A Spectrum Smoothing Method for Speaker Verification

Zhaofeng Zhang*, Jing Deng†, Longbiao Wang*, Xiong Xiao#
* Nagaoka University of Technology, Nagaoka, Japan
E-mail: {s147002@stn, wang@vos}.nagaokaut.ac.jp
†ZingTech. Co. Ltd, Beijing, China
E-mail: dengjing02@tsinghua.org.cn
Nanyang Technological University, Singapore
E-mail: xiaoxiong@ntu.edu.sg

***Abstract*—In speech processing, speech signal is usually processed frame by frame due to the non-stationary characteristic of speech. In this paper, a frequency-domain averaging based frame smoothing method is proposed. Besides the conventional frame shift, we introduce a short time shift to create several frames around current frame. Then we take the average of power spectrum for these frames. The average will be treated as a new frame instead of current frame. The new frame is considered to retain more integrated phonetic information than conventional frames. An experiment on speaker verification task showed that this method could improve the performance of speaker verification. The evaluation tasks performed on the NIST SRE 2008 database showed that our proposed method could achieve a better verification performance when compared with the conventional framing methods.**

Index Terms: speaker verification, frame smoothing, PLDA, i-vector

I. INTRODUCTION

The speaker verification techniques have been researched for many years. Appropriate and effective feature selection affects the accuracy of speaker verification. Several features such as spectral feature [1], glottal waveforms [2], phase information [3, 4, 5], and prosodic feature [6] have been researched. Typically, spectral feature are used in most tasks due to their robust performance in various environments.

For spectral feature extraction, the first thing is to frame signal to short segments with given analysis window. This framing method is devoted to capturing a stable representation of phonetic information. A phoneme usually lasts for 40 to 400 ms. In a typical analysis window method, the frame length was set as 20-30 ms, which is based on assumption that the speech signal can be assumed to be quasi-stationary within a frame. However, this assumption does not hold all the time because the actual portion of the inherent heterogeneous speech signal that can be deemed stationary depends critically on the underlying speech sound characteristics. Hence the frame shift was introduced to create the overlap between each frame. The overlap between two frames will compensate the phonetic information which is not accurately captured by one frame. The length of frame shift is usually set by half of frame length. Previous research [7] showed the performance of speaker verification could be increased with shorter frame shift. While this performance increasing was not unlimited, the performance dropped after a certain point [8]. Meanwhile, we considered that simply decreasing the frame shift to increase frame number may add redundant information to new features, and additional frames bring more computation costs. This drawback motivates us to develop an improved method in this paper.

In previous research, a phonetic information based frame selection method has been proposed [9]. This method varied the frame shift according to phonetic information, aimed to capture quasi-stationary phonetic information in each frame. As feature selection computations were necessary here, it would take additional computation cost.

Recently, a multi-taper based feature extraction method was successfully applied on speaker verification task [10]. Spectrum estimated using one window periodogram usually has a large variance [11]. MFCC feature computed from this spectrum also has high variance. In multi-taper window periodogram method, multi-taper means that the spectrum is estimated by using multiply time-doming window functions or tapers than a single one in regular [12]. Then the spectrums with different taper are used to construct a new spectrum for MFCC extraction.

In this paper, we propose a method using the current and "neighboring frames" to construct a new frame. The "neighboring frames" means frame taken near current frame with a short time shift. The new frame used the power spectrum mean of current frame and neighboring frames. Then the new frame can be treated as the same way of conventional frame without additional computational cost.

One possible reason make this method work is that new frame will achieve more stable phonetic information representation than conventional method. Meanwhile the phonetic information of frame remains valuable, because speaker-distinctive information varies according to phonetic information in speaker verification tasks [13]. Therefore a stable representation of phonetic information can achieve a better speaker recognition performance. Our method was evaluated on the NIST SRE 08 [14] dataset with the i-vector [15] PLDA [16] based back-end processing. The speaker verification achieved a better performance than regular framing method in all test conditions.

