CANDIDATE SELECTION BASED ON SIGNIFICANCE TESTING AND ITS USE IN NORMALISATION AND SCORING

Ji-Hwan Kim[1], Gil-Jin Jang, Seong-Jin Yun and Yung-Hwan Oh

Korea Advanced Institute of Science and Technology,
Department of Computer Science,
Kusong, Yusong, Taejon, Korea(ROK) (305-701)
{jhkim, jangbal, sjyun, yhoh}@bulsai.kaist.ac.kr

ABSTRACT

Log likelihood ratio normalisation and scoring methods have been studied by many researchers and have improved the performance of speaker identification systems. However, these studies have disadvantages: the recognised distorted speech segments are different for each speaker. Also the background model in log likelihood ratio normalisation is changed in each speech segment even for the same speaker. This paper presents two techniques. Firstly, candidate selection based on significance testing, which designs the background speaker model more accurately. And secondly, the scoring method, which recognises the same distorted speech segments for every speaker. We perform a number of experiments with the SPIDRE database.

1. Introduction

In text independent speaker identification in a real environment, mismatches occur between the speaker model and input speech because of difficulties in collecting sufficient amounts of data, the differences between training and testing environments and the distortion from several kinds of noise [2, 6]. In order to reduce these mismatches, scoring methods, which eliminate segments with lower log likelihood, have been studied [3].

For an input utterance $X = x_1, x_2, ..., x_T$, the probability of a speaker model λ given the input utterance is:

$$P(\lambda|X) = \frac{p(X|\lambda)p(\lambda)}{p(X)} \tag{1}$$

Assuming each speech segment is independent of the other, the log likelihood for the input speech is the sum of each segment's log likelihood.

$$logL(\lambda|X) = \sum_{t=1}^{T} logL(\lambda|x_t) \tag{2}$$

In equation (1), $p(X)$ is a static factor within a given utterance, but will vary from utterance to utterance. As a result, the likelihoods returned by the speaker model for each λ are not absolute measures, but relative measures with respect to $p(X)$, and therefore not directly comparable. So, methods, which make the comparison of likelihoods between segments meaningful are required. The general approach is to apply likelihood normalisation to input utterance using the speaker model λ and background speaker model λ_B [5, 1]. For simplicity, we define the normalised log likelihood ratio(LLR) as follows:

$$LLR(X) = log\frac{p(X|\lambda)}{p(X|\lambda_B)} \tag{3}$$

Studies have been performed in speaker identification and verification on the scoring method and the design of background speaker models and have shown good performance [3, 1]. However they have disadvantages: the recognised distorted speech segments are different for each speaker since each speaker has different selected segments. Also, the background speaker model is changed in each speech segment. As the number of enrolled speaker becomes large, a lot of overlap between speaker subspaces occurs and speaker model correctness decreases. Therefore, the scoring method which recognises the same distorted speech segments for every speaker and the candidate selection method, which designs the background speaker model more accurately, are needed.

The rest of the paper is arranged as follows. We first give details of the candidate selection method based on significance testing. We then describe our work on likelihood ratio normalisation and the scoring method with selected candidates, which recognises the same distorted speech segments for every speaker. This is followed by a number of experiments that show the effect of the method.

2. Candidate Selection Using a Confidence Measure

The underlying assumption for correct identification is that the difference in log likelihood between the best and the second best speaker models during correct identification is generally larger than the difference during incorrect identification.

Figure 1 shows the histograms of log likelihood ratio be-

[1]Ji-Hwan Kim is currently at Cambridge University Engineering Department. E-mail: jhk23@eng.cam.ac.uk

