Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2015) xxx–xxx
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

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

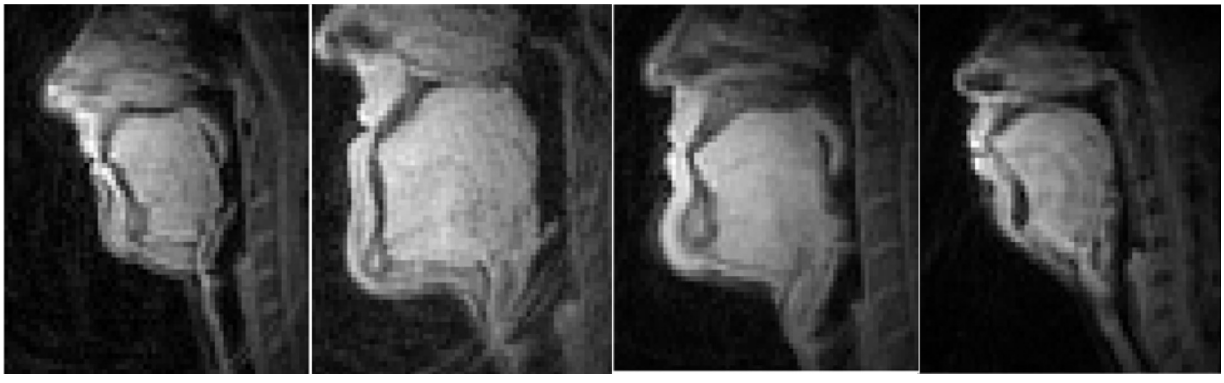

Fig. 1. Vocal tract rtMRI images from four different subjects in the USC-TIMIT corpus (Narayanan et al., 2014).

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

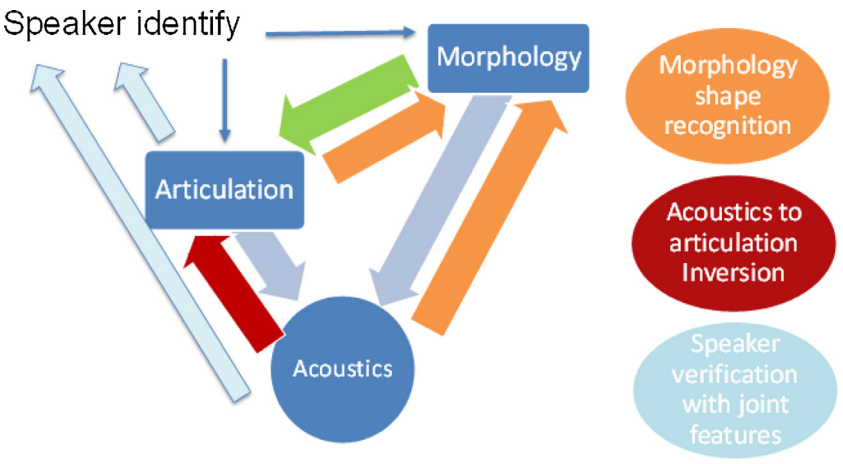

Fig. 2. Relation between speaker identity, vocal tract morphological patterns, articulatory trajectories and speech acoustics in this work. Arrows with different colors represent different processes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

morphological variations as the cue for capturing speaker characteristics in speaker verification applications. In order to obtain the vocal tract morphological shape information, direct measurements using magnetic resonance imaging or electromagnetic articulography are desirable. Our earlier work (Li et al., 2013) explored the possibility of automatically characterizing hard palate and pharyngeal wall morphological shape patterns directly from speech acoustics. However, indications are that those morphological differences may not be abundantly evident in the acoustics because speakers adjust their lingual articulation in compensation (Brunner et al., 2007; Lammert et al., 2013a), making estimation of these characteristics from acoustics a challenging task. But on the other hand, the details of the associated articulation can itself carry useful speaker-specific information.