This paper is organized as follows. Section II introduces our

proposed method. Evaluation experiment is shown in Section III. At last, Section IV summarizes this paper.

II. General Instructions

The speech signal continuously changes due to articulatory movements; hence we broke the signal down to short frames to recognize the periodicity changing then to extract meaningful information from them.

In speech processing, short-term spectral feature can capture the speech phonetic information during 20-30ms, and a frame shift is fixed about half of the frame length to make sure the overlap between two frames. Even though a phoneme usually lasts longer than frame length, the analysis window of frame setting contributes to capture some quasi-stationary information of phoneme. Only with that, the spectrum of a frame can capture a specific characteristic of a phoneme or speaker. While this setting may not hold effective for all the time, because in certain cases, a quasi-stationary information can span the transition between two frames, or last longer than a typical frame length, blurring the spectral properties of the two frames, which may lead to poor discrimination in pattern recognition problems [9]. Hence a more stable representation of frame is necessary.

Conventional research indicated that simply decrease of the frame shift to get more frames can improve the recognition performance in speaker verification tasks. Some variable scale frame selecting method also can achieve a better recognition performance [13, 17]. While these methods need to use more frames for back-end processing or use some data driven tactics to choose appropriate frames, more computation cost is needed in frame estimation stage, which is less attractive for practical speaker recognition applications.

In this paper, we simply used the power spectrum average of several neighboring frames as a new frame. These "neighboring frames" means taking frames which is near current frame with a short time shift. They use the same frame length.

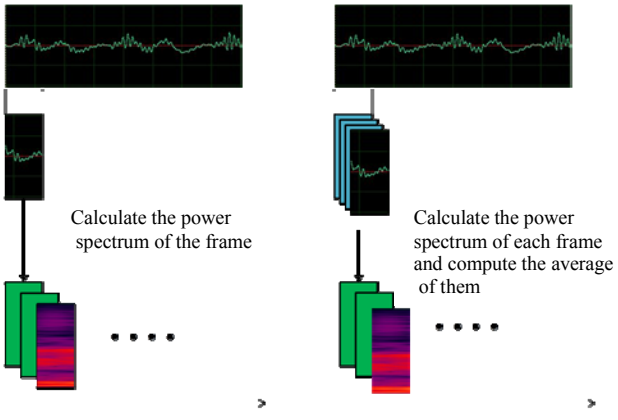

Figure 1: ***The comparison of conventional and proposed frame method. In conventional frame method (left), signal after pre-processing is broken into frame with analysis window and power spectrum is calculated. While in our proposed method (right), many frames with a short time shift is used for power spectrum computing, and new frame is the average of them.***

We consider one possible function of this method is these neighboring frames take different parts of phoneme duration. Even though some of them will take an unstable spectrum part, as the problems mentioned above, most of them will take a quasi-stationary frame. The average of these frames shall mitigate the variance occurred by unstable frame and guarantee quasi-stationary information can be captured in each new frame. Effectiveness of phoneme representation can affect the speaker characterization presentation [8]. Hence the phonetic information is valuable and can affect the performance in speaker verification tasks.

The difference between our proposed frame method and regular processing method is introduced in Fig. 1. The Fig. 2 presents the processing from speech signal input to MFCC extraction.

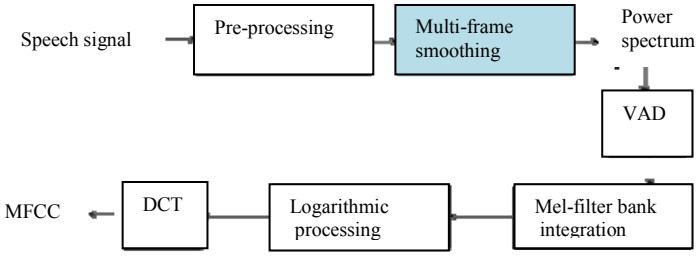

Figure 2: ***A block diagram for multi frame smoothing based MFCC extraction. The blue block is our proposed framing method. The detail is presented in Fig. 1***

In a regular speech segmenting method, the analysis window of each frame is estimated with given frame length L and frame shift K. Then the speech can be segmented to many frames. With pre-processing and FFT performing, we can achieve the power spectrum of these frames $F_1, F_2, ..., F_I$.

