SECONDARY CLASSIFICATION FOR GMM BASED SPEAKER RECOGNITION

Jason Pelecanos, Dan Povey, Ganesh Ramaswamy

Conversational Biometrics Group
IBM T.J. Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY 10598
{jwpeleca, dpovey, ganeshr}@us.ibm.com

ABSTRACT

This paper discusses the use of a secondary classifier to reweight the frame-based scores of a speaker recognition system according to which region in feature space they belong. The score mapping function is constructed to perform a likelihood ratio (LR) correction of the original LR scores. This approach has the ability to limit the effect of rogue model components and regions of feature space that may not be robust to different audio environments, handset types or speakers.

Prior information available from tests on a development data set can be used to determine a log-likelihood-ratio mapping function that more appropriately weights each speech frame. The computational overhead for this approach in online mode is close to negligible for significant performance gains shown for the NIST 2004 Speaker Recognition Evaluation data.

1. INTRODUCTION

In adverse speaker recognition environments there is the problem of accurately generalizing the speaker's statistics based on information available from usually a single audio session. Consequently, a speaker model will not only represent statistics of the speaker, but also the speech content, channel and handset influences, and the audio environment. A number of techniques such as speaker model synthesis and feature mapping [1, 2], Bayesian and other constrained channel adaptation [3], attempt to accommodate for channel variation artifacts. These successful techniques, encompass prior information to enable a better generalization of the speaker's characteristics, but do not directly account for the importance of each speech observation towards the global decision.

A technique is presented that is capable of effectively weighting the importance of different observations in different regions of the feature space. This weighting function may be combined in concert with various forms of other techniques [1, 2, 3]. The score mapping approach utilizes development data to determine the weighting for each partition according to the partition's discriminative ability. The score mapping approach encompasses a secondary classifier (or output mapping function) that transforms the hypothesized single-session frame-based Log-Likelihood Ratios (LLRs) into log-likelihood ratios that take into consideration intersession mismatch.

A framework is proposed, whereby if the feature space can be decomposed into different partitions, and for each frame a log-likelihood ratio or other figure of merit can be extracted, then the scores can be mapped appropriately. Thus, a log-likelihood ratio determined from statistics available from a single session can be mapped to a log-likelihood ratio that is derived based on a generalization of the reliability of each of the feature space regions. In this paper, the framework is implemented in a simple form using an adapted Gaussian Mixture Model (GMM) structure [4]. Here, each mixture component of the GMM becomes a natural candidate for partitioning the feature space.

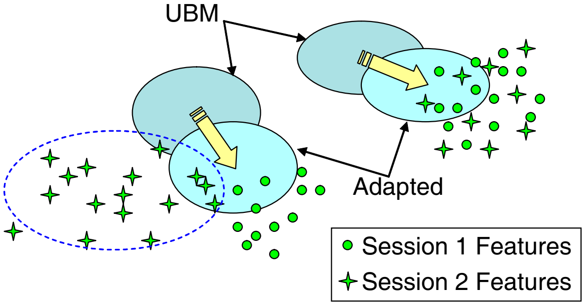

Fig. 1. *MAP adaptation for incomplete data. This example demonstrates features that do not obey the identically distributed assumption across audio sessions.*

In this work, a GMM based LLR architecture is adopted, where a target speaker model is derived by adapting a Universal Background Model (UBM) to data from a single session of a target speaker. A UBM is a GMM with a large number of mixture components trained on features from many speakers. The diagram in Figure 1 indicates the scenario for when mixture component secondary classifier rescoring becomes important. Figure 1 shows two mixture component pairs, where the target speaker mixture components are adapted towards the data from the first session. If all sessions following on from this matched the distribution of the first session (identically distributed assumption), then there would be no mismatch introduced and no need for rescoring. However, it is well known that there is significant variation across sessions attributed to a number of identified artifacts. If it is such that this variation across speaker sessions is significantly less for some mixture components than others, then it would prove useful to weight them appropriately. In addition, the figure shows that the second session of speech data from the same target speaker may indicate that the mixture component pair in the upper right is in the region of space that is more consistent across multiple sessions.

