Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 47 (2018) 259–271

www.elsevier.com/locate/csl

Incorporating pass-phrase dependent background models for text-dependent speaker verification☆

Achintya Kumar Sarkar*, Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Aalborg, Denmark

Received 19 November 2016; received in revised form 12 June 2017; accepted 31 July 2017
Available online 18 August 2017

Abstract

In this paper, we propose pass-phrase dependent background models (PBMs) for text-dependent (TD) speaker verification (SV) to integrate the pass-phrase identification process into the conventional TD-SV system, where a PBM is derived from a text-independent background model through adaptation using the utterances of a particular pass-phrase. During training, pass-phrase specific target speaker models are derived from the particular PBM using the training data for the respective target model. While testing, the best PBM is first selected for the test utterance in the maximum likelihood (ML) sense and the selected PBM is then used for the log likelihood ratio (LLR) calculation with respect to the claimant model. The proposed method incorporates the pass-phrase identification step in the LLR calculation, which is not considered in conventional standalone TD-SV systems. The performance of the proposed method is compared to conventional text-independent background model based TD-SV systems using either Gaussian mixture model (GMM)-universal background model (UBM) or hidden Markov model (HMM)-UBM or i-vector paradigms. In addition, we consider two approaches to build PBMs: speaker-independent and speaker-dependent. We show that the proposed method significantly reduces the error rates of text-dependent speaker verification for the non-target types: target-wrong and impostor-wrong while it maintains comparable TD-SV performance when impostors speak a correct utterance with respect to the conventional system. Experiments are conducted on the RedDots challenge and the RSR2015 databases that consist of short utterances.

Keywords: Pass-phrase dependent background models (PBMs); GMM-UBM; HMM-UBM; I-vector; Text-dependent; Speaker verification

1. Introduction

Speaker verification (SV) (Bimbot et al., 2004; Reynolds et al., 2000) is the process of authentication of a person's claimed identity by analyzing his/her speech signal. It is a binary pattern recognition problem where a SV system makes the decision by calculating the log-likelihood ratio (LLR) between the claimant and background models (also called alternative/negative hypothesis) for the test signal. If the LLR value is greater than a pre-defined threshold, the claimant is accepted and otherwise it is rejected.

☆ This paper has been recommended for acceptance by Roger Moore.

* Corresponding author.

E-mail address: akc@es.aau.dk (A.K. Sarkar), zt@es.aau.dk (Z.-H. Tan).

http://dx.doi.org/10.1016/j.csl.2017.07.010
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

Speaker verification systems are broadly divided into two categories: text-independent (TI) and text-dependent (TD). In TI-SV, speakers are free to speak any sentences, i.e. phrases, during the enrollment as well as the test phases. It does not impose any constraint that enrollment and test utterances are to be the same phrase. However, TD-SV systems require speakers to speak within pre-defined sentences, i.e. fixed pass-phrases during the speaker enrollment and test phases.

In real-life applications, we need a speaker verification system that is accurate on short utterances. In this regard, TD-SV systems are the ideal choice. Since speakers use the same pass-phrase during both the enrollment and test phases, it provides a well matched phonetic content between the enrollment and test phases. Therefore, TD-SV systems are more accurate compared to their TI-SV counterparts. Over the last decades, many techniques have been introduced in literature to improve the performance of TD-SV on short utterances. Examples are deep neural network (DNN) (Lei et al., 2014; Scheffer and Lei, 2014; Liu et al., 2015), i-vector (Scheffer and Lei, 2014; Dehak et al., 2011), hierarchical multi-Layer acoustic model (HiLAM) (Larcher et al., 2012), phone-dependent hidden Markov model (HMM) (Kajarekar and Hermansky, 2001; Auckenthaler et al., 1999), domain adaptation (Aronowitz and Rendel, 2014) and phonetic higher level maximum likelihood linear regression (MLLR) super-vector based features (Stolcke et al., 2005). In Lei et al. (2014) and Scheffer and Lei (2014), phonetic information is incorporated into an i-vector system by accumulating statistics from speech with respect to a pre-defined phonetic class through an DNN based automatic speech recognition (ASR) system. In Liu et al. (2015), the intermediate output of the DNN layers are used to vectorize characterization of speech data. HiLAM builds a HMM model by concatenating the speech segment-wise adapted models from the Gaussian mixture model- universal background model (GMM-UBM) (Reynolds et al., 2000). In domain adaptation (Aronowitz and Rendel, 2014), the mismatch between the text-independent and the text-dependent data is reduced by transforming the text-independent data to better match the text-dependent task (using the a priori transcription knowledge of the text-dependent data). In conventional HMM based TD-SV systems (Kajarekar and Hermansky, 2001; Auckenthaler et al., 1999), phoneme (context) dependent speaker models are built using the knowledge of speech transcriptions. In Stolcke et al. (2005), a speech signal is represented by a super-vector concatenation of MLLR transformations estimated with respect to a pre-defined phonetic class (e.g. vowel and consonant) using automatic speech recognition (ASR).

All of these techniques need a background model as an alternative/negative hypothesis for TD-SV. A single *text-independent background model* (either gender dependent or independent) is commonly used in literature where target speakers are represented by models (say in GMM-UBM framework) derived from the background model. In Kenny et al. (2014a; 2014b) and Sarkar and Umesh (2012), a multiple background model concept is proposed to improve the performance of the conventional speaker verification system, by training background models (BMs) based on the vocal tract length (VTL) characteristic of target speakers as in Sarkar and Umesh (2012) and pass-phrases as in Kenny et al. (2014a; 2014b) for text-independent and text-dependent speaker verification, respectively. During enrollment, target speaker models are derived from the BM based on VTL in Sarkar and Umesh (2012) and pass-phrase of target data in Kenny et al. (2014a; 2014b). In the test phase, a test utterance is scored against the claimant and background models specific for the claimant (defined during the enrollment phase) in order to calculate the log-likelihood ratio. However, this (Kenny et al., 2014a; 2014b; Sarkar and Umesh, 2012) does not incorporate the pass-phrase identification process to address the following two non-target types: target-wrong and impostor-wrong. Recently, in Kinnunen et al. (2016), the authors proposed a fusion system which combines the score/decision of an utterance verification system with a conventional SV system to improve the performance of the TD-SV system against target/impostor-wrong non-target trials.

In this paper, we propose pass-phrase dependent background models (PBMs) for TD-SV to integrate the utterance identification process (without an extra separate system) into the conventional SV system, aiming at rejecting more non-target types: target-wrong and impostor-wrong than the conventional SV system while maintaining the performance for the impostor-correct non-target type. In the proposed method, PBMs are derived from the *text-independent background model* by pooling the training data for a particular pass-phrase from many speakers. During enrollment, pass-phrase specific target speaker models are derived from the particular PBM with maximum a posteriori (MAP) adaptation by using the training data for the respective target model. In the test phase, the best PBM is selected for *a particular test utterance* in the maximum likelihood (ML) sense and *the best selected PBM* is used as an alternative hypothesis for the log-likelihood ratio calculation with respect to the claimant model, which differs from Kenny et al. (2014a; 2014b) and Sarkar and Umesh (2012). Furthermore two strategies are considered for building the PBM: one is called *speaker-independent (SI)* and the other is *speaker-dependent (SD)* again in contrast