Minimising Speaker Verification Utterance Length through Confidence Based Early Verification Decisions

Robbie Vogt and Sridha Sridharan

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

Abstract. This paper presents a novel approach of estimating the confidence interval of speaker verification scores. This approach is utilised to minimise the utterance lengths required in order to produce a confident verification decision. The confidence estimation method is also extended to address both the problem of high correlation in consecutive frame scores, and robustness with very limited training samples. The proposed technique achieves a drastic reduction in the typical data requirements for producing confident decisions in an automatic speaker verification system. When evaluated on the NIST 2005 SRE, the early verification decision method demonstrates that an average of 5–10 seconds of speech is sufficient to produce verification rates approaching those achieved previously using an average in excess of 100 seconds of speech.

1 Introduction

A number of practical issues inevitably arise in the process of deploying a speaker verification system. Typically these difficulties involve determining system parameters such as the required quantities of speech for adequately trained models and for accurate verification trials, as well as deciding an appropriate decision threshold to achieve the required verification error rates. Despite the importance of such decisions, very limited speaker verification research has been published that specifically address these issues. This work focuses on the issue of test utterance length.

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, accurately estimating the prior probability of a true trial is problematic due to the difficulties in identifying and quantifying the non-target class, and secondly,

* 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.

scores produced by verification systems would need to be representational of true likelihoods ratios, which is rarely the case for automatic speaker recognition systems.

Prompted by the importance of presenting meaningful results in forensic applications, recent work has begun to address the production of accurate likelihood ratios [1] and the interpretation of scores that are *not* likelihood ratios [2]. Also, the analysis and evaluation of speaker verification systems based on the accuracy of output likelihood ratios is also a topic of recent interest [3]. Regardless, speaker verification systems do not in general produce scores that should be interpreted as true 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 interval $\Lambda_S = a \pm b$ 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. The Early Verification Decision (EVD) method, first proposed in [4], exploits this verification score confidence interval to make confident verification decisions with minimal speech based on a specified threshold. This paper expands substantially on [4] and additionally investigates operating at the minimum DCF threshold and the interaction of the EVD method with Z-norm score normalisation.

The following section describes the baseline speaker verification system used in this paper and explores the effect on performance of reducing the available test data. Section 3 then presents the EVD method for minimising test utterance length by estimating confidence intervals on the speaker verification score. Several methods of estimating the verification score confidence interval are then developed including an extension to incorporate Z-norm score normalisation. Experimental evaluation of these estimates are presented in Section 4.

2 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,

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

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

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

System	No Normalisation		Z-Norm Normalisation	
	Min. DCF	Act. DCF	Min. DCF	Act. DCF
Reference	.0293	.0293	.0249	.0249
20 sec	.0391	.0422	.0368	.0406
10 sec	.0489	.0601	.0482	.0636
5 sec	.0616	.0976	.0626	.1031
2 sec	.0794	.1770	.0810	.1851

estimating the session conditions of the testing utterance. Additionally, Z-Norm score normalisation [6] was applied to this system.

Experiments were conducted on the 2005 NIST SRE protocol using conversational telephony speech drawn from the Mixer corpus [7]. The focus of these results is on the 1-side training, common evaluation condition of this corpus.

2.1 The Effect of Short Verification Utterances

While researchers typically prefer as much data as possible to make the most reliable verification decision possible, system designs desire utterances to be as short as possible to minimise the inconvenience for the user. Compromise is usually necessary. Thus, an understanding of the impact of limiting verification utterance lengths is important. Table 1 assesses this impact for the baseline system. 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 [8].

Table 1 presents both the minimum DCF value as well as the actual DCF value if the optimal threshold of the reference system is chosen. The substantial difference between the minimum and actual detection costs can be seen to be increasing as the utterance length is reduced, to the extent that it is more costly to use the 2-second system with the best threshold for the reference system than reject every verification claim *a priori* (this gives a DCF of 0.1). These numbers also highlight the difficulty of choosing a suitable threshold as this choice is evidently affected by the choice of utterance length.

