interspeech 2008 incorporating SST 08

Making Confident Speaker Verification Decisions with Minimal Speech

Robbie Vogt, Sridha Sridharan and Michael Mason

Speech and Audio Research Laboratory,
Queensland University of Technology, Brisbane, Australia.
{r.vogt, s.sridharan, m.mason}@qut.edu.au

Abstract

Drastic reductions in the typical data requirements for producing confident decisions in an automatic speaker verification system are demonstrated through the application of a novel approach of score confidence interval estimation. The confidence estimation procedure is also extended to produce robust results with very limited and highly correlated frame scores. The early verification decision method evaluated on the 2005 NIST SRE protocol demonstrates that an average of 2–10 seconds of speech is sufficient to produce verification results approaching those achieved previously using an average of over 100 seconds of speech.

Index Terms: automatic speaker verification, confidence measures, verification decision confidence.

1 Introduction

Deploying a speaker verification system is a difficult task for several reasons. Typically these difficulties involve determining system parameters such as the required amount of speech for sufficiently accurate enrolment and for sufficiently accurate verification trials. Additionally, there is always the problem of estimating a threshold for acceptance and rejection and the eventual error rate to expect from such a threshold. This is particularly difficult in the presence of a significant mismatch between the development database and the anticipated deployment conditions. This set of related issues is also notably absent from published research in speaker recognition.

Ideally, a verification system would produce a verification confidence from a trial, as this is the most useful and usable result from a system designer perspective: Knowing that there is a 96% probability that an utterance was produced by speaker s makes it easy for a designer to employ Bayesian logic to produce the best possible system. There are two distinct impediments to this: Firstly it is essentially impossible to accurately estimate the prior probability of a true trial due to the difficulties in identifying the non-target class[1], and secondly, scores produced by verification systems would need to be representational of true likelihoods, which is rarely the case given the rudimentary statistical models, the difficulty in modelling the non-target class and score normalisation processes.

Work addressing the production of accurate likelihood ratios [1, 2] and the interpretation of scores that are *not* considered likelihood ratios [3] has been prompted by the importance of presenting confident and meaningful results in forensic applications. The analysis and evaluation of speaker verification systems based on the accuracy of output likelihood ratios is also an emerging topic of recent interest [4], but speaker verification systems do not in general produce scores that should be interpreted as likelihood ratios.

Given these difficulties with determining an accurate verification confidence, an alternative approach pursued in this work is to determine a method by which one can state that the "true" verification score for a trial lies within the range $\Lambda_S = a \pm b$ at, for example, the 99% confidence level. Here the "true" verification score is defined as the score that the verification system would produce given an infinite quantity of testing speech.

Using this approach, this work presents an initial attempt to address the issue of the quantity of speech required for sufficiently accurate verification results, by employing confidence measures on the verification score to determine the minimum speech required to make a confident verification decision at a specific threshold.

The following section describes the assumptions made on the nature of the score produced by current speaker verification systems and the effect of reducing the available test data has on the performance of such a system.

Section 3 presents the concept and applications of confidence measures for the speaker verification score. To account for the specific issues encountered in speaker verification several methods of estimating the verification score variance are then developed. This variance statistic provides the fundamental tools required to estimate the confidence of a verification decision. Experimental evaluation of these estimates are presented in Section 4 for the application of making confident verification decisions with as little data as possible.

2 Background

2.1 Baseline System and Experimental Setup

The verification system used in this study is a GMM-UBM system with inter-session variability modelling, as described in [5]. The verification score used for this system is the expected log-likelihood ratio of the target speaker to the UBM. The expectation is taken over the individual frame-based log-likelihood ratios for the test utterance,

$$\Lambda_S = \frac{1}{T}\sum_{t=1}^{T} \ell_S(t) = \frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{p(\boldsymbol{x}_t|\lambda_S)}{p(\boldsymbol{x}_t|\lambda_{ubm})}\right) \quad (1)$$

where, $p(\boldsymbol{x}|\lambda)$ is the standard GMM density.

This system uses explicit inter-session variability modelling [5] in the training procedure to mitigate the effects of mismatch, however session variability was not considered during testing. This configuration was chosen to have performance representative of the current state-of-the-art but avoiding the complication of estimating the session conditions of the testing utterance.