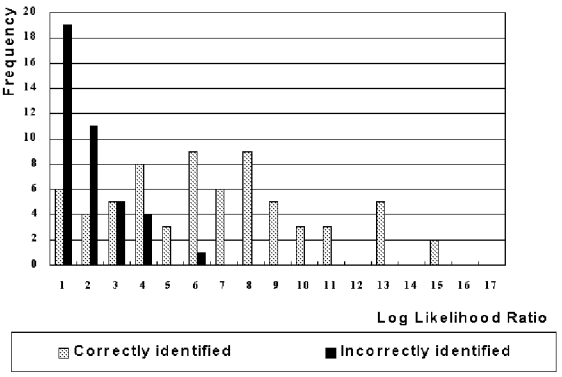

Figure 1: Histograms of log likelihood ratio

tween the best and the second best speaker models. The distribution differs greatly depending on whether or not the speaker is correctly identified. The distribution of the log likelihood ratio for incorrectly identified results has an exponential probability-like distribution. However, it is difficult to say whether or not the distribution of correctly identified has a specific distribution. Measuring confidence based on the significance testing was presented [3]. Let C_F, r, and $f_F(r) = f(r|C_F)$ denote the class of incorrect identifications, the log likelihood ratio between the best and second best speaker models and the distribution of log likelihood ratio for incorrectly identified results respectively. The significance confidence measure, denoted by $Conf(r)$, is defined as follows:

$$Conf(r) = 1 - \int_r^{\infty} f_F(x)d(x) \tag{4}$$

The higher the confidence measure, the more we believe that the log likelihood ratio is too high to have been generated by a misclassification.

Figure 2 illustrates the estimation of a confidence measure. The log likelihood ratio's distribution of incorrectly

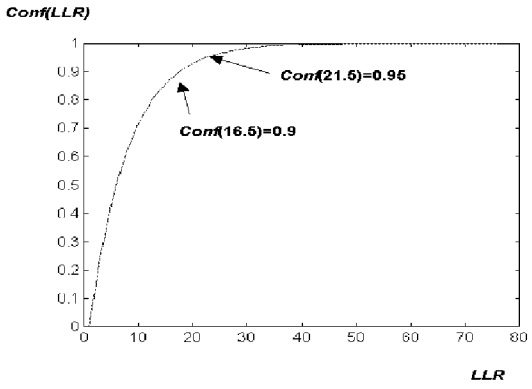

Figure 2: Estimated confidence measure

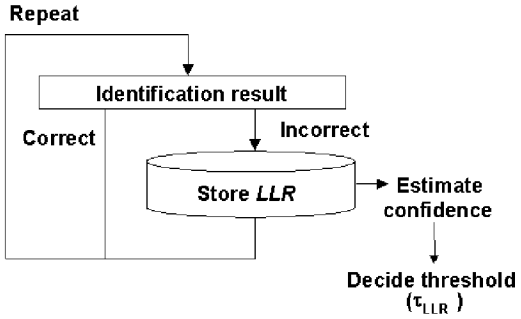

Figure 3: Procedures in estimating confidence

identified results may be estimated by an exponential distribution [4]. From the confidence measure, if a log likelihood ratio, r, is acquired, we believe the identification result with a $Conf(r)$ confidence. Based on the definition of the confidence measure in the significance method, as r increases, so too does the degree of confidence, $Conf(r)$. Using this confidence measure, the log likelihood ratio which gives 0.95 confidence is estimated at 21.5.

The procedures for confidence estimation in this research are described as follows: First, speaker models are trained using training data. Then, the distribution of log likelihood ratios for the development data, which is not included in training data, is found. Using this distribution, the degree of confidence function is estimated. Finally, a threshold of log likelihood ratio is selected which satisfies the desired degree of confidence. Figure 3 shows these procedures.

Figure 4 explains how to select candidates. In the testing environment, the log likelihood for each speaker model is calculated. Then the log likelihoods are sorted in decreasing order and a rank to each speaker is given. Then, the method steps through the ranked speakers one-by-one, calculating the confidence level from the current ranked

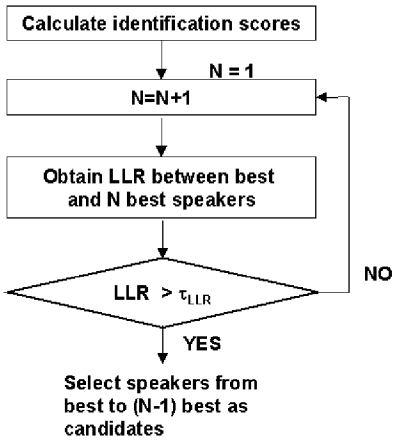

Figure 4: Procedures in candidate selection

speaker to the best ranked speaker. It finds the lowest rank N which exceeds the required confidence level. Then, it selects candidates from the best speaker to the $(N-1)$th best speaker. As a result, the system selects candidate speakers whose log likelihood ratio with the best speaker is smaller than the pre-defined threshold.