The present study is based on the working hypothesis that articulatory movements reflect the morphological variability of speakers. For example, it is known that speakers with flat palates exhibit less articulatory variability during vowel production than speakers with highly domed palates (Perkell, 1997; Mooshammer et al., 2004; Brunner et al., 2005, 2009). Articulation of coronal fricatives is also influenced by palate shape, including influencing apical vs. laminal articulation of sibilants (Dart, 1991), as well as jaw height and the positioning of the tongue body (Honda et al., 2002; Thibeault et al., 2011). Inter-speaker differences in vocal tract morphology characteristics hence can lead to differences in speech articulation patterns. In Fig. 2, we can see that there is a clear relation between the speaker's vocal tract morphology and articulatory behavior. Finally, both morphology and articulation jointly influence the generation of the acoustic signal. This motivates us to perform speaker verification by using both the articulatory and the acoustic features (light blue arrows in Fig. 2).

We find that concatenating articulatory features obtained from measured speech production data with conventional Mel-frequency cepstral coefficients (MFCCs) improves the speaker verification performance dramatically. Fig. 3 shows some examples of articulatory measurements and the corresponding speech waveform. However, since measuring articulatory movement during speech production is not practical for real world applications, experiments are performed where the measured articulatory features are replaced with estimated articulatory features obtained using acoustic-to-articulatory inversion techniques. In this way, the inter-speaker variations could be projected into the intra-speaker variabilities of the exemplar speaker assuming he/she is asked to mimic different speakers' pronunciations. Since some components of the articulatory trajectories are highly correlated, we apply the principal component analysis (PCA) for dimension reduction and used multiple exemplar speakers in our experiments to enhance the speaker verification performance. Specifically, we show that augmenting MFCCs with features obtained from subject-independent acoustic-to-articulatory inversion techniques achieves promising results against the MFCC baseline and significantly improves the performance after score level fusion. In this work, we applied one exemplar-based speaker independent acoustic-to-articulatory inversion methods based on Ghosh and Narayanan (2011) and one deep neural network (DNN) based approach based on Uria et al. (2011) to generate the estimated articulatory signals. It is worth noting that other types of acoustic-to-articulatory mapping, such as CCA (Bharadwaj et al., 2012; Arora and Livescu, 2013), Kernel CCA

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

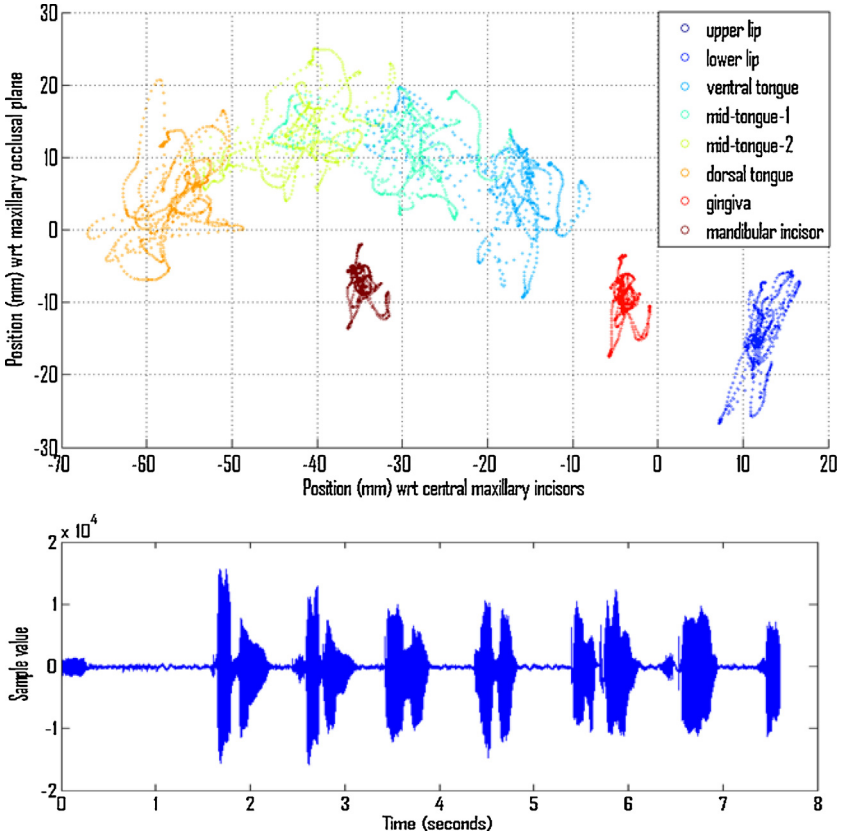

Fig. 3. Articulatory trajectories and the corresponding speech waveform of speaker JW 43 session 1 in the Wisconsin X-ray Microbeam (XRMB) database.

