An Investigation into Better Frequency Warping for Time-Varying Speaker Recognition

Linlin Wang, Xiaojun Wu, Thomas Fang Zheng* and Chenhao Zhang
Center for Speech and Language Technologies, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
E-mail: {wangll, wuxj, zhangchh}@cslt.riit.tsinghua.edu.cn
*Corresponding Author: fzheng@tsinghua.edu.cn, Tel/Fax: +86-10-62796393

Abstract— **Performance degradation has been observed in presence of time intervals in practical speaker recognition systems. Researchers usually resort to enrollment data augmentation, speaker model adaptation, and variable verification threshold to alleviate the time-varying impact. However, in this paper, efforts have been made in the feature domain and an investigation into better frequency warping for the target task has been done. Two methods to determine the discrimination sensitivity of frequency bands are explored: an energy-based F-ratio measure and a performance-driven one. Frequency warping is performed according to the discrimination sensitivity curves of the whole frequency range. Experimental results show that the proposed features outperform both MFCCs and LFCCs, and to some extent, alleviate the time-varying impact on speaker recognition.**

I. INTRODUCTION

Speaker recognition, also known as voiceprint recognition, is one kind of biometric authentication technology that can be used to automatically recognize a speaker's identity by using speaker-specific information contained in speech waves. This technology enables access control of various services by voice [1-2]. In all these typical situations, voiceprint enrollment (speaker model training) and identity verification (utterance recognition) processes are usually separated by some period of time, which poses a possible threat to speaker recognition systems.

Although pioneer researchers in speaker recognition field believed identifiable uniqueness did exist in each voice, they, at the same time, put forward the question: whether voice changed significantly with time [3]. Similar ideas were expressed in [1][4], as they argued that a big challenge to uniquely characterize a person's voice was that voice changes over time. Performance degradation has also been observed in presence of time intervals in practical speaker recognition systems [5-7].

It is a generally acknowledged phenomenon that speaker recognition performance degrades with time varying. From a machine learning point of view, more enrollment data lead to more representative models. Therefore, some researchers resorted to several training sessions over a long period of time to help cope with long-term variability of speech [8]. In [9], the best recognition performance was obtained when 5 sessions successively separated by at least 1 week were used to define the training set. In [10-11], the authors used a similar technique called data augmentation. This approach means, at a point when a positive identification of the candidate speaker is made, extra data is appended to original enrollment data to provide a more universal enrollment model for the candidate. This approach requires original data to be maintained for re-enrollment. Thus, a remedy is to use MAP adaptation to adapt from the original model to a new model considering new data at hand [10-11]. Both approaches received promising results. Other speaker-adaptation techniques, such as MLLR-based adaptation, can also be used to reduce the effects of model aging [12]. After adapting the speaker models on data from the intervening session, EER on the last two sessions is reduced to 1.7% from 2.5%. In these above approaches, obtaining the necessary data may require a long enrollment procedure. Therefore, together with its high efficiency, the shortcoming is also evident, as it is costly, user-unfriendly and sometimes may be unrealistic in real applications.

Apart from the efforts in enrollment data-model domain, there are also efforts in verification score domain. Some researchers observed that verification scores of genuine speakers decreased progressively as the time span between enrollment and verification increased, while impostor scores were less affected [13-14]. Thus a stacked classifier method of introducing an ageing-dependent decision boundary was applied, significantly improving long-term verification accuracy.

All the above approaches have yielded better results in face of the time-varying issue. However, they did not solve the central problem to such a typical pattern recognition task: the feature [15]. Speech signal includes features from many aspects of which not all are important for speaker identity discrimination. An ideal feature to the target task should "have large between-speaker variability and small within-speaker variability, …, not be affected by … long-term variations in voice." [16-18] Therefore, we aim to extract exact speaker-specific and time-insensitive (i.e. stable across time-varying sessions) information as features. Since acoustic features are closely related to attributes of frequencies of speech signals, efforts are made in frequency band level in this paper [19].

