P-VALUE SEGMENT SELECTION TECHNIQUE FOR SPEAKER VERIFICATION

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

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]UNSW Asia, 1 Kay Siang Road, Singapore 248922
[3]School of Engineering, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
m.nosratighods@student.unsw.edu.au, ambi@ee.unsw.edu.au, j.epps@unsw.edu.au, m.carey@bham.ac.uk

ABSTRACT

This paper presents a segment selection technique for discarding portions of speech that result in poor discrimination ability in speaker verification tasks. Theory supporting the significance of a frame selection procedure for test segments, prior to making decisions, is also developed. This approach has the ability to reduce the effect of the acoustic regions of speech that are not accurately represented due to a lack of training data. Compared with a baseline system using both CMS and variance normalization, the proposed segment selection technique brings 24% relative reduction in error rate over the entire testing data of the 2002 NIST Dataset in terms of minimum DCF. For short test segments, i.e. less than 15 seconds, the novel frame dropping technique produces a significant relative error rate reduction of 23% in terms of minimum DCF.

***Index Terms*—** Speaker Verification, Segment Selection, Null Hypothesis

1. INTRODUCTION

Speaker verification tasks using Gaussian Mixture Models (GMMs) rely 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 [1]. However, the score varies significantly across the frames, depending on whether it can be attributed to the speech content or environmental artifacts such as channel and handset effects, resulting in poor discrimination between true and impostor models. This variation of Log-Likelihood Ratios (LLR) across all frames is illustrated in Figure 1, where the matching scores for a target speaker model are plotted as heavy lines and the scores against the five closest impostor speaker's models are shown as dots. Although the target speaker models usually give the highest scores among all models, this is not always the case. The reason for this is partly because not all of the areas of acoustic features are equally adapted from the background model. The rate of change of the score distributions reveals that the relevant mixture distributions are updated according to the availability of training data for that particular speech sound. Furthermore, the channel, handset and environmental mismatch between training and testing conditions also results in variability of the scores across frames. Several techniques such as feature mapping [2], speaker model synthesis [3] and CDF matching [4] accommodate this mismatch by increasing the feature robustness to environmental artifacts, whereas some other methods, such as discarding the unvoiced segments of speech [5], rely on the noise robust sections of the speech at the input.

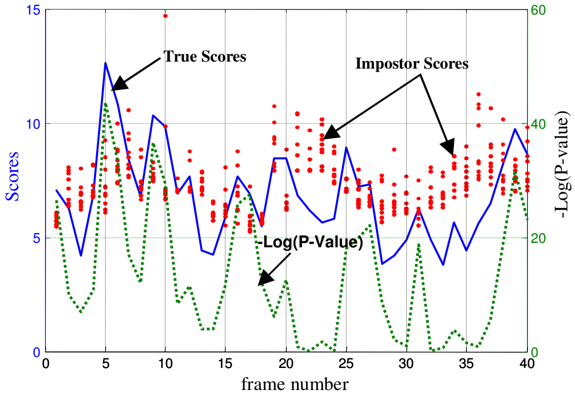

Figure 1. Frame-based scores from speaker and impostor models and its corresponding frame-based *P-value* for first forty frames of a male target test speech segment from the NIST 2002 Dataset

An early study [6] addressed the score variability, selecting reliable frames by setting a speaker-independent threshold. This issue can also be addressed in a different manner [7] by de-emphasizing the contribution of unreliable mixture components and emphasizing discriminative regions.

Following our previous investigations we introduce a technique which selects the most reliable and discriminative parts of speech without any *a priori* assumptions about the distributions of impostor and true scores. If frames with low discriminative ability can be detected, and a log-likelihood ratio can be extracted for each frame, then the frame can be discarded. Statistical hypothesis testing is used to detect the non-discriminative frames.

It has been shown empirically [6] that for frames 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 in section 2 can be implemented by making minor changes to the decision-making section of the existing speaker verification systems. Since it uses the same impostor scores employed in score normalization, it does not impose additional overhead on the system.

Section 2 presents the algorithm for detecting the non-discriminative frames based on Null hypothesis theory. Then we

1-4244-0728-1/07/$20.00 ©2007 IEEE

describe the system setup (Section 3), and experiments supporting the theory are reported in Section 4.

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. 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.

