ON THE USE OF VOTING METHODS FOR SPEAKER IDENTIFICATION BASED ON VARIOUS RESOLUTION FILTERBANKS

Bong-Jin Lee, Sung-Wan Yoon, Hong-Goo Kang, Dae Hee Youn

DSP Lab. Dep. of EE, Yonsei University
Biometrics Engineering Research Center
Seoul, Korea

ABSTRACT

This paper proposes a novel speaker identification system based on score fusion of various resolution filterbanks. The proposed system uses multiple features which are extracted from filterbanks having various spectral resolutions. Each speaker model is constructed by independent feature set, but the system makes final decision by combining the outcome of each model. We introduce several well-known voting methods for decision. Simulation results using TIMIT database show that the proposed score fusion method significantly improves speaker identification performance compared to single model one. Especially, 59.28% of relative improvement is achieved by using a product rule.

1. INTRODUCTION

There have been lots of studies to improve speaker recognition performance, and varieties of new methods are proposed to obtain good performance. One of the methods is adopting multiple speaker models constructed by various resolution filterbanks. We have also proposed a speaker identification system based on various resolution filterbanks to improve the performance [1].

The idea of the system is started from the observation of the inconsistence of speaker identification error patterns depending on the type of filterbank structure used. By applying various resolution filterbanks to system, we can choose a filterbank that has the best performance. However, there is no analytic method to find generalized optimal filterbank which gives best performance in all condition. Thus, we choose a method of combining the results of multiple filterbanks instead of finding optimal one [1].

The proposed system combines the outcome of each system by multiplying the likelihood values of each model. In addition to the multiplication rule in our previous study, there are several other ways on combining outcomes to make a final decision. In this paper, we adopt several well-known voting methods for combination such as plurality, majority, sum, product, and Borda count [5] and analyze the performance of each voting method.

The speaker identification experiments using TIMIT database is performed to analyze the performance. The experimental results show that all the combination methods except the sum method improve the performance of the proposed system. In addition, the product rule gives the best performance among them though the plurality and majority rules also provide comparable performance. These methods can be used together if the system is working in high secure condition.

In section 2, we briefly describe our previous system using various resolution filterbanks again because it was not been published yet. In section 3, voting methods that are main contribution of in this paper are described. Section 4 gives experimental results and analysis. Finally in section 5, we summarize the study.

2. VARIOUS RESOLUTION FILTERBANKS

It is very important to extract suitable speaker discriminative information for successful speaker recognition because speaker information varies with the feature types used in general. In the system using filterbank based features, the type of filterbank structure affects the representation of speaker information, thus different filterbank structures result in performance variation in speaker recognition systems [3]. Though, many researchers have tried to find the optimal filterbank structure [4], it is difficult to find generalized and optimized filterbank structure. Gravier & et al. concluded that the optimal filterbank was varied depending on the duration of test segment [3]. Moreover, in our preliminary experiments, it was also shown that error patterns of speaker identification test were somewhat inconsistent depending on the type of filterbank warping function. In other words, recognition of each speaker by varying filterbank warping function causes different types of identification error. To show the inconsistency of error patterns in speaker recognition systems, speaker identification using TIMIT database is performed by changing filterbank structure. In the experiment, the filterbank is generated

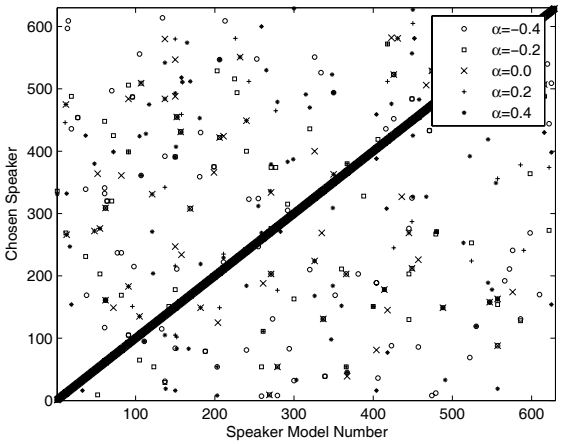

Fig. 1. Error pattern of different filterbank structure.

using following equation [3]:

$$\omega' = \omega + 2\arctan\left[\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)}\right], \tag{1}$$

where ω and ω' are the original and warped frequency in radian, and α is a parameter which controls spectral resolution of filterbank. The parameter α can be varied in $\alpha \in (-1, 1)$. Positive α gives higher spectral resolution in low frequencies and vice versa.

