A Novel Technique for the Selection of Speech Segments for Speaker Verification

Mohaddeseh Nosratighods[1,2], Eliathamby Ambikairajah[1,2], Julien Epps[2,1] and Michael Carey[3]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
[3]School of Engineering,
The University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

m.nosratighods@student.unsw.edu.au, ambi@ee.unsw.edu.au, julien.epps@nicta.com.au, m.carey@bham.ac.uk

Abstract

This paper presents a novel selection technique to discard portions of speech that result in poor discrimination ability in speaker verification tasks. Theory supporting the significance of a sub-segmental selection procedure for test segments, prior to making decisions, is also developed. This approach has the ability to reduce the effect of the regions of the feature space that are not fully modeled by the speaker adaptation algorithm. The proposed technique utilises the frame-based score of the claimed and impostor speakers to select the most discriminative parts of the test segment. The frame selection technique, together with score normalisation, is evaluated on male and female speaker populations separately. Compared with a baseline system using both CMS and variance normalization, the proposed segment selection technique brings 6% , 18% relative reductions in error rate for female speakers, while for male speakers a more significant relative error rate reduction of 10%, 20% is achieved, in terms of EER and minimum DCF respectively.

1. Introduction

Speaker verification determines whether a given speaker is who they claim to be, based on a score comparing the likelihood of the observed speech given the claimant speaker against the likelihood of the same segment given the general population background model. A problem arises when the speaker's score varies widely in some frames, such that the speaker cannot be categorized as a true or impostor speaker. This variation of Log-Likelihood Ratios (LLR) across all frames is illustrated in Figure 1. The matching score for target (Figure 1(a)) and non-target (Figure 1(b)) speaker models are plotted as heavy lines whereas the dots are scores against 15 closest impostor speaker's models. Although the target and non-target speaker models usually give the highest and lowest scores respectively among all models, that is not always the case. Variation from this is due to the fact that all of the areas of acoustic features are not equally updated from the background model. As a result, the lack of available training data to accurately adapt the background model to the claimant speaker results in poor discrimination in some frames. In other words, the rate of change of the score distributions reveals that phonetic content of the unknown speech is updated according to the availability of training data for that particular phoneme.

An early study of the importance of selecting the more discriminative partitions of the feature space based on their frame-based LLRs is introduced by Li (Li & Porter, 1988). Li and Porter selected the reliable frames to set a speaker-independent threshold. This issue is addressed in a different manner (Pelecanos, Povy & Ramaswamy, 2006) by de-emphasising the contribution of unreliable mixture components and emphasizing discriminative regions. Pelecanos *et al*. also introduced a score mapping approach that utilises development data to determine the weighting for each partition according to its discriminative ability. To avoid the assumption of Gaussian distributions for the target and impostor scores, Pelecanos *et al.* modeled the scores as a function of training soft counts.

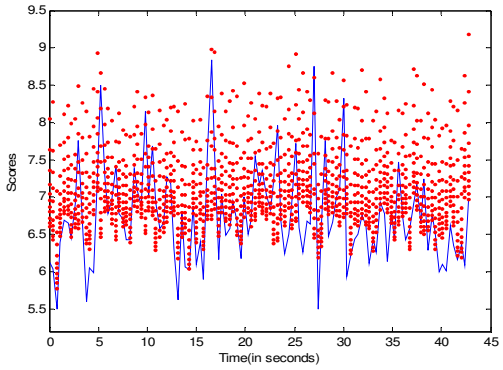
Succeeding the previous investigations, our proposal introduces a novel technique to select the most reliable and discriminative parts of speech without any assumption on the distribution of impostor and true scores. A framework is proposed, whereby if the sub-segments with low discrimination ability can be detected, and for each frame a log-likelihood ratio can be extracted, then the sub-segmental dropping can be successfully carried out. Statistical hypothesis testing is used to detect the non-discriminative sub-segments.

It has been shown empirically (Li & Porter, 1998) that the sub-segments with low target scores, the LLR of the observed speech given the target speaker, and the low variance impostor scores result in poor discrimination and the overall performance would be improved greatly if they were left out in making the final decision. This result is supported by the theory presented in this paper.