A problem arises when the matching score of true and impostor models varies across the frames. Figure 1 shows this variability across the frames, for true and impostor models for one test segment. It can be seen that setting a fixed threshold on raw scores or taking an average of scores does not guarantee a reliable decision, since the averaging of some low scores might cause false rejection. Poor representation of speakers can be mainly attributed to the score variability across all frames. MAP adaptation [8], 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 [9]. 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 the final decision, due to channel, handset, and noise artifacts. Thus, removing frames with poor discrimination ability after score matching 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 score segmentation method.

2.2. Theory Development and algorithm Description

The frame selection technique discussed in this section compares the following two hypotheses: H_0: The null hypothesis is that the frame x_T does not contain discriminative information. H_1: The alternative hypothesis is that the frame x_T does contain discriminative information. We call frame x_T indiscriminative iff its likelihood, given true or impostor models is insufficient to classify it as a true or impostor speaker; the likelihood of that frame belongs to impostor models is equal to the likelihood of its belonging to the target model. Therefore, the null hypothesis and the alternative hypothesis are defined respectively as follows:

$$H_0 : E\{\log(p(x_T \mid \lambda_{\mathrm{Im}\,p}))\} = \log(p(x_T \mid \lambda_{True}))$$
or (1)

$$H_0 : \eta = \eta_0$$
$$H_1 : \eta \neq \eta_0 \qquad (2)$$

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

A *P-value* is a measure of how much evidence we have against the null hypothesis. The smaller the *P-value*, the more evidence we have against H_0. It is also a measure of how likely we are to get a certain sample result or a result "more extreme," assuming H_0 is true. In other words, the *P-value* measures consistency by calculating the probability of observing the results from a sample of data or a sample with results more extreme, assuming the null hypothesis is true, as seen in Figure 2**.** The smaller the *P-value* is, the greater is the inconsistency [10].

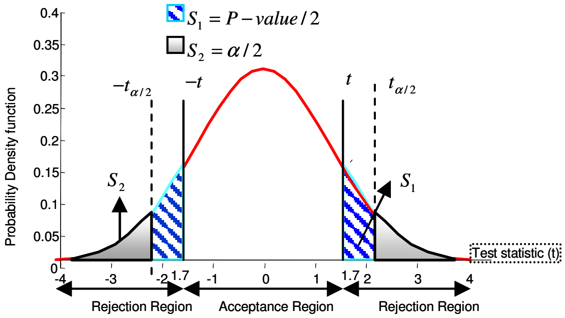

Figure 2. t-distribution and *P-value* for null hypothesis.

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 [10]. 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 a large sample size is made to allow the use of the properties of the central limit theorem (CLT). In fact the CLT allows us to use these test methods even if the population of interest is not normal [10].

In performing a large sample *t*-test, for the population $X_1,...,X_n$ with corresponding sample means $\bar{x}$, true means η and sample variance S , the test statistic and the rejection region for a specific significance level of test are as follows:
Test statistic value:

$$t = \frac{\bar{x} - \eta}{S\sqrt{\frac{1}{n}}} \qquad (3)$$

which has a t distribution with n-1 degrees of freedom, and where S^2 is the pooled estimator of the common variance σ^2 [10]. The rejection regions for level α test (Figure 2) are:

$$t \geq t_{\alpha/2,n-1} \text{ or } t \leq -t_{\alpha/2,n-1} \qquad (4)$$
or
$$P - value < \alpha \qquad (5)$$

Further, important to recognize is the fact that

$$\bar{x} \rightarrow N\left(\eta, \sqrt{\frac{1}{n}}\sigma\right) \qquad (6)$$

according to CLT [10].

Substituting values in equation (3) for $\bar{x}$ (the sample mean of impostor likelihoods over n impostor models given the frame x_T), η (the target likelihood given frame x_T), and S (the sample variance of impostor likelihoods over n impostor models given frame x_T) allows the test statistic value to be evaluated.