[1] In a forensic situation, deductive logic and other evidence may help in this regard.

Table 1: The effect of shortened test utterances on speaker verification performance.

System	EER	Min. DCF	Act. DCF
Reference	6.34%	.0293	–
20 sec	8.87%	.0391	.0422
10 sec	12.15%	.0489	.0601
5 sec	16.99%	.0616	.0976
2 sec	23.89%	.0794	.1770

Experiments were conducted on the 2005 NIST SRE protocol using conversational telephony speech drawn from the Mixer corpus [6]. The focus of these results is on the 1-side training, common evaluation condition of this corpus.

2.2 The Effect of Short Verification Utterances

From a researcher's perspective it is preferable to have as much speech as possible available for each verification to make the most accurate decision. This is the exact opposite of a system designer's preference to put the least possible demand on the end user. Compromise is usually necessary. To this end, it is important to have an understanding of the impact of limiting the verification utterance length. The impact of restricted test utterance length for a GMM-UBM system is presented in Table 1. These results demonstrate that utterance length, predictably, has a significant effect on overall system performance in the range that is typically of interest for a system designer, as previously observed [7].

3 Confidence Measures

Having recast the desire to estimate decision confidence as the desire to determine the confidence that a score produced lies within a given bound from the "true" score, there are a number of ways which such information could be used. Of specific interest in this work is the ability to use such information in order to shortcut a verification trial when we are confident that the "true" verification score is above or below a particular threshold. Other useful applications of this information come in the ability to; estimate the upper and lower bounds or errors for verification, estimate the level of confidence for which the verification score is above or below a threshold or shortcut a verification trial when we are confident the "true" score lies within a particular interval of the current score.

Assuming a verification score is a random variable drawn from a Gaussian distribution with a mean of the "true" verification score, the main difficulty arises because the variance is unknown and must be estimated. The variance of a trial score distribution is usually dependent on many factors including whether a trial is a genuine or impostor trial (which we obviously do not know *a priori*), the length of a particular verification utterance and the noise levels and other environmental conditions of the recording. These factors lead to the conclusion that the variance must be estimated for each trial *individually*. The observed frame scores of a trial are used as the fundamental statistics for estimating this variance. This estimation forms the basis of the presented techniques and is addressed in the next section.

3.1 Early Verification Decision Method

The aim of early verification decision methods is to minimise the amount of speech required to make a verification decision. This is achieved by making a verification decision as soon as we are confident the "true" verification score is above or below the specified threshold based on the confidence interval of the current estimated score.

The crux of confidence-based methods for verification is therefore the ability to estimate confidence intervals based on the observed sequence of frame scores. This ability in turn relies on estimating the variance of the mean estimate distribution from the sequence of frame scores. To do this, it is assumed that the observed verification score is a random *process* that evolves over time. It is assumed that this random process is Gaussian at time t, has a fixed mean (the "true" score) and a time-dependent variance, that is

$$\Lambda_S(t) \sim \mathcal{N}\left(\mu_S, \sigma_S^2(t)\right). \tag{2}$$

Presented below are several methods for estimating $\sigma_S^2(t)$.

Naïve Variance Estimate: As can be seen from (1), the verification score is the sum of the log-likelihood ratios of individual frames. The central limit theorem states that a sum of random variables will exhibit a Gaussian distribution. Furthermore, it is assumed for now that the feature vectors $\boldsymbol{x}_t$ and, by consequence, the frame log-likelihood ratios $\ell_S(t)$ are independent and identically distributed (iid) random variables. Thus, if $\ell_S(t)$ has sample mean m_ℓ and variance s_ℓ^2, the ELLR verification score will have a mean and variance approximated by

$$\mu_S = m_\ell \qquad \sigma_S^2 = \frac{s_\ell^2}{T-1} \tag{3}$$

Thus, for any sequence of frames $\boldsymbol{X}$ it is possible estimate the mean and variance of the ELLR score.

Using these estimates of the ELLR score statistics, a confidence interval for the "true" score can be calculated using a confidence level and the Gaussian cumulative density function.