We try to identify frequency bands that reveal high

discrimination sensitivity for speaker-specific information, while low discrimination sensitivity for time-varying session-specific information. Then during the feature extraction procedure, proper frequency warping can be done to ensure that more information is extracted from target frequency bands. Determining discrimination sensitivity is essential for the proposed approach and two ways are explored in this paper. The first one is based on the F-ratio measure, which makes use of energy information of frequency bands, while the other one based on purely experimental results of recognition performance of emphasizing each frequency band respectively. By this means, the resulting frequency warping from this proposed approach could be one step closer to an ideal one for time-varying speaker recognition.

The remainder of this paper is organized as follows. The proposed approach is detailed in Section II. A brief description of the highly-controlled voiceprint database specially designed for the time-varying issue is presented in Section III [18]. Experimental results and analysis are given in Section IV. Finally in Section V, conclusions are drawn and future research directions are suggested.

II. THE FREQUENCY WARPING APPROACH

Nowadays, Mel-Frequency Cepstral Coefficients (MFCCs) are the state-of-the-art acoustic features for two totally different speech-related tasks: speaker recognition and speech recognition. The Mel scale emphasizes lower frequencies, which are generally believed to contain more linguistic information, while suppresses higher parts, which are generally believed to contain more speaker-specific information. Therefore, it is a long-term debate whether MFCC serves as a proper frequency warping method for speaker recognition. In [20], the authors compared performance of MFCCs with that of Linear Frequency Cepstral Coefficients (LFCCs) from many aspects and their results suggested that LFCC should be more often used, which places equal emphasis both on lower and higher frequencies. Finding an optimal frequency warping method is the goal in this paper and its key point is how to determine the discrimination sensitivity for each frequency band. Two approaches are proposed below.

A. Energy-based F-ratio Measure

The F-ratio measurement [21], widely used as a criterion of feature selection in pattern recognition, has been employed in [21] to process the power spectrum of utterances and to determine the speaker discriminative score in each of frequency bands. Similarly, in our target time-varying speaker recognition task, there exist two kinds of F-ratios. One is speaker-related F-ratio (denoted as F_ratio_spk), which gives the discrimination sensitivity for speaker-specific information for each frequency band just as the F-ratio used in [19]. The other one is time-varying session-related F-ratio (denoted as F_ratio_ssn), which gives the discrimination sensitivity for time-varying session-specific information for each frequency band. Thus the aim is to identify frequency bands with higher F_ratio_spk and lower F_ratio_ssn .

Suppose that the whole frequency range is divided into K frequency bands uniformly, and there are M speakers and S time-varying sessions for F-ratio calculation. Then for a given frequency band k, the two kinds of F-ratios are illustrated in the following equations:

$$\begin{cases} F\_ratio\_spk^k = \left(\prod_{s=1}^{S} \frac{\sum_{i=1}^{M} \left(\mu_{i,s}^k - \mu_s^k\right)^2}{\sum_{i=1}^{M} \frac{1}{N_{i,s}} \sum_{j=1}^{N_{i,s}} \left(x_{i,s}^{k,j} - \mu_{i,s}^k\right)^2} \right)^{\frac{1}{S}} \\ F\_ratio\_ssn^k = \left(\prod_{i=1}^{M} \frac{\sum_{s=1}^{S} \left(\mu_{i,s}^k - \mu_i^k\right)^2}{\sum_{s=1}^{S} \frac{1}{N_{i,s}} \sum_{j=1}^{N_{i,s}} \left(x_{i,s}^{k,j} - \mu_{i,s}^k\right)^2} \right)^{\frac{1}{M}} \end{cases}, \quad (1)$$

where $x_{i,s}^{k,j}$ is the energy of k in frame j of the speaker i in session s, $N_{i,s}$ is the frame number of speaker i in session s, and $\mu_{i,s}^k$, μ_s^k, μ_i^k are corresponding averages calculated by the following equations:

$$\begin{cases} \mu_{i,s}^k = \frac{1}{N_{i,s}} \sum_{j=1}^{N_{i,s}} x_{i,s}^{k,j} \\ \mu_s^k = \frac{1}{M} \sum_{i=1}^{M} \mu_{i,s}^k \\ \mu_i^k = \frac{1}{S} \sum_{s=1}^{S} \mu_{i,s}^k \end{cases}. \quad (2)$$

