Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **2**, March **2017**

A Long-Term Feature Selection Based on Heuristic Strategy for Speaker Diarization

Xixiang Zhang, Huan Zhao*

School of Information Science and Engineering
Hunan University
Changsha, P.C 410082 - China
Corresponding Author: hzhao@hnu.edu.cn

Received July, 2016; revised December, 2016

Abstract. *The goal of speaker diarization is to determine who spoke when? in a speech clip. Recent studies on speaker diarization have shown that many long-term features can improve the diarization result. However, some selected features do not have obvious speaker discriminability and some of them are redundant which may obscure diarization. In this paper, we propose a feature selection based on heuristic strategy, which is possible to generate long-term features with high speaker discrimination. Results of experiments on AMI meeting corpus demonstrate that proposed method can effectively select long-term features while improving the performances of the state-of-the-art speaker diarization system LIUM. The average diarization error rate (DER) is reduced by almost 4.8% relative to the baseline feature set.*

Keywords: Speaker diarization, Long-term feature, Features selection, Heuristic strategy

1. **Introduction.** Speaker diarization has appeared as a considerable domain of speech research. The aim of speaker diarization is to judge who spoke when? The speaker diarization task consists of segmenting a long speech containing multi-speakers into the segments which contain only one speaker, and clustering together all the divisional speech segments that accord to the same speaker. The pre-processing step of various speech processing tasks needs to group all speech from one special speaker. The common application instances for speaker diarization include audio analysis, speaker verification, audio retrieval and automatic speech recognition. Therefore, it is possible to gradually improve speaker diarization performance.

Recent studies have shown that a lot of speech long-term features can afford important information for speaker discrimination[1]. Using a combination of the long-term features combined with traditional acoustic features leads to enhancement in terms of the diarization error rate. In the research domain of speaker diarization, long-term features have been effectively utilized in combination with Mel Frequency Cepstrum Coefficients (MFCCs). Previous studies have investigated different long-term features for the speaker diarization task. However, many selected long-term features have not high score of speaker discriminability[2]. There are some studies selected lots of long-term feature, but did not specify what the speaker discriminability they are[3]. This paper aims at giving a list of candidate long-term features with high discriminability for speaker diarization. We proposed a long-term feature generate method using the heuristic strategy. Furthermore, we extend our analysis so as to estimate the candidate long-term features and to obtain more detailed information about the discrimination in speaker diarization system.

The rest of the paper is organized as follows. In Section 2, we introduce the related word in speaker diarization. Section 3 introduces the baseline speaker diarization system, and discusses methods for long-term feature selection. Section 4 presents the proposed method. Section 5 discusses the experiments in LIUM speaker diarization system[4]. Section 6 describes the conclusion and future work.

2. **Related work.**

Prosodic and other long-term features can indicate personal characteristics of the speakers voices, such as intonation, timing, and loudness. Some long-term features have high computational complexity for extraction, while others are hard to solely extraction from acoustics. Therefore, higher-level speech long-term features have more and more consideration in recent years [2]. In the related field of speaker recognition, higher-level long-term features have been effectively utilized in combination with traditional acoustic features such as MFCCs[5]. In[6], authors have proposed a method to improve the short-term spectral feature based overlap detector by fusing information from long-term conversational features in the form of speaker change statistics. Studies on speaker diarization have shown that long-term features are one of the signicant ways of optimization the DER. In [7], authors have discussed the probability of using long-term prosodic features for the exploration of overlapping speech for speaker diarization. The authors in [2] investigated the speaker discriminability of 70 different long-term features and then, selected the top 10 long-term features with short-term acoustic features to increase the performance of speaker diarization. They provide evidence that lots of long-term features can offer effective information (between/within) for speaker distinction. In [3], 12 top-ranked long-term acoustic features such as formants, pitch and harmonics are conjunctively used for improving the diarization performance. Reference [8] shows that good accuracy is obtained with an appropriate selection of prosodic, temporal and basic signal features extracted from speech clips after speaker segmentation. It also discusses the use of 12 long-term features. All the above-mentioned studies are based on long-term features, such as pitch, formants, energy, long-term average spectrum and HNR. For each feature (e.g., pitch), a lot of statistical properties are estimated: mean, median, minimum, maximum, difference, standard deviation and the slope of the curve. However, some chosen features do not have a high speaker discriminability and many of them are redundant and may obscure diarization. It is more desired to investigate long-term features with high discriminability for speaker diarization.

