NORMALIZED MINIMUM-REDUNDANCY AND MAXIMUM-RELEVANCY BASED FEATURE SELECTION FOR SPEAKER VERIFICATION SYSTEMS

Chi-Sang Jung [1,3], Moo-Young Kim [2,3], and Hong-Goo Kang [1]

[1]Department of Electrical & Electronic Engineering, Yonsei University, Seoul, Korea
[2]Department of Information & Communications Engineering, Sejong University, Seoul, Korea
[3]Biometrics Engineering Research Center (BERC), Seoul, Korea

[jtoctos@dsp.yonsei.ac.kr[1], mooyoung@sejong.ac.kr [2], and hgkang@yonsei.ac.kr[1]]

ABSTRACT

In this paper, an information theoretical approach to select features for speaker recognition systems is proposed. Conventional approaches having a fixed interval of analysis frames are not appropriate to represent dynamically varying characteristics of speech signals. To maximize the speaker-related information varied by the characteristics of speech signals, we propose an information theory based feature selection method where features are selected to have minimum-redundancy with in selected features but maximum-relevancy to training speaker models.

Experimental results verify that the proposed method reduces the error rates of speaker verification systems by *27.37 %* in NIST 2002 database.

Index Terms— feature selection, minimum-redundancy, maximum-relevancy, speaker verification systems

1. INTRODUCTION

In pattern recognition tasks, it is well-known that finding an effective set of features both in training and test stages is an important step to improve the performance of recognition systems [1, 2]. Especially, it is crucial in speaker recognition systems since feature characteristics vary dynamically due to the difference of articulation on each phoneme. In addition, since the relative ratio and distinctiveness of each phoneme to recognition is not equivalent, a method to select relevant features among many candidates becomes an important research topic recently [3, 4].

In speaker recognition systems, a frame shift length is generally fixed to a half of analysis frame length to reduce computational complexity. It is also known that the performance of speaker recognition systems improves as the length of input data or the number of features to be trained and tested increases [5]. However, the performance can be even dropped if redundant features that do not clearly represent speaker-related information dominantly affect to either a training or testing process. For example, we may not need to continuously extract features in steady-state vowel regions, while we had better frequently capture features in dynamically varying regions such as consonants or transition [6]. Since an approach needs a classification step which is normally complicated, it may not be a good idea to rely on phonetically motivated information.

The mutual information defined in information theory has been used to measure the amount of speaker-related information embedded in features because the recognition accuracy is proportional to the mutual information between speaker models and test features [4]. In other words, a minimum error can be achieved by maximizing statistical dependency between a training model and matched test data [2]. One of the most popular approaches to realize the maximum dependency is extracting features having maximum-relevance or maximizing mutual information. As we mentioned above, however, utilizing features that are tightly coupled with a training model does not lead to the best performance because the training model is over fitted to the characteristics of redundant features [2, 4]. Therefore, we also need to include a criterion to reduce the redundancy between already extracted feature vectors.

Considering these two facts, this paper proposes a new feature selection method based on the normalized minimum-redundancy and maximum-relevancy (NmRMR) criterion, which minimizes the redundant information between selected features but maximizes the mutual information between training speaker models and test features. At first, to train speaker models reduced redundant feature vectors, we choose training features using the minimum-redundancy criterion. In addition, to select features used for a test process, three criteria, i.e. maximum-relevancy, minimum-redundancy, and NmRMR criterion, are applied. Compared to conventional feature selection methods, the proposed NmRMR method improves the performance by *27.37 %* in NIST 2002 database [7].

Section 2 describes an idea of implementing the proposed NmRMR algorithm by controlling a feature selection interval. In section 3, we compare the performance of the proposed algorithm with conventional approaches using speaker verification systems. Conclusions follow in section 4.

2. PROPOSED FEATURE SELECTION ALGORITHM

This section describes an efficient way to represent the dynamic characteristics of spectral features based on the criterion of mutual information theory.

2.1. Normalized minimum-redundancy and maximum-relevancy

The information theoretical approach is a powerful method that has proven useful to analyze feature selection in speaker recognition systems [4]. The mutual information is measured by following definition:

$$\begin{aligned} I(f;c) &= H(c) - H(c \mid f) \\ &= \left(-\sum_{c} p(c) \log p(c) \right) \\ &\quad - \left(-\sum_{f} p(f) \sum_{c} \frac{p(f \mid c)}{\sum_{c'} p(f \mid c')} \log \frac{p(f \mid c)}{\sum_{c'} p(f \mid c')} \right), \end{aligned} \tag{1}$$

where f and c denote test features and speaker models, respectively. $H(c)$ and $H(c \mid f)$ denote the entropy of c and conditional entropy of c given the test features f. The mutual information between features and speaker classes can be computed by, the probability $p(c)$, $p(f)$ and the likelihood $p(f \mid c)$ which are already known.