Results including Z-Norm score normalisation are included in the rightmost columns of Table 1. This application shows a clear advantage for the reference system with 15% reduction in DCF. This advantage, however, is less apparent when shortened utterances are used. This is particularly apparent in the case of the actual DCF results using 2 and 5 second utterances where the application of Z-Norm has a detrimental effect.

3 The Early Verification Decision Method

The aim of the Early Verification Decision (EVD) method 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

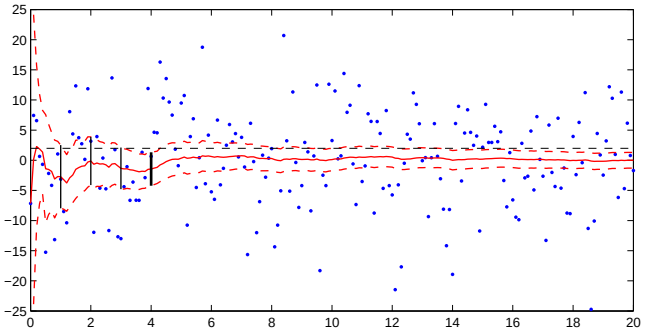

Fig. 1. Example verification trial using the early decision method. After observing only 4 seconds of speech a *reject* decision can be made.

score is above or below the specified threshold based on the *confidence interval* of the current estimated score.

The current verification score estimate is assumed to be a random variable drawn from a distribution with the "true" score as mean. To determine the confidence interval it is thus necessary to determine the variance of this distribution. This variance is usually dependent on many factors such as whether a trial is a genuine or impostor trial (which we can not know *a priori*), the length of a particular verification utterance and the noise and other environmental conditions of the recording. Consequently, the variance must be estimated individually for each verification trial using the observed sequence of frame scores as the fundamental statistics for this estimation. This estimation forms the basis of the EVD method and is addressed in the next section.

An example of the early verification decision process is presented in Fig. 1. In this figure, the samples (frame scores) used to estimate the distribution are represented as dots, the evolving mean verification score estimate is shown as a thick red line and the 99% confidence interval of this estimate depicted with dashed lines above and below the estimate. The verification threshold is shown as a horizontal line through the centre of the figure. After a couple of seconds of the trial the estimate of the verification score is quite eratic, which is reflected in the wide confidence interval, but looks to be converging to a point below the threshold. By four seconds the estimate seems to be more stable as more samples become available and the width of the confidence interval has narrowed to be entirely below the threshold. At this point, *after only four seconds*, we can be confident that the verification score will continue to lie below the threshold and thus make a reject decision for this trial. The subsequent part of the trial confirms that the verification score does in fact continue to lie below the threshold and the confidence interval continues to narrow, even though the entire confidence interval does not necessarily lie below the threshold at all times.

3.1 Variance Estimation Approaches

As detailed in [4], the crux of confidence-based EVD method is the ability to estimate confidence intervals on the ELLR score. This ability in turn relies on

estimating the variance of the ELLR estimate distribution from the sequence of observed 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 in [4] were several methods for estimating $\sigma_S^2(t)$ of this process. These methods are summarised here.

Firstly, the *Naïve* approach exploits the central limit theorem and the fact that the verification score is a sum of the frame scores, which in this case are assumed to be i.i.d. 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

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

The *Decorrelated* variance estimate attempts to compensate for the high level of correlation between consecutive acoustic feature vectors and, consequently, frame scores. This compensation is achieved through a transformation approach 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

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

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

Finally, for the EVD approach to be effective, it is particularly important to robustly estimate the variance of the frame scores with a very limited number of samples. This issue is also exacerbated by the correlated nature of these scores. In this work a more robust variance estimate is produced through Bayesian estimation and introducing *a priori* information. This *With Prior* estimate is 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 hyper-parameters 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 using either (3) or (5) above.

