Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 79 (2016) 14–29

www.elsevier.com/locate/specom

Improving speaker verification performance against long-term speaker variability

Linlin Wang[a,b,1], Jun Wang[a,b], Lantian Li[a,b], Thomas Fang Zheng[a,b,*], Frank K. Soong[c]

[a] *Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, PR China*

[b] *Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

[c] *Microsoft Research Asia, Beijing, China*

Received 18 March 2015; received in revised form 16 February 2016; accepted 18 February 2016
Available online 3 March 2016

Abstract

Speaker verification performance degrades when input speech is tested in different sessions over a long period of time chronologically. Common ways to alleviate the long-term impact on performance degradation are enrollment data augmentation, speaker model adaptation, and adapted verification thresholds. From a point of view in features of a pattern recognition system, robust features that are speaker-specific, and invariant with time and acoustic environments are preferred to deal with this long-term variability. In this paper, with a newly created speech database, CSLT-Chronos, specially collected to reflect the long-term speaker variability, we investigate the issues in the frequency domain by emphasizing higher discrimination for speaker-specific information and lower sensitivity to time-related, session-specific information. *F*-ratio is employed as a criterion to determine the figure of merit to judge the above two sets of information, and to find a compromise between them. Inspired by the feature extraction procedure of the traditional MFCC calculation, two emphasis strategies are explored when generating modified acoustic features, the pre-filtering frequency warping and the post-filtering filter-bank outputs weighting are used for speaker verification. Experiments show that the two proposed features outperformed the traditional MFCC on CSLT-Chronos. The proposed approach is also studied by using the NIST SRE 2008 database in a state-of-the-art, i-vector based architecture. Experimental results demonstrate the advantage of proposed features over MFCC in LDA and PLDA based i-vector systems.

Keywords: Speaker verification; Long-term speaker variability; Discriminability emphasis; Frequency warping; Outputs weighting.

1. Introduction

Speaker verification is a biometric authentication technology that can automatically verify a speaker's identity with speaker-specific information embedded in speech. Similar to other pattern recognition systems, it consists of a training process (to obtain speaker models from training data) and a recognition process (to verify whether a claimed identity is correct or not). This technology enables access control of various services by voice, including: voice dialing, telephone banking, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access of computers Furui (1997). Apart from the above commercial applications, it also has applications in forensics Künzel (1994). In all applications, training and recognition processes are usually separated chronologically, which makes the short-term and long-term speaker variability an unavoidable issue in maintaining a decent performance in speaker verification.

1.1. The long-term speaker variability issue

Some pioneering researchers believed the identifiable uniqueness does exist in voice as fingerprints, but questions still remained to be answered at the same time Kersta (1962): Does the voice of an adult change significantly with time? If so, then how to alleviate or eliminate them? In 1997, Furui

[*] Corresponding author at: Room 3-411, Information Science and Technology Building, Tsinghua University, Beijing, 100084, China.

E-mail addresses: linlinwang.cn@gmail.com (L. Wang), fzheng@tsinghua.edu.cn (T.F. Zheng).

[1] Linlin Wang is currently a research associate in University of Cambridge

http://dx.doi.org/10.1016/j.specom.2016.02.004

summarized advances in automatic speaker recognition in decades and raised an open question about the way to deal with long-term variability in people's voices Furui (1997). It was conjectured whether there is any systematic long-term variation that can update speaker models to cope with gradual long-term changes. A similar question was raised in Bonastre et al. (2003), where the authors argued that a major challenge to uniquely characterize a person's voice is to harness voice change over time.

Performance degradation has been observed in separated time intervals for practical systems. Soong et al. (1985) concluded from experiments that the longer the separation between training and testing recordings, the worse the performance. Kato and Shimizu (2003) also reported a significant loss in accuracy between two sessions separated by three months and they conjectured that ageing was considered to be the cause Hébert (2008).

1.2. Overview of existing approaches

It is generally acknowledged that speaker verification performance degrades with time separation between enrollment and testing. To some extent, this speaker variability issue might be seen as part of the more general session variability issue in speaker verification, which could be typically solved nowadays by joint factor analysis (JFA) and i-vector approaches Dehak et al. (2009, 2011); Kenny et al. (2005, 2007a, 2007b, 2008). However, researchers have also proposed several specific approaches with respect to long-term speaker variability.

From a machine learning point of view, more training data leads to more representative models. Therefore, some researchers resorted to several training (enrollment) sessions over a long period of time to cope with the long-term variability in speech Bimbot et al. (2004); Soong et al. (1985). In Markel and Davis (1979), the best speaker verification performance was obtained when 5 sessions, where adjacent sessions are separated by at least 1 week apart were used to define the reference (training) set. In Beigi (2009, 2010), authors explored two adaptation techniques: data augmentation and MAP adaptation Gauvin and Lee (1994). The data augmentation approach is to augment positively identified data to the enrollment data of a speaker to retrain a more robust enrollment model for the speaker. This approach required the original data to be stored for re-enrollment. An alternative way is to use MAP adaptation to adapt the original model to a new model by considering the new data just augmented. Both approaches yield promising results. Other speaker-adaptation techniques, such as MLLR-based adaptation Leggetter and Woodland (1995), can also be used to reduce the effects of model aging. In Lamel and Gauvin (2000), after adaptation of the speaker models on data from the intervening session, the equal error rate (EER) of the last two sessions can be reduced from 2.5% to 1.7% on a French telephone corpus.

Different from the adapting the enrollment data or the speaker models, there are also studies on the verification scores. Researchers observed that verification scores of genuine speakers decrease progressively with the time separation between training and verification sessions, while impostor scores are less affected Kelly et al. (2012a, 2012b, 2013); Kelly and Harte (2011). A stacked classifier method of introducing an age-dependent decision boundary can be applied, and significant improvement against long-term variation can be obtained.

While more training data or gradually updated speaker models from extra adaptation data does yield performance improvement, however, these of approaches either require a longer speaker registration process, or need a sophisticated risk-benefit analysis to determine whether an utterance could be used to update the speaker model. Thus, together with efficiency, the shortcoming is also obvious, as it is costly, user-unfriendly and sometimes may be unrealistic for real applications. Also, simply by updating a speaker's model from the more recent data leads to little basic understanding of the aging issue. Conversely, the age-dependent score in a threshold approach makes use of the fact that verification score changes over time, which tends to be more meaningful in dealing with the long-term speaker variability.

1.3. Efforts in the feature domain

The foresaid approaches do not cover the features' role in speaker verification Huang et al. (2001). Speech signal includes many features, which are unequally distributed in their relative importance in speaker discriminability. An ideal feature should have large between-speaker variability and small within-speaker variability, not be affected by long-term variation in voice Kinnunen and Li (2010); Rose (2002); Wolf (1972). Therefore, we aim at addressing the long-term speaker variability issue in the feature domain, i.e., to extract more exact speaker-specific and time-insensitive (i.e. stable across different sessions) information. Since acoustic features are closely related to speech signal frequencies, effort is made in different frequency bands in this paper. We try to identify frequency bands that reveal higher discrimination for speaker-specific information and lower sensitivity with respect to different sessions. Thus during the feature extraction, more emphasis should be placed on those focused frequency bands. Through this kind of discriminability emphasis, the resultant features can be more robust against the long-term speaker variability for speaker verification systems.

The rest of this paper is organized as follows. In Section 2, a new speech database (CSLT-Chronos), specifically designed for investigating the long-term speaker variability issue is described in detail. Based on our observations, the proposed approach is systematically presented in Section 3. Algorithms of the two problems related to the approach are presented in Sections 4, 5. Experimental results are given in Section 6. In Section 7, conclusions are drawn and future research directions are suggested.

Table 1
A summary of major characteristics of different databases.

Database	Spkrs	Sessions	Time span	Speaking style	Recording environment	Channel conditions	Samp. rate
YOHO Campbell and Higgins (1994)	138	14	3 months	Reading	Office	Microphone	8kHz
CSLU Cole et al. (1998)	91	12	2 years	Repeating and free speech	Various locations	Telephone	8kHz
Greybeard Brandschain et al. (2010)	172	mostly 12	mostly 2–4 years	conversational	Various locations	Telephone	8kHz
MARP Lawson et al. (2009a)	32	21	34 months	Conversational	Anechoic room	Microphone	8kHz
Used in Markel and Davis (1979)	17	10	3 months	Interview	IAC sound room	Microphone	6.5kHz
Used in Beigi (2009, 2010)	22	3	5 months	Question responses	-	Various channels	8kHz
NTT-VR Matsui and Furui (1992)	36	5	10 months	reading	-	Microphone	16kHz
AWA-LTR Kuroiwa and Tsuge (in press)	6	once a week	2–10 years	Reading	Soundproof room	Microphone	16kHz
TCDSA Kelly and Harte (2011)	26	4-35	28–58 years	Broadcasts	Various locations	Various channels	8kHz
Used in Lamel and Gauvin (2000)	100	35	2 years	Reading and spontaneous	Various locations	Telephone	8kHz
CSLT-Chronos	**60**	**14**	**2 years**	**Reading**	**Lab**	**Microphone**	**8kHz**

2. CSLT-Chronos: the speech database

2.1. Known databases

A proper speech database collected chronologically to reflect the aging effects in speakers' voices is essential for this study but challenging. There are currently several speech resources, some of which are available through the LDC (Linguistic Data Consortium), such as the YOHO Speaker Verification Database Campbell and Higgins (1994), the CSLU Speaker Recognition Corpus Cole et al. (1998), the Greybeard Corpus Brandschain et al. (2010) (used in NIST SRE 2010), and the MARP corpus Lawson et al. (2009a, 2009b). However, these databases were not well designed for the research on the long-term speaker variability issue in speaker recognition. The following gives a general analysis of these known databases in four aspects: the time span, the number of recording sessions, the number of speakers, and the impact from factors other than time intervals (such as recording environments, speaking styles, and so on).