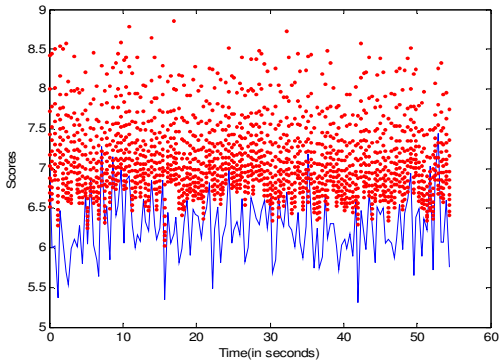
The technique proposed herein can be implemented by making minor changes to the decision-making section of the existing speaker verification system. Since it uses the same impostor scores employed in score normalisation, it does not impose additional overhead to the system (Auckenthaler, Carey & Lloyd-Thomas, 2000).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review



(a)



(b)

Figure 1: 300ms sub-segmental scores from speaker and impostor models for male (a) target (b) non-target test speech segments, from the NIST 2002 Dataset

2. Speech Segment Selection

2.1. Problem Formulation

Decision-making is the final processing stage of the speaker verification system, preceded by feature extraction and speaker modeling, as shown in Fig. 2. The decision-making process compares the LLR resulting from the claimed speaker model and the general population model (UBM) for a given test segment with a decision threshold.

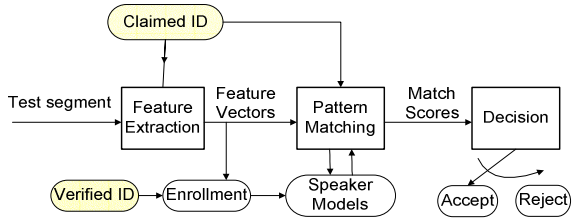


Figure 2: Schematic diagram of a speaker verification System (Campbell, 1997)

A problem arises when the matching score of true and impostor models varies across the frames. Fig. 1 shows this variability across the frames, for target and non-target true and impostor models for one test segment. It can be seen that setting a fixed threshold on raw scores or making an average of scores, does not guarantee a reliable decision, since the averaging of some low scores might cause false rejection, as in Fig. 1 (a). Poor representation of speakers can be mainly attributed to the score variability across all frames. MAP adaptation (Gauvain & Lee 1994), which has been widely used to model the characteristics of a specific speaker, was proposed as a solution for applications with sparse training data, such as speaker verification. However, the assumption that the background model is representative of the acoustic regions of the feature space that are not accurately updated, due to a lack of training data, is not always valid. Furthermore, the variability of the feature vector distribution from session to session makes some speech frames less reliable in making the final decision, due to channel, handset, and noise artifacts. Thus, dropping or removing frames with poor discrimination ability reduces the miss detection error and consequently improves the overall performance of speaker verification system. Frame-based processing of likelihood ratios with these considerations in mind motivated the following new score segmentation method.

2.2. Theoretical Basis for Speech Segment Selection

Since selecting specific segments of test data with poor discrimination ability is a key task, the problem is formulated as a hypothesis test.

2.2.1. Hypothesis Testing and Statistical Definitions

In hypothesis testing we are interested in testing between two mutually exclusive hypotheses, called the null hypothesis (denoted H_0) and the alternative hypothesis (denoted H_1). H_0 and H_1 are complementary hypotheses in the following sense: if the parameter being hypothesized about is θ, and the parameter space (i.e., possible values for θ) is Θ, then the null and alternative hypotheses form a partition of Θ:

$$H_0 : \theta \in \Theta_0 \subset \Theta$$
$$H_1 : \theta \in \Theta_0^c \subset \Theta \quad (1)$$

Θ_0^c is the set of all test statistic values for which H_0 will be rejected. This region is called the rejection region. A test statistic, similarly to an estimator, is just some real-valued function $T_n \equiv T(X_1,...,X_n)$ of the data sample $X_1,...,X_n$. Clearly, a test statistic is a random variable. A test is a function mapping values of test statistic into {0, 1}, where