Table 2. Results at the Actual DCF operating point for the EVD method.

System	Act. DCF	Trial Length		Shortcut Errors	
		Median	Mean	Impostor	Target
Reference	.0293	103.4	103.4	–	–
Naïve					
90% Conf.	.1032	2	2.9	7.2%	22.1%
99% Conf.	.0600	3	5.4	2.9%	13.2%
99.9% Conf.	.0427	4	8.4	1.3%	7.7%
Decorrelated	$N = 10$				
90% Conf.	.0701	2	4.4	3.9%	15.6%
99% Conf.	.0369	5	11.3	0.7%	4.8%
99.9% Conf.	.0314	9	17.7	0.2%	1.4%
With Prior	$\tau = 100, \kappa^2 = 0.25$				
90% Conf.	.0583	3	5.4	2.7%	12.9%
99% Conf.	.0325	7	13.1	0.3%	3.0%
99.9% Conf.	.0302	11	19.8	0.1%	0.9%

3.2 Verification Score Normalisation

Typically, raw scores output by speaker verification systems are further processed to normalise for factors such as the quality of the trained speaker model, mismatch between the training and testing conditions and the linguistic content in the test utterance. Z-Norm [6] is an example of a score normalisation technique that normalises the verification score by the mean and variance of the speaker model's response to a set of impostor trials.

It is straight forward to apply Z-Norm to the applications described above as it can be characterised as a simple linear transform of the frame-based scores. If the Z-Norm statistics are given by μ_Z and σ_Z then the normalised ELLR score is given by,

$$\Lambda_Z(s) = \frac{\Lambda(s) - \mu_Z(s)}{\sigma_Z(s)} = a\Lambda(s) + b \tag{7}$$

where $a = 1/\sigma_Z(s)$ and $b = -\mu_Z(s)/\sigma_Z(s)$. As the ELLR score is a scaled sum of the frame scores, this transform can alternatively be applied directly to the individual frame scores,

$$\ell'_S(t) = a\ell_S(t) + b; \qquad \Lambda_Z(s) = \frac{1}{T}\sum_{t=1}^{T} \ell'_S(t). \tag{8}$$

It is then straightforward to apply any of the estimates in (3) through (6) above using the transformed frame scores.

4 Experimental Results

The results of using the EVD scoring approach are presented in Table 2 for the *Naïve*, *Decorrelated* and *With Prior* variance estimates with the threshold set to minimise the NIST Detection Cost Function (DCF) for the reference system. Performance is shown at three confidence levels, 90%, 99% and 99.9%, which are the minimum confidence with which the "true" verification score must be above or below the DCF threshold for the system to make an early verification decision. Also included are the results for the reference system, replicated from

Table 1. These results do not include Z-norm score normalisation at this stage. The performance of the systems in Table 2 are measured by the *actual* DCF value achieved at the specified threshold. Also included are measures of the average utterance length for each system, measured by both the mean and median[1] statistics.

It can be seen from the actual DCF results that the performance of the EVD approach drops behind that of the reference system in all cases—sometimes dramatically so—but this drop is both expected and actually quite small when the utterance lengths are also taken into consideration. This can be readily seen by comparing the results of Tables 1 and 2. For example, the *Decorrelated 99.9%* system shows a 7% relative drop in actual DCF but achieves this performance with less than a tenth of the utterance length for most trials (median length of 9 seconds). In contrast, using a similar but fixed utterance length of 10 secnods results in an actual DCF of .0601 (Table 1); this is more than a 100% increase in DCF.

Further analysing the utterance length statistics for the EVD systems, there is a consistent discrepancy between the mean and median statistics as the mean length are considerably longer in each case. This indicates a signifcantly skewed distribution of utterance lengths and that the *mean* test utterance lengths are dominated by a relatively small number of long trials. For the *Naïve* EVD systems, the majority of trials provide a result within 2, 3 or 4 seconds, with increasing confidence level, as indicated by the *median* trial lengths in Table 2.