The databases used in Beigi (2009, 2010); Markel and Davis (1979), as well as the YOHO Speaker Verification Database, only have a time span of three to five months, which is not long enough. The NTT-VR database Matsui and Furui (1992) has a time span of ten months, but only contains five recording sessions. Databases with longer time span usually have a smaller number of speakers. For example, the AWA-LTR database in NII-SRC Kuroiwa and Tsuge (in press) has six speakers, while the TCDSA Database Kelly et al. (2012a, 2012b, 2013); Kelly and Harte (2011) has 26. The CSLU Speaker Recognition Corpus, the Greybeard Corpus, and the database used in Lamel and Gauvin (2000) collected speech samples through phone calls. Thus the background noise and phone channels were uncontrollable. The same problem also exists in the TCDSA Database, for its speech samples were from various sources: broadcasts, TV interviews or public speeches. Furthermore, the MARP corpus, the CSLU Speaker Recognition Corpus, the Greybeard Corpus, and the TCDSA Database are in a form of free-flowing conversations (or interviews). In this case, speech contents were not fixed, and the speakers' emotions, speaking styles, or engagement level could be easily influenced by his/her partner in the conversation or audience; all these were superfluous variability in this research targeted at the long-term variability in speaker verification. A summary of major characteristics of these databases is listed in Table 1.

2.2. The CSLT-Chronos

With the aim to examine solely the impact of long-term speaker variability on speaker verification, a speech database with a suitable size has been created, named as CSLT-Chronos, which contains 14 recording sessions within a time span of approximately two years. Since long-term speaker variability is the only focus of CSLT-Chronos, other factors such as recording equipments, software, conditions and environment are kept as constant as possible throughout all recording sessions.

Two major factors were well considered, the prompt texts and the time intervals.

Speakers were requested to utter in a reading style, predefined fixed prompted texts instead of free-style conversations. Prompt texts were designed to remain unchanged throughout all recording sessions for all speakers to avoid or to reduce the impact of speech contents on speaker verification performance. The prompt texts were made up of 100 Chinese sentences selected from The Peoples Daily using the selection algorithm proposed in Xiong et al. (2003). The length of

Table 2
Acoustic coverage of prompt texts.

Acoustic unit	Number covered in prompt texts	Total number	Percentage (%)
Initials	23	23	100
Finals	38	38	100
di-IFs	1183	1523	78

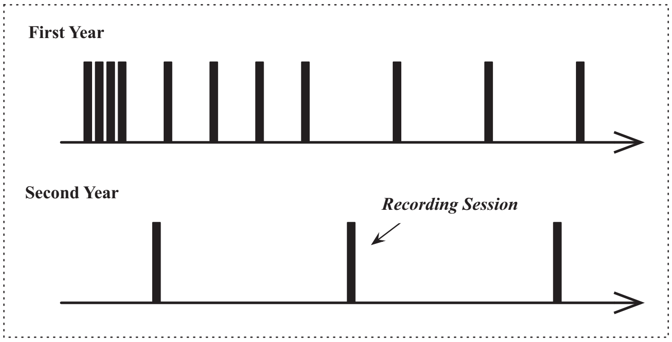

Fig. 1. An illustration of the timeline of recording sessions.

each sentence ranges from 8 to 30, with an average of 15, Chinese characters.

Chinese is a syllabic language with an Initial/Final structure where there are 21 Initials and 38 Finals Zhang et al. (2001). Pronunciations of these Initials and Finals (IF) are strongly influenced by their contexts, so the di-IF modeling (similar to that of diphones) is used in prompt texts selection Xiong et al. (2003). Due to characteristics of Chinese syllables, there are altogether 1523 di-IFs. The acoustic coverage of the designed prompt texts is listed in Table 2.

Since there exists no precedent reference of time interval design and it is costly, perhaps unnecessary, to record in a fixed-length time interval resolutions more than ten times needed, gradient time intervals were used in this database. To be concrete, the first four sessions were recorded in an interval, separated by approximately one week, the following four sessions in an interval of approximately one month, with the length of intervals increasing gradually, based on a hypothesis that the speaker verification performance degrades drastically in the beginning, and not so much when the time difference between testing and training becomes longer. The timeline of recording sessions is illustrated in Fig. 1.

Sixty university students were recruited for this project, with 30 males and 30 females, and they all speak standard mandarin Chinese fluently. An ordinary room (about 3.5 m long, 2 m wide and 2.5 m high) in the laboratory was used for recording with a table-mounted omni-directional microphone, where there was no burst noise (such as printers, phone ringing, or background speakers) with only the ambient noise at a low level (such as ventilation noise). Sampling was performed at a rate of 8kHz using a USB sound card (Creative SB X-Fi Go). In the first session, all speakers were carefully trained on how to make recordings. We did not apply calibration tones before recordings.

In short, the special design of prompt texts, time intervals, and recording setup makes the CSLT-Chronos a suitable one for studying the long-term speaker variability in speech perception, speech production, speaker verification and speaker-dependent speech recognition.

2.3. Observations on CSLT-Chronos: the long-term effect

Observations and experiments were done over the newly created database CSLT-Chronos in this study.

Since speakers were trying to familiarize with the recording procedure in the first recording session which makes the first session not be of expected quality (pauses in the recorded utterances, speaking rates change, volume fading, or style change, etc.), following experiments in this paper were based on utterances from the other 13 recording sessions (from the second session to the fourteenth session) with a time span of approximately 2 years.

A 1024-mixture Gaussian Mixture Model - Universal Background Model (GMM-UBM) speaker verification system Reynolds et al. (2000) was adopted, where 16-dimensional MFCCs and their first derivatives were used as acoustic features Xiong et al. (2006) to evaluate the long-term effect. The UBM was trained using another speech corpus of 4-h, 84-speaker microphone data recorded in the laboratory with 42 male and 42 female. The speakers uttered the sentences in a reading style and the reading materials are from newspapers and they were different for each speaker.

We consider the second session (regarded as day 0 in the experiments) as the training session and all sessions as verification sessions. That is, data from the second session were used to train speaker models. Specifically, speaker models were trained with 3 utterances randomly selected from the entire 100 utterances with a length of about 10 s from the second session, and all other utterances from all sessions were used for verification, with each utterance ranging from 2 to 5 s. Then, a list of EERs was obtained corresponding to each session, with each EER calculated after approximately 360,000 ($= 60 \times 100 \times 60$) verification trials. This list of EERs can demonstrate how the performance of the speaker verification system changes with time elapse as shown by the black line with solid dots in Fig. 2. Here, all sessions are distributed along the horizontal axis according to their time intervals from the second session (day 0) in days.

To further evaluate the long-term impact, the third session (around day 10) and the seventh session (around day 120) were also taken as the training session, respectively. Then another two lists of EERs were obtained, corresponding to the red line with hollow dots and the blue line with stars in Fig. 2. Finally, similar experiments were done with all other sessions as training sessions as well, and a surface plot of these 13 lists of EERs is shown in Fig. 3.

The general trend of these lines clearly demonstrated the long-term speaker variability effect. We confirm the assumption that the speaker verification performance degrades drastically in the beginning, and gradually flattens out as time goes on.

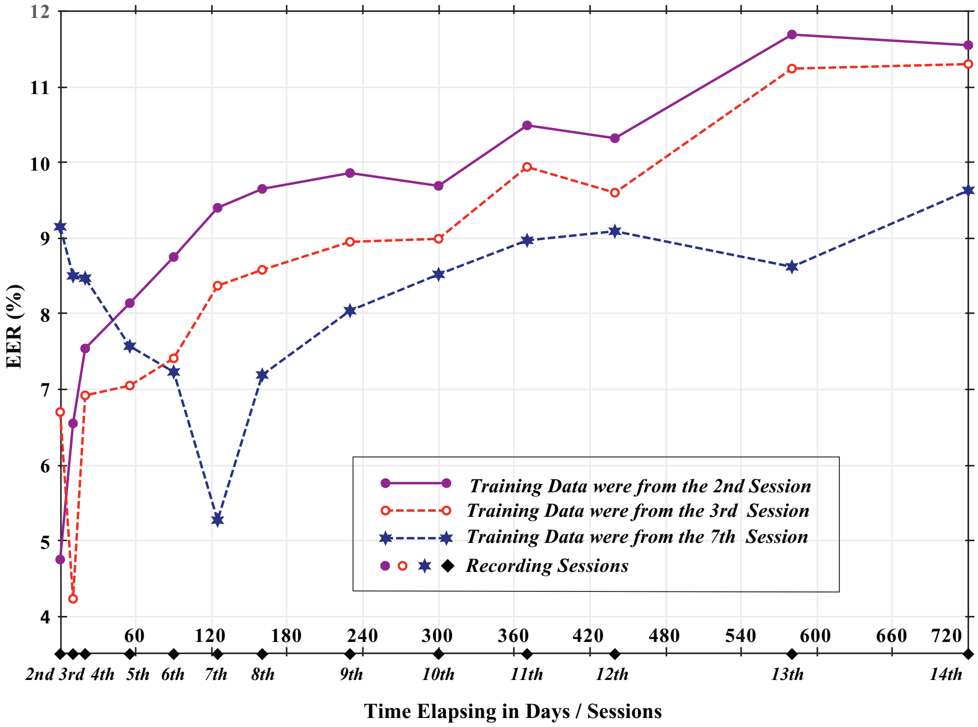

Fig. 2. Three lists of EERs corresponding to three different training sessions.