In our proposed method, the new frame is calculated by the mean power spectrum of current and neighboring frames. The neighboring frames are taken by a small time shift s which is shorter than K. Given s, we can calculate the power spectrum of F_i's neighboring frame. The frame length of these neighboring frames are set as same as current frame F_i for computational convenience. Totally N neighboring frames are taken. Then we can compute the power spectrum of them and achieve the frame series $f_{i1}, f_{i2}, .., f_{iN}$ given current frame F_i. The mean of the frame series and current frame are used for reconstructing new frame $\tilde{F}_i$ as shown in Eq. (1).

$$\tilde{F}_i = \frac{1}{N+1}\left(\sum_{n=1}^{N} f_{in}(t+ns) + F_i\right) \quad (1)$$

Where the $f_{in}(t+ns)$ is the power spectrum of neighboring frame of F_i at time $t+ns$, and t is the start time of frame F_i. All $F, \tilde{F}, f$ keep the same frame length of L.

The new frame $\tilde{F}_i$ can be applied for the back-end processing of speaker recognition task.

The new frame $\tilde{F}_i$ smoothed the regular frame F_i with the same frame length. The total number of frames for back-end processing (processing after power spectrum estimation) does not change a lot. The additional computation cost is only occurred at power spectrum estimation stage. Unlike some variable scale frame selecting method [13, 17] the proposed method use the fixed length and shift, hence prior processing and training for frame length estimation is not needed. These merits mean that our proposed method can be easily implemented in a speaker recognition system for practical applications.

III. EXPERIMENTS

A. Development experiment

In order to determine the frame setting of our proposed method, we design development experiments to estimate the parameters. We need to estimate two parameters: **1**. How many frames shall be selected for smoothing (frame number n in Eq. (1)) and **2**. How long the short time shift (S in Eq. (1)) shall be set. The two parameters are determined empirically and two experiments are performed jointly.

In the first experimental setting, the frame number n for smoothing is fixed at 12. We change the time shift s from 3.125 ms to 18.750 ms with a step of 3.125 ms. In second experiment, we adjust the frame number n to achieve the best performance. The frame number of n was set from 2 to 47 with a varied step. The time shift s was fixed at the best value of the first experiment. The parameters of development experiments are shown in 2nd column of Table 1 and 2.

With fixed parameter of smoothing, we can obtain the power spectrum of new frame. It can be used for the following-up processing.

An energy based voice active detection was performed. We extracted MFCC with 36 dimensions ($12+\Delta+\Delta\Delta$) as the feature for back-end phase. In addition, the cepstral mean subtraction (CMN) method was applied to remove the channel effect.

I-vector [15] and PLDA [16] based speaker verification system was performed to get the final results. In an i-vector-PLDA based speaker verification system, after we trained the universal background model (UBM), total variance matrix (T matrix), probability linear discriminant analysis model (PLDA) by training data sets. The utterance set consist of enrollment-test need to be tested whether be the same speaker or not.

The same frame estimation method is performed both on training and test data sets. The training set contains speech of NIST Speaker recognition evaluation (NIST SRE) 2004 2005 and 2006 database [14], which was used for training UBM [1], total variance matrix and PLDA model. After we train these models, development experiment was performed on NIST SRE 2008 male sets with the enrollment condition of *short2* and test condition of *short3* (*short2-short3*) [18]. Only *telephone-telephone* trial is used. The equal error rate (EER) and Minimum decision cost function (MinDCF) are used for evaluation. The result of the first experiment is shown in Table 1.