Estimate with Correlation Compensation: Acoustic features commonly used for speaker verification, such as MFCC features, exhibit high levels of correlation between consecutive observation frames. This is due to the short-term spectra and cepstra calculated for consecutive frames sharing typically two-thirds of their waveform samples and the delta cepstra explicitly averaging over a number of frames. The mechanics of speech production also limits the rate at which vocal tract shape can change which causes correlations, a fact exploited by techniques such as RASTA filtering [8]. This correlation obviously voids the commonly cited assumption of statistically iid feature vectors.

Given the invalidity of the iid assumption, the estimated ELLR variance is invalid and empirical evidence shows that it is often underestimated, particularly with short sequences. For this reason, it is necessary to develop an alternative estimate to reduce the effect of this correlation.

In this research a transformation approach was adopted to reduce the correlation by producing a series of ELLR estimates $\boldsymbol{y}_S$ from short, fixed-length, non-overlapping frame sequences,

$$y_S(i) = \frac{1}{N} \sum_{t=Ni}^{N(i+1)-1} \ell_S(t) \tag{4}$$

where N is the length of the short frame sequences. If N is sufficiently large, the correlation between successive $y_S(i)$ drops to a negligible level.

From $\boldsymbol{y}_S$, it is then possible to estimate the overall ELLR mean and variance as

$$\mu_S = m_y \qquad \sigma_S^2 = \frac{s_y^2}{T/N-1} \tag{5}$$

where m_y and s_y^2 are the sample mean and sample variance of y_S respectively.

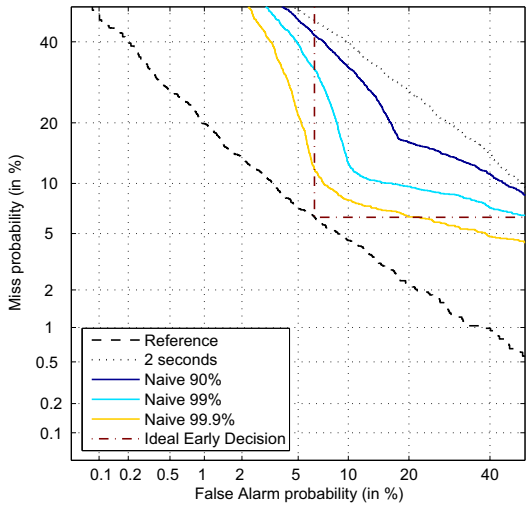

Figure 1: DET plot using the naïve method at the EER operating point.

Robustly Estimating the Sample Variance: For these techniques to be effective, it is important to robustly estimate the variance of the frame log-likelihood ratios with a very limited number of samples. This issue is also exacerbated by the correlated nature of these scores. One possible method to produce a more robust estimate of this variance is to introduce *a priori* information to the estimation, with the resulting estimate given by

$$\hat{s}^2 = \frac{\tau\kappa^2 + (M-1)s^2}{\tau + (M-1)}, \tag{6}$$

where s^2 is unbiased sample variance from M samples and κ^2 and τ are hyperparameters of the prior distribution, which takes the form of a Dirichlet distribution [9].

This estimate can then be used to produce more robust estimates of the ELLR variance, as estimated in (3) and (5) above.

4 Results

Fig. 1 shows the performance of a system employing early decision scoring using the naïve frame-based estimate in (3) with the threshold set for the equal error rate operating point at three confidence levels, 90%, 99% and 99.9%. These confidence levels are the minimum confidence with which the "true" verification score must be above or below the EER threshold for the system to make an early verification decision. Also shown is the DET curve for the baseline reference system using all available speech and a system using a fixed 2-second utterance length (dotted curve) as a "worst case" system as a 2 sec minimum length is also imposed on the early decision method.

As can be seen in Fig. 1 there is a significant drop in performance compared to the reference system due to the shortcut stopping criterion however there are some interesting aspects to this plot. First, the degradation in performance is actually quite modest as the reference system typically used *at least 6 times* the amount of speech to make a verification decision, as described in Table 2. This point will be addressed further below.

Second, a higher confidence level provides a better EER; Table 2 supports this with the *Naïve 99.9%* system showing an EER 8.9% lower than at the 90% confidence level.