Lu and Dang have shown in their previous research that the speaker-specific information is distributed non-uniformly across different frequency bands Lu and Dang (2007, 2008), and they adopted Fisher's F-ratio to measure the importance of speaker information. Although the F-ratio criterion has theoretical limitations in situations where the classes have same means or are not unimodal as demonstrated in Campbell (1997), it was shown to give similar trends as those of mutual information measurements but easier to use in non-uniform subband filters design Lu and Dang (2007, 2008), and remains a popular approach in speaker discrimination measurement Gallardo et al. (2014a, 2014b); Hyon et al. (2012); Kinnunen (2002, 2003); Lei and Gonzalo (2009); Orman and Arslan (2001). A similar experiment was performed in our database, also with the F-ratio criterion to determine the discrimination-sensitivity for the speaker-specific information in different frequency bands, hereinafter referred to as F_ratio_spk.

In Lu and Dang (2007, 2008), the authors divided the whole frequency range (16 kHz) into 60 frequency bands uniformly. In our case, the whole frequency range (8 kHz) was also divided into 30 frequency bands uniformly, and linear-scaled triangular filter banks were also used to process the corresponding power spectrum. The output of filter banks after taking the logarithm was seen as power of corresponding frequency bands. Suppose there were I speakers and S time-spaced sessions for F-ratio calculation.

In each recording session s, for each frequency band k, an F-ratio value, denoted as $F\_ratio\_spk_{s,k}$, was obtained from Eq. (1):

$$F\_ratio\_spk_{s,k} = \frac{\sum_{i=1}^{I} \left(\mu_{i,s,k} - \mu_{s,k}\right)^2}{\sum_{i=1}^{I} \frac{1}{N_{i,s}} \sum_{j=1}^{N_{i,s}} \left(x_{i,s,j,k} - \mu_{i,s,k}\right)^2}, \tag{1}$$

where $x_{i,\ s,\ j,\ k}$ was power of the frequency band k in frame j of speaker i in session s, $N_{i,\ s}$ was the total frame number of speaker i in session s, and, $\mu_{i,\ s,\ k}$ and $\mu_{s,\ k}$ were corresponding averages calculated as follows.

$$\mu_{i,s,k} = \frac{1}{N_{i,s}} \sum_{j=1}^{N_{i,s}} x_{i,s,j,k}. \tag{2}$$

$$\mu_{s,k} = \frac{1}{I} \sum_{i=1}^{I} \mu_{i,s,k}. \tag{3}$$

An illustration of F_ratio_spk curves of five sessions separated from each other by approximately half a year (to be concrete, the second, eighth, eleventh, thirteenth, and fourteenth sessions) is shown in Fig. 4.

It can be seen that the lower frequency bands (below 0.3 kHz) and higher ones (above 2.5 kHz) exhibit more speaker discriminative power than middle ones, which was similar with findings on other databases in previous literature Auckenthaler and Mason (1997); Besacier and Bonastre (1997); Kinnunen (2003); Lei and Gonzalo (2009); Lu and Dang (2007, 2008), in spite of different languages.

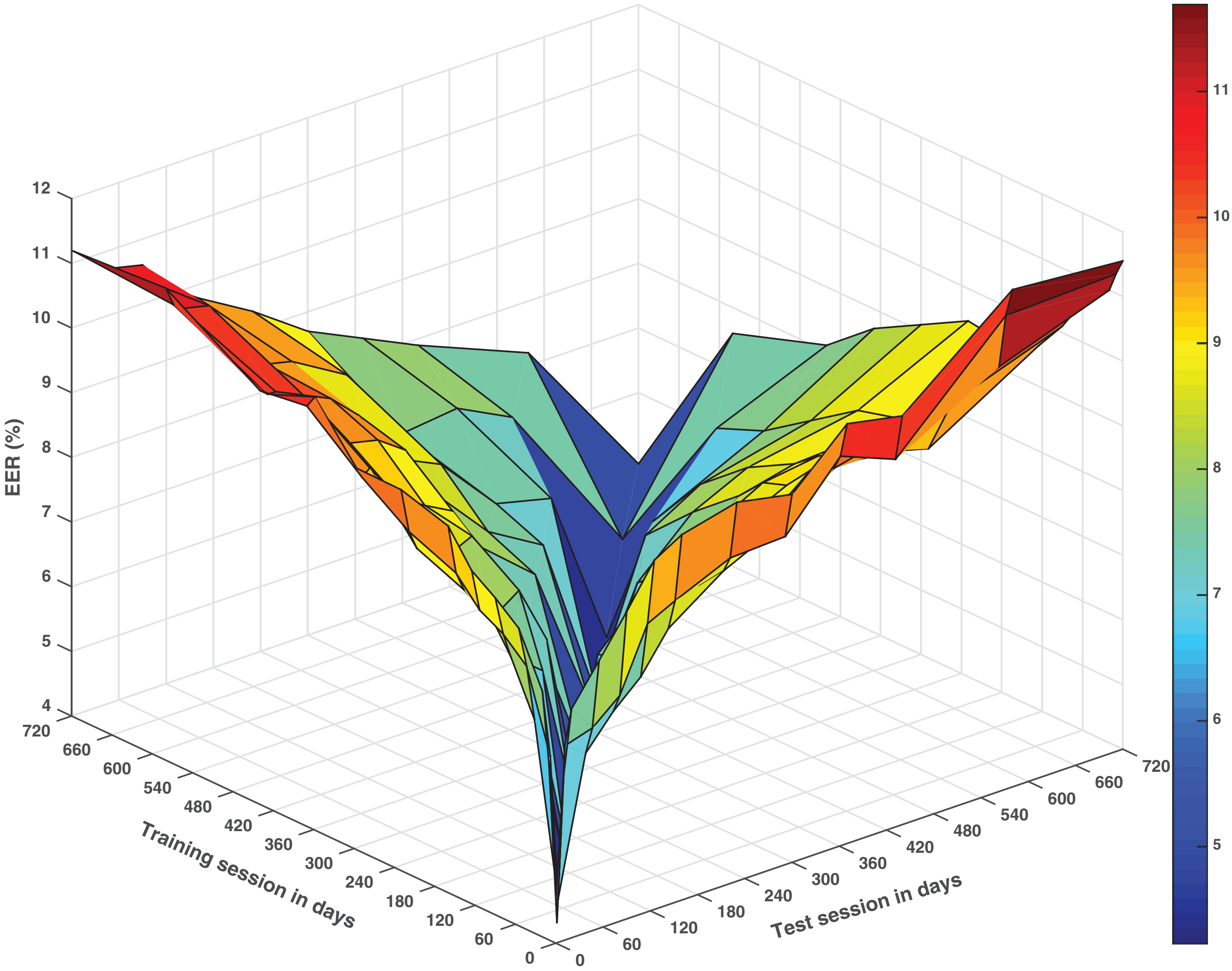

Fig. 3. The long-term speaker variability effects on performance of speaker verification systems.

The basic trend of each F_ratio_spk curve along frequency bands is more or less the same among the five recording sessions, while the values change to varying degrees, with larger variation especially in the higher frequency. This variability in discrimination-sensitivity can be viewed as an indicator of existence of the long-term effect of the speaker-specific information, as all other factors were kept as constant as possible during recording. Therefore, a possible assumption is that the long-term variable part of the speaker-specific information is also distributed non-uniformly among frequency bands, which brings about the possibility to make efforts in the frequency band level to extract more speaker-specific and time-invariant information as acoustic features.

3. The discriminability emphasis method

3.1. Focusing on critical frequency bands

The vocal cord (source) and the vocal tract (filter) are two important components in speech production, and properties of them, such as length, elasticity, and shape, are different from one speaker to another. Thus, it is generally believed that they contribute to the speaker-specific information for speaker recognition.

Lu and Dang (2007, 2008) have investigated the relationships between the frequency components and the vocal tract based on speech production, and found that speaker information is not uniformly encoded in different frequency bands. It has also been confirmed by us in prior section on our database. Statistical methods were employed to quantify the dependencies between frequency components and speaker identities to improve the speaker recognition performance. The idea of analyzing the contribution of different frequency bands will shed light on the long-term variability issue in speaker verification.