- "0" implies accept the null hypothesis $H_0 \Leftrightarrow$ reject the alternative hypothesis H_1
- "1" implies reject the null hypothesis $H_0 \Leftrightarrow$ accept the alternative hypothesis H_1

In this paper, we refer to a test as a combination of both (i) a test statistic; and (ii) the mapping from realizations of the test statistic to {0, 1}. Normally, we start with the research hypothesis and "set up" the null hypothesis to be the opposite of what we hope to prove. We then try to show that, in the light of the collected data, the null hypothesis is false. We do this by calculating the probability of the data if the null hypothesis is true. A test with significance level α is one for which the probability of rejecting H_0 when it is actually true, is controlled at a specified level (Devore, 1995; Papulis, 1991).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

In parameter estimation, an interval of plausible values for the parameter being estimated is called a confidence interval (Wackerly, Mendelhall & Scheaffer, 1996). Usually, we use the term confidence interval (CI) to refer to a combination of an interval estimate, along with a measure of confidence (such as the confidence coefficient). Hence, a confidence interval is a statement like "θ is between 1.5 and 2.8 with probability 80%." This interval is found using pivotal quantity given a confidence coefficient. This quantity is a random variable which is a function of the parameter in question and the random variables $X_1,...,X_n$ but whose distribution is independent of that parameter. When we create CI's by inverting tests, the relevant pivotal quantity is the test statistic.

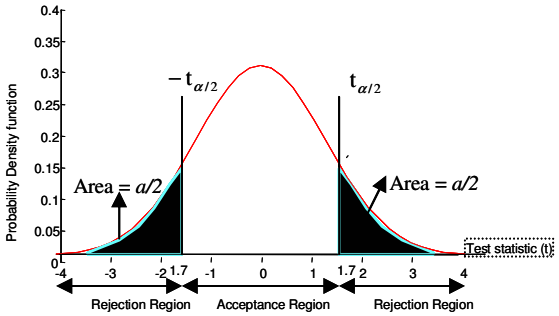


Figure 3: t-distribution and critical regions for null hypothesis

Table 1: Level of significance for a two-tailed test with degree of freedom (df) = 28

Level of significance for a two-tailed test (α)				
.05	.025	.01	.005	.0005
Corresponding test statistics value(t)				
t=1.7	t=2.05	t=2.47	t=2.76	t=3.67

2.2.2. Segment Selection based on hypothesis testing

The sub-segment selection discussed in this section compares the following two hypotheses: H_0: The null hypothesis is that the segment x_T does not contain discriminative information. H_1: The alternative hypothesis is that the segment x_T does contain discriminative information. We call segment x_T discriminative iff its likelihood, given true or impostor models is enough to classify it as a true or impostor speaker. The LLR of segment x_T, L_0, is defined as:

$$L_0 = E\{\log(p(x_T \mid \lambda_{True}))\} - E\{\log(p(x_T \mid \lambda_{\mathrm{Im}\,p}))\} = \mu_1 - \mu_2 \quad (2)$$

where $p(x_T \mid \lambda_{True})$, $p(x_T \mid \lambda_{\mathrm{Im}\,p})$ are the likelihoods of data segment x_T given the true and impostor models, λ_{True} and $\lambda_{\mathrm{Im}\,p}$, respectively.

In real problems, it is virtually always the case that the values of the population variances are unknown. For large sample sizes, the sample variance is used in place of population variance in the test procedure. The assumption of large sample size is made to use the properties of the central limit theorem (CLT) (Proakis & Manolakis, 1996). In fact the CLT allows us to use these test methods even if the two populations of interest are not normal (Devore, 1995).

In performing a large sample *t*-test, for the two populations $X_1,...,X_n$ and $Y_1,...,Y_m$ with corresponding sample means $\bar{x}$, $\bar{y}$ and true means μ_1, μ_2 and a common sample variance S_p, the null hypothesis, the test statistic, the alternative hypothesis and the rejection region for a specific significance level of test will be as follows:

Null hypothesis:

$$H_0 : \mu_1 - \mu_2 = L_0 \quad (3)$$

where in this case $L_0 = 0$, since this likelihood ratio results in poor discrimination when the two likelihoods are very close.