Third, and more interestingly, the DET curves for these systems veer away from the reference system the farther they are from the EER operating point, both in the low false alarm and low miss regions. This characteristic is a direct consequence of the shortcut method as the system is only interested in the performance at the specified threshold and essentially trades performance in other areas for shorter test utterances.

Table 2: Verification results at the EER operating point for the early verification decision method.

System	EER	Trial Length		Shortcut Errors	
		Med.	Mean	Imp.	Target
Reference	6.34%	103.4	103.4	–	–
Naïve					
90% Conf.	17.64%	2	3.4	15.8%	14.4%
99% Conf.	11.26%	4	8.9	7.5%	7.6%
99.9% Conf.	8.73%	6	15.6	3.6%	4.3%
Decorrelated $N = 10$					
90% Conf.	12.62%	3	6.8	9.7%	9.1%
99% Conf.	7.74%	9	21.4	1.9%	2.6%
99.9% Conf.	6.66%	18	33.3	0.6%	1.0%
With Prior $\tau = 100, \kappa^2 = 0.25$					
90% Conf.	10.89%	4	8.6	7.2%	7.5%
99% Conf.	7.04%	12	24.5	1.0%	1.6%
99.9% Conf.	6.48%	21	36.0	0.2%	0.5%

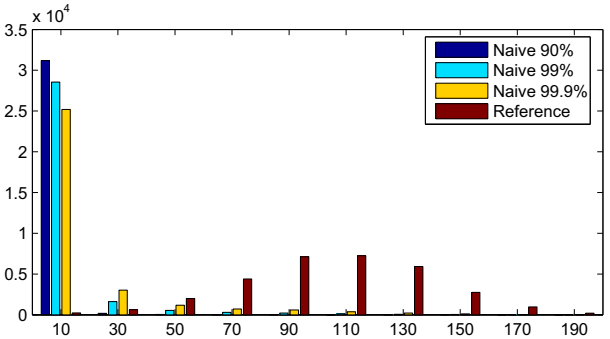

Figure 2: Histogram of the test utterance length using the naïve variance estimate method with the EER operating point.

In the ideal case the system would *only* provide performance at the specified threshold and trade all other performance for shorter trials (the curve labelled "Ideal" in Fig. 1). Using the equal error rate as the criterion, an ideal system would provide identical performance to the reference system.

By comparing the Tables 1 and 2 it can be seen that the shortcut method is effective in trading performance at a specific operating point for shorter trials. Comparing the results of the fixed 5 sec system to the 99% confidence level—with a median utterance length of 4 seconds—the EER improves from 16.99% to 11.26%, halving the gap to the reference with a shorter test utterance length on average.

Additionally, the *mean* test utterance lengths are dominated by a relatively small number of long trials with the majority of trials providing a result within 2, 4 or 6 seconds respectively for the systems in Table 2, as indicated by the *median* trial lengths for the Naïve system.

This last point has an astonishing implication: For the majority of trials a text-independent speaker verification system will produce the same decision with only 2 *seconds* of speech that it will with almost 2 *minutes* of speech. A better understanding of the distribution of trials lengths can be taken from the histogram in Fig. 2.

Presented in the two rightmost columns of Table 2 are the rates of errors introduced by the early decision criteria for impostor and target trials, respectively. These represent the trials for which the reference system and the early decision system have produced differing decisions. This is the loss introduced by the early decision methods and, if the distribution assumptions and estimates are accurate, should closely match the confidence levels specified.

It can be seen from these results that the error rates for the naïve system do not match the specified confidence levels well,

particularly as the confidence is increased. The fact that the error rates don't reflect the desired confidence levels suggests two possible issues. Firstly, the naïve variance estimates are not sufficiently accurate particularly when based on a small number of frames. Secondly, the assumption of a Gaussian frame score distribution is invalid. Observations of frame score distributions show that this is in fact a valid assertion as they exhibit significant third and fourth order statistics. This could particularly have an effect with very short utterances where there is not a sufficient number of observations for the central limit theorem to be valid.

