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.