Test statistic value:

$$t = \frac{\bar{x} - \bar{y} - L_0}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \quad (4)$$

which has a *t* distribution with *m*+*n*-2 degrees of freedom, and where S_P^2 is the pooled estimator of the common variance σ^2 (Devore, 1995).

Alternative hypothesis:

$$H_a : \mu_1 - \mu_2 \neq L_0 \quad (5)$$

And rejection regions for level α test (Figure 3 and Table 1):

$$t \geq t_{\alpha/2,m+n-2} \text{ or } t \leq -t_{\alpha/2,m+n-2} \quad (6)$$

In cases where the variances of the two populations are not equal, the following procedure (called the Smith-Satterthwaite test) is known to be an approximately level α test but the probability of accepting H_0 when it is not true in this test has proved difficult to study.

Test statistic value:

$$t' = \frac{\bar{x} - \bar{y} - L_0}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \quad (7)$$

where S_1 and S_2 are the sample variances of the populations.

2.3. Algorithm Description

The aim of this algorithm is to detect and remove the non-discriminative or misleading sub-segments as has been discussed in section 2.2.2. If a sub-segment has the same probability of being uttered by the true or impostor speakers, it results in poor discrimination. But the above statement is valid when the likelihood of the test-segment given the impostor models does not change from one impostor to another, which means the variance of these likelihoods is small across different impostors. The smaller the variance, the more confident we are to discard a frame. How small it should be is the issue which has been addressed by introducing the critical regions to discard a sub-segment in section 2.2.2. The rejection region to discard a sub-segment is the interval wherein the sub-segment is significant

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

corresponding to a significant level α. Hence, only significant sub-segments are kept and the rest are discarded.

By substituting L_0 =0 in Equation 7, substituting values for $\bar{x}$ and $\bar{y}$ (the sample mean of target and impostor likelihood over a sub-segment of test data), s_1 and s_2 (the sample variance of target and impostor likelihood over the same sub-segment of test data), and *m* and *n* (the number of samples of target and impostor likelihoods over that same sub-segment of test data) allows the test statistic value to be evaluated.

According to equation (6) and Figure 3, if the test segment was in the rejection regions of level α, the null hypothesis is false with α confidence, i.e. the probability that the current sub-segment is discriminative equals α. The segment selection algorithm can be summarized as in Figure 4.

1. Divide the test segment into *M* consecutive sub-segments, select a set of frames in the first sub-segment: $x_i, i = 1...T$, where x_i is a frame of the current sub-segment and *T* is the number of frames in a sub-segment.
2. Compute the frame-based log likelihood of the test sub-segment given the true model, $\log(\ p(x_i \mid \lambda_{True}\ ))$, $i = 1...T$, where x_i is a frame of the test sub-segment and *T* is the number of frames in a sub-segment.
3. Compute the frame-based log-likelihood of the test sub-segment given the impostor models. $\log(p(x_i \mid \lambda_{\mathrm{Im}\, p_l}), i = 1...T, l = 1...N$, where *N* is the number of impostor models.
4. Compute the mean and variance of the impostor and true log-likelihoods over a sub-segment of *T* frames as below:

$$\bar{y} = \frac{1}{m}\sum_{l=1}^{N}\sum_{t=1}^{T}\log(\ p(x_t \mid \lambda_{\mathrm{Im}\, p_l}\ ))$$

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{T}\log(\ p(x_t \mid \lambda_{True}\ ))$$

$$s_1 = \frac{1}{m^2}\sum_{l=1}^{N}\sum_{t=1}^{T}(\log(\ p(x_t \mid \lambda_{\mathrm{Im}\, p_l}) - \bar{y})^2$$

$$s_2 = \frac{1}{n^2}\sum_{t=1}^{T}(\log(\ p(x_t \mid \lambda_{true}\ ) - \bar{x})^2$$