3. **Feature Selection for Speaker Diarization.**

3.1. **Methods for Speaker Diarization.**

A typical speaker diarization system is composed of the following steps: pre-processing module, feature extraction module, speaker-based segmentation module, speaker clustering module, and speech labeling module[1]. Most of present state-of-the-art speaker diarization systems fit into one of two categories: the bottom-up and the top-down approaches[9].

1) Speech pre-processing generally includes speech activity detection (SAD) and acoustic beamforming. SAD identifies the labeling of speech and non-speech. Acoustic beamforming technology is mainly used for speaker diarization in multiple distant microphone (MDM) condition.

2) Feature extraction can concern the diarization system performance obviously. Features extracted from the acoustic signal are intended to distinguish each speaker. Mel-frequency cepstral coefficients (MFCCs), MFCCs first or second derivatives, short-time

energy (STE), zero-crossing rate (ZCR), Pitch, Spectrum Magnitude, Line Spectrum Pairs (LSPs) are the most common features.

3) Speaker segmentation designs to splitting the audio stream into segments of same speaker, alternatively, explores speaker turns. Segmentation approaches are generally categorized in the following four types: model-based segmentation, silence detection based methods, distance based methods and hybrid speaker segmentation.

Distance based speaker segmentation approaches do not require any prior knowledge on the information of speaker identities or the number of speakers. Bayesian information criterion (BIC) was used in this paper, which is the most popular distance based segmentation. BIC can distinguish which of the models indicates speech segments best. This criterion looks for speaker turns points within a detecting window using a penalized likelihood ratio test of whether the speech in the detecting window is better represented by a single distribution (no speaker turns) or two different distributions (speaker turns). Assume sample x_i is n dimensional feature vectors. In order to determine whether or not a speaker turns point appears at t_j, two neighboring analysis windows X and Y are next to time t_j are considered. Suppose $Z = X \bigcup Y$, between two hypotheses H_0 and H_1, then need to compute a penalized likelihood ratio test. Under H_0, there is no speaker turns at time t_j. This value of θ_z implies the data samples in Z are represented by a multivariate Gaussian probability distribution function. The log likelihood L_0 is computed as follows:

$$L_0 = \sum_{i=1}^{n_X} \log p\left(x_i|\theta_z\right) + \sum_{i=1}^{n_Y} \log p\left(y_i|\theta_z\right). \tag{1}$$

where n_X are the number of speech samples in detecting windows X, and n_Y are the number of speech samples in detecting windows Y. Under H_1, a speaker turn occurs at time t_j. The windows X and Y are accords with two multivariate Gaussian densities, which are denoted by h_X and h_Y, respectively. The log likelihood L_1 is given by:

$$L_1 = \sum_{i=1}^{n_X} \log p\left(x_i|\theta_X\right) + \sum_{i=1}^{n_Y} \log p\left(y_i|\theta_Y\right) \tag{2}$$

The metric between the two neighboring windows X and Y is computed by ΔBIC criterion:

$$\Delta BIC = L_1 - L_0 - \frac{\lambda}{2}\left(d + \frac{d\left(d+1\right)}{2}\right)\log n_z \tag{3}$$

where n_z is the number of frames in window Z, λ is a penalty factor. If $\Delta BIC > 0$, time t_j is considered to be a speaker turns point; if $\Delta BIC < 0$, there is no such point.

4) Speaker clustering obtains one cluster for each speaker with this speakers all speech snippets. Then speech labeling indexes a unique label for each cluster. Speaker clustering are categorized two main groups bottom-up approach and top-down approach. BIC is a common bottom-up clustering method, $BIC_{i,j}$ for grouping two clusters is computed as follows:

$$BIC_{i,j} = \frac{n_i + n_j}{2} log\left|\Sigma\right| - \frac{n_i}{2} log\left|\Sigma_i\right| - \frac{n_j}{2} log\left|\Sigma_j\right| - \lambda P \tag{4}$$

$$P = \frac{1}{2}\left(d + \frac{d\left(d+1\right)}{2}\right) + \log\left(n_i + n_j\right) \tag{5}$$

where Σ is the covariance matrix, and d is the dimension of the feature vectors. If two clusters is best represented by a single full covariance Gaussian, implying only one speaker, the $BIC_{i,j}$ will be a low value; whereas if there are two separate distributions, implying two speakers, the $BIC_{i,j}$ will be high.