(Rudzicz, 2010; Arora and Livescu, 2013), Gaussian Mixture Model (GMM) (Ghosh and Narayanan, 2013; Ozbek et al., 2011; Özbek et al., 2012), attributes classification (Leung et al., 2004; Zhang et al., 2007; Siniscalchi et al., 2013, 2012) and articulatory phonological code (Zhuang et al., 2009), etc., could also be applied here. The reason to choose the exemplar-based speaker independent acoustic-to-articulatory inversion methods is that we can directly compare the performance against the real articulatory trajectories measurement to find out the gap which shows the potential for better speaker aware inversion techniques. Our future work includes investigating the effects of different speaker independent acoustic-to-articulatory mapping methods in terms of speaker verification performance with training data from multiple exemplar speakers, especially in the direction of highlighting inter-speaker variations.

Although the inverted articulatory features are also generated from speech signals, we can show that adding this new information (articulation-acoustics mapping learned from the exemplar data) on top of MFCCs can still enhance the speaker verification performance. Theoretical support from machine learning fields is provided in Pechyony and Vapnik (2010), Vinyals et al. (2012). Previously, this concatenation based speech-articulatory feature level fusion has been reported to increase the performance of ASR (Toutios and Margaritis, 2003; King et al., 2007; Ghosh and Narayanan, 2011b) significantly. In this work, we show that by utilizing information from both speech and inverted articulation, the equal error rate of speaker verification system is also reduced.

Furthermore, we demonstrate that the articulatory level constraints introduced by the inverted articulatory features also help the TDSV system to reject wrong password trials and therefore improve the performance. In Hébert (2008), text dependent speaker verification is defined as a speaker verification task in which the lexicon used in the test phase is a subset of the lexicon pronounced by the speaker during the enrollment. By constraining the text of enrollment and testing utterances to be the same (verbal password), higher accuracy with shorter utterances can be achieved (Larcher et al., 2014b; Novoselov et al., 2014; Kenny et al., 2014; Variani et al., 2014). In Larcher et al. (2014b), the hierarchical multi-layer acoustic model (HiLAM) was shown to outperform the conventional i-vector approach, since the latter

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

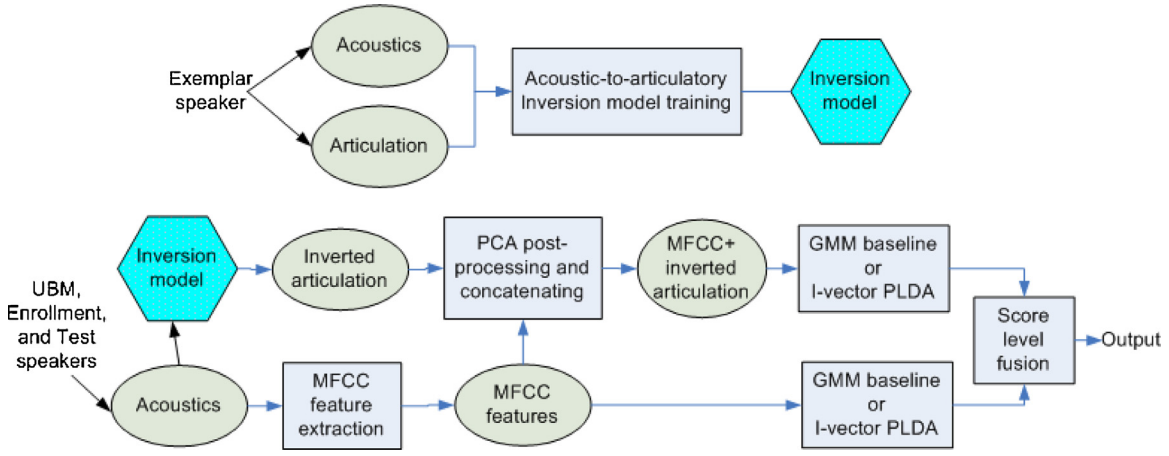

Fig. 4. System overview of the proposed speaker verification system with a single exemplar speaker.

