Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 196–211

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Speaker verification based on the fusion of speech acoustics and inverted articulatory signals☆

Ming Li[a,b,c,*], Jangwon Kim[d], Adam Lammert[d], Prasanta Kumar Ghosh[e], Vikram Ramanarayanan[d], Shrikanth Narayanan[d]

[a] *Sun Yat-Sen University Carnegie Mellon University Joint Institute of Engineering, Sun Yat-Sen University, China*
[b] *Sun Yat-Sen University Carnegie Mellon University Shunde International Joint Research Institute, Shunde, China*
[c] *School of Mobile Information Engineering, Sun Yat-Sen University, China*
[d] *Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles, USA*
[e] *Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India*

Received 3 July 2014; received in revised form 6 May 2015; accepted 14 May 2015
Available online 22 May 2015

Abstract

We propose a *practical*, feature-level and score-level fusion approach by combining acoustic and estimated articulatory information for both text independent and text dependent speaker verification. From a practical point of view, we study how to improve speaker verification performance by combining dynamic articulatory information with the conventional acoustic features. On text independent speaker verification, we find that concatenating articulatory features obtained from measured speech production data with conventional Mel-frequency cepstral coefficients (MFCCs) improves the performance dramatically. However, since directly measuring articulatory data is not feasible in many real world applications, we also experiment with estimated articulatory features obtained through acoustic-to-articulatory inversion. We explore both feature level and score level fusion methods and find that the overall system performance is significantly enhanced even with estimated articulatory features. Such a performance boost could be due to the inter-speaker variation information embedded in the estimated articulatory features. Since the dynamics of articulation contain important information, we included inverted articulatory trajectories in text dependent speaker verification. We demonstrate that the articulatory constraints introduced by inverted articulatory features help to reject wrong password trials and improve the performance after score level fusion. We evaluate the proposed methods on the X-ray Microbeam database and the RSR 2015 database, respectively, for the aforementioned two tasks. Experimental results show that we achieve more than 15% relative equal error rate reduction for both speaker verification tasks.

Keywords: Text independent speaker verification; Text dependent speaker verification; Speech production; Articulatory features; Acoustic-to-articulatory inversion

☆ This paper has been recommended for acceptance by R.K. Moore.

* Corresponding author at: Sun Yat-Sen University Carnegie Mellon University Joint Institute of Engineering, Sun Yat-Sen University, China.
E-mail addresses: liming46@mail.sysu.edu.cn (M. Li), jangwon@usc.edu (J. Kim), lammert@usc.edu (A. Lammert), prasantg@ee.iisc.ernet.in (P.K. Ghosh), vramanar@usc.edu (V. Ramanarayanan), shri@sipi.usc.edu (S. Narayanan).
URLs: http://jie.sysu.edu.cn/~mli/ (M. Li), http://sail.usc.edu/~jangwon/ (J. Kim), http://www-scf.usc.edu/~lammert/ (A. Lammert), http://www.ee.iisc.ernet.in/new/people/faculty/prasantg/ (P.K. Ghosh), http://sail.usc.edu/~vramanar/ (V. Ramanarayanan), http://sail.usc.edu/shri.php (S. Narayanan).

http://dx.doi.org/10.1016/j.csl.2015.05.003
0885-2308/

1. Introduction

The goal of a speaker verification system is to determine automatically whether a given segment of speech is indeed spoken by the claimed speaker. It can be further divided into text independent speaker verification (TISV) and text dependent speaker verification (TDSV) depending on whether we constrain the speech content during verification.