3.2. **Long-Term Features List.** Long-term features can capture unique characteristics of each speaker, which cannot be obtained by short-term acoustic features. Therefore, we can combine commonly short-term features with long-term features to improve the diarization results. In this paper, we investigate five different categories long-term features: pitch, energy, formants, harmonics-to-noise-ratio, and long-term average spectrum.

1) Pitch. The range of pitch depends on length and shape of the speakers throat vocal cords, effective frequency range of males was between 87 and 425 Hz, while females between 184 and 880 Hz (see Fig. 1 (2)). The pitch of the maximum value(Max), the minimum value(Min), the median value(Mean), the standard deviation value(Stdev), the slope of the curve value(Swoj) and the differential value(Diff) are extracted for each speech processing section.

2) Energy: The energy feature is mainly relevant to the tonic accent and the expression emotion and which is a intensity contour based on dB value, such as the threshold value which is relevant to human auditory is 1kHz (see Fig. 1 (3)). The energy of the Max, the Min, the Mean, the Stdev, the Swoj and the Diff are extracted for each speech processing section as well.

3) Formant: Formant is the convergence domain of the acoustic energy which the particular frequency interval is about 1000Hz (see Fig. 1 (4)). The formants from 1 order to 5 order are respectively taken as $F1 - F5$, every order formant of the Max, the Min, the Mean, the Stdev, the Swoj and the Diff are extracted as well.

4) Harmonics-to-Noise-Ratio (HtNR): Harmonics to Noise Ratio. The quantization of HtNR feature is associated with the noise which is attached to the speech signal. HtNR is represented by dB, if there is 99% of the periodic signal energy and 1% of the noise, then 0dB HtNR means that the energys between harmonic and non-harmonic are equal (see Fig. 1 (5)). The HtNR of the Max, the Min, the Mean, the Stdev, the maximum frequency value(Fmax), the minimum frequency value(Fmin), the Diff are extracted for each speech processing section.

5) Long-Term Average Spectrum (LTAS): In order to obtain the LTAS, the spectral energy band of 100Hz is related to the features of speech (see Fig. 1(6)). The LTAS of the Mean, the Stdev, the Diff, the Slope value(Slope), the peek-high value(Lph) are extracted for each speech processing section.

4. **Heuristic Feature Selection Method.** The long-term features extracted in this work are pitch, energy, first five formants, HtNR, and LTAS. Statistical properties are estimated for each long-term feature. We evaluate the discrimination of long-term features on TIMIT dataset [10] for feature selection task. This train dataset contains 462 speakers, which are including nearly two-thirds men and one-third women. Each speaker have 10 utterances, total 4620 audio files. Because each utterance contains only one speaker, it easily calculates the long-term feature from speech. The extraction has been completed with Praat (version 5.3.68), one of the most popularly applied speech processing tools [14].

According to Kinnunen and Li [11], appropriate long-term features for speaker modeling and discrimination should have large between-speaker variability and small within-speaker variability. Through the above analysis, we can see that the most long-term features follow a Gaussian distribution. Therefore, we employ Fisher discriminant analysis (FDA)