5. Compute the *t*-statistics value from equation (7)
6. If $-t_{\frac{\alpha}{2},m+n-2} < t < t_{\frac{\alpha}{2},m+n-2}$ discard the sub-segment
7. Is it the last sub-segment of the test segment?
   - yes : finish
   - no : continue from the step 1 for the next sub-segment

Figure 4: Segment selection algorithm

This work is different from Pelecanos *et al.'s* proposal in the following aspects:

- It discards the non-discriminative frames rather than modifying them.
- Not limiting the non-discriminative frames to the ones with 1 to 10 Gaussian counts of trainings data (Reynolds, Quatieri & Dunn, 2000). Instead, using null-hypothesis testing to detect all frames would result in poor discrimination.
- Working with sub-segments rather than frames allows use of the Central Limit Theorem (CLT) to estimate the mean and variance of target and impostor model rather than utilizing Spline functions to estimate the target and impostor statistics.

The same speech scores as shown in Figure 1 are plotted again after applying the segment selection technique with $\alpha = 10^{-4}$ for male speakers, in Figure 5. The improvement after applying this algorithm is apparent in the reduced low true speaker scores in the target trial (a) and the reduced relative spread of true speaker scores in the non-target trial (b).

3. System Setup

3.1. Database

Speaker recognition experiments were conducted on cellular telephone conversational speech from the switchboard corpus, the set defined by NIST for the 1-speaker cellular detection task in the 2002 Speaker Recognition Evaluations (SRE). The 2002 set contains 330 targets (139 males and 191 females) and 3570 trials (1442 males and 2128 females) with a majority of CDMA codec utterances; these are scored against roughly 10 gender-matched impostors and the true speaker. The 60 development speakers (2 minutes of speech for each of 38 males and 22 females), 174 target speakers (2 minutes of speech for each of 74 males and 100 females) from NIST-2001 were used to train the background model of NIST-2002 system.174 NIST-2001 target speakers were also used as the impostor data for the NIST-2002 evaluation system.

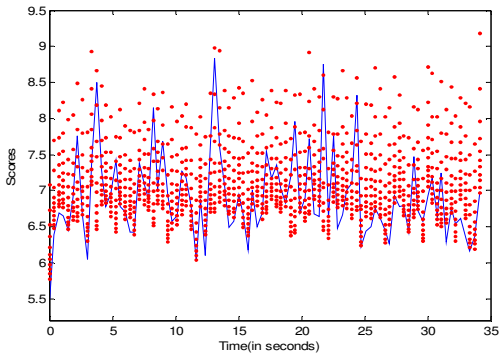
3.2. Baseline System

The feature set consisted of 15 Mel-PLP cepstrum coefficients (Gauvain, Lamel & Adda, 2002; Barras & Gauvain, 2003) 15 delta coefficients plus the delta-energy estimated on the 0-3.8kHz bandwidth. Cepstral mean subtraction and variance normalization were applied to each speech file during training and testing. The speech detector discarded the 15-20% of the lower energy speech frames before the extraction process.

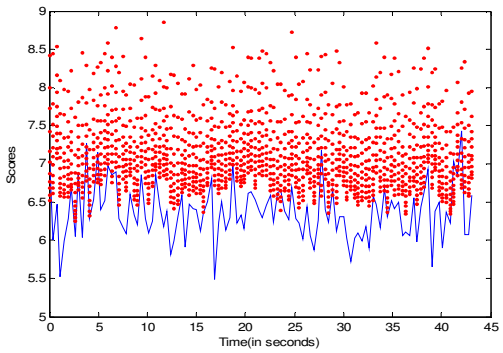
The speaker modeling is based on a GMM-UBM approach. The UBM consisted of two-gender dependent models with 512 Gaussians which were trained on 112 male and 122 female speakers from the training portion of development and evaluation datasets of NIST 2001, and about 6 hours of data in total. For each target speaker, a speaker-specific GMM with diagonal covariance matrices was trained using the speaker training data via maximum a posteriori (MAP) (Gauvain & Lee, 1994) adaptation of the Gaussian means with 5 iterations of the EM algorithm.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review



(a)



(b)

Figure 5: Sub-segmental scores from speaker and impostor models for the same (a) target (b) non-target test segment in Figure1, after applying segment selection technique.