Table 2 also presents the performance of the early decision method using the decorrelated distribution estimates from (5). This method is assessed with a short frame sequence length of $N = 10$ for it's ability to reduce the degree of correlation in the samples used to estimate the ELLR score distribution. With a typical frame rate of 100 frames per second, a value of $N = 10$ averages the frame scores over the period of a tenth of a second of active speech.

It can be seen from these results that decorrelating the samples used to estimate the ELLR score distribution does in fact reduce the proportion of errors introduced by the early decision scoring method (the two rightmost columns of Table 2), producing performance closer to that of the reference system. The best performing configuration drops only 0.32% at the EER operating point.

The errors introduced by the decorrelated early decision approach also produces errors at a rate much closer to the specified confidence level. While the rate at 99.9% confidence is still almost an order of magnitude too high, this result at least demonstrates that the variance estimated is more accurate with the data correlations diminished.

There is unfortunately an increase in both the mean and median utterance length associated with the decorrelated estimation method, however, despite this increase the median utterance lengths required are still very short at 3–18 seconds.

By incorporating *a priori* information in the variance estimate it is possible to reduce the performance discrepancy between the reference system and the early decision version to be insignificant. This improved performance unfortunately comes at the cost of longer verification utterances both in terms of the mean and median length statistics, as presented in last three rows of Table 2. Prior information was incorporated into the decorrelated variance estimate system with $N = 10$. The hyperparameter τ was fixed at the equivalent of 1 sec while a value of $\kappa^2 = 0.25$ was determined empirically.

Particularly noticeable for these results is the consistency between the specified confidence level and the rate of shortcut errors introduced by making early verification decisions at that confidence level.

Fig. 3 graphically summarises the performance of the early verification decision approach by comparing the EER to the median utterance length. Also presented are the fixed utterance-length systems as a reference. It is evident that the early verification decision method demonstrates consistently and significantly superior performance compared to specifying a fixed utterance length.

5 Summary

This paper introduced a novel method for estimating the confidence interval for the expected log-likelihood ratio scoring method used in speaker verification based on estimating the variance of individual frame scores. Several enhancements to this estimate were proposed to increase its robustness and accuracy for the peculiarities of GMM-based speaker verification.

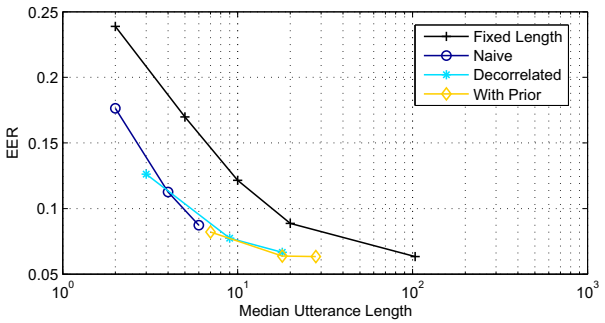

Figure 3: Median utterance length versus EER for the fixed short utterance and early verification decision systems.

One particular application for this information was explored to determine the minimum quantity of speech required to confidently make a verification decision based on a given threshold. This early verification decision method demonstrated that as little as 2–10 seconds of active speech on average was able to produce verification results approaching that of using an average of over 100 seconds of speech. Moreover, the performance loss incurred by making an early decision can be controlled by adjusting the confidence required in the resultant decision.

6 Acknowledgements

The authors would like to acknowledge the collaborative contribution of Torqx Pty. Ltd. on this research.

This research was supported by the Australian Research Council Discovery Grant No DP0877835.

7 References

[1] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 159–191, 2006.

[2] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 331–355, 2006.

[3] W. M. Campbell, K. J. Brady, J. P. Campbell, R. Granville, and D. A. Reynolds, "Understanding scores in forensic speaker recognition," in *Odyssey: The Speaker and Language Recognition Workshop*, 2006.

[4] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[5] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[6] A. Martin, D. Miller, M. Przybocki, J. Campbell, and H. Nakasone, "Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004," in *International Conference on Language Resources and Evaluation*, 2004, pp. 587–590.

[7] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluation—an overview," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 1–18, 2000.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[9] J.-L. Gauvain and C.-H. Lee, "Bayesian adaptive learning and MAP estimation of HMM," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. Soong, and K. Paliwal, Eds. Boston, Mass: Kluwer Academic, 1996, pp. 83–107.