To minimize recognition errors, selected features should be statistically dependent with a reference model [2]. Since highly-correlated test features do not always lead to good recognition performance [8], it also needs to reduce the redundancy between selected features. The rationale behind of our idea is extracting feature vectors to minimize the redundancy among selected features, but maximize the relevancy between speaker's reference model and test features.

From these theoretical backgrounds, we propose a new feature selection algorithm called the normalized minimum-redundancy and maximum-relevancy (NmRMR) criterion as follows:

$$F_{TR} = \min_{f_{n,i} \in F_n} \underbrace{\left[I(f_{n,i}; f_S) \right]}_{\text{redundancy}}, \tag{2}$$

$$F_{TE} = \max_{f_{n,i} \in F_n} \left[\underbrace{\frac{I(f_{n,i};c) - \mu_{I(f_i;c)}}{\sigma_{I(f_n;c)}}}_{\text{normalized relevancy}} - \underbrace{\frac{I(f_{n,i};f_S) - \mu_{I(f_n;f_S)}}{\sigma_{I(f_n;f_S)}}}_{\text{normalized redundancy}} \right], \tag{3}$$

where F_n and $f_{n,i}$ are a candidate feature set and the i-th candidate feature at n-th segment, respectively, f_S is a selected feature set, and c denotes a speaker model. Selected feature sets in training utterances and test utterances, F_{TR} and F_{TE}, are obtained in each segment. The criterion is normalized by considering the standard deviation and mean of relevancy term and redundancy term in each segment. $\mu_{I(f_n;c)}$, $\sigma_{I(f_n;c)}$, $\mu_{I(f_n;f_S)}$ and $\sigma_{I(f_n;f_S)}$ denote the mean and standard deviation of relevancy term and redundancy term at n-th segment, respectively. The normalization process in each segment removes the deviation of dynamic range and the variance between relevancy and redundancy terms. Regardless of the types of database and phonetic characteristics of segment, the normalization process compensates for the difference between relevancy and redundancy terms.

The relevancy term $I(f_{n,i};c)$, defined in equation (3) needs features that have the largest relevancy with the speaker model c. The relevancy is measured by mutual information between a candidate feature and a speaker model. It can be calculated by equation (4), which indicates that the mutual information between features and speaker class can be computed by $p(c)$, $p(f_{n,i})$, and the likelihood $p(f_{n,i} \mid c)$ which have been already known. On the contrary, the redundancy factor $I(f_{n,i};f_s)$ should be minimized to maximize equation (3). It is approximated by the mutual information between selected features and a candidate feature as shown in equation (5) where $N(f_{n,i} \mid \mu, \Sigma)$ is the probability mass function of $f_{n,i}$

$$\begin{aligned} I(f_{n,i};c) &= H(c) - H(c \mid f_{n,i}) \\ &= \left(-\sum_{c} p(c) \log p(c) \right) - \left(-\sum_{F_n} p(f_{n,i}) \sum_{c} p(c \mid f_{n,i}) \log p(c \mid f_{n,i}) \right) \\ &= \left(-\sum_{c} p(c) \log p(c) \right) - \left(-\sum_{F_n} p(f_{n,i}) \sum_{c} \frac{p(f_{n,i} \mid c)}{\sum_{c'} p(f_{n,i} \mid c')} \log \frac{p(f_{n,i} \mid c)}{\sum_{c'} p(f_{n,i} \mid c')} \right). \end{aligned} \tag{4}$$

$$\begin{aligned} I(f_{n,i};f_s) &= H(f_s) - H(f_s \mid f_{n,i}) \\ &= \left(-\sum_{f_s} p(f_s) \log p(f_s) \right) - \left(-\sum_{F_n} p(f_{n,i}) \sum_{f_S} p(f_S \mid f_{n,i}) \log p(f_S \mid f_{n,i}) \right) \\ &\approx \left(-\sum_{f_s} p(f_s) \log p(f_s) \right) - \left(-\sum_{F_n} p(f_{n,i}) \sum_{m=1}^{M_S} \frac{w_m N\left(f_{n,i} \mid \mu_m, \Sigma_m\right)}{\sum_{m'=1}^{M_S} w_{m'} N\left(f_{n,i} \mid \mu_{m'}, \Sigma_{m'}\right)} \log \frac{w_m N\left(f_{n,i} \mid \mu_m, \Sigma_m\right)}{\sum_{m'=1}^{M_S} w_{m'} N\left(f_{n,i} \mid \mu_{m'}, \Sigma_{m'}\right)} \right), \end{aligned} \tag{5}$$