does not explicitly take advantage of the temporal structure of the text dependent speech utterances. However, the HiLAM approach requires composing a specific acoustic model for each known password text content which may not work well on accented speech, dialect or out-of-vocabulary words. In the present study, we enhance the robustness of the i-vector representation against variation in the lexicon contents by adding articulatory features into the generalized i-vector framework. The proposed feature level fusion approach reduces the error rate of those wrong lexicon trials. However, it can also make the system vulnerable to those trials where imposter speakers utter the same password as the target speaker. The solution in this work is to fuse the i-vector baseline and the proposed feature-level fusion system together at the score level to achieve performance improvement for all three types of trials.

The remainder of the paper is organized as follows. The baseline and the proposed algorithms are explained in Section 2. The databases are described in Section 3. Experimental results and discussion are presented in Section 4. Conclusions and future work directions are provided in Section 5.

2. Methods

In this section, the baseline and the proposed methods are presented separately for each task. The overview of the proposed TISV and TDSV systems is shown in Fig. 4.

2.1. TISV task

2.1.1. Subject-independent acoustic-to-articulatory inversion

We adopted two different acoustic-to-articulatory inversion methods in this work, namely the generalized smoothness criterion (GSC) by Ghosh and Narayanan (2010) and the DNN framework by Uria et al. (2011). In the present paper, both techniques were performed under a subject-independent setting (Ghosh and Narayanan, 2011) since the exemplar speakers are excluded from our training/testing data and the testing speaker's identity is unknown.

The GSC method estimates articulatory parameters given acoustic features so that the estimated parameters are optimal solution which satisfies two conditions jointly: (1) the estimated trajectories are smooth and slowly varying and (2) the difference between the estimated and original articulatory parameters is minimum. The subject-independent inversion setting uses a probability feature vector (PFV) for acoustic features. PFV is a normalized likelihood score of the conventional acoustic feature vector, i.e. MFCCs, to the 40 clusters of a general acoustic model (GAM) (Ghosh and Narayanan, 2011). The general acoustic model represents the variabilities in acoustic space, which was created here with TIMIT data (Garofolo et al., 1993).

In subject-independent acoustic-to-articulatory inversion, MFCCs of an arbitrary test subject are converted to a PFV which is then used to find the closest PFV from the chosen exemplar whose articulatory data is used for training the inversion mapping. It is expected that the PFV reflects the acoustic sound produced by the test subject irrespective of the speaker, i.e., the PFVs corresponding to a sound recorded from different speakers including the exemplar should

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

be similar to each other so that the speaker variability is eliminated in the inversion. The quality of this speaker variability elimination solely depends on the generalizations of the GAM used to compute the probability feature vector. Note that the GAM used in this work is built using the TIMIT training corpus whereas the articulatory inversion is performed on XRMB corpus which may have different acoustic characteristics from TIMIT. This may result in poor elimination of speaker variability during inversion. This in turn gets reflected in the estimated articulatory features which provides inter-speaker discrimination in addition to MFCCs when used for the TISV task. In this case, inter-speaker variations could be projected onto the intra-speaker variability space of the training speaker assuming he/she is asked to mimic different speakers' pronunciations. Furthermore, the same MFCC features could be employed for both ASR and TISV, and therefore information regarding inter-speaker variation may also leak into the estimated articulatory signals through MFCCs. Thus the TISV performance improvement in this work may results from the non-linear mapping between acoustic and articulatory spaces and the residual speaker specific information present in the probability features computed during the subject-independent acoustic-to-articulatory inversion.