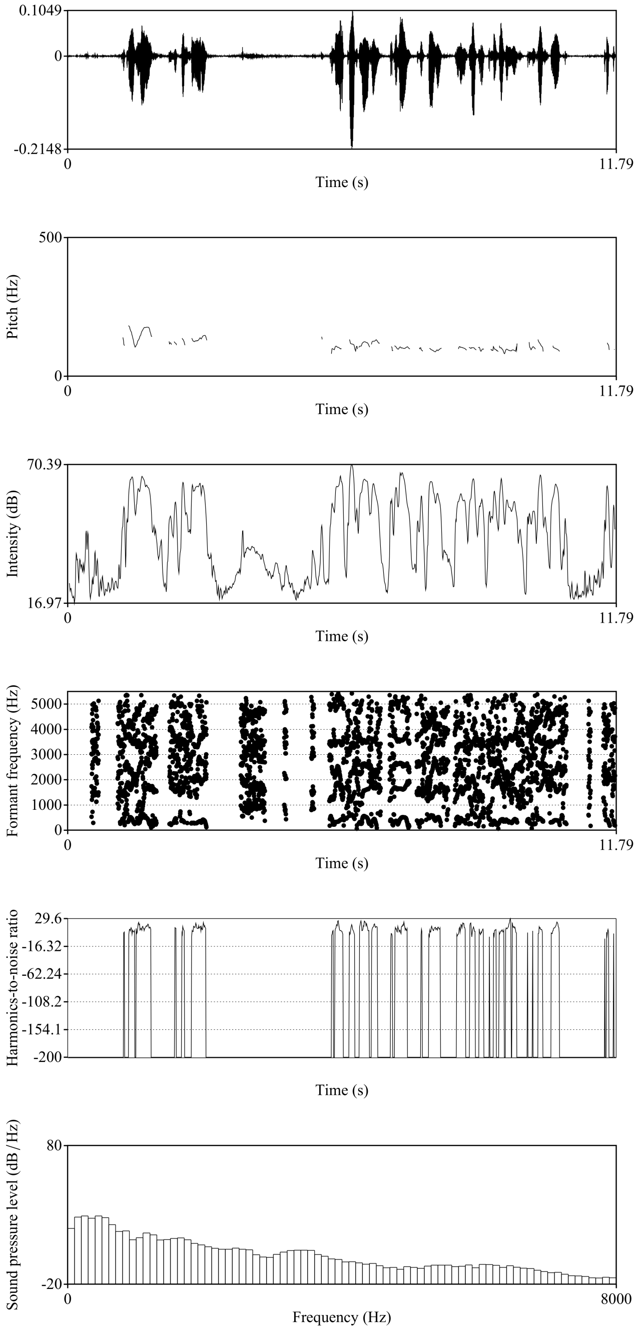


Figure 1. Example waveform (1), a pitch track (2), an energy contour (3), a formant analysis (4), the harmonics-to-noise ratio (5), and long-term average spectrum (6).

approach. For each entire speech fragment, we extract one value per feature. The speaker distinction degree was assessed assuming that features achieve more effective when the rate between inter-speaker and intra-speaker variance is higher. The measure score of the features is defined as:

$$Score = \frac{\sum_{i=1} \sum_{j=1} (\mu_i - \mu_j)(\mu_i - \mu_j)^T}{\sum_i \sum_{j:y_j=i} (x_j - \mu_i)^2} \tag{6}$$

where x represents a sample features, μ is a particular feature mean for speaker i or j, y_j represents the speaker index for the jth sample.

A heuristic algorithm is an attempt at searching an answer to a problem. A common method of applying heuristic strategy is to state the problem, record a number of possible solutions, and then eliminate those that are least likely to be correct [13]. At first, all long-term features are extracted from the corpus. The heuristic feature selection is generated a set of possible solutions and used to determine which new features have higher score of speaker discriminability. The algorithm is iterated until the desired search depth is attained. The proposed heuristic selection as described in Algorithm 1.

The heuristic algorithm works as follows: starting from normal long-term features and their statistical features, the algorithm can generate new forms that have strong speaker discriminability with the normal features in which the FDA score is used to quantify the discrimination degree of forms. And then the generated forms can be further used in finding more new high discriminability features. The top- ranked new features will be added into the candidate feature list for the next iteration. The process can be run iteratively until the search depth is met.

Algorithm 1 Heuristic Feature Selection for Speaker Discriminability

```
Input. Audio corpus including clips of every speaker.
Output. The list of feature ranking result.
• for each $f_i$ in candidate features
•     initiation factor $F_i = f_i$;
•     calculate baseline Score, backup features mean of every speaker;
•         for 1 to max search depth
•             for each $f_j$ in candidate features
•                 $F_i$ combination with factor $f_j$, calculate $Score$;
•             select the form of best Score
•             recovery features mean of every speaker;
```

The outputs are compared with the conventional long-term features and their statistical features in order to obtain overall discriminability as discussed in Section 5. Figure 2 illustrates the flow chart of proposed heuristic selection algorithm.