In order to limit these effects a mapping function in the form of a secondary classifier can be applied. The classifier models the frame-based log-likelihood ratio scores derived from the target and the background model. The LLR scores are modelled, for each coupled Gaussian pair of the GMM, and for target speaker and impostor speaker tests. The initial proposal involves modelling the score distribution of the target and impostor scores using a Gaussian for each. In order to better model the statistics using a Gaussian assumption, this approach was further extended to model the scores as a function of speaker model training soft counts.

This proposal saves on the need for complex training algorithms because the mapping is applied a Gaussian pair at a time. It however does not take into consideration dependencies on other Gaussian components but it also means that fewer classes and a smaller quantity of data are required to train it.

Section 2 presents an approach for evaluating the importance of each feature sub-space towards the overall classification decision. Section 3 extends the work to include splines into the solution for incorporating training parameter trends. This is followed by the system description (Section 4) and the corresponding experiments reported in Section 5. Section 6 concludes with the outcomes of this work.

2. SECONDARY CLASSIFICATION

The purpose of the secondary classifier (or secondary scoring) is to re-weight decisions made by the primary classifier (using the acoustic GMMs) that may not be made on information representative of the speaker. One of a number of studies [5] indicated that not all phonemes are equally discriminative, but in addition, this also implies that not all areas of the acoustic feature space are equally discriminative. The use of a secondary classifier presents a simple solution for evaluating the weight of evidence from the speaker's score and the consideration of the discriminative power of each partitioned region of the feature space. A mapping of this form is able to encompass both the discriminative ability present for a region and the score of the primary classifier in a joint manner.

The strength of the secondary classifier is that it can deemphasize the contribution of unreliable mixture components and emphasize discriminative regions. Let us consider a Gaussian Mixture Model that has its component means adapted to the target speaker's speech from a background model. It is assumed that for each feature vector being tested, only the most significant scoring mixture component from the background model and its corresponding target model component will be evaluated. Given this mixture component pair, the standard approach is to take the log-likelihood ratio of the probability densities.

In performing this calculation, a number of assumptions are made of which two are identified here. That is, the prior information sustained by the background model, provided to the target model in the adaptation process is representative of the target speaker properties. In addition, the scoring procedure then requires that either the target or background probability density functions are identically distributed.

Many of these problems are attributed to the lack of available training data from each speaker to accurately adapt the background model to the speaker. In addition, the speaker's features across sessions are not identically distributed with each other. Hence the i.i.d. assumption would be a poor assumption. Under i.i.d. the log-likelihood ratio would be an optimal test but under intersession mismatch a modification can be introduced to limit its effect. If the mismatch across speaker sessions is different for different feature space regions, then the secondary classifier can incorporate and allow for these mismatches in its log-likelihood ratio calculation.

For a coupled target and background model setup, a log-likelihood ratio, Λ, for each frame of speech, $\boldsymbol{x}$, may be extracted. The LLR is approximated using the coupled Gaussian pair with the highest scoring background mixture component. The log-likelihood ratio may be calculated for a Gaussian pair with the same diagonal covariance matrix $\boldsymbol{\Sigma}$, a target mean vector of $\boldsymbol{\mu}$ and a background component mean vector of $\bar{\boldsymbol{\mu}}$.

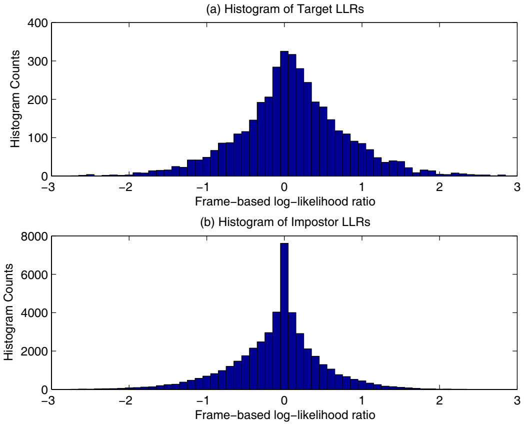

Fig. 2. *Distribution of target and impostor log-likelihood ratio scores for a single Gaussian pair across many speakers. The result was determined over the entire NIST 2003 data set.*

$$\Lambda = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) + \frac{1}{2}(\bar{\boldsymbol{\mu}}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \qquad (1)$$