Fig. 1 shows the identification result. In the figure, x-axis denotes speaker model number, y-axis denotes chosen speaker, and the points marked by five symbols are speaker identification results. Thus, marks at diagonal points mean that identification is successful and ones at off diagonal is not. According to the results, different α causes different error pattern. Therefore, even if the optimal filterbank is found, it can still bring errors which are not presented in other filterbank structures. To compensate these kinds of errors and improve the performance, various resolution filterbanks approach is proposed.

The structure of the proposed system is depicted in Fig. 2. The proposed system extracts multiple features from several filterbanks which have various spectral resolution; $\mathbf{FB}_1$, $\cdots$,$\mathbf{FB}_N$. To generate multiple filterbanks, the warping function defined in Eq. (1) is used. Feature vectors from each of the filterbank structure are enrolled to each speaker model; $\mathbf{M}_1, \cdots, \mathbf{M}_N$. In the test procedure, the likelihood values from each structure are calculated independently. The decision logic takes the likelihood values and makes final decision. Several ways of score fusion and voting rules can be used for the decision procedure.

The conventional speaker identification rule [2]:

$$\hat{S} = \arg\max_{1\le k\le S} \log p(\mathbf{X}|\lambda_k), \tag{2}$$

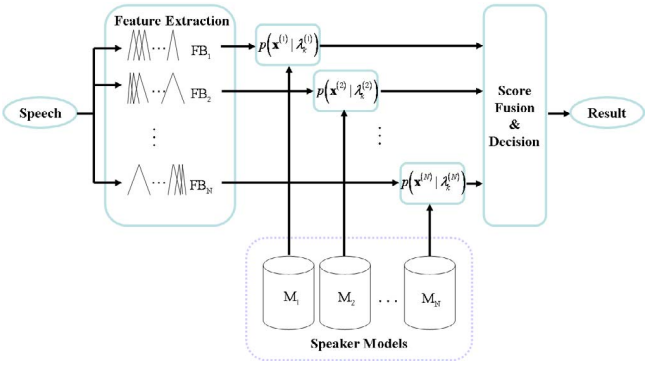

Fig. 2. Structure of the proposed system.

where S is a group of speakers, λ_k is a statistical model of speaker k, and $\mathbf{X}$ is a sequence of feature vectors, is changed to the proposed decision criterion defined as follows:

$$\hat{S} = \arg\max_{1\le k\le S} \Gamma\left(\log p(\mathbf{X}^{(\alpha_1)}|\lambda_k^{(\alpha_1)}), \right. \\ \left. ..., \log p(\mathbf{X}^{(\alpha_{N_A})}|\lambda_k^{(\alpha_{N_A})})\right), \tag{3}$$

where A is a set of chosen filterbank warping parameters, N_A is the number of chosen α. $\Gamma(\cdot)$ is the combination function of decision block, $\mathbf{X}^{(\alpha_n)}$ is a sequence of feature vectors generated from various warped filterbanks with $\alpha = \alpha_n$, and $\lambda_k^{(\alpha_n)}$ is the GMM of speaker k which is generated from the feature vectors with $\alpha = \alpha_n$.

3. VOTING METHODS

There is a variety of score combination and decision rules in pattern recognition tasks which uses multiple classifiers. Choosing a combination method is one of the important issues in a multiple classifier system because it affects the performance of the system directly. The proposed various resolution filterbank system can be also regarded as a multiple classifier system, and the combination function $\Gamma(\cdot)$ varies to the voting method. In this section, several voting methods are described in brief. Voting methods are classified into three types [5]; unweighted voting methods, confidence voting methods, and ranked voting methods. This section reviews some of these voting methods and apply them to the proposed system [1].