5. Experiment Verification.

The experiments were carried out in the framework of the LIUM speaker diarization[4]. LIUM_SpkDiarization is a software dedicated to speaker diarization. LIUM is composed of acoustic BIC segmentation followed with BIC hierarchical clustering. Viterbi decoding is performed to adjust the segment boundaries. The speaker diarization performance is also evaluated by speaker clustering method according to corpus annotation. The long-term features extraction has been performed with Praat (version 5.3.68) and normalized.

5.1. Dataset and Evaluation.

We use AMI (Augmented Multiparty Interaction) corpus in our experiments [15]. AMI contains a set of audio and video meeting about almost 100 hours. Each one is made up of 4 to 5 persons. People attending the meeting play different roles in the process, discussing about a products development and promotion. Each one's speech is pre-designed, which we call text-dependent. During discussion, every participant wears a headset microphone and a lapel microphone to record their speech.

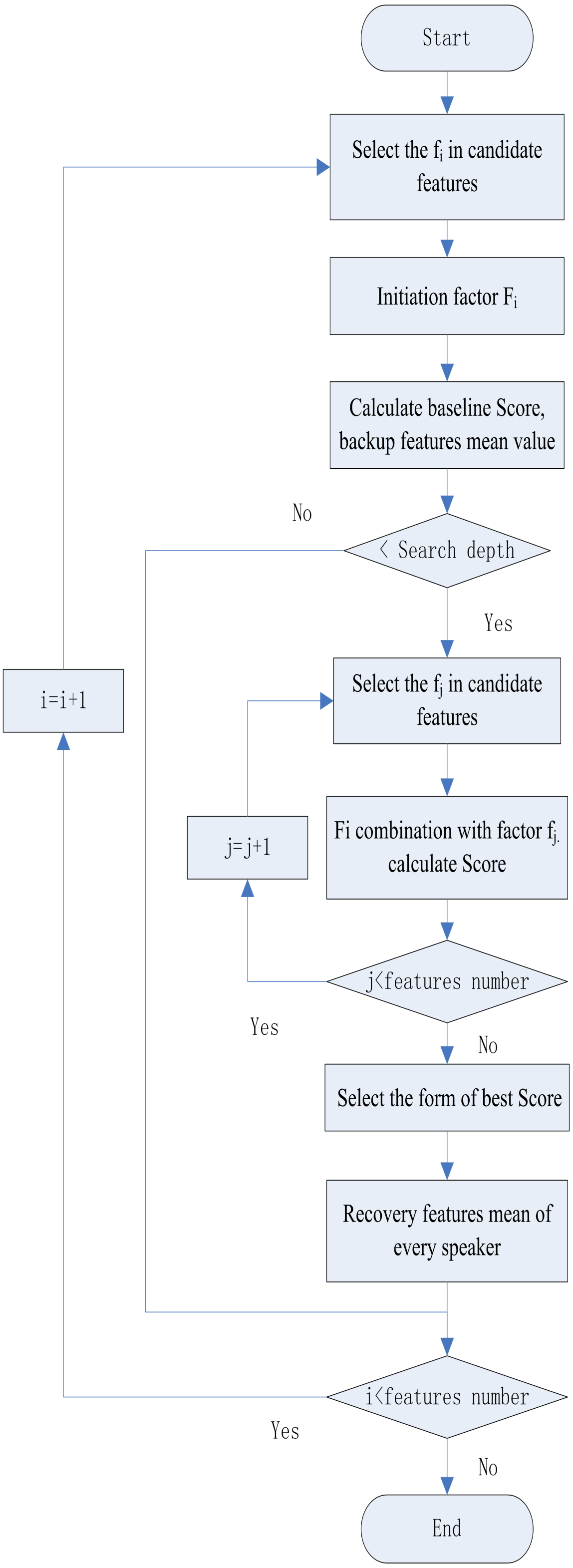

Start
Select the f_i in candidate features
Initiation factor F_i
Calculate baseline Score, backup features mean value
No
< Search depth
Yes
i=i+1
Select the f_j in candidate features
j=j+1
Fi combination with factor f_j. calculate Score
j<features number
Yes
No
Select the form of best Score
Recovery features mean of every speaker
i<features number
Yes
No
End

Figure 2. The flow chart of proposed heuristic selection algorithm.