According to equation (4) and Figure 2, 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 frames is discriminative equals α. The frame selection algorithm can be summarized as follows:

- Select the impostor models from development dataset
- Calculate the frame-based LLR for the claimant and impostor models
- Calculate the test statistics value, *t*, for each frame from equation 3.
- Discard the frames whose corresponding *t*-values are not in rejection regions (equation 4) or whose *P-values* are more than significant level α (equation 5)
- Calculate the new LLR by averaging the remaining frame scores

P-values calculated using this technique for a male test segment have been shown with dashed lines in Figure 1. It can be seen that the frames with smaller true scores and smaller impostor score variances correspond to higher *P-values* or smaller negative logarithm of *P-values*. On the contrary, the smaller *P-values*, larger negative logarithm of *P-values*, correspond to the frames with high true scores. Therefore, the smaller the *P-value*, the more discriminative the frame is.

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. The same target speakers were also used as the impostor data for the NIST-2002 evaluation system. 2038 evaluation test segments (850 males and1188 females) of NIST-2001 were used to find the optimum value for significant level α.

3.2. Baseline System

The feature set consisted of 15 Mel-PLP cepstrum coefficients [11], 15 delta coefficients plus the delta-energy estimated on the 0-3.8 kHz 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 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, 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 GMM with diagonal covariance matrices was trained using the speaker training data via maximum a posteriori (MAP) [8] adaptation of the Gaussian means with 3 iterations of the EM algorithm.

Table1. Segment Selection Technique Results

System	Duration(in seconds)		
EER and min DCF x 1000	0-15+	46-65+	All segments
Baseline	18.05	10.81	11.47
	78	43.9	49.2
Baseline + T-Norm	19.38	11.3	11.46
	66.7	34.4	41.1
Segment selection + T-Norm	**18.44**	**9.2**	**10.76**
	59.6	**33.3**	**37.2**

4. EXPERIMENTAL RESULTS

The experiments reported in this section examined the benefit of the proposed segment selection technique to discard frames with poor discrimination based on their target and impostor LLRs. These investigated the performance improvement after applying the technique on the entire NIST 2002 testing dataset in terms of Equal Error Rate (EER) and minimum Detection Cost Function (DCF) [13]. The effects of the proposed technique on two extreme categories of test segment duration (i.e. long and short) were also examined.

T-Norm [12] 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 exactly the same impostor speakers used in the T-Norm.

Table 1 presents the Equal Error Rate (EER) and minimum Detection Cost Function (DCF) results for segment selection-based system with T-Norm and the baseline with and without T-Norm for the two categories of test segment durations. The results show that while the baseline system fails in verifying short test segments, the frame selection algorithm improves the miss detection rate significantly. 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 $\alpha = 10^{-5}$ for all speakers, is reported here. Also, maximum frame dropping rate was 20% of frames in each test segment.

The segment selection technique improves the minimum DCF and EER at least 10% and 6% over the T-norm and 24% and 6% over the baseline system for all test segments respectively. This improvement was still significant for test segments less than 15 seconds (Table 1) bringing at least 10% improvement over the T-norm and 23% over the baseline system in terms of minimum DCF. On the other hand, for test segments more than 45 seconds the improvement is more significant in EER operating point while they are benefited 18.58% and 14.89% relative error reduction compared to T-Norm and Baseline.

Figures 3 and 4 plot the Detection Error Tradeoff (DET) curves for the baseline, baseline plus T-Norm and optimum segment selection technique plus T-Norm for the entire test segments and then on test segments less than 15 seconds respectively. It can be clearly seen (Figure 3) that the new segment selection technique with T-Norm performs better than T-Norm alone in terms of EER operating point and particularly in the area of minimum DCF over entire test segments, whereas all three systems exhibit similar performance in low miss-rate areas. For test segments less than 15 seconds, the improvement is significant in minimum DCF and low miss-rate areas while the EER operating point for all systems is almost the same. Therefore, the experiments support the theory (section 2.2) that discarding the non-discriminative frames reduces the miss detection rate.

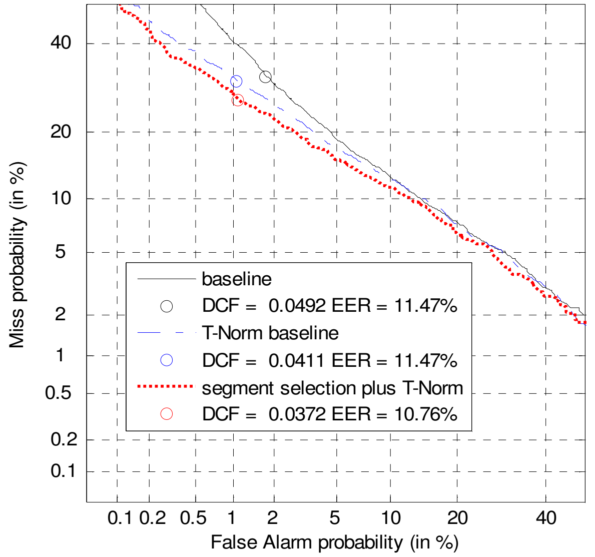

Figure 3. DET plot for the baseline and segment selection systems with and without T-Norm, for the entire NIST 2002 dataset.