Furthermore, It is known that the fundamental frequency reflects characteristics of vocal cords, while “formants” or spectral envelope reflect those of vocal tracts. Studies in physiology have shown that the fundamental frequency of both males and females decreases with age after adulthood Rhodes (2011); Stathopoulos et al. (2011). Reubold *et al.* launched a long-term study of the possible changes in adult speech,

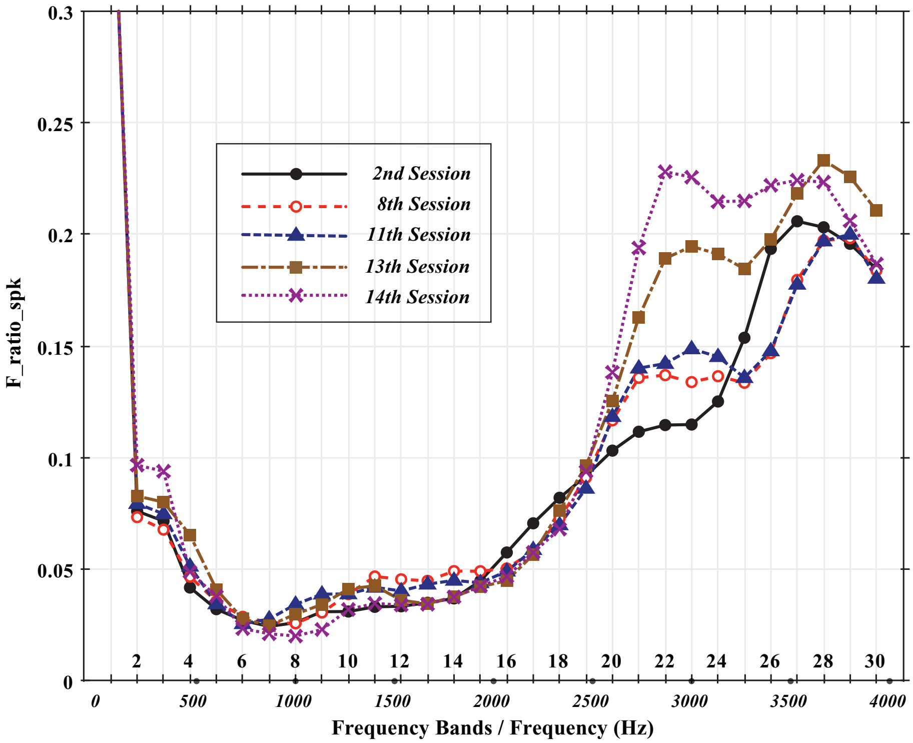

Fig. 4. *F_ratio_spk* curves of the second, eighth, eleventh, thirteenth, and fourteenth recording sessions. The curves all peak at frequency band 1, with values of 0.38, 0.41, 0.42, 0.49, and 0.50, respectively, which are not shown in the figure to allow higher resolution in other frequency bands.

and found that the changes of the first formant during this long-term observation are roughly the same as those of the fundamental frequency Reubold et al. (2010).

Thus, broadly speaking, the slowly declining trends in both the fundamental frequency and the first formant, can be seen as a change of the distribution of the speaker-specific information among different frequency bands of speech signals as time goes on. In other words, in view of the long-term speaker variability issue, the distribution of this speaker-specific information among frequency bands, can be split into two parts: the more invariable and less invariable ones over time, as indicated in Fig. 4.

In speaker verification, frequency bands with more speaker-specific information should be emphasized for feature, while in considering of the long-term speaker variability, frequency bands with less time-related information should be emphasized, which is the starting point of the proposed discriminability emphasis method.

3.2. The discriminability emphasis method

The discriminability emphasis proposed is to place different weights on frequency bands according to their discrimination capability of time-invariant acoustic features for speaker verification.

Obviously, the proposed method should deal with two core issues: how to determine the discrimination-sensitivity in frequency bands for the target task, and how to place different emphasis among frequency bands. The two issues are further investigated in the following two sections.

4. Discrimination-sensitivity determination

In this section, a strategy based on the *F*-ratio criterion is proposed to determine the discrimination-sensitivity in different frequency bands. Then the overall discrimination-sensitivity is compromised to find an constrained optimal balance.

4.1. F-ratio as an intermediary criterion

For discrimination in machine learning, *F*-ratio has broadly served as a criterion of feature selection Wolf (1972), as shown in Eq. (4),

$$F\_ratio_{\text{grouping}} = \frac{\text{between-group variability}}{\text{within-group variability}}. \tag{4}$$

A higher *F*-ratio value indicates more appropriate feature selection for the target grouping. That is to say, the selected feature with a higher *F*-ratio possesses higher discriminability against the target grouping. As mentioned before, many researchers have made use of *F*-ratio to determine the importance of information in different frequency bands for speaker identification. Similarly, it is also adopted here to quantify the importance of frequency bands for speaker verification in terms of long-term speaker variability and our approach differs from theirs in way of grouping.

In our case, each frequency band makes feature selection, and the speaker verification task across time-separated sessions makes the target grouping. There exist two different groupings: by speaker for each time-separated session and by session for each speaker. The first kind of grouping is just the one usually employed in the traditional speaker verification task, as it is covered in Section 2, while the second kind of grouping is the special one in terms of long-term speaker variability.

As discussed above, more emphasis should be placed on frequency bands that reveal higher discrimination for the speaker-specific information (a higher F-ratio value when grouping by speaker, denoted as F_ratio_spk), and lower sensitivity for the time-separated session-specific information (hereinafter referred to as F_ratio_ssn). Therefore, the final overall discrimination-sensitivity of each frequency band should have a positive correlation with its F_ratio_spk, while negative with its F_ratio_ssn.

Since the energy is the most important attribute of a frequency band that is closely related to the resulting cepstra, the power spectrum is used as the distance measure in F-ratio calculation. The equations of this calculation and the determination of the final overall discrimination-sensitivity are given in detail below.

4.2. F-ratio based discrimination-sensitivity score definition

Assume that the whole frequency range is uniformly divided into K frequency bands, and there are I speakers and S time-separated sessions for the corresponding F-ratio calculation, with the same configuration as in Section 2.

For each frequency bank k, the averaged $F\_ratio\_spk_k$:

$$F\_ratio\_spk_k = \left(\prod_{s=1}^{S} F\_ratio\_spk_{s,k} \right)^{\frac{1}{S}}. \tag{5}$$

Similarly, the second kind of F-ratio, denoted as F_ratio_ssn, is illustrated by Eq. (6):

$$F\_ratio\_ssn_{i,k} = \frac{\sum_{s=1}^{S} \left(\mu_{i,s,k} - \mu_{i,k}\right)^2}{\sum_{s=1}^{S} \frac{1}{N_{i,s}} \sum_{j=1}^{N_{i,s}} \left(x_{i,s,j,k} - \mu_{i,s,k}\right)^2}, \tag{6}$$

where $F\_ratio\_ssn_{i,k}$ denotes the corresponding F-ratio value of frequency band k of speaker i, and $\mu_{i,\,k}$ is the average calculated as follows.

$$\mu_{i,k} = \frac{1}{S} \sum_{s=1}^{S} \mu_{i,s,k}. \tag{7}$$

For each frequency band k, the averaged $F\_ratio\_ssn_k$:

$$F\_ratio\_ssn_k = \left(\prod_{i=1}^{I} F\_ratio\_ssn_{i,k} \right)^{\frac{1}{I}}. \tag{8}$$

In this way, two F-ratio values are obtained for each frequency band.

4.3. Determining the overall discrimination-sensitivity

Frequency bands with higher $F\_ratio\_spk_k$ reveal higher discrimination for the speaker-specific information, while frequency bands with lower $F\_ratio\_ssn_k$ reveal lower sensitivity for the time-separated session-specific information. Frequency bands with high values should have higher overall discrimination-sensitivity in speaker verification. Thus for each frequency band k, an overall discrimination-sensitivity score $Discrim\_score_k$ can be defined by Eq. (9).

$$Discrim\_score_k = \log\left(\frac{F\_ratio\_spk_k}{F\_ratio\_ssn_k}\right). \tag{9}$$

Actually, the overall discrimination sensitivity score without the logarithmic operation has also been compared in experiments, but did not function so well as the one with the logarithmic operation shown in Eq. (9). A comparison of them in the verification performance is illustrated in Section 6.

As a result more emphasis should be placed on frequency bands with higher $Discrim\_score_k$.

In this section, F-ratio has been employed as a criterion to determine the discrimination-sensitivity of different frequency bands in our data-driven approach. Although it is theoretically based on a single Gaussian distribution assumption, this simple approach will be shown later experimentally that it is effective for frequency band selection.

5. Discriminability emphasis during feature extraction

Nowadays, cepstral coefficients are still widely used as acoustic features in speaker verification applications, and among them, MFCC is still the dominant one. To extract cepstral coefficients, different emphasis among frequency bands can be implemented: pre-filtering frequency warping and post-filtering weighting. By performing frequency warping, the filter bank resolution can be changed according to the overall discrimination-sensitivity score. Higher resolution means more information can be extracted from those corresponding frequency bands. Furthermore, outputs weighting of filter banks is also a straightforward method to increase the proportion of effects from those corresponding frequency bands in generating the final acoustic features. The two methods are presented in more details below.

5.1. Frequency warping

The Mel scale, taking into account human auditory characteristics, is a frequency warping used extensively in speech applications. It employs higher resolutions in lower frequencies, while lower resolutions in higher frequencies. Thus more detailed information is extracted in low frequencies. With this method, lower frequency bands are emphasized, which are generally believed to contain more information, and higher frequency bands are suppressed, which are generally believed to contain more speaker-specific information. Thus it is interesting to argue whether MFCC serves as a proper frequency warping method for speaker verification and the authors in Zhou et al. (2011) suggest that LFCC (Linear Frequency Cepstral Coefficients) should be used for speaker recognition tasks.