3. Scoring Method with Candidates

The underlying assumption for the scoring method is that in distorted segments, candidates have generally worse ranks than ranks in undistorted segments.

In the conventional method, which is presented in [3], the normalised log likelihood ratio for the speaker model λ_n given the speech segment x_t is evaluated as follows:

$$LLR(\lambda_n|x_t) = log\frac{p(x_t|\lambda_n)}{max\ p(x_t|\lambda_m)} \quad (5)$$

The denominator is the maximum probability over all models not belonging to speaker n. Therefore, the background speaker model is different in each speech segment even when originating from the same speaker. In the proposed method, the background model for the candidate c_n is designed taking into account all candidates. Hence, the normalised log likelihood ratio becomes:

$$LLR(\lambda_{c_n}|x_t) = log\frac{p(x_t|\lambda_{c_n})}{p(x_t|\lambda_B)} = log\frac{p(x_t|\lambda_{c_n})}{\sum p(x_t|\lambda_{c_m})} \quad (6)$$

The denominator is the sum of the candidates' likelihood.

In the proposed scoring method, the average rank of the candidates for each frame is calculated. The method then sorts these average ranks in increasing order and selects the best T frames. After the selection of frames, only the normalised log likelihood ratios of the selected frames are credited to each candidate. Then, the speaker identification system identifies the speaker with the highest normalised score among the different candidates. Table 1 summarises the differences between the conventional and the proposed scoring method.

Conventional	Proposed
Background model	
Best speaker m except speaker n	All candidates
Selection method	
Best LLR	Average rank of candidates
Speech distortion assumption	
As the normalised score decreases	As the average rank becomes poor (value increases)
Extracted frames	
Differ among the speakers	The same for all speakers

Table 1: Comparison between the conventional and the proposed scoring methods

Database	SPIDRE
No. of Speakers	45(23 males and 22 females)
Included Noise	Environmental noise, cross talk transmission noise, etc
Speaker Model	HMVQM
Feature parameter	Bark cepstrum delta Bark cepstrum Energy

Table 2: Experimental setup

Training Data	Exp.1: 45 speakers × 30 sec Exp.2: 45 speakers × 4 min Exp.3: 15 speakers × 30 sec
Development Data	Exp.1: 45 speakers × 30 sec Exp.2: 45 speakers × 4 min Exp.3: 15 speakers × 30 sec
Testing Data	Exp.1: 45 speakers × 30 sec Exp.2: 45 speakers × 4 min Exp.3: 15 speakers × 30 sec

Table 3: Summary of experiments

4. Experiments and Results

The experiments used the SPIDRE database, which is widely used in speaker recognition. Speaker models were implemented using Hidden Markov VQ-codebook Models(HMVQM); these have shown good results in small amounts of training data in speaker identification [6]. All input speech was divided into 20ms frames with 10ms overlap. The parametrisation used was 14 dimensional bark cepstrum, 14 dimensional delta bark cepstrum, energy and delta energy were used as input parameters. Models were built with 1,2 and 4 emitting states. Table 2 summarises the experimental setup.

We performed three different experiments, varying the number of speakers and the amount of training data. Experiment 1 is performed with 45 speakers and 30 seconds of training data for each speaker. Experiment 2 was performed with 45 speakers and 4 minutes of training data per speaker. Experiment 3 was performed with 15 speakers and 30 seconds of training data per speaker. Table 3 summarises these 3 experiments.