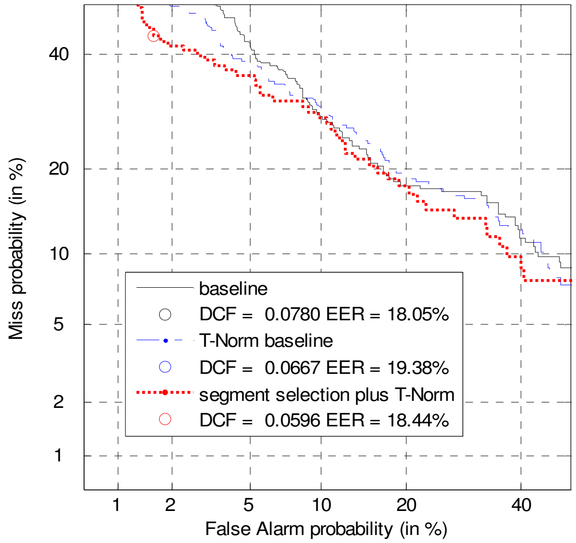

Figure 4. DET plot for the baseline and segment selection systems with and without T-Norm, for test segments with duration less than 15 seconds

5. CONCLUSION

This paper has reported 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 likelihood ratios. The results indicate a consistent equal error rate and minimum DCF reduction compared with the baseline across all experiments conducted. A relative reduction in error rate averaged all test segments, of 24% and 6% in terms of min DCF and EER respectively was obtained using the proposed segment selection technique.

ACKNOWLEDGEMENT

Funding for this research was fully provided by the National ICT Australia (NICTA).

6. REFERENCES

[1] M. J. Carey, E. S. Parris and J. S. Bridle, “A Speaker Verification System Using Alpha-Nets,” in *Proc. ICASSP*, pp.397-400, Toronto, 1991.

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

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

[4] J. Pelecanos, and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. Odyssey*, pp. 213-218, 2001.

[5] N. Fakotakis, and G. Kokkinakis “Speaker Verification System based on probabilistic neural networks,” *NIST Speaker Evaluation Workshop*, USA , 2002

[6] K.-P Li, and J. E. Porter, “Normalization and selection of speech segments for speaker recognition scoring,” in *Proc. ICASSP*, Vol. 1, pp. 595-598, 1988.

[7] J. Pelecanos, D. Povey, and G. Ramaswamy, “Secondary Classification for GMM Based Speaker Recognition,” in *Proc. ICASSP*, Vol. I, pp. 109-112, France, 2006.

[8] J. L. Gavain, and C. Barras, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chain ,” *IEE Trans. On Speech and Audio Processing*, Vol. 2, no. 2, pp. 291-298, 1994.

[9] M. J. Carey, E. S. Parris, S. J. Bennett and H. Lloyd-Thomas, “A Comparison Of Model Estimation Techniques For Speaker Verification,” in Proc. *ICASSP,* pp.1083-1086, Munich, 1997.

[10] J. L. Devore, *Probability and statistics for engineering and the sciences*, Brooks/Cole Publishing Company, 1995.

[11] J. L. Gavain, L. Lamel, and G. Adda, “The LIMSI broadcast news transcription system,” *Speech Communication*, vol. 37, no. 1-2, pp. 89-108, 2002.

[12] R. Auckenthaler, M. J. Carey, H. Lloyd-Thomas, “Score normalization for text independent speaker verification system,” *Digital Signal Processing*, Vol.10, pp.42-54, 2000.

[13] M. Przybocki, A. Martin, “NIST Speaker Recognition Evaluation Chronicles,” *Proc. Odyssey*, *the Speaker and Language Recognition Workshop,* 2004.