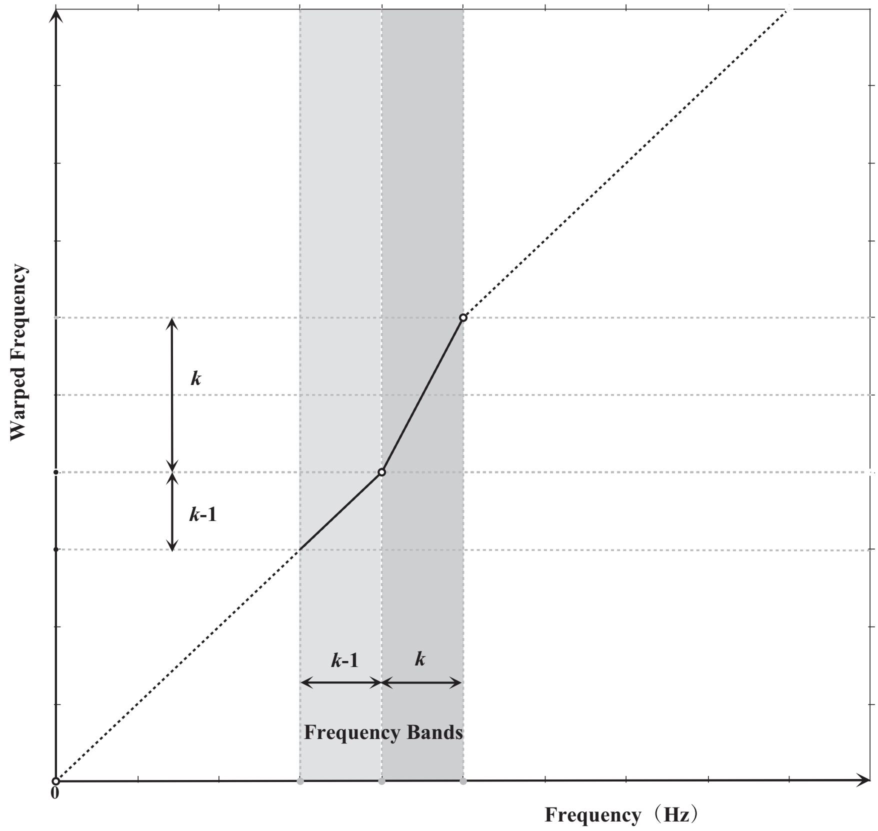

Fig. 5. An illustration of the frequency bands k and $k-1$ in two frequency domains.

Therefore, in the target task, the frequency resolution of each frequency band is determined according to its overall discrimination-sensitivity score as calculated in Section 4. For example, suppose that $Discrim\_score_k$ is twice of $Discrim\_score_{k-1}$, their corresponding frequency bands in the warped frequency domain are shown in Fig. 5 with the frequency resolution of frequency band k twice of that of $k-1$.

In this way, a relationship between the two frequency domains is established, and the warping algorithm is described in Algorithm 1.

5.2. Filter-bank outputs weighting

Given the log-energy spectrum after the triangular filtering, denoted as $S(k)$, we can define the weighted energy spectrum $Weighted\_S(k)$ as in Eq. (10),

$$Weighted\_S(k) = Discrim\_score_{k+1} \cdot S(k). \tag{10}$$

The equation of the following DCT goes as Eq. 11.

$$Cepstrum(n) = \sum_{k=0}^{K-1} Weighted_S(k) \cdot \cos\left(\frac{\pi n(k+0.5)}{K}\right)$$
$$= \sum_{k=0}^{K-1} Discrim\_score_{k+1} \cdot S(k) \cdot \cos\left(\frac{\pi n(k+0.5)}{K}\right). \tag{11}$$

Algorithm 1 Design of a frequency warping table according to discrimination scores of frequency bands to maximize speaker recognition.

```
**Inputs:** original frequency before warping $orig\_freq$, discrimination score vector of frequency bands $discrim\_score$
**Output:** warped frequency $warped\_freq$
$num \leftarrow$ size of $discrim\_score$
$sum \leftarrow$ sum of $discrim\_score$
$freq\_range \leftarrow$ frequency range of all frequency bands
$freq\_start \leftarrow$ starting frequency of those frequency bands
$width \leftarrow freq\_range/num$
$index \leftarrow (orig\_freq - freq\_start)/width$
$rest \leftarrow (orig\_freq - freq\_start) \bmod width$
$acc \leftarrow 0$
**for** $k = 0$ to $index - 1$ **do**
    $acc \leftarrow acc + discrim\_score(k)$
**end for**
$rest \leftarrow rest/width * discrim\_score(index)$
$warped\_freq \leftarrow (acc + rest) * freq\_range/sum$
**return** $warped\_freq$
```

In the resulting cepstra, the portion of the effect of each frequency band is emphasized according to its overall discrimination-sensitivity score.

6. Experiments and results

6.1. Experimental setup

Since the proposed approach is feature-based, a GMM-UBM system described in Section 2 was adopted as the experimental system to verify its effectiveness, with the benefit of faster computation and fewer hyperparameters over other sophisticated techniques like JFA Kelly et al. (2012a). The baseline MFCC features, LFCC features Zhou et al. (2011) and new features generated from the proposed approaches shared the same configuration: 16-dimensional cepstral coefficients and their first derivatives.

For each recording session in CSLT-Chronos, we divided the sentence set into two equal subsets: one for the overall discrimination-sensitivity determination (development data set) and the other one for training and verification. Also, for the latter part, 3 utterances from the second session were chosen for speaker model training and other utterances from all 13 sessions were used for verification. Experimental details are shown in the following subsections.

6.2. The overall discrimination sensitivity of frequency bands

The whole frequency range was divided into 30 frequency bands uniformly as described in Section 2.

The $F\_ratio\_spk_k$ and $F\_ratio\_ssn_k$ values calculated through Eq. (5) and Eq. (8) for each frequency band are plotted in Fig. 6, respectively.

The curve of F_ratio_spk values goes smoothly from 500Hz to 2000Hz (frequency bands 5–16). After that, the curve climbs up and reaches two local peaks around 2800Hz and 3700Hz (frequency bands 21 and 28). However, the curve of F_ratio_ssn values goes up after 700Hz (frequency band 6) with an almost consistent positive slope. There exist two local peaks, which are located around 1400Hz and 3400Hz (frequency bands 11 and 26).

The overall discrimination-sensitivity score $Discrim_score_k$ of each frequency band is calculated by Eq. (9), and shown in Fig. 7.

The curve in Fig. 7 is a compromise of the two curves in Fig. 6. For example, higher frequency bands should be emphasized, but not that much as in the F_ratio_spk curve, because higher frequency bands also have worse (i.e., higher sensitivity) F_ratio_ssn values. A similar situation also exists for lower frequency bands.

Omitting the logarithmic operation in Eq. (9), we could obtain another series of the overall discrimination-sensitivity scores, the trend of which is just the same as shown in Fig. 7.

6.3. Experimental results

Experiments on the proposed discriminability emphasis method were done based on the *Discrim_score* curve in Fig. 7. Acoustic features were extracted in two ways: frequency warping and filter-bank output weighting, were denoted as Warping Features and Weighting Features. A comparison of the acoustic features (MFCC features, LFCC features, Warping Features, and Weighting Features) for each recording session is shown in Fig. 8, and EER(%) values are specified in Table 3.

Table 3
A comparison of the four kinds of features in EER of each recording session.

Recording	MFCC	LFCC	Proposed approach	
session			Warping	Weighting
2nd	4.5	4.7	**4.0**	4.1
3rd	6.4	6.6	**6.1**	6.5
4th	7.4	7.3	**6.6**	7.2
5th	8.1	**7.2**	7.5	7.8
6th	8.7	**7.2**	**7.2**	7.8
7th	8.9	8.6	**8.4**	9.0
8th	9.3	8.7	**8.5**	9.1
9th	9.9	8.4	**7.9**	8.7
10th	9.6	8.5	**8.1**	8.5
11th	10.0	9.3	**8.9**	10.0
12th	9.7	9.8	**8.8**	9.7
13th	11.1	11.0	**9.7**	10.1
14th	11.0	11.2	**9.4**	9.7

Table 4
A comparison of different features in the mean and standard deviation of EERs across recording sessions.

Features	Performance		Relative reduction	
	Mean	StDev	Mean	StDev
MFCC (baseline)	9.18	1.38	–	–
LFCC	8.65	1.48	5.77	−7.24
Warping	**8.09**	**1.09**	**11.87**	**21.01**
Weighting	8.68	1.15	5.45	16.67
Warping_NoLog	8.87	1.22	3.38	11.59

As shown in Fig. 8, the MFCC and the LFCC features achieved comparable performance in the first three sessions and in the last three sessions, however in other sessions the LFCC features greatly outperformed the MFCC features. The proposed Weighting Features gave the overall performance with higher recognition accuracy than the MFCC features in most recording sessions, but did not show superiority in error rates over the LFCC features. The proposed Warping Features consistently outperformed the baseline MFCC features and the proposed Weighting Features for every recording session and also outperformed the LFCC features in most recording sessions, with the fifth session as an exception, where the LFCC features gave a slightly lower error rate.