The score distribution for each Gaussian pair for target and impostor tests can now be modelled. For a series of target score observations, $\{\Lambda_{tar}\}$, and non-target scores, $\{\Lambda_{non}\}$, the solution that maximizes the likelihood of the models given the corresponding observations may be determined. Although this may be regarded as a strong assumption, the first proposal will model the underlying score distributions as single Gaussians. A common Gaussian variance constraint is also introduced to provide a monotonic scoring response. The mapping amounts to a shift and scale applied to the LLR which is specific to each Gaussian in the UBM. The shift and scale parameters are learned from held out data. An equal weight to both the target (tar) and non-target (non) classes is imposed. ie. the constraint that observations from both classes will weight the parameter estimates with equal contributions. Here, μ_{tar} and μ_{non} represent the target and impostor score Gaussian means while σ^2_{tar} and σ^2_{non} represent the corresponding Gaussian variances.

$$\begin{aligned} \mu_{tar} &= E\{\Lambda_{tar}\} \\ \sigma^2_{tar} &= E\{\Lambda^2_{tar}\} - E\{\Lambda_{tar}\}^2 \\ \mu_{non} &= E\{\Lambda_{non}\} \\ \sigma^2_{non} &= E\{\Lambda^2_{non}\} - E\{\Lambda_{non}\}^2 \qquad (2) \end{aligned}$$

For the constraint that each Gaussian has the same variance, the tied variance (σ^2_{tied}) is the following result.

$$\sigma^2_{tied} = \frac{1}{2}(\sigma^2_{tar} + \sigma^2_{non}) \qquad (3)$$

For an N component GMM, there will be N sets of the μ_{tar}, μ_{non} and σ^2_{tied} statistics giving a total of $3N$ parameters. The parameter estimates for each Gaussian pair are used to estimate the

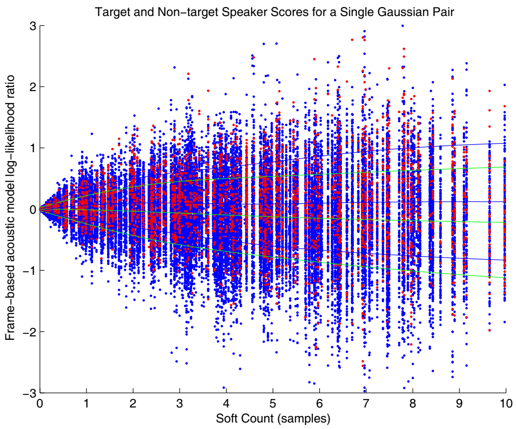

Fig. 3. *Distribution of log-likelihood ratio scores for a single Gaussian pair for different Gaussian adaptation amounts. The means and +/- one standard deviation for the target and impostor classes are indicated. (The result is extracted from the NIST 2003 speaker recognition database.)*

corrected frame-based log-likelihood ratio. In contrast to the basic LLR approach (with the target adapted GMM versus the background model producing the log-likelihood ratio estimate) there is a secondary classification level that corrects these single session LLR scores. Let $\hat{\Lambda}$ be the corrected log-likelihood ratio estimate. Equation 4 (acoustic space) is simply the one dimensional form of Equation 1 (score space).

$$\hat{\Lambda} = \frac{1}{\sigma^2_{tied}} \left\{ \Lambda(\mu_{tar} - \mu_{non}) + \frac{1}{2}(\mu^2_{non} - \mu^2_{tar}) \right\} \qquad (4)$$

The underlying assumption is that the target and impostor score distributions are Gaussian. By observing Figure 2, it is apparent that this is not the case. In addition, the impostor histogram tends to exhibit a slight negative skew while the target histogram presents a marginally positive skew. The following section extends the solution to lessen the impact of this issue.

3. SPLINE BASED MODELS

As identified in Figure 2 the component score distributions are not Gaussian and are less Gaussian for impostor trials than target trials. The benefit with using a single mean and variance pair for each Gaussian is the lower complexity. It is proposed that the log-likelihood ratio score distribution of the Gaussian pairs is also a function of the quantity of adaptation data for that mixture component, which is also termed the mixture component soft count (see [4] for background information). Figure 3 presents the target and impostor frame based scores for a single mixture component for a range of Gaussian component counts. For a single component count cross-sectional slice, the target and impostor score distributions are more Gaussian-like than the distributions that ignore the component count. Thus, the score distributions would be better represented if the mean and standard deviation of the target and impostor score distributions are estimated as a function of the model component count.