This last point has an amazing implication: For the majority of trials a text-independent speaker verification system will produce the same decision with only *a few seconds* of speech that it will with almost 2 *minutes* of speech.

Fig. 2 is a DET plot of the *Naïve* EVD systems at differing confidence levels. 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.[2] For all systems the operating point at the specified threshold is highlighted with a circle.

Interestingly, the DET curves for these systems veer away from the reference system the farther they are from the DCF operating point. The performance curves of the early decision systems drop back toward the 2-second worst-case system in these areas. This effect is even more dramatic at the DCF operating point than for the EER, as explored in [4]. This characteristic is a direct consequence of the EVD 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.

By comparing the Tables 1 and 2 it can be seen that the EVD method is effective in trading performance at a specific operating point for shorter trials. It is also evident increasing the required confidence level provides an improved DCF

[1] The median utterance length for the EVD systems always falls on a whole-second increment as the EVD implementation used in these experiments only tests the stopping criteria at 1-second intervals.

[2] The EVD systems were restricted to a 2 sec minimum utterance length.

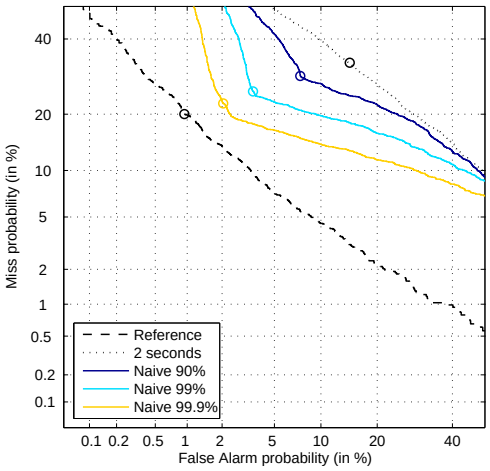

Fig. 2. DET plot using the naïve method at the minimum DCF threshold.

for each of the EVD methods, demonstrating that setting the confidence level is a viable method of controlling the trade-off between verification performance and utterance length.

The two rightmost columns of Table 2 quantify the errors introduced by the early decision criteria for impostor and target trials, respectively. These represent the trials for which the reference system and the EVD system have produced differing decisions. This is the approximate 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 that the *Naïve* variance estimates are not sufficiently accurate, particularly when based on a small number of frames.

It is also evident that, unlike in [4], the errors introduced by the EVD method are not evenly distributed between the target and impostor trials at the DCF operating point, with the target trial errors far outweighing the low rate of impostor trial errors. It is hypothesised that this discrepancy is due to the threshold lying much closer to the centre of the target trial score distribution (at approximately 20% miss rate) compared to near the tail of the impostor scores distribution (approvimately 1% false alarms) at this threshold. Hence it is simpler to dismiss a larger proportion of the impostor trials due to the increased distance of the score to the threshold.

This situation is improved significantly with the introduction of the *Decorrelated* variance estimation. With $N = 10$ and a typical frame rate of 100 frames per second, this method averages the frame scores over approximately 0.1 seconds 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, resulting

Table 3. Results with Z-Norm score normalisation at the Actual DCF operating point for the EVD method.

System	Act. DCF	Trial Length		Shortcut Errors	
		Median	Mean	Impostor	Target
Reference	.0249	103.4	103.4	–	–
Naïve					
90% Conf.	.1078	2	2.9	7.7%	20.8%
99% Conf.	.0585	3	5.5	3.0%	12.5%
99.9% Conf.	.0407	4	8.7	1.4%	7.3%
Decorrelated	$N = 10$				
90% Conf.	.0701	2	4.5	4.1%	14.7%
99% Conf.	.0331	5	11.7	0.7%	4.4%
99.9% Conf.	.0271	10	18.4	0.2%	1.6%
With Prior	$\tau = 100, \kappa^2 = 0.25$				
90% Conf.	.0565	3	5.5	2.8%	12.1%
99% Conf.	.0283	7	13.6	0.3%	2.5%
99.9% Conf.	.0257	12	20.5	0.1%	0.8%