3.1. Unweighted Voting Methods

In unweighed voting methods, each speaker recognition system with a warping factor of $\alpha = \alpha_n$ gives just one vote of speaker. In other words, each of the system makes its own decision and the proposed system chooses the most voted speaker among them. These methods have less complexity

[1] Reference of this section: [5].

than other methods because they require only the speaker information that has the best likelihood value. However, these methods can cause lots of tied results due to integer counting.

Plurality The speaker who receives the highest number wins.

Majority This method is almost same as the plurality method, but the speaker who receives more than half of the votes wins. If there are many speakers and classifiers, the probability of majority win becomes low. It gives no result if majority is not achieved.

3.2. Confidence Voting Methods

In confidence voting methods, the degree of preference for a speaker can be expressed. Since the likelihood value is used for voting in this paper, these methods have more information than unweighted voting methods. However, the absolute scale of the likelihood values between each classifier can be varied. Therefore, before using these methods, the scale of probability should be checked. In this paper, we suppose that the likelihood values of each classifier are within the comparable range.

Sum rule The likelihood values of each filterbank structure are added for each speaker. The speaker with the highest sum is chosen.

Product rule The likelihood values of each filterbank structure are multiplied for each speaker. The speaker with the highest product is chosen. In this method, if there is a very low value, the speaker who receives that value may lose the chance of winning.

3.3. Ranked Voting Methods

In ranked voting methods, each classifier expresses the ranking of the speakers. These methods have less information than the confidence voting methods. However, the scaling is not required because these methods contain just the ranking of the speakers, not the score of speakers.

Borda count This method is developed by Jean-Charles de Borda [6]. The ranking from all classifier are averaged, and the speakers are reranked using the averaged ranking. The speaker who is ranked first is chosen.

4. EXPERIMENTS AND ANALYSIS

4.1. Database Description

TIMIT database is selected for the experiment. TIMIT consists of 630 speakers; 438 male speakers and 192 female speakers. There are 10 sentences for each speaker and the length of each sentence is approximately 3 seconds. In the experiments, five sentences are used for training and remaining five sentences are used for testing.

4.2. Speech Analysis

In the experiment, most of the procedure is same as conventional filterbank-based feature extraction procedure [2]. In the procedure, the filterbank warped by the warping function Eq. (1) is used instead of mel-scale filterbank. The number of filterbank is 23 and first 20 coefficients are used as a feature vector for the experiment.

4.3. Experimental Results

Table 1 shows the result. In the table, $\alpha = \{-n, \cdots, n\}$ means that the classifiers which have the filterbank with $\alpha = \{-n, -n+0.1, \cdots, n-0.1, n\}$ are combined. There are *best* and *worst* in the plurality, majority, and Borda count methods. The error rate in the *best* columns is the error rate with the assumption that whenever the methods cannot decide who the speaker is, it is regarded as correct decision. This happens when two speakers have tied result in the plurality and Borda count or no one receives majority in the majority methods. In these cases, the identification system may request additional utterance to the speaker and run the recognition system again. Thus, it can be the minimum bound of the systems. On the other hand, the error rate in the *worst* columns is the maximum bound of the systems.

In the table, the *plurality* method has small difference between *best* and *worst* and the difference goes to smaller as the number of α values is increased because the use of more classifiers also reduces tied result. In the table, the minimum error rate of the plurality method is 0.73% and it is slightly better than the result of the product method. Moreover, the *worst* probability is also not so bad. Thus, if the system requirements include less complex and error rate, the plurality method can be adopted for the application.

The *majority* method has tighter constraint than the plurality method for a decision. As shown in the table, the minimum error rate is 0.13% because most of the ambiguous trials -may be the errors- are detected. It means that, by correcting all of the undecided identification trials, the majority method can achieve the best performance among the tested methods. However, this method has a drawback that the undecided event happens more frequently comparing to the *plurality* method.

The *sum* method achieves poor result. The performance is improved until $\alpha = \{-0.4, \cdots, 0.4\}$ but it turns to degrade after that. It seems that the reason of performance degradation is the poor identification performance of each system using α outside $\alpha = \pm 0.4$.