Therefore, the discrimination sensitivity of the given frequency band in time-varying speaker recognition task can be illustrated as in Equ. (3).

$$ds^k = \log F\_ratio\_spk^k - \log F\_ratio\_ssn^k. \quad (3)$$

B. Performance-driven Measure

The above measure makes purely use of the power spectrum to calculate F-ratio values. However, in actual speaker recognition systems, cepstra are in use. Is the discrimination sensitivity of the power spectrum consistent with that of its resulting cepstra? In [22], the authors calculated F-ratio values of the whole frequency range just as what had been done in [21], and similar trends were observed. Higher F-ratio values were obtained generally in higher frequency bands, unlike the Mel scale. However, MFCCs and LFCCs produced the best overall results, contrary to the observations that implied a-MFCCs (anti-MFCCs) would be better. So a performance-driven measure to determine the discrimination sensitivity may be a more direct one.

Suppose the whole frequency range is divided into K frequency bands uniformly, and there are S time-varying sessions for discrimination sensitivity calculation. Then a set of K x S experiments should be done, with each corresponding to emphasizing one single frequency band in frequency warping using utterances from one single time-varying session. To be specific, frequency warping for each experiment is done by doubling frequency resolution of the

target frequency band, as shown in Fig. 1 with frequency band k being the target band.

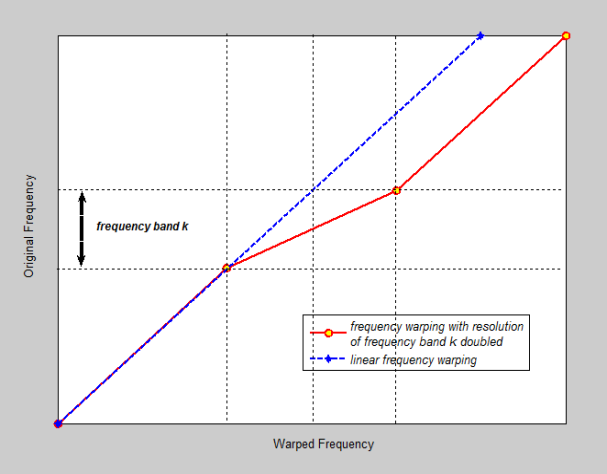

Fig. 1. The relationship between original frequency and warped frequency in two kinds of frequency warping methods

The Equal Error Rate (EER) of each experiment is denoted as eer_s^k for frequency bank k in time-varying session s, which is considered to be a probable guidance towards discrimination sensitivity of the given frequency bank. For each frequency bank k, there exists a set of S EERs, corresponding to each time-varying session. The mean and standard deviation of these EERs are denoted as μ^k and σ^k, which serve as performance indicators of speaker recognition and time varying impact, separately. For the target time-varying speaker recognition task, frequency bands with lower μ^k (more speaker-specific) and lower σ^k (less time-varying session-specific) are more preferable.

Therefore, the discrimination sensitivity of the given frequency band in time-varying speaker recognition task can be illustrated as in Equ. (4).

$$ds^k = -\log \mu^k - \log \sigma^k. \quad (4)$$

III. The Database

The time-varying voiceprint database [20] aims to examine solely the time-varying impact on speaker recognition performance. Thus recording equipments (microphone channel), software, conditions and environment are kept as constant as possible to avoid mismatches other than time-related variability, and speakers (30 female and 30 male university students) are requested to utter in a reading way with fixed prompt texts (100 Chinese sentences with varied lengths per recording session) instead of free-style conversations throughout 16 designated recording sessions in a period of approximately 3 years. Recording sessions are of gradient time intervals where initial ones are of shorter intervals and following ones of longer and longer intervals. Experiments were performed on an 8 kHz sampling rate microphone data from the first 10 sessions, and the last one was recorded nearly 1 year away from the first one.

IV. Experimental Results

A. Experimental Setup

Following experiments were based on the state-of-the-art 1024-mixture Gaussian Mixture Model – Universal Background Model (GMM-UBM) speaker recognition system. 16 dimensional MFCCs and their first derivatives were taken as acoustic features in the baseline system. The proposed frequency warping approach also used the same configuration: 16 dimensional cepstral coefficients and their first derivatives.