3.3. Score Normalization and Segment Selection

The T-norm was calculated using impostor models from 62 male and 89 female speakers from the evaluation portion of the NIST2001 dataset, trained in a similar manner to the target models. The impostor models for segment selection algorithm are chosen from the 80% closest distant impostor speakers used in the T-norm.

4. Experimental Results

The experiments reported in this section examined the benefit of incorporating the proposed segment selection technique to discard frames with poor discrimination based on their target and impostor LLRs. The experiments investigated the performance improvement after applying this technique on male and female speakers separately in terms of Equal Error Rate (EER) and minimum Detection Cost Function (DCF) (Przybocki & Martin, 2004).

4.2 Gender-Specific Segment Selection

Tables 1 and 2 present the NIST2002 speaker recognition results for the proposed segmentation techniques on male speakers and female speakers respectively. The segment-selection technique was evaluated with different values of the significant level α ; but only the best result corresponding to the optimum value of $\alpha = 10^{-12}$ for male and $\alpha = 10^{-14}$ for female speakers are reported. The sub-segments should be small enough to track the score changes in different portions of a test segment but at the same time the assumption of large sample size to use the properties of the central limit theorem (section 2.2.2). These considerations led us to use 300ms (30 frame) non-overlapped sub-segments to implement the segment-selection technique.

The results shown in Table 2 reveal that this method, together with score normalization, provides a significant improvement in terms of minimum DCF and EER. The segment selection technique improves the minimum DCF and EER at least 3% and 4% over the T-norm and 18% and 6% over the baseline system for female speakers respectively. This improvement was more significant for male speakers (Table 3) as it brings at least 6% and 7% improvement over the T-norm and 20% and 10% over the baseline system.

Table 2: Segment Selection results on Male Speakers

System	EER	DCF (×1000)
Baseline	10.99	44.7
T-Norm Baseline	10.58	38.4
Segment selection plus T-Norm	**9.83**	**35.9**

Table 3: Segment Selection results on Female Speakers

System	EER	DCF (×1000)
Baseline	11.82	52.7
T-Norm Baseline	11.42	45
Segment selection plus T-Norm	**11.1**	**43**

Figures 6 and 7 plot the Detection Error Tradeoff (DET) curve for the baseline, baseline plus T-Norm and optimum segment selection plus T-Norm on male and female speakers respectively. It can be clearly seen in both Figures that segment selection technique with T-Norm performs better than T-Norm alone in EER operating point and in the area of minimum DCF. So, the experiments support the theory (section 2.2.2) that discarding the non-discriminative frames reduces the miss detection rate. All three systems perform the same in low miss-rate areas on male speakers whereas these systems behave quite differently in this area on female speakers; both T-Norm and segment selection are better than baseline for low miss-rate areas.

Generally, this technique was more successful on male speakers rather than female population. The reason might be attributed to the fact that the duration of test segments varies a lot for female compared with male speakers. So, using a fixed significance level for all durations is not as effective for female as for male speakers. Furthermore, in short test segments, the number of frames are limited, and discarding a discriminative frame wrongly could affect the results more than longer duration tests which are the majority of the male speaker utterances.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

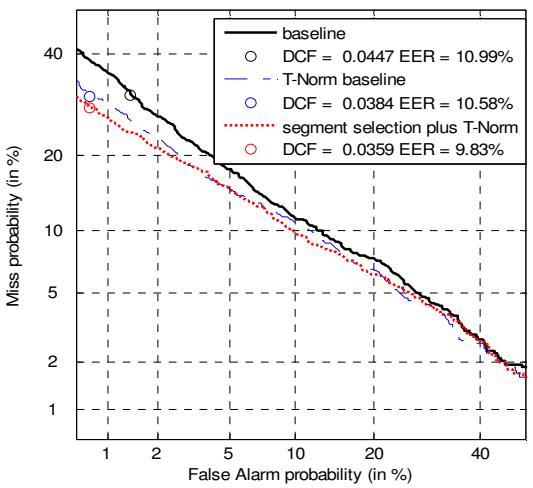


Figure 6: DET plot for the baseline and segment selection systems with and without T-Norm, for male speakers from the NIST 2002 dataset.