Table 4 shows the identification rate of the baseline system(without using normalisation and scoring methods). The identification rate shown is that for data which surpasses a 0.95 threshold for the degree of confidence. In all experiments, the identification rate for data exceeding the 0.95 confidence level is much better than that for the conventional method. Therefore, we can conclude that accepting or rejecting the identification result using the proposed confidence measure is effective. Table 4 also shows the percentage of data which exceeds the 0.95 confidence level. In experiment 1 with 1 state, the percentage of the data exceeding the 0.95 confidence level is 31.5

Table 5 shows the probability that the input speaker ex-

	Identification rate		Percentage
	Baseline	Over 0.95	over 0.95 thr.
Experiment 1			
1 state	63.0	97.7	31.5
2 state	59.3	97.1	26.9
4 state	55.6	89.7	26.9
Experiment 2			
1 state	77.8	96.7	56.5
2 state	74.1	98.2	51.9
4 state	74.1	97.0	30.6
Experiment 3			
1 state	70.0	100.0	33.3
2 state	66.7	100.0	30.0
4 state	63.3	85.7	23.3

Table 4: Effectiveness of confidence measure

	$Pr\_Ex(0.95)$	$Av\#\_Cand(0.95)$
Experiment 1		
1 state	70.3	3.23
2 state	77.0	3.50
4 state	76.4	4.28
Experiment 2		
1 state	89.4	2.91
2 state	76.9	2.50
4 state	84.0	4.14
Experiment 3		
1 state	73.9	2.33
2 state	71.4	2.40
4 state	78.3	3.43

Table 5: The probability that the input speaker exists among the set of candidates under the 0.95 confidence level($Pr\_Ex(0.95)$), and the average number of candidates in each set($Av\#\_Cand(0.95)$)

ists among the set of candidates under the 0.95 confidence level. In all cases, the rate of existence of true input speakers among candidates is higher than the identification rate for the baseline system in table 4. This shows that even when the confidence level of the identification result does not reach the predefined level, there is still the possibility that we can obtain a enhanced result. Table 5 also shows the average number of candidates in each set. The smaller the number, the better.

Table 6 shows the identification rates of the baseline system, the conventional method and the proposed method. This table verifies the effectiveness of normalisation. Identification rates were improved using normalisation and scoring methods compared to the baseline system which does not use them. In addition, the proposed method shows better or equal performance when compared to the conventional method. As the number of enrolled speakers becomes large, candidate selection has a great effects. Also, as the amount of training data reduces, there is more room for enhancement using normalisation.

	Baseline	Conventional	Proposed
Experiment 1			
1 state	63.0	67.6	70.4
2 state	59.3	62.0	65.7
4 state	55.6	57.4	59.3
Experiment 2			
1 state	77.8	80.6	80.6
2 state	74.1	75.9	77.8
4 state	74.1	74.1	74.1
Experiment 3			
1 state	70.0	70.0	70.0
2 state	66.7	70.0	70.0
4 state	63.3	66.7	66.7

Table 6: Identification rates of the baseline system, the conventional method and the proposed method

5. Conclusions

This paper has described a candidate selection method using a confidence measure based on significance testing, likelihood ratio normalisation using candidates as a background model, and a scoring method with candidates. Our results show that we could select candidates well using the proposed confidence measure. Performance was also enhanced in the selection of candidates with likelihood ratio normalisation and a scoring method. As the number of enrolled speakers increases and the amount of training data per speaker decreases, the performance gain becomes greater.

6. Acknowledgements

Thanks to Phil Woodland, Nathan Smith and Gavin Smith for careful review of the paper and comments on it.

7. REFERENCES

1. A. E. Rosenberg Chin-Hui Lee, J. DeLong. The Use of Cohort Normalized Scores for Speaker Verification. In *Proc. ICSLP*, pages 599–602, 1992.
2. Sadaoki Furui. An Overview of Speaker Recognition Technology. In *ESCA workshop on Automatic Speaker Recognition Identification Verification*, pages 1–9, 1994.
3. H. Gish. Text Independent Speaker Identification. In *IEEE Signal Processing Magazine*, pages 18–32, 1994.
4. Ji-Hwan Kim. Improvement of Speaker Identification Systems Using Candidate Selection and Likelihood Ratio Normalisation. Master's thesis, KAIST, 1998.
5. Jack E. Poter Kung-Pu Li. Normalizations and Selection of Speech Segments for Speaker Recognition Scoring. In *Proc. ICASSP*, pages 595–598, 1988.
6. Seong-Jin Yun. Performance Improvement of Speaker Recognition System for Small Training Data. Master's thesis, KAIST, 1994.