Table 1. ***The result of 1st development experiment: Find the appropriate short time shift by changing S , the S1, S2.. mean different parameter settings***

Experiment Description	Setting of short time shift *s* in 1st experiment (ms)	EER (%)	MinDCF
Regular method	0	6.80	0.0710
S1	3.125	6.41	0.0608
S2	6.250	**5.90**	**0.0565**
S3	9.375	6.18	0.0643
S4	12.500	5.95	0.0582
S5	15.625	5.95	0.0582
S6	18.750	5.95	0.0589

Table 2. ***The result of 2nd development experiment: Find the appropriate frame number by changing n , the N1, N2.. mean different parameter settings***

Experiment Description	Setting of frame number *n* in 2nd experiment	EER (%)	MinDCF
Regular method	0	6.80	0.0710
N1	2	6.18	0.0610
N2	5	6.41	0.0606
N3	11	**5.90**	**0.0565**
N4	17	6.18	0.0602
N5	23	5.94	0.0579
N6	47	7.07	0.0643

When the short time shift S is set at 6.250 ms, we achieve the minimum EER. This setting will be used for next experiment. In next experiment, we fix the frame shift to find best frame number. The result of this experiment is shown as Table 2. We obtain the best result of frame number n of 11 and frame shift S of 6.250ms.

From the development experiment, we find that the EER of each condition is not a monotonic changing along with parameter changing. Some setting even achieves a worse result than regular method. This may due to the smoothing method can capture more quasi-stationary part of each frame, while it also may decreases the frequency resolution of the analysis at the same time. How to trade off them shall be investigated in the future. In this study, we use empirical based method to determine them.

B. Evaluation experiments

With using parameters determined in develop experiment, where the short time shift s in equ. (1) is 6.250ms and frame number n is 11, we further test the proposed method on data set NIST SRE 08. The tasks of *10sec-10sec*, *short2-short3* and *short2-10sec* are tested on both male and female set.

Table 3. ***The speaker verification task on NIST 08 with conventional and proposed spectral frame estimation method.***

Test condition	Conventional Method		Proposed Method	
	EER (%)	Min DCF	EER (%)	Min DCF
	Male set			
10sec-10sec	22.41	0.2181	**21.03**	**0.2080**
short2-short3	6.80	0.0071	**5.90**	**0.0061**
short2-10sec	12.27	0.1174	**11.81**	**0.1065**
	Female set			
10sec-10sec	24.76	**0.2016**	**22.53**	0.2055
short2-short3	7.81	0.0652	**6.58**	**0.0601**
short2-10sec	14.81	0.1518	**12.13**	**0.1273**

The training set is the same as that for development experiment. Data set of NIST SRE 2004, 2005, 2006 are used to train UBM and i-vector PLDA model. We use 1024 mixtures of UBM with diagonal covariance matrix and an i-vector extractor of 400 dimensions. The UBM and PLDA model are trained separately on male and female sets. The verification system can automatically judge the gender of object speech by UBM and using corresponding speaker model to compute verification scores.

We compare our proposed method with regular spectral frame estimation method. In regular frame estimation setting, the frame shift and frame length is set by 10ms and 25ms, and no power spectrum mean is performed. The result of evaluation experiment was shown as Table 3.

Note that the extent of improvement is different for each task. The error reduction rate for female sets is better than that for male sets. For test trails of *short2-short3*, the error reduction rate is better than the other two groups (*10sec-10sec*, *short2-10sec*). It could be considered that our proposed method is more efficient in long time speaker verification tasks.

The proposed method improved speaker verification performance in all experiment settings. The reason for this improvement is not proved in theory yet. One hypothesis is that our proposed frame smoothing method could not only capture the current spectral, with an additional short time shift and spectral average, the new frame can capture more quasi-stationary part of each short time speech, it can be a phoneme, event sound and others which is helpful for speaker verification. However the frequency resolution will decrease at the same time which goes against to our goal. How to trade off them is decided by parameter setting.