Total variability i-vector modeling has gained significant attention in speaker verification due to its excellent performance, compact representation and small model size (Dehak et al., 2011a). In this framework, first, zero-order and first-order Baum-Welch statistics are calculated by projecting the acoustic level Mel-frequency cepstral coefficients (MFCC) features onto universal background model (UBM) components using the occupancy posterior probability. Second, in order to reduce the high dimension of the concatenated statistics supervectors, a single factor analysis is adopted to generate a low dimensional total variability space which jointly models language, speaker and channel variabilities all together (Dehak et al., 2011). The factor analysis can also be extended to a simplified and supervised version to enhance the performance and reduce the computational cost (Li and Narayanan, 2014). Within this i-vector space, variability compensation methods, such as within-class covariance normalization (WCCN) (Hatch et al., 2006), linear discriminative analysis (LDA) and nuisance attribute projection (NAP) (Campbell et al., 2006) are performed to reduce the variability for subsequent scoring methods (e.g., cosine similarity (Dehak et al., 2011a), support vector machine (SVM) (Cumani et al., 2011), probabilistic linear discriminant analysis (PLDA) (Prince, 2007; Matejka et al., 2011), deep belief networks (Cumani et al., 2011), etc.). Several types of phonetics-aware generalized i-vectors have also been recently proposed for better performance (Lei et al., 2014; D'Haro et al., 2014; Li and Liu, 2014).

In addition to the aforementioned state-of-the-art modeling methods, various features have also been proposed for speaker verification (e.g. short-term spectral features, voice source features, spectral-temporal features, prosodic features and high-level features) (Kinnunen and Li, 2010). Based on these multiple sets of features, both feature-level and score-level fusion approaches have been shown to enhance the overall system performance (Kinnunen and Li, 2010; Kim and Stern, 2012; Shao and Wang, 2008; Wang and Johnson, 2014). Specifically, by fusing the phonetic level tandem features and the acoustic level MFCC features together at the feature level, more than 40% relative error reduction is achieved (Li and Liu, 2014; D'Haro et al., 2014; Wang et al., 2013). In this work, our goal is to examine the use of speech production oriented features for the speaker verification task.

The ability to understand sources of inter-speaker variability in speech production and to predict those sources of variability from the acoustic signal can afford a variety of advantages. Several studies have shown that an important source of inter-speaker variability in speech acoustics lies in the variability in the vocal tract morphology across various speakers. Morphological variability could result from the differences in the vocal tract length (Peterson and Barney, 1952; Fant, 1960; Lee et al., 1999; Stevens, 1998), or the morphology of the hard palate and the posterior pharyngeal wall (Lammert et al., 2011, 2013b,a). Fig. 1 shows magnetic resonance images of the vocal apparatus of four different subjects from the USC-TIMIT corpus (Narayanan et al., 2014) illustrating this variability. Since vocal tract length is closely related to the formant frequency (Stevens, 1998; Fant, 1960), change in vocal tract length scales the spectral envelope for voiced sounds. This has been extensively used for vocal tract length normalization (VTLN) (Eide and Gish, 1996; Lee and Rose, 1996) in automatic speech recognition (ASR). Unlike normalization, we focus on exploiting

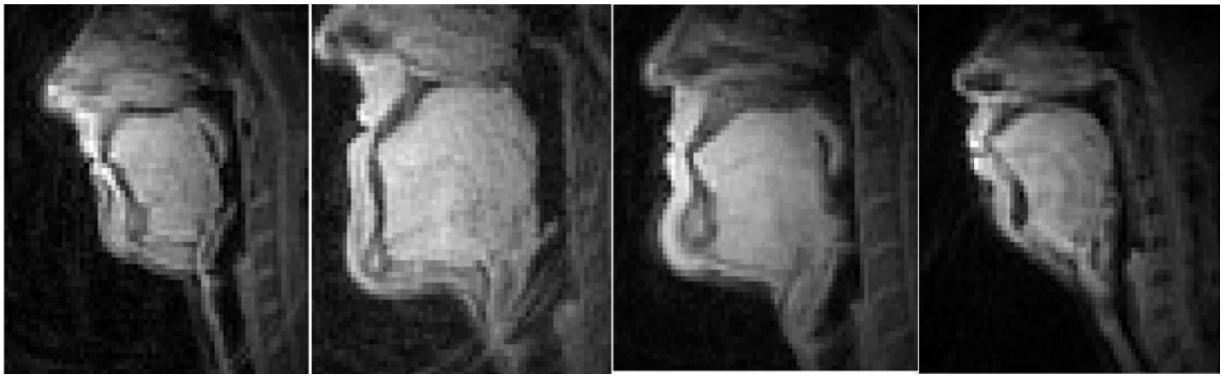

Fig. 1. Vocal tract rtMRI images from four different subjects in the USC-TIMIT corpus (Narayanan et al., 2014).