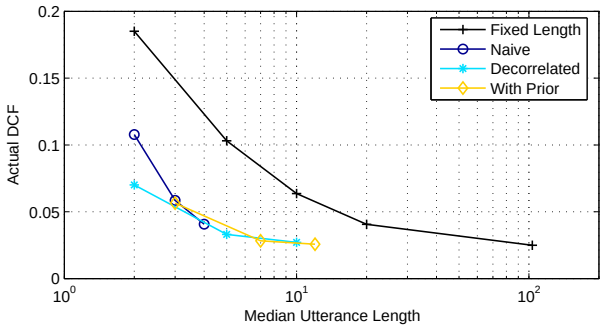

Fig. 3. Median utterance length versus Actual DCF for the fixed short utterance and EVD systems.

in performance closer to that of the reference system. The *Decorrelated* approach also produces errors at a rate much closer to the specified confidence level. While the rate at 99.9% confidence is still an order of magnitude too high for target trials, this result at least demonstrates that the variance estimated is more accurate with the data correlations diminished. There is 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 less than 10 seconds even at the 99.9% confidence level.

By incorporating a prior 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 (last three rows of Table 2). The hyperparameter τ was fixed at the equivalent of 1 sec while a value of $\kappa^2 = 0.25$ was determined empirically for this system.

Table 3 reproduces the results of Table 2 with the combination of the EVD method and Z-Norm score normalisation by applying the transform described in Section 3.2. These results demonstrate that the EVD method is just as effective with the application of Z-norm, showing much the same trends as described

above. As with the fixed length systems, it can be seen that the efficacy of Z-Norm with the EVD method is reduced with shorter utterance lengths. Notably, though, this effect is not as severe with the EVD approach as only the *Naïve 99.9%* system is degraded through the application of Z-Norm.

Fig. 3 graphically summarises the performance of the early verification decision approach by comparing the actual DCF to the median utterance length. Also presented are the fixed utterance-length systems as a reference. All systems have Z-Norm score normalisation applied. It is evident that the EVD 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 speaker verification scores 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. The Early Verification Decision (EVD) method, based on this confidence interval estimates, demonstrated that as little as 5–10 seconds of active speech on average was able to produce verification results approaching that of using an average of over 100 seconds of speech.

References

1. Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., Ortega-Garcia, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech & Language **20**(2-3) (2006) 331–355
2. Campbell, W.M., Brady, K.J., Campbell, J.P., Granville, R., Reynolds, D.A.: Understanding scores in forensic speaker recognition. In: Odyssey: The Speaker and Language Recognition Workshop. (2006)
3. Brümmer, N., du Preez, J.: Application-independent evaluation of speaker detection. Computer Speech & Language **20**(2-3) (2006) 230–275
4. Vogt, R., Sridharan, S., Mason, M.: Making confident speaker verification decisions with minimal speech. In: Interspeech. (2008) 1405–1408
5. Vogt, R., Sridharan, S.: Explicit modelling of session variability for speaker verification. Computer Speech & Language **22**(1) (2008) 17–38
6. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing **10**(1/2/3) (2000) 42–54
7. Martin, A., Miller, D., Przybocki, M., Campbell, J., Nakasone, H.: Conversational telephone speech corpus collection for the NIST speaker recognition evaluation 2004. In: International Conference on Language Resources and Evaluation. (2004) 587–590
8. Martin, A., Przybocki, M.: The NIST 1999 speaker recognition evaluation—an overview. Digital Signal Processing **10**(1-3) (2000) 1–18
9. Gauvain, J.L., Lee, C.H.: Bayesian adaptive learning and MAP estimation of HMM. In Lee, C.H., Soong, F., Paliwal, K., eds.: Automatic Speech and Speaker Recognition: Advanced Topics. Kluwer Academic, Boston, Mass (1996) 83–107