Data from the time-varying voiceprint database are evenly divided into two parts (balanced in gender): one for research on discrimination sensitivity and the other one for training and verification. Speaker models were enrolled by using 3 sentences randomly selected from the 2nd session with length of about 10 seconds, and sentences (each ranging from 2 to 5 seconds) from all the 10 sessions were used for verification.

The whole frequency range (from 0 Hz to 4,000 Hz) was divided into 30 frequency bands uniformly.

B. Experimental Results of Energy-based F-ratio Measure

The discrimination sensitivity plot for all 30 frequency bands are shown in Fig. 2.

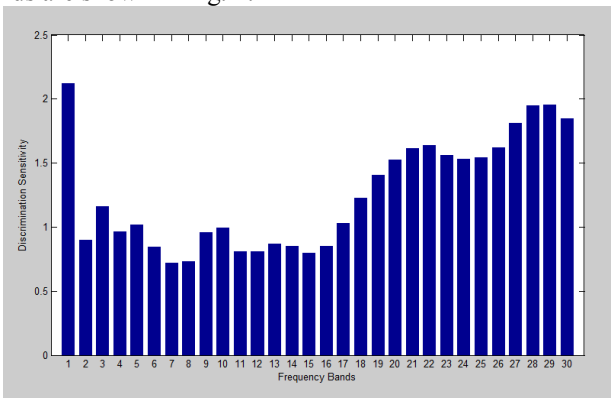

Fig. 2. The discrimination sensitivity plot obtained from the energy-based F-ratio measure

Frequency warping was done according to the above plot using the frequency resolution emphasis method depicted in Fig. 1. Experimental results of the first proposed approach are shown in Table I (Proposed Approach I).

TABLE I
A COMPARISON OF THE PERFORMANCE OF FOUR FREQUENCY WARPING METHODS

	2nd Session	All 10 sessions		*Reduction Rate (%)*	
	EER (%)	*Mean*	*Standard deviation*	*Mean*	*Standard deviation*
MFCCs	4.75	8.07	1.73	--	--
LFCCs	4.37	7.27	1.32	9.91	23.70
Proposed Approach I	4.16	6.92	1.33	14.25	23.12
Proposed Approach II	3.72	6.32	1.24	22.80	28.32

It can be seen from the above table that the proposed frequency warping approach yielded a relative reduction of 14.25% and 4.81% in average EER of all 10 sessions

compared with MFCCs and LFCCs, respectively. However, with respect to the time-varying issue, it did not outperform LFCCs.

C. Experimental Results of Performance-driven Measure

The discrimination sensitivity plot for all 30 frequency bands are shown in Fig. 3. A value of 2.5 is added to all ds^k s to ensure that they are positive numbers.

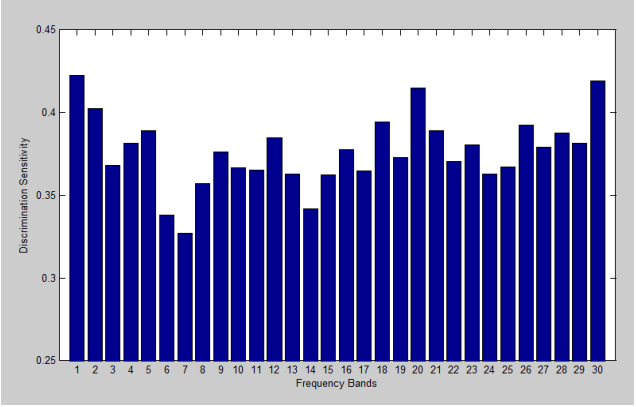

Fig. 3. The discrimination sensitivity plot obtained from the performance-driven measure

Frequency warping was also done according to the above plot using the frequency resolution emphasis method depicted in Fig. 1. Experimental results are shown in Table I (Proposed Approach II).