An empirical study of whether or not to take the logarithm in Eq. (9) when calculating the overall discrimination-sensitivity was also conducted. We take the frequency warping approach as an example. By omitting the logarithmic operation in Eq. (9), another kind of acoustic features can be obtained, denoted as Warping_NoLog Features. The overall performance between Warping Features and Warping_NoLog Features is also compared in Table 4. The Warping Features yielded higher recognition accuracy than the Warping_NoLog Features, which outperformed the MFCC features. Therefore,

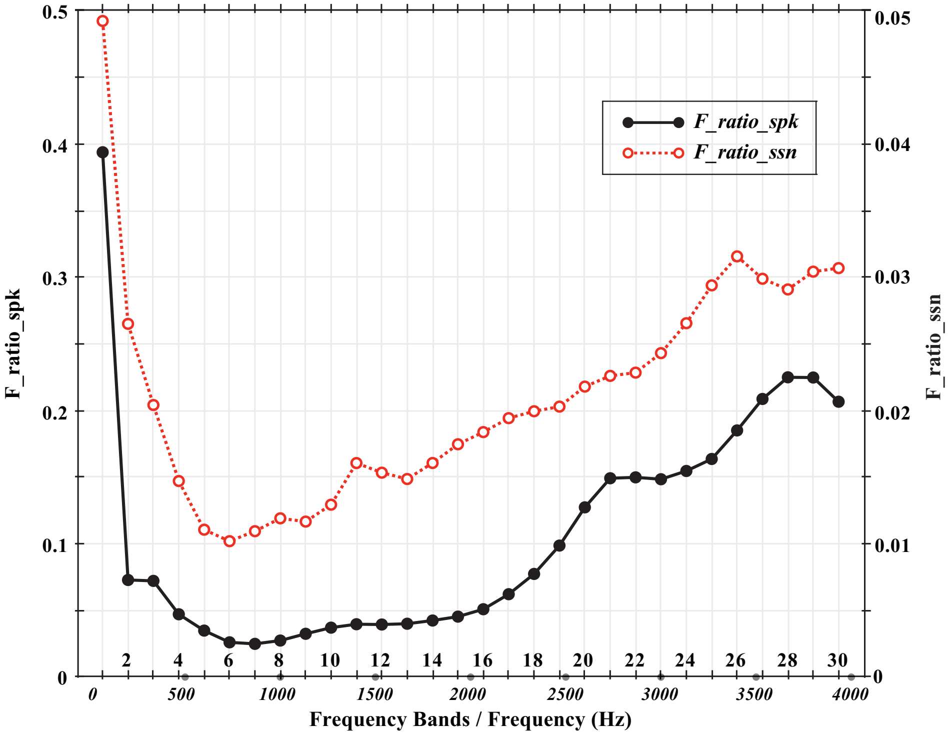

Fig. 6. Curves of *F_ratio_spk* and *F_ratio_ssn* values for each frequency bands, respectively.

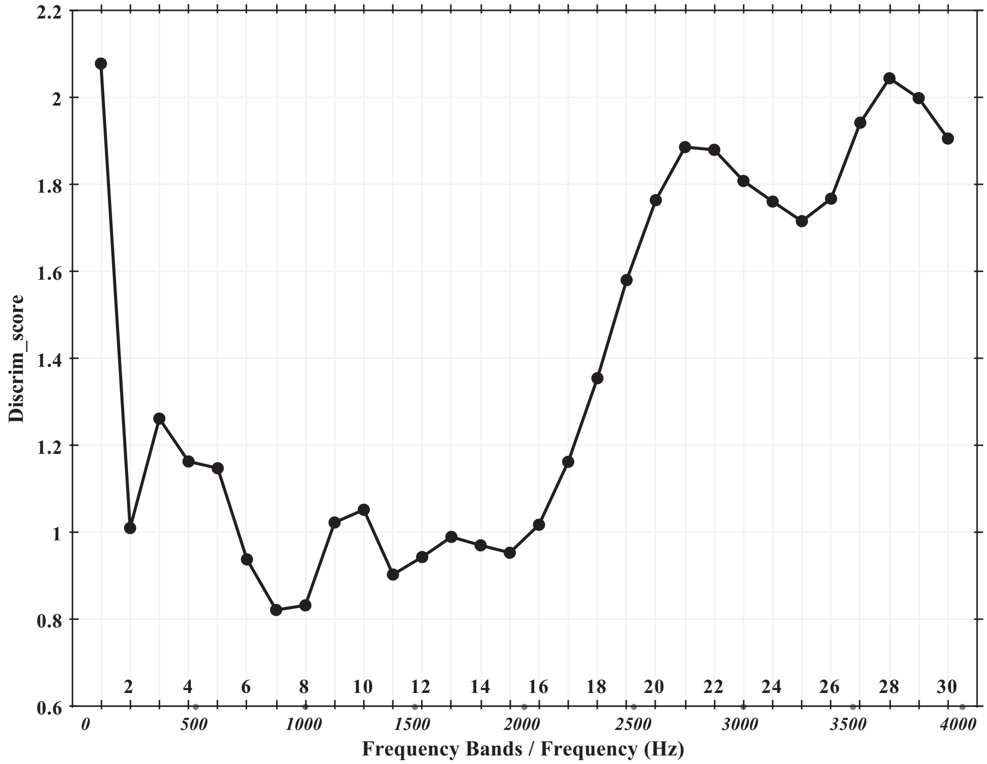

Fig. 7. *Discrim_score* for each frequency band.

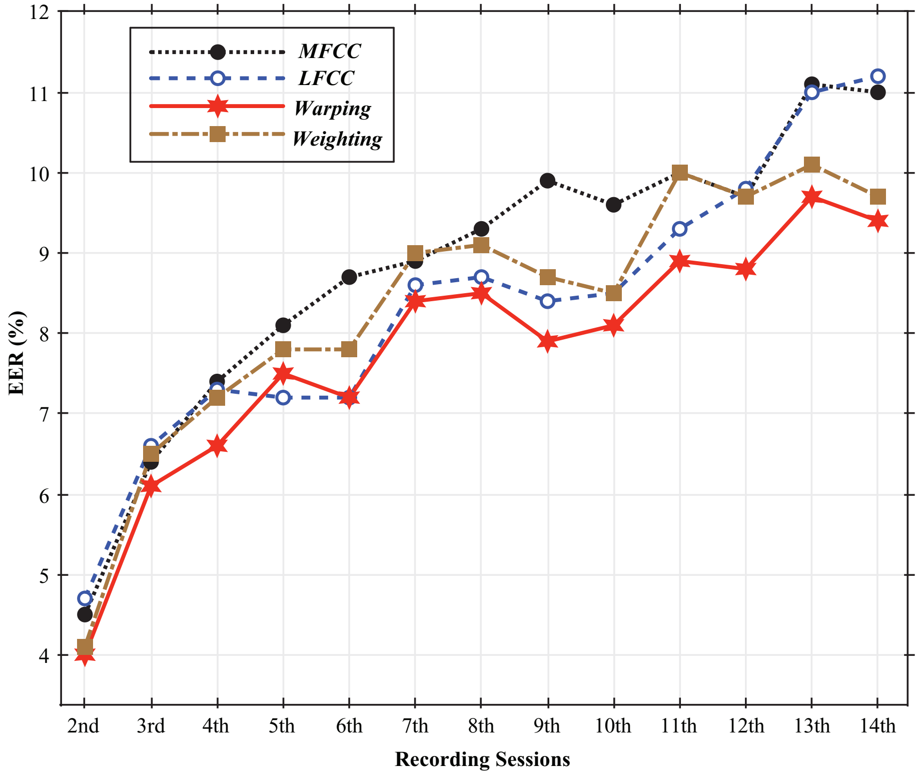

Fig. 8. A comparison of the four kinds of features in EER of each recording session.

in the proposed method, the overall discrimination-sensitivity score was calculated with the logarithm.

6.4. Robustness regarding long-term speaker variability

EER is commonly used as a one number measure in evaluating the overall performance of a speaker verification system, which is a cross-over point where values, the false acceptance rate (FAR) and the false rejection rate (FRR), are equal.

For the target speaker verification task across time-separated sessions, a series of EERs can generally be obtained, as listed in Table 3. When comparing the performance of two acoustic features, we compared two arrays of EERs. The acoustic features with an array of consistently lower EERs or more slowly changing EERs are preferred, corresponding to both aspects of the target task: more speaker-specific and more time-insensitive, respectively, as mentioned before. Thus, it is natural to use statistics of the array of EERs to evaluate the overall performance of acoustic features, such as the mean and standard deviation.

The mean of the array of EERs indicates the averaged performance of speaker verification for all recording sessions, while the standard deviation serves as an indicator of robustness across time-separated sessions. Then Table 4 shows another comparison of the four kinds of acoustic features in the mean and the standard deviation of EERs across sessions. Since training data were from the second session, only the remaining 12 sessions (the third to fourteenth sessions) were considered in this section to avoid possible bias.

Table 5
p-values of the Student's t-test for the "null hypothesis" that the two proposed kinds of features perform similarly as the MFCC features.