One approach is to determine the trend of the distributions for any particular count based on an underlying derived relation. A difficulty is that there are typically a number of assumptions that must be made, but there is the benefit of a terse number of parameters to describe the statistical trend. Alternatively, a spline may be used to describe the trend function of the mean and variance statistics of the target and impostor class scores. Similar work [6] was performed using splines to estimate the model response characteristics for the purpose of utterance length compensation.

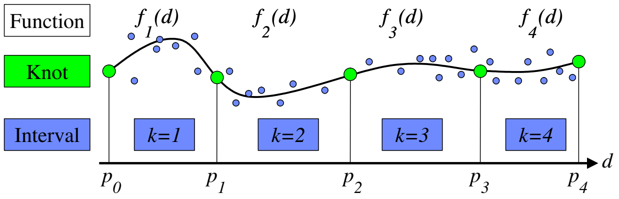

Fig. 4. *Spline based construction. The parameters* $\{p, k, f(d)\}$ *represent the spline knot positions, intervals and polynomial functions respectively.*

The advantage of a spline is that through the use of piecewise polynomials (see Figure 4), it can approximate a parameter value at any point within a bounded region. The spline can be trained using a least squares criterion. The piecewise polynomial approximation also produces a compact representation of the underlying data. This work uses a single cubic spline with six intervals to describe the target speaker score mean (for a single Gaussian component pair) as a function of the mixture component count for that model. The same procedure is performed for calculating the impostor spline function. The variance trend spline function is determined by taking the square of the difference between each target or impostor observation from the result determined by the corresponding spline mean estimator. The tied variance spline is estimated directly from the (merged target and non-target) squared difference statistics. Thus, there are three spline functions describing the statistics for each Gaussian pair. These spline functions were determined from frame scores with corresponding Gaussian counts between 1 and 10 samples. In testing, all frame scores were mapped according to bounded Gaussian counts.

4. GMM SYSTEM DESCRIPTION

The speaker recognition system uses 38 dimensional features comprising 19 Mel-Frequency Cepstral Coefficients and their corresponding deltas. Feature warping is also applied over a three second window for all features to reduce the effect of slowly varying additive noise and channel variability.

This GMM system is based on the work outlined in [4, 7]. The basic concept is that a universal GMM is trained on a large quantity of speech data from a wide variety of acoustic conditions. This model serves as the background model from which all other speaker models are adapted. In this work only the mixture component means are adapted using a single iteration of the Expectation-Maximization MAP algorithm. Once the speaker model is trained, speaker testing is performed by calculating the expected frame-based log-likelihood ratio between the target speaker and the universal background models. Scoring of the mixture components utilizes only the top-1 scoring mixture component as determined by the 2048 mixture component background GMM.

The target speaker scores are also normalized by the mean and standard deviation estimates of a set of held out impostor speakers.

Table 1. *Score Mapping Results.*

	Common Condition		Core Condition	
System	DCF (x1000)	EER (%)	DCF (x1000)	EER (%)
LLR (No T-Norm)				
Baseline	54.0	15.0	**54.5**	16.0
Gaussian Mapping	**53.8**	14.3	56.6	15.1
Spline Mapping	54.1	**13.1**	56.1	**13.5**
(Spline Self Test)⇒	(52.5)	(12.4)	(53.6)	(12.3)
T-Norm				
Baseline	48.7	14.5	50.9	15.4
Gaussian Mapping	49.3	14.1	51.0	15.4
Spline Mapping	**46.9**	**12.7**	**47.5**	**13.1**
(Spline Self Test)⇒	(45.6)	(11.7)	(45.6)	(11.9)