Recall and precision are the most commonly used to evaluate speaker classification results, defined as follows:

$$\text{Recall} = \frac{\text{Number of truly detected speaker boundaries}}{\text{Number of actual speaker boundaries}} \tag{7}$$

$$\text{Precision} = \frac{\text{Number of truly detected speaker boundaries}}{\text{Number of detected speaker boundaries}} \tag{8}$$

$$\text{F-measure}=\frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall+Precision}} \tag{9}$$

The higher Recall Precision and F-measure are, the better performance is. In this paper, we take the global average of them.

5.2. **Evaluation of Heuristic Feature Selection.** In this section, we attempt to evaluate the proposed heuristic feature selection. Figure 3 provides the speaker discrimination measure of the long-term features on TIMIT, compared to the candidate features that performed best in [2] (P&LTF).

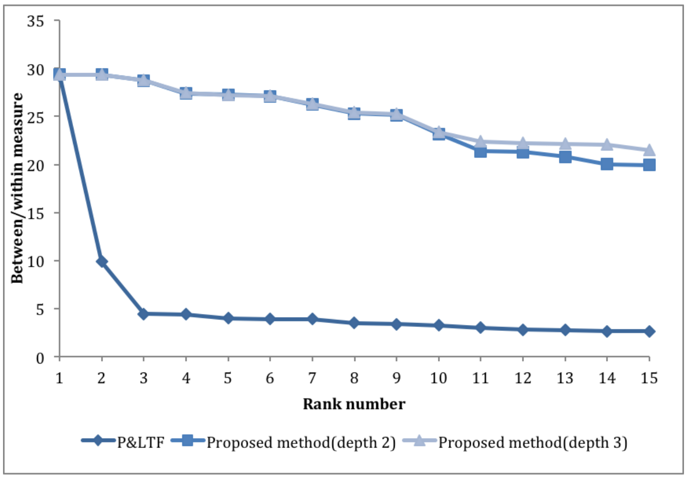


Figure 3. The comparison of the ratio of inter-speaker variance and intra-speaker variance.

Feature ranking results of P&LTF bases on the ratio of between-speaker and within-speaker variability, and the ratio is obtained from Fisher discriminant analysis. Feature ranking results of proposed method bases on heuristic strategy.

The P&LTF method selects the features with the highest score, only Pitch Median, Pitch Mean have a high score. The speaker discrimination measure values are lower after rank 3 (Long-Term Average Spectrum Mean). It can be observed that proposed approach improves the speaker discrimination measure value. When rank feature index > 3, the measure value does not reduce rapidly for the proposed method. Note that the results of depth 3 are more according to expectations.

In the case of TIMIT test dataset, these experiments show that the proposed approach that uses Algorithm 1 compared to the P&LTF approach. When the number of selected features is 5, Precision and Recall of our method is lower than P&LTF method. When the number of selected features is 10, the Precision and Recall of the two methods are similar. When the number of selected features is equal or greater than 15, our method should work better than P&LTF method. Therefore, we suggest a long-term feature subset of top-15 that is more suitable to detect a speaker change. Figure 4 illustrates the comparison of F-measure.

5.3. **Evaluation of Speaker Diarization.** (1) Supervised speaker clustering analysis. When conducting speaker segments clustering, as a result of that the length of the speaker classification audio is shorter, the samples are uneven. Therefore, some of the speaker classification error rate is very high while using the candidate feature sets. This paper