IV. Conclusion and Future Works

In this paper, we proposed a simple but efficient frame smoothing method. This method uses the power spectrum average of current and neighbor frames to construct a new frame. We suppose that new frame can capture more quasi-stationary short time sound information which is helpful for speaker verification task. We evaluated our proposed method on NIST SRE 08 speaker verification task. Our method outperforms conventional method in all selected tasks. Meanwhile, additional computational cost is only occurred before spectrum computation.

However, there are still many unsolved problems in this topic. We assume that the method of conventional frame may lead to some distortion for capturing phonetic information in one frame, while our method can only mitigate this distortion, not preventing it. The parameter setting for smoothing is determined empirically. We do not quantify that how much distortion of phonetic information is prevented either. Some data driven technique can be applied to find the best parameters in future work.

V. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 15K16020 and a research grant from the Telecommunications Advancement Foundation (TAF), Japan.

References

[1] D. Reynolds, T. Quatieri and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing, vol. 10, no. 1-3, pp. 19-41, 2000.

[2] M. Plumpe, T. Quatieri and D. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification”, IEEE Transactions on Speech and Audio Processing, vol. 7, no. 5, pp. 569-586, 1999.

[3] L. Wang, S. Nakagawa and S. Ohtsuka, “High improvement of speaker identification and verification by combining MFCC and phase information”, Proc. Of IEEE ICASSP 2009, pp. 4529-4532, Apr. 2009

[4] L. Wang, K. Minami, K. Yamamoto and S. Nakagawa, “Speaker identification by combining MFCC and phase information in noisy environments”, Proc. Of IEEE ICASSP 2010, pp. 4502-4505, Mar. 2010.

[5] S. Nakagawa, L. Wang and S. Ohtsuka, “Speaker identification and verification by combining MFCC and phase information”, IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, No. 4, pp. 1085-1095, 2012.

[6] A. Adami, “Modeling prosodic differences for speaker recognition”, Speech Communication, vol. 49, no. 4, pp. 277-291, 2007.

[7] H. Gish and M. Schmidt, “Text-independent speaker identification”, IEEE Signal Process. Mag., vol. 11, no. 4, pp. 18-32, 1994.

[8] Chi-Sang Jung, Moo Young Kim and Hong-Goo Kang, “Selecting Feature Frames for Automatic Speaker Recognition Using Mutual Information”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 6, pp. 1332-1340, 2010.

[9] V. Tyagi, H. Bourlard and C. Wellekens, “On variable-scale piecewise stationary spectral analysis of speech signals for ASR”, Speech Communication, vol. 48, no. 9, pp. 1182-1191, 2006.

[10] T. Kinnunen, R. Saeidi, F. Sedlak, K. Lee, J. Sandberg, M. Hansson-Sandsten and H. Li, “Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 7, pp. 1990-2001, 2012.

[11] F. Harris, “On the use of windows for harmonic analysis with the discrete Fourier transform”, Proc. IEEE, vol. 66, no. 1, pp. 51-83, 1978.

[12] K. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation", IEEE Transactions on Signal Processing, vol. 43, no. 1, pp. 188-195, 1995.
[13] Y. Kim and J. Chung, "Pitch synchronous cepstrum for robust speaker recognition over telephone channels", Electron. Lett., vol. 40, no. 3, pp. 207, 2004.
[14] NIST Speaker Recognition Evaluation: http://www.itl.nist.gov/iad/mig/tests/spk/.
[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, 2011.
[16] Y. Jiang, K. Lee and L. Wang, "PLDA in the i-supervector space for text-independent speaker verification", EURASIP Journal on Audio, Speech, and Music Processing, 2014:29, pp. 1- 13, 2014.
[17] S. Lee, D. K. Hee, and S. K. Hyung, "Variable time-scale modification of speech using transient information," Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., IEEE, vol. 2, pp. 1319-1322, 1997.
[18] The NIST Year 2008 Speaker Recognition Evaluation Plan: http://www.itl.nist.gov/iad/mig/tests/spk/2008/sre08_evalplan_release4.pdf