Features pairs	*p*-Values
MFCC-Warping	1.62×10^{-5}
MFCC-Weighting	9.35×10^{-3}

From the statistics in Table 4, especially the relative reduction of the standard deviation of EERs, it can be concluded that the two proposed acoustic features both yielded higher recognition accuracy than the MFCC features for the target issue in speaker verification. The LFCC features achieved lower error rate on average recognition performance, but not that robust as the MFCC features. Also, in this *F*-ratio based discrimination-sensitivity scenario, the Warping Features gave higher recognition accuracy and more robust performance than the Weighting Features.

In order to further examine the statistical significance of the experimental results as shown in Table 3, tests of normality were first performed on the three lists of EERs corresponding to MFCC, Warping and Weighting features, and they all passed the Jarque–Bera test with a significance level of 0.01 Jarque and Bera (1987). Then *p*-values were calculated through the paired Students *t*-test for the "null hypothesis" Devore (1995) that the two proposed kinds of features perform similarly as the MFCC features, respectively, and shown in Table 5. (The null hypothesis is rejected if *p*-values are smaller than a certain significance level, traditionally 0.05 or 0.01.)

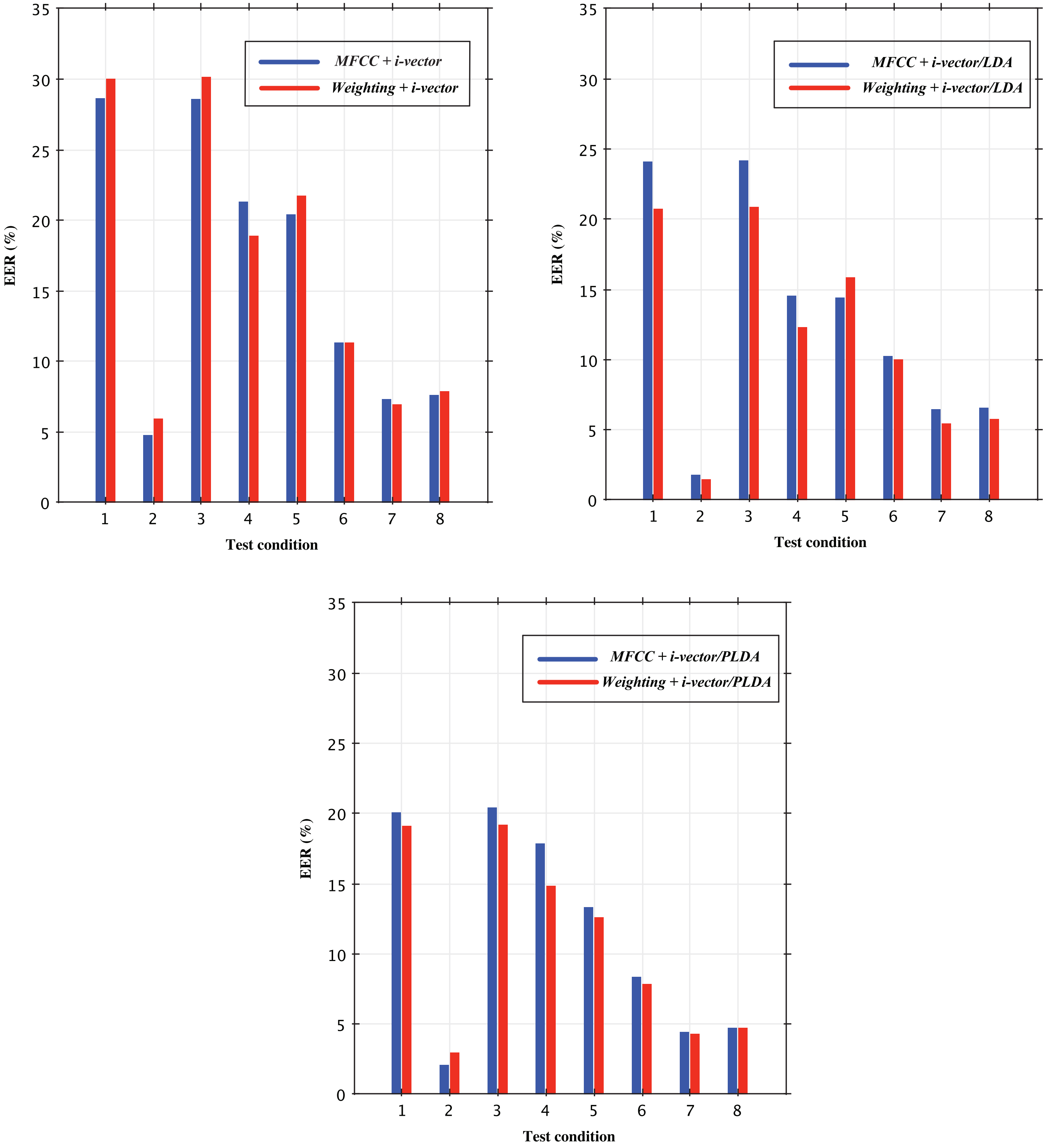

Fig. 9. Experimental results of i-vector systems.

Table 5 clearly demonstrates the statistical significance of the two proposed kinds of features over the MFCC features as the *p*-values were smaller than 0.01, especially the Warping Features.

6.5. Generalization of the proposed approach

The proposed approach worked well on the specially-created CSLT-Chronos in the conventional GMM-UBM framework. However, we would like to verify the improvement of proposed features to other databases, e.g. the widely-used NIST SRE (Speaker Recognition Evaluation) databases. Take the state-of-the-art, i-vector based systems, the loading matrix T_c in i-vector models is trained through unsupervised learning, which might affect the performance of the proposed approach aiming at discriminability emphasis based on F-ratio. Furthermore, i-vectors are usually fed into discriminative models to achieve further performance, such as LDA Dehak et al. (2011) or PLDA Prince and Elder (2007) models, which also make use of similar linear discriminant criteria in training their projection matrices.

Therefore, the baseline MFCC features and the proposed Weighting Features are tested on the NIST SRE 2008 database NIS (2008) within the i-vector framework, based on the *Discrim_score* curve obtained from CSLT-Chronos as shown in Fig. 7.

1997 female speakers were selected from the core evaluation data set (short2-short3) of NIST SRE 2008, and 59,343 trials were made (including 47,184 impostor trials).

Apart from the standard i-vector system with the simple cosine-distance scoring, i-vector/LDA and i-vector/PLDA systems were also implemented. 7196 female speakers from the Fisher corpus (English speech) were selected to train the loading matrix T_c for i-vector extractor (400 dimensions) and the projection matrix G for LDA/PLDA (150 dimensions for the speaker subspace). A 2048-mixture gender-dependent UBM was also trained using utterances from 4000 randomly-selected female speakers in the Fisher corpus. Experimental results of i-vector systems with both the baseline MFCC features and the proposed Weighting Features are shown in Fig. 9.

It can be seen that, in most test conditions except for 4 and 7, Weighting Features showed no advantage over MFCC features in the standard i-vector system, probably due to suppression of discriminability emphasis in parameter training as discussed before. However, when LDA or PLDA models were applied, Weighting Features outperformed MFCC features in most test conditions with the exception of test condition 5 in the i-vector/LDA system and test condition 2 in the i-vector/PLDA system. This indicated that the advantage of Weighting Features could be recovered with aid of those discriminative models, which lead to significant performance improvement. Meanwhile, it also verified that the *Discrim_score* curve is well generalizable: the parameters derived from a small database can be extended successfully to other i-vector based systems trained with a large multi-channel database.

7. Conclusions and future work

In this paper, we studied how to find more appropriate acoustic features for speaker verification in terms of long-term performance. Emphases are made among the frequency band selection. A strategy based on the F-ratio criterion is proposed to determine the overall discrimination-sensitivity of frequency bands by considering both the speaker-specific information and the session-specific variability information. Different emphasis is placed upon different frequency bands during feature extraction through pre-filtering frequency warping or post-filtering filter-bank output weighting. Experimental results have shown that the two proposed acoustic features have both yielded higher and more robust recognition accuracy than the MFCC features, especially the Warping Features, which outperforms MFCC significantly.

While this paper explored the frequency warping and the filter-bank output weighting separately, how to combine the two methods to achieve further performance improvement deserves more careful studies.

Due to the lack of available speech resources for the purpose of the target research, experiments were performed on a specially-created CSLT-Chronos and a common database, NIST SRE 2008 database, which is publicly available to check its effectiveness. We hope to evaluate the proposed approach on more databases in the future.

Although theoretically, a higher F-ratio value means higher discrimination-sensitivity for the target grouping, it does not lead to higher accuracy in speaker verification, as the final system has a combined effect of features and models. Thus, discriminability emphasis based feature extraction through a data-driven approach will be studied further in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61271389 and the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

References

Auckenthaler, R., Mason, J.S., 1997. Equalizing sub-band error rates in speaker recognition. In: Proceedings of the Eurospeech ’97. Rhodes, Greece, pp. 2303–2306.

Beigi, H., 2009. Effects of time lapse on speaker recognition results. In: Proceedings of the 16th International Conference on Digital Signal Processing. Santorini-Hellas, Greece, pp. 1–6.

Beigi, H., 2010. Fundamentals of speaker recognition. Springer, New York, USA.