On the other side, the *product* method still gives performance improvement when α values outside $\alpha = \pm 0.4$ are used. The reason is that the product rule is highly subjective to low probability [5]. Actually, it is known as the drawback of the product rule because a low probability can ruin the chance of winning. It means that a speaker who is not the speaker of given speech has relatively large chance of re-

Table 1. Speaker identification error rate of various voting methods

System	Unweighted voting methods				Confidence voting methods		Ranked voting methods	
	Plurality		Majority		Sum	Product	Borda count	
	best	worst	best	worst			best	worst
$\alpha = 0.0$	1.94	1.94	1.94	1.94	1.94	1.94	1.94	1.94
$\alpha = \{-0.1, \cdots, 0.1\}$	1.08	1.62	1.08	1.62	1.24	1.27	1.37	1.46
$\alpha = \{-0.2, \cdots, 0.2\}$	0.95	1.37	0.83	1.49	1.14	1.11	1.24	1.46
$\alpha = \{-0.3, \cdots, 0.3\}$	0.95	1.33	0.69	1.59	1.24	1.17	1.46	1.46
$\alpha = \{-0.4, \cdots, 0.4\}$	0.89	1.30	0.54	1.65	1.05	1.08	1.37	1.43
$\alpha = \{-0.5, \cdots, 0.5\}$	0.98	1.24	0.41	1.71	1.11	1.05	1.30	1.44
$\alpha = \{-0.6, \cdots, 0.6\}$	0.92	1.05	0.32	1.65	1.59	1.11	1.49	1.49
$\alpha = \{-0.7, \cdots, 0.7\}$	0.76	0.98	0.25	1.75	2.41	0.98	1.30	1.37
$\alpha = \{-0.8, \cdots, 0.8\}$	0.73	0.89	0.19	1.81	5.24	0.86	1.24	1.30
$\alpha = \{-0.9, \cdots, 0.9\}$	0.73	0.95	0.13	2.00	26.38	0.79	1.59	1.59
Relative Improvement(%)	62.37	51.03	93.29	-3.09	-1260	59.28	18.04	18.04

ceiving a low probability in each filterbank system. Thus, the probability gap between correct speaker and other speakers grows by multiplying the likelihood values.

Borda count gives ordinary performance. Its tendency of performance is similar to that of the sum method. However, the performance degradation is much less than the sum method because it replaces the probabilities to the ranks, which gives similar effect to normalization.

The results show that plurality, majority, and product methods give reasonable performance. Plurality method has less complexity and small bound of identification error. Majority method has relatively large bound of error comparing to the plurality method but it can be applied for more secure systems than others. Product method is more complex than other methods but it gives best performance on average. Based on the observations, flexible voting methods can be designed depending on the application areas.

5. CONCLUSION

In this paper, a speaker identification system based on score fusion of various resolution filterbanks was proposed to improve the performance. By applying several well-known voting methods, the performance improvement of the proposed system was verified. In the experiments using TIMIT database, the product method achieves 59.28% of relative improvement comparing to the single filterbank system. Moreover, by using plurality and majority methods, further improvement of identification performance was expected. Even if the proposed system has higher complexity than conventional single filterbank system, it can be useful for the speaker identification systems requiring high security.

6. ACKNOWLEDGEMENT

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University.

7. REFERENCES

[1] B. J. Lee, H. G. Kang, and D. H Youn, "On the Use of Various Resolution Filterbanks for Speaker Identification," submitted to *Speech Communication*.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72-83, 1995.

[3] G. Gravier, and et al., "Model Dependent Spectral Representations for Speaker Recognition," *Proc. of Eurospeech 1997*, pp. 2299-2302, 1997.

[4] C. Miyajima, and et al., "Speaker recognition based on discriminative feature extraction - optimization of mel-cepstral features using second-order all-pass warping function," *Proc. of Eurospeech 1999*, pp. 779-782, 1999.

[5] M. V. Erp, and et al., "An Overview and Comaprison of Voting Methods for Pattern Recognition," In *proc. of IWFHR*, pp. 195-200, 2002.

[6] J. C. d. Borda. *Memoire sur les Elections au Scrutin.* Histoire de l'Academie Royale des Sciences, Paris, 1781.