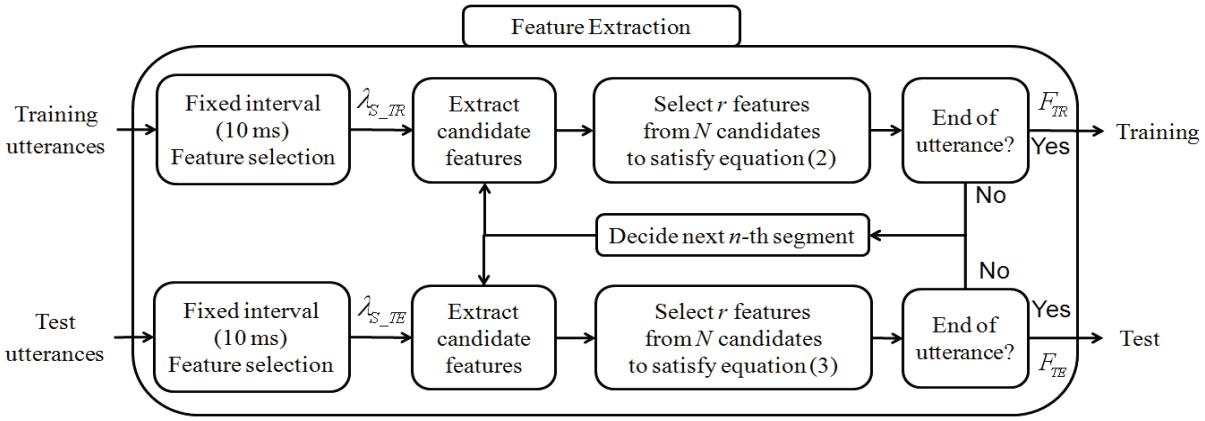

Fig. 1 Flow chart of the NmRMR based feature selection algorithm

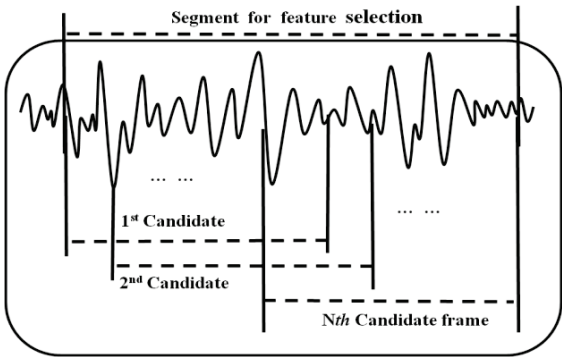

Fig. 2 Segment for the NmRMR based feature selection algorithm

given a Gaussian distribution λ_S. λ_S denotes a Gaussian mixture model of initially selected feature set *S*, with weight, mean, covariance parameters and the number of mixture is set to M_S.

$$\lambda_S = (w_m, \mu_m, \Sigma_m), \quad m = 1, ..., M_S. \tag{6}$$

Equation (5) denotes that the redundancy value, which is the mutual information between selected features and a candidate feature, is approximated by the conditional probability of candidate feature given the Gaussian mixture model λ_S.

2.2. NmRMR based feature selection method

Using the proposed NmRMR algorithm and derived equations, we perform a feature selection process. Fig. 1 shows a flow chart of the proposed NmRMR based feature selection algorithm. In a training procedure, an initial speaker model λ_{S_TR} for computing redundancy values is first trained using a shift interval of 10ms. Then features for training speaker models are re-selected to reduce redundancy between features. We choose *r* features among *N* features such that the mutual information between initial speaker model λ_{S_TR} and candidate features is minimal. Fig. 2 shows that the segment consists of *N* candidate features with the frame adopted to choose *r* features in each segment until the segment is set to the end of utterance. Finally, the selected feature set F_{TR} in training utterances are used for training speaker models.

A process to select distinctive features for a test is similar to the process to select features for the training. An initial speaker model λ_{S_TR} is trained similar to the training method. We also choose *r* features in every *N* candidate features that satisfy the NmRMR equation (2). The feature selection algorithm for a training step utilizes only the redundancy term because the relevancy value cannot be considered, while the feature selection algorithm for a test uses not only redundancy term but also relevancy term between the features and speaker models. When the iteration of the feature selection comes to an end of the test utterance, the selected feature set F_{TE} in the utterance is used for test features. Following these processes, we finally select features to satisfy the normalized minimum-redundancy maximum- relevancy criterion.

We compare the speaker verification performance of three information theoretical criteria, i.e. maximum-relevancy, minimum-redundancy and NmRMR with that of the conventional fixed frame length method.