The frequency warping method based on the performance-driven measure outperformed LFCCs in standard deviation of all 10 sessions' EERs. Comparing the two discrimination sensitivity curves, we can see that the latter one did not emphasize that much over higher frequency bands. From a physiological point of view, higher frequency bands should be emphasized, but clearly, they should not be overemphasized. An optimal frequency warping may serve as a tradeoff between MFCCs and LFCCs.

V. Conclusions

This is an endeavor into finding better frequency warping for time-varying speaker recognition. Two approaches to determine discrimination sensitivity of frequency bands are proposed and evaluated. Their resulting frequency warping methods outperform both MFCCs and LFCCs, and, to some extent, alleviate the time-varying impact in speaker recognition.

Finding an ideal frequency warping function for target task is no easy work, and further experiments are needed to test data dependency. Also, other commonly used features in speaker recognition tasks, such as LPCCs and PLPs, and their stability over time will be studied in our future research.

References

[1] S. Furui, "Recent advances in speaker recognition," *Pattern Recognition Letters*, vol. 18, iss. 9, pp. 859-872, September 1997.

[2] H. J. Kunzel, "Current approaches to forensic speaker recognition," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 135-141, 1994.

[3] L. G. Kersta, "Voiceprint recognition," *Nature*, no. 4861, pp. 1253-1257, December 1962.

[4] J. Bonastre, F. Bimbot, L. Boe, *et al*., "Person authentication by voice: a need for caution," *Proc. of Eurospeech 2003*, pp. 33-36, Geneva, 2003.

[5] F. Soong, A. E. Rosenberg, L. R. Rabiner, *et al*., "A vector quantization approach to speaker recognition," *Proc. of ICASSP 1985*, vol. 10, pp. 387-390, Florida, 1985.

[6] T. Kato and T. Shimizu, "Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence preserving connected digit patterns," *Proc. of ICASSP 2003*, Hong Kong, 2003.

[7] M. Hebert, "Text-dependent speaker recognition," *Springer Handbook of Speech Processing*, Springer-Verlag: Berlin, 2008.

[8] F. Bimbot, J. Bonastre, C. Fredouille, et al., "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, iss. 4, pp. 430-451, 2004.

[9] J. Markel and S. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-27, No. 1, pp. 74-82, February 1979.

[10] H. Beigi, "Effects of time lapse on speaker recognition results", *Proc. of 16th International Conference on Digital Signal Processing*, pp. 1-6, 2009.

[11] H. Beigi, "Fundamentals of speaker recognition", New York Springer, 2010.

[12] L. Lamel and J. Gauvin, "Speaker verification over the telephone", *Speech Communication*, Volume 2000, Issue 31, pp. 141-154, 2000.

[13] F. Kelly and N. Harte, "Effects of long-term ageing on speaker verification", *Biometrics and ID Management*, Volume 6583 of *Lecture Notes in Computer Science*, pp. 113-124, Springer Berlin/Heidelberg, 2011.

[14] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification with long-term ageing data", *Proc. of 5th IAPR International Conference on Biometrics*, New Delhi, 2012.

[15] X. Huang, A. Acero, and H. Hon, "Spoken language processing: a guide to theory, algorithm and system development", pp. 419-426, Prentice Hall, New Jersey, 2001.

[16] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors", *Speech Communication*, Volume 2010, Issue 52, pp. 12-40, 2010.

[17] P. Rose, "Forensic speaker identification", Taylor & Francis London, 2002.

[18] L. Wang and T. F. Zheng, "Creation of time-varying voiceprint database," *Proc. of O-COCOSDA 2010*, Kathmandu, 2010.

[19] X. Lu and J. Dang, "Physiological feature extraction for text independent speaker identification using non-uniform subband processing," *Proc. of ICASSP 2007*, pp. 461-464, 2007.

[20] X. Zhou, D. Garcia-Romero, R. Duraiswami, *et al*., "Linear versus Mel frequency cepstral coefficients for speaker recognition," *Proc. of ASRU 2011*, pp. 559-564, Hawaii, 2011.

[21] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *Journal of Acoustic Society of America*, vol. 51, no. 6, pp. 2044-2056, 1972.

[22] H. Lei and E. Lopez, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," *Proc. of Interspeech 2009*, pp. 2323-2326, Brighton, 2009.