Besacier, L., Bonastre, J.-F., 1997. Subband approach for automatic speaker recognition: optimal division of the frequency domain. In: Proceedings of the AVBPA ’97. Crans-Montana, Switzerland, pp. 195–202.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrir-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., Reynolds, D.A., 2004. A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. 2004, 430–451.

Bonastre, J.-F., Bimbot, F., Boë, L., Campbell, J.P., Reynolds, D.A., Magrin-Chagnolleau, I., 2003. Person authentication by voice: a need for caution. In: Proceedings of the Eurospeech ’03. Geneva, Switzerland, pp. 33–36.

Brandschain, L., Graff, D., Cieri, C., Walker, K., Caruso, C., 2010. Greybeard - voice and aging. In: Proceedings of the LREC 2010. Malta, pp. 2437–2440.

Campbell, J., Higgins, A., 1994. YOHO speaker verification. Linguistic Data Consortium (LDC), Philadelphia. http://catalog.ldc.upenn.edu/LDC94S16.

Campbell, J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85 (9), 1437–1462.

Cole, R.A., Noel, M., Noel, V., 1998. The CSLU speaker recognition corpus. In: Proceedings of the ICSLP ’98. Sydney, Australia, pp. 3167–3170.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P., 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proceedings of the Interspeech '09. Brighton, UK, pp. 1559–1562.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech Lang. Process. 19 (4), 788–798.

Devore, J.L., 1995. Probability and Statistics for Engineering and the Sciences. Thomson Brooks/Cole Publishing Company.

Furui, S., 1997. Recent advances in speaker recognition. Pattern Recogn. Lett. 18 (9), 859–872.

Gallardo, L.F., Wagner, M., Moller, S., 2014a. Advantages of wideband over narrowband channels for speaker verification employing mfccs and lfccs. In: Proceedings of the Interspeech '14. Singapore, pp. 1115–1119.

Gallardo, L.F., Wagner, M., Moller, S., 2014b. Spectral sub-band analysis of speaker verification employing narrowband and wideband speech. In: Proceedings of the Speaker Odyssey '14. Joensuu, Finland, pp. 81–87.

Gauvin, J.L., Lee, C.-H., 1994. Maximum a posterior estimation for multivariate gaussian mixture observations of markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Hébert, M., 2008. Text-dependent speaker recognition. In: Springer Handbook of Speech Processing. Springer Berlin Heidelberg, Berlin, Germany, pp. 743–762.

Huang, X.-D., Acero, A., Hon, H.-W., By-Reddy, R.F., 2001. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, New Jersey, USA.

Hyon, S., Wang, H., Wei, J., Dang, J., 2012. An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean f-ratio contribution. In: Proceedings of the APSIPA ASC '12. Hollywood, USA, pp. 1–4.

Jarque, C.M., Bera, A.K., 1987. A test for normality of observations and regression residuals. Int. Stat. Rev. 55 (2), 163–172.

Kato, T., Shimizu, T., 2003. Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence preserving connected digit patterns. In: Proceedings ICASSP '03. Hong Kong, pp. 57–60.

Kelly, F., Drygajlo, A., Harte, N., 2012a. Speaker verification with long-term ageing data. In: Proceedings 5th IAPR International Conference on Biometrics. New Delhi, India, pp. 478–483.

Kelly, F., Dyrgajlo, A., Harte, N., 2012b. Compensating for ageing and quality variation in speaker verification. In: Proceedings of the Interspeech '12. Portland, OR, USA, pp. 498–501.

Kelly, F., Dyrgajlo, A., Harte, N., 2013. Speaker verification in score-ageing-quality classification space. Comput. Speech Lang. 27 (5), 1068–1084.

Kelly, F., Harte, N., 2011. Effects of long-term ageing on speaker verification. In: Biometrics and ID Management: Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Germany, pp. 113–124.

Kenny, P., Boulianne, G., Dumouchel, P., 2005. Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio Process. 13 (3), 345–354.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007a. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio, Speech Lang. Process. 15 (4), 1435–1447.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007b. Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio, Speech .Lang. Process. 15 (4), 1448–1460.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., 2008. A study of inter-speaker variability in speaker verification. IEEE Trans. Audio, Speech Lang. Process. 16 (7), 980–988.

Kersta, L.G., 1962. Voiceprint identification. Nature 196 (4861), 1253–1257.

Kinnunen, T., 2002. Designing a speaker-discriminative adaptive filter bank for speaker recognition. In: Proceedings of the Interspeech '02. Denver, USA, pp. 2325–2328.

Kinnunen, T., 2003. Spectral features for automatic text-independent speaker recognition. Ph.D. thesis. University of Joensuu.

Kinnunen, T., Li, H.-Z., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. 52 (1), 12–40.

Künzel, H.J., 1994. Current approaches to forensic speaker recognition. In: Proceedings of the ASRIV-1994. Martigny, Switzerland, pp. 135–142.

Kuroiwa, S., Tsuge, S., AWA long-term recording speech corpus (AWA-LTR). NII Speech Resources Consortium (NII-SRC). http://research.nii.ac.jp/src/en/AWA-LTR.html.

Lamel, L.F., Gauvin, J.L., 2000. Speaker verification over the telephone. Speech Commun. 31 (2–3), 141–154.

Lawson, A.D., Stauffer, A.R., Cupples, E.J., Wenndt, S.J., Bray, W.P., Grieco, J.J., 2009a. The multi-session audio research project (MARP) corpus: goals, design and initial findings. In: Proceedings of the Interspeech '09. Brighton, UK, pp. 1811–1814.

Lawson, A.D., Stauffer, A.R., Smolenski, B.Y., Pokines, B.B., Leonard, M., Cupples, E.J., 2009b. Long-term examination of intra-session and inter-session speaker variability. In: Proceedings of the Interspeech '09. Brighton, UK, pp. 2899–2902.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. Comput. Speech Lang. 9 (2), 171–185.

Lei, H., Gonzalo, E.L., 2009. Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition. In: Proceedings of the Interspeech '09. Brighton, UK, pp. 2323–2326.

Lu, X.-G., Dang, J.-W., 2007. Physiological feature extraction for text-independent speaker identification using non-uniform subband processing. In: Proceedings of the ICASSP '07. Honolulu, Hawaii, USA, pp. 461–464.

Lu, X.-G., Dang, J.-W., 2008. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Commun. 50 (4), 312–322.

Markel, J., Davis, S., 1979. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. IEEE Trans. Audio, Speech Signal Process. 27 (1), 74–82.

Matsui, T., Furui, S., 1992. Comparison of text-independent speaker recognition methods using VQ distortion and discrete/continuous HMMs. In: Proceedings of the ICASSP '92. San Francisco, CA, USA, pp. 157–160.

Orman, D., Arslan, L., 2001. Frequency analysis of speaker identification. In: Proceedings of the Speaker Odyssey '01. Crete, Greece, pp. 219–222.

Prince, S.J.D., Elder, J.H., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the ICCV 07. Rio de Janeiro, Brazil, pp. 1–8.

Reubold, U., Harrington, J., Kleber, F., 2010. Vocal aging effects on f0 and the first formant: a longitudinal analysis in adult speakers. Speech Commun. 52 (7), 638–651.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10 (1), 19–41.

Rhodes, R., 2011. Changes in the voice across the early adult lifespan. In: Proceedings of the IAFPA '11. Vienna, Austria.

Rose, P., 2002. Forensic Speaker Identification. Taylor & Francis, London, UK.

Soong, F., Rosenberg, A.E., Rabiner, L.R., Juang, B.-H., 1985. A vector quantization approach to speaker recognition. In: Proceedings of the ICASSP '85. Florida, USA, pp. 387–390.

Stathopoulos, E.T., Huber, J.E., Sussman, J.E., 2011. Changes in acoustic characteristics of the voice across the life span: measures from individuals 4–93 years of age. J. Speech, Language, Hearing Res. 54 (4), 1011–1021.

Wolf, J.J., 1972. Efficient acoustic parameters for speaker recognition. J. Acoust. Soc. Am. 51 (6), 2044–2056.

Xiong, Z.-Y., Zheng, T.F., Song, Z.-J., Soong, F., Wu, W.-H., 2006. A treg-based kernel selection approach to efficient Gaussian mixture model - universal background model based speaker identification. Speech Commun. 48 (10), 1273–1282.

Xiong, Z.-Y., Zheng, T.F., Wu, W., Li, J., 2003. An automatic prompting texts selecting algorithm for di-ifs balanced speech corpus. In: Proceedings of the NCMMSC ’03. Xiamen, China, pp. 252–256.

Zhang, J.-Y., Zheng, T.F., Li, J., Luo, C.-H., Zhang, G.-L., 2001. Improved context-dependent acoustic modeling for continuous chinese speech recognition. In: Proceedings of the Eurospeech ’01. Aalborg, Denmark, pp. 1617–1620.

Zhou, X.-H., Garcie-Romero, D., Duraiswami, R., Espy-Wilson, C., Shamma, S., 2011. Linear versus mel frequency cepstral coefficients for speaker recognition. In: Proceedings of the ASRU ’11. Hawaii, USA, pp. 559–564.

The, 2008 NIST Speaker Recognition Evaluation (SRE-08).2008. http://www.itl.nist.gov/iad/mig//tests/sre/2008/.