3. EXPERIMENTS AND RESULTS

We perform a one-speaker detection task using the NIST evaluation task database 2002 [7]. We build a GMM-based speaker verification system with mel frequency cepstrum coefficients (MFCC) and their delta values [9]. Twelfth order MFCCs are extracted with a 20ms analysis window, their delta-MFCCs are additionally used and cepstral mean subtraction (CMS) is applied. Speaker models and universal background models (UBMs) include 128 Gaussians. For an NmRMR based feature selection, we set the interval of frame to 1ms for extracting candidate features. The number of candidate features, *N,* and the number of selected features, *r,* for each segment is set to 10 and 6, respectively. The determined parameter values *N* and *r* show the best performance in our intensive experiments

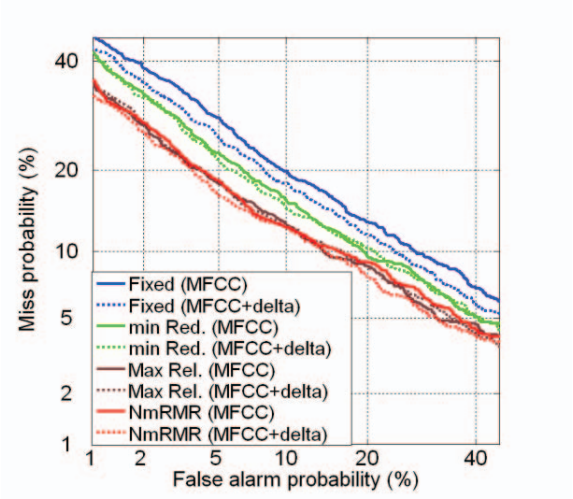

Fig. 2 DET curves for various feature section methods in NIST 2002 database

TABLE I
EER (%) for various feature selection methods in NIST 2002 database

Feature Selection Method	EER (%)	
	MFCC	MFCC + delta-MFCC
Fixed	14.69	13.67
Min-Redundancy	12.67	12.28
Max-Relevancy	11.01	10.78
NmRMR	10.67	10.44

In addition, to separately analyze the effect of relevancy and redundancy terms on a test process, we use three types of criteria given in equation (3), i.e. using only minimum-relevancy, only maximum-relevancy or the combination of two terms which is NmRMR. On the other hand, a unique equation (2) is used for the feature selection criterion in a training process because it is impossible to measure the relevancy of features in a training process.

Fig. 2 and Table I indicate the equal error rates (EER) and detection error tradeoff (DET) curves for the one-speaker detection task. Experiments consist of verifications with MFCCs only and MFCCs with delta-MFCCs. Experimental results show that the proposed NmRMR method using only MFCCs has a relative improvement of 27.37 % to conventional fixed method in terms of equal error rate. In the case of combination with delta-MFCCs, the proposed NmRMR method also shows the best performance among all feature selection methods.

The maximum-relevancy or minimum-redundancy only criterion also shows some improvement. Please note that the performance of the maximum-relevancy term only shows similar performance to the NmRMR criterion. It means that it is important to choose test features which maximize relevancy with speaker models because we have already trained speaker models by removing redundant features in equation (2). However, the combination of two criteria, i.e. NmRMR, still shows the best performance, which confirms the efficacy of the proposed NmRMR method in speaker verification systems.

4. CONCLUSION

In this paper, we have proposed a feature selection method based on the NmRMR algorithm. The conventional fixed frame interval based feature selection method is not a good choice if the characteristics of speech signal change dynamically. The proposed method selects distinctive features not only to have high mutual information between feature sets and speaker models, but also to have minimum-redundancy between selected features. Experimental results showed the superiority of the proposed method.

ACKNOWLEDGEMENTS

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University. (R112002105070040 (2008))

REFERANCE

[1] H. C. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min- redundancy," *IEEE trans. Pattern Analysis and Machine Intelligence*, vol 27, pp. 1226-1238, Aug. 2005.

[2] D. P. W. Ellis and J. A. Bilmes, "Using mutual information to design feature combinations," *in Proc. Int. Conf. on Spoken Language Processing*, pp. 79–82, 2000.

[3] S. Kim, T. Eriksson, H. G. Kang, and D. H. Youn, "A pitch synchronous feature extraction method for speaker recognition," *in Proc. Internat. Conf. Acoust. Speech Signal Processing*, vol. 1, pp. 405–408, 2004.

[4] T. Eriksson, S. Kim, H. G. Kang, and C. Y. Lee, "An Information-Theoretic Perspective on Feature Selection in Speaker Recognition," *IEEE Signal Processing Letters*, vol. 12, no. 7, July 2005

[5] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, no. 2. 1995.

[6] P. Scanlon, D. Ellis, R. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 15, pp. 803–812, Mar. 2007.

[7] "The NIST year 2002 speaker recognition evaluation plan," 2002, http://www.nist.gov/speechltests/spk/ZOOZdoc

[8] S. Cang and H. Yu, "A new approach for detecting the best feature set," *in Proc. IEEE Networking, Sensing and Control*, pp. 74-79, Mar. 2005.

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