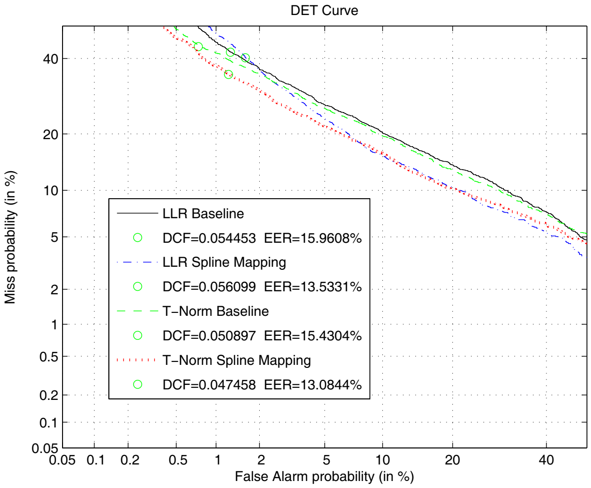

Fig. 5. *DET plot for the baseline and secondary classifier systems with and without T-Norm (core condition).*

This procedure is known as T-norm [8].

5. RESULTS

The systems were evaluated on the NIST 2004 Speaker Recognition database [9]. The NIST 2003 data set was used as the held out set for estimating the GMM component-wise statistics. The interesting contrast between these two databases is that the NIST 2003 speech database consists predominantly of cellular phone calls, while the NIST 2004 data set is comprised of a significant quantity of landline telephone calls. This mismatch will indicate how well the mapping generalizes.

These experiments examine the benefit of incorporating a secondary classifier layer that maps the acoustic model log-likelihood ratio scores to a corrected LLR score domain. The results constitute a table of minimum Detection Cost Function (DCF) and Equal Error Rate (EER) statistics and a Detection Error Tradeoff (DET) plot (see [9]). The minimum DCF is the minimum cost of an optimal tradeoff of weighted miss and false alarm probabilities.

Table 1 presents the NIST 2004 speaker recognition results for the common and core conditions. There are three score mapping systems evaluated in addition to the baseline. The Gaussian mapping is the mapping technique identified at the end of Section 2. The spline base mapping is the score mapping approach which is dependent on the model soft count. The spline self test item is the optimistic result when the spline parameters are trained on data from the NIST 2004 dataset. For the basic LLR system without T-Norm applied, the effect of LLR mapping tends to degrade the performance marginally for the DCF measurements but provides significant improvements at the equal error rate operating point. For both evaluation conditions, with or without T-Norm, the spline mapping improves the EER by at least 12% relative over the baseline. With T-Norm applied to the spline approach, both the EER and DCF are consistently enhanced.

Figure 5 plots the Detection Error Tradeoff (DET) performance for the basic GMM LLR system, the spline based LLR mapping and the corresponding T-Norm equivalents.

6. CONCLUSIONS

This paper identified the importance of log-likelihood ratio correction and demonstrated the potential improvements attributed to normalizing the frame based log-likelihood ratio scores according to the robustness of feature space sub-regions. In this work a mapping based on the training counts for each mixture component was proposed. Normalization was performed as a score shift and scale specific to each mixture component and its corresponding training count. The spline based mapping results indicated a consistent equal error rate reduction over the baseline across the measured conditions. Further work may examine mappings that are specific to handset and gender in training and testing.

7. REFERENCES

[1] L. Heck and M. Weintraub, “Handset-dependent background models for robust text-independent speaker recognition,” *ICASSP*, vol. 2, pp. 1071–1074, 1997.

[2] D. Reynolds, “Channel robust speaker verification via feature mapping,” *ICASSP*, vol. 2, pp. 53–56, 2003.

[3] P. Kenny, et al, “Factor analysis simplified,” *ICASSP*, vol. 1, pp. 637–640, 2005.

[4] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[5] J. Eatock and J. Mason, “A quantitative assessment of the relative speaker discriminating properties of phonemes,” *ICASSP*, vol. 1, pp. 133–136, 1994.

[6] J. Pelecanos, U. Chaudhari, and G. Ramaswamy, “Compensation of utterance length for speaker verification,” *Odyssey, The Speaker and Language Recognition Workshop*, pp. 161–164, 2004.

[7] G. Ramaswamy, J. Navratil, U. Chaudhari, and R. Zilca, “The IBM system for the NIST 2002 cellular speaker verification evaluation,” *ICASSP*, vol. 2, pp. 61–64, 2003.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[9] National Institute of Standards and Technology, “NIST speech group website,” http://www.nist.gov/speech, 2005.