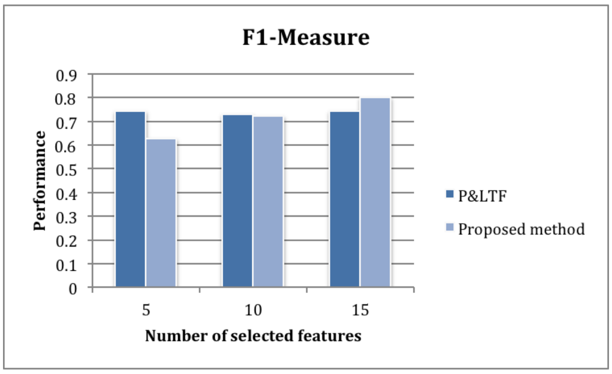
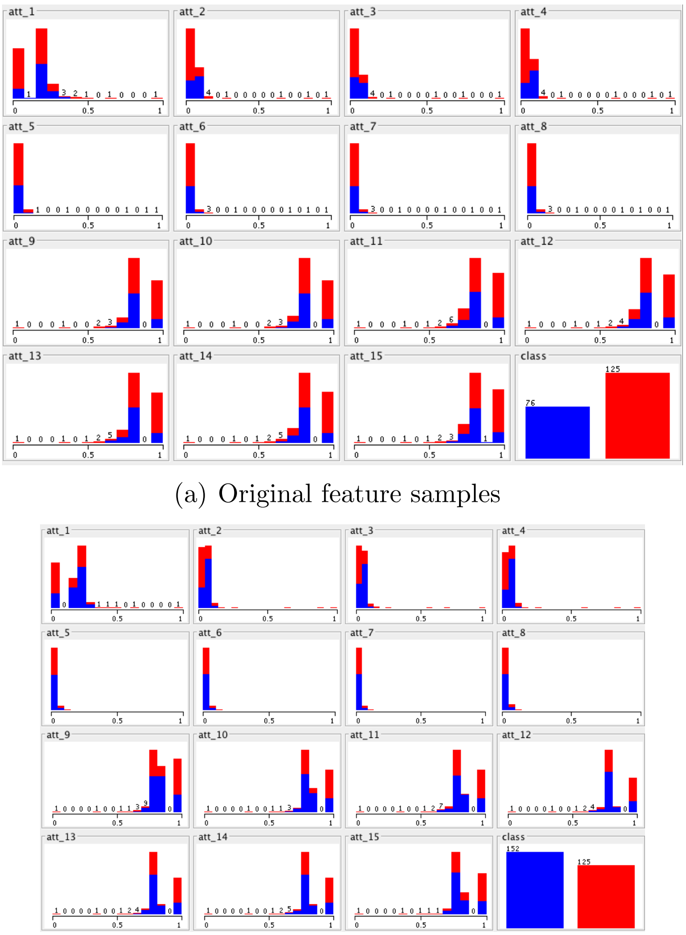


Figure 4. Comparison of F-measure on the test data set.

chooses synthetic minority over-sampling technique (SMOTE) to process the classification audio features and takes advantage of interpolation method to change the distribution of the train sample data, which improves the imbalance of sample data and the accuracy of subsequent classification.



(a) Original feature samples

(b) Feature samples using SMOTE

Figure 5. Data set preprocessing using synthetic minority over-sampling technique

When the number of candidate long-term features is 15, we analyze each speaker by one-vs-all. That is to say, we compare the top-15 long-term features of a speaker with others to verify that whether the features can distinguish each speaker or not. The

Table 1. Comparison of results on DER.

Method	AMI Meeting ID			
	ES2012a	IS1008a	TS3012a	Ave.
P&LTF[2]	26.23%	25.25%	27.08%	26.19%
Zelenak[17]	25.83%	26.92%	25.61%	26.12%
Proposed method	24.73%	24.53%	25.67%	24.98%

samples processed by SMOTE are showed in Figure 5. ROC graphs are another way besides confusion matrices to examine the performance of speaker diarizaiton. The area beneath an ROC curve can be used as a measure of accuracy. The area is improved from 0.736 to 0.820. It can be seen that the processed feature set of each speaker makes the sample set more uniform and effective, which is of important guiding significance for the supervision classifier learning and improves the performance of diarization.

(2) Unsupervised speaker diarization. We have performed the experiments with English meeting data from the AMI evaluation and using the LIUM speaker diarization system. The long-term features are combined with 12-order MFCCs along with their first derivatives. The average Diarization Error Rate (DER) is reduced by almost 4.8% relative to baseline features from 26.19% to 24.98%. Table 1 shows the information provided by the long-term features with high speaker discrimination value is quite useful in speaker diarization.