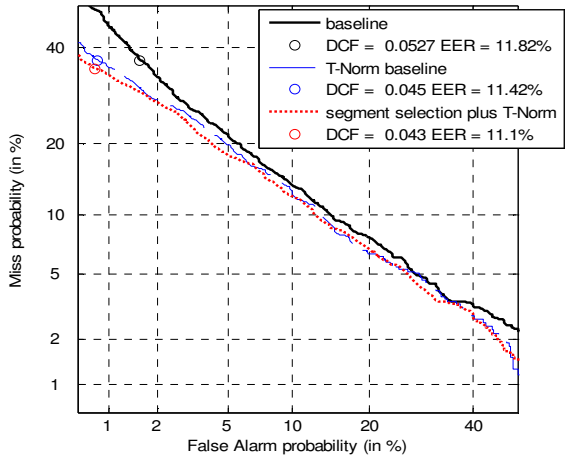


Figure 7: DET plot for the baseline and segment selection systems with and without T-Norm, for female speakers from the NIST 2002 dataset

As a result, choosing an adaptive significant level α corresponding to the test segment durations especially for female speakers can improve the performance of the proposed method, greatly. Also, having a higher frame rate in short test segments can avoid dropping the discriminative sub-segments.

5. Conclusion

This paper has investigated the importance of selecting specific portions of a test segment to enhance the efficacy of the decision-making stage in speaker verification systems. A segment selection algorithm has been proposed to discard the non-discriminative parts of the test utterance based on their target and impostor log-likelihood ratios. This frame selection technique can be gender-specific in training and testing. The results indicate a consistent equal error rate reduction compared with the baseline across all experiments conducted. A relative reduction in error rate of 19% and 8% averaged over all test speakers in terms of min DCF and EER was obtained using the proposed segment selection technique.

Future work may examine the optimum α values for test segments specific to their duration, handset, and modulation types.

6. References

Auckenthaler R., Carey M. & Lloyd-Thomas H. (2000). Score normalization for text independent speaker verification system. *Digital Signal Processing*, Vol.10, pp.42-54.

Barras, C. & Gauvain J.L., (2003). Feature and score normalization for speaker verification of cellular data. in Proc. ICASSP, Vol2. pp.49-52.

Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, vol. 85, pp. 1437-1462.

Devore, J. L. (1995), Probability and statistics for engineering and the sciences. Duxbury Press, Clifornia.

Guvain, J.L., Lamel, L. & Adda, G., (2002). The LIMSI broadcast news transcription system. *Speech Communication*, vol.37, no.1-2, pp.89-108.

Gauvain J.L. & Lee, C.H. (1994). Maximum a postriori estimation for multivariate Gaussian mixture observations of markov chain. *IEE Trans. On Speech and Audio Processing*, Vol.2. no.2, pp-291-298.

Li, K.-P. Porter, J.E. (1988). Normalization and selection of speech segments for speaker recognition scoring. *In Proc. ICASSP*, Vol.1, pp.595-598.

Papulis A. (1991). Probability, Random Variables and Stochastic Processes, McGraw-Hill, New York, pp.265-272.

Pelecanos, J., Povey, D. & Ramaswamy, G. (2006). Secondary Classification for GMM Based Speaker Recognition. *In Proc. ICASSP*, Vol.I, pp.109-112.

Proakis, J. G., Manolakis, D.G (1996). *Digital signal processing principles, algorithms and applications*, Third edition, Prentice Hall.

Przybocki, M., Martin, A. (2004). NIST Speaker Recognition Evaluation Chronicles*., in Proc. Odyssey*, *the Speaker and Language Recognition Workshop.*

Reynolds, D., Quatieri, T. & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*. Vol.10, no.1/2/3, pp.133-136.

Therrien, C.W. (1992). *Discrete random signals and statistical signal processing*, Prentice Hall.

Wackerly, D., Mendelhall, W. & Schaeffer, R. (1996). *Mathematical statistics with application,* Wadsworth Publishing company.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review