weight(s) and another 50% are used for EER calculation. Two alternatives are used for average results. Figure 2. gives SV EER over four different methods. The results from curve SYM-SW is same as in Figure 1. The results show that the asymmetric normalisation overall outperforms the symmetric approach significantly. The asymmetric factor gives some compensation on using high order of sorted cohort scores. For example, it reduces EER from 1.92% to 1.41% using normalisation with all cohort scores. The individual weight approach can significantly improve SV performance for the

![Figure 2. SV performance with different weighting strategy](image2)
symmetric normalisation, particularly on the high order. However it has little effect on the asymmetric normalisation. This suggests that the asymmetric normalisation approach with shared weight on top $K$ cohort scores is quite optimal for Equation 2. Thus only a single weight (shared) needs to be determined. The optimal point for $K$ remains 30% of cohort size for all weight strategies. The best EER 1.32% is obtained for ASYM-EW, same for ASYM-IW.

### 4.4 Comparison Results

**Figure 3.** ROC curves using on cohort size 40

![ROC curves](image)

**Figure 4.** SV EER using different normalisation approaches over various cohort size

![SV EER](image)

Figure 3. shows four ROC curves for some significant points in Figure 2. Four curves represent normalisation using the overall average approach (ALL-AVG), the top 1 approach (TOP-1), the average of 30% most competitive cohort scores (30%MC-AVG) and the asymmetric approach with 30% most competitive cohort scores (30%MC-ASYM). The 30%MC-AVG gives an overall improvement from both TOP-1 and ALL-AVG. The asymmetric normalisation approach gives a further overall improvement. For example, at the false rejection rate 2.0% the false acceptance rate is reduced from 1.81% (TOP-1), 1.82% (ALL-AVG) to 1.19% (30%MC-AVG) and 1.04% (30%MC-ASYM).

Figure 4. summerises some comparisons with the baseline $LLR$ methods on different cohort sizes. The results demonstrate that $LLR$, ALL-AVG and TOP-1 methods perform closely. Improvements are obtained with using the 30% most competitive cohort scores and further improvements are obtained with asymmetric normalisation method. Overall in comparison with the baseline $LLR$ measurement SV ERRs are reduced by 28.5%, 31.3% and 30.6% on cohort size 20, 40 and 80 respectively.

### 5. CONCLUSIONS

In this paper we propose a measurement with the cohort model approach based on a criterion of competitiveness for speaker verification. The measurement is evaluated on YOHO database. Different weight strategies are experimented. The comparison with conventional $LLR$ is also given. The results show that using the competition-based measurement improves SV performance significantly. The evaluation results demonstrate that SV EER is reduced by about 20% using most competitive cohort scores and further about 10% by the asymmetric normalisation scheme. Different cohort sizes are used for the experiment. Improvements are obtained with all different cohort sizes.

### 6. ACKNOWLEDGEMENT

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

### 7. REFERENCES