6. **Conclusions.** In this paper, we proposed a long-term feature selection using heuristic strategy for speaker diarization. It is used to generate new long-term features with high speaker discrimination. Our method has greatly improved the ratio of inter-speaker variance and intra-speaker variance of long-term features, which is compared with the measure baseline method of Fisher discriminant analysis. We have computed the feature ranking measure on TIMIT corpus. Furthermore, Experiments on the AMI corpus revealed that the improved long-term features in combination with MFCCs increase the accuracy of the LIUM speaker diarization system. The DER is reduced 4.8% relative to the baseline long-term features from 26.19% to 24.98%.

In the future work, we will investigate the non-linear characteristics of long-term features because Fisher discriminant analysis and the proposed feature selection based on heuristic strategy are the linear approach. Also, we will explore novel methods to combine estimates long-term features.

Acknowledgment. This work was sponsored by National Nature Science Foundation of China (61173106).

REFERENCES

[1] M. H. Moattar, M. M. Homayounpour, A Review on Speaker Diarization Systems and Approaches, *[J]. Speech Communication*, vol. 54, no. 10, pp. 1065-1103, 2012.

[2] G. Friedland , O. Vinyals , Y. Huang Prosodic and other Long-Term Features for Speaker Diarization, *[J]. IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 985-993, 2009.

[3] D. Imseng, G. Friedland, Tuning-Robust Initialization Methods for Speaker Diarization, *[J]. IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2028-2037, 2010.

[4] S. Meignier, T. Merlin, LIUM SpkDiarization: An Open Source Toolkit For Diarization, *[C]// In proceeding CMU SPUD Workshop. CMU*, pp. 1-7, 2010.

[5] C. L. Huang, C. Hori, H. Kashioka, Speaker Clustering using Vector Representation with Long-term Feature for Lecture Speech Recognition, *[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Press*, pp. 532-3536, 2013.

[6] S. H. Yella, H. Bourlard , Improved overlap speech diarization of meeting recordings using long-term conversational features, *[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE Press,*pp. 7746-7750, 2013.

[7] M. Zelenak, J. Hernando, Speaker Overlap Detection with Prosodic Features for Speaker Diarization,*[J]. IET Signal Processing*, vol. 6, no. 8, pp. 798-804, 2012.

[8] B. Bigot, I. Ferrane, Pinquier, Detecting Individual Role using Features Extracted From Speaker Diarization Results, *[J]. Multimedia tools and Application*, vol. 60, no. 2, pp.347-369, 2012.

[9] X. Anguera,S. Bozonnet , N. Evans, Speaker Diarization: A Review of Recent Research,*[J]. IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no.6, pp. 356-361, 2013.

[10] J. S. Garofolo , L. F. Lamel, W. M. Fisher, The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM, *National Inst. Standards Technol.,* NISTIR 4930, 1993.

[11] T. Kinnunen, H. Li, An Overview of Text-Independent Speaker Recognition: From Features to Supervectors,*[J]. Speech Communication*, vol. 52, no. 1, pp.1240, 2010.

[12] O. R. Duda, P. E. Hart, D. Stork, Pattern Classification Second Edition*[M]. Wiley*, pp. 67-83, 2001.

[13] Z. Michalewics, D. B. Fogel, How to Solve It: Modern Heuristics, *[M]. Springer-Verlag,*pp. 141-169, 2000.

[14] Praat, Doing Phonetics by Computer (version 5.3.68). http://www.praat.org/

[15] J. Carletta, S. Ashby, S. Bourban, The AMI Meeting Corpus: A Pre-announcement. Machine Learning for Multimodal Interaction, *[J]. Lecture Notes in Computer Science,* vol. 3869, pp.28-39, 2006.

[16] D. Gregory, P. Robi, Incremental Learning of Concept Drift from Streaming Imbalanced Data, *[J]. IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp.2283-2301, 2013.

[17] M. Zelenak, H. Schulz,J. Hernando, Speaker Diarization of Broadcast News in Albayzin 2010 Evaluation Campaign, *[J]. EURASIP Journal on Audio, Speech, and Music Processing*, vol. 1, pp.1-9, 2012.