In the DNN framework (Uria et al., 2011), articulatory features are estimated from acoustic features using the deep neural network (DNN) regression model. This model was selected due to its promising estimation accuracy for acoustic-to-articulatory inversion in a previous study (Uria et al., 2011), but with a single speaker's data, i.e., the MNGU0 corpus (Richmond, 2011) in a speaker-dependent task, and with different feature pairs (line spectral frequencies and EMA sensor trajectories). In this work, first, a deep belief network (DBN) (Hinton et al., 2006) is constructed using a Gaussian-Bernoulli restricted Boltzmann machine (RBM) in the bottom layer and Bernoulli-Bernoulli RBMs in the other hidden layer(s). Each RBM is stacked one by one during a greedy layer-wise pre-training stage, where pre-training parameters, e.g., learning rate, momentum, and the number of epochs, are roughly tuned in the range of [0.002, 0.01], [0.002, 0.01] and [5, 50], respectively. Next, a simple linear regression is added on the top layer of the DBN in order to perform a regression task. The final model (i.e., DNN model) is fine-tuned using the well-known back-propagation algorithm. Tuning parameters, e.g., learning rate, momentum and the number of epochs, are roughly tuned in the same range as the pre-training stage. The optimal numbers for neurons and hidden layers of the DBN are determined in the range of 6–300 and 1–5, respectively, since improvement of estimation performance near the edges of the ranges was not observed. We used a freely available MATLAB toolbox (Palm, 2012) for this model training and testing.

The tract variable (TV) features were normalized by z-scoring, where the mean and the standard deviation were computed on training set. Note that the acoustic probability features are in the range of 0–1. Nine frames (90 ms) of the acoustic probability features computed by the GAM model in Ghosh and Narayanan (2011) were used as the input of aforementioned DNN model. The normalized TV features of a single frame (corresponding to the centered frame of the acoustic probability features) were used as the output of the DNN model. This setup is chosen to incorporate context information in the estimation process, which was also similarly used in the previous inversion experiment by Uria et al. (2011).

In Section 4, we show that the outputs of both the above mentioned inversion methods carry useful information about the inter-speaker variation, which could help boost the speaker verification performance.

2.1.2. *Inversion training data and the inverted articulatory features from multiple exemplars*

We used five exemplar speakers from three different sources to demonstrate the performance with multiple exemplars, but the inversion model is still trained separately for each exemplar speaker. The first two exemplar speakers come from the multichannel articulatory (MOCHA) database (Wrench, 1999) that contains electromagnetic articulography (EMA) data for 460 utterances (20 min) read by a female (fsew0) and a male (msak0) talker of British English. We refer to these subjects as exemplar 1 and exemplar 2, respectively. The second source of parallel articulatory-acoustic data comes from the EMA data collected at the University of Southern California (USC) from a male talker of American English (exemplar 3) as a part of a Multi-University Research Initiative (MURI) project (Silva et al., 2007; Ghosh and Narayanan, 2011b). In contrast to the read speech in the MOCHA database, the articulatory data in the MURI database were collected when the subject was engaged in a spontaneous conversation (50 min) with an interlocutor. The third database for inversion model training is the electromagnetic articulography database (Kim et al., 2013) collected at USC which includes speech audio spoken by two native female speakers (exemplar 4 and 5) of American English and the parallel articulatory data. These two speakers were asked to read 460 English sentences (approximately 69 min) identical to the sentences of MOCHA TIMIT database (Wrench, 1999). We used tract variables for articulatory parameters as in a previous study (Ghosh and Narayanan, 2011). The tract variables include nine articulatory parameters for exemplar 4 and 5, such as lip aperture (LA), lip protrusion (PRO), jaw opening (JAW_OPEN), the constriction degree (CD) and

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

constriction location (CL) of tongue tip (TT), tongue blade (TB), and tongue dorsum (TD). The constriction location parameter for each tongue sensor is the distance from a fixed point on the palatal line, which is manually chosen by visual inspection, to the projected point of each sample to the palatal line. We followed the definitions in the previous study (Ghosh and Narayanan, 2011) for the other parameters, such as LA, PRO, JAW_OPEN, CDs. For the other three exemplars (1,2 and 3), we adopt six tract variables, namely, LA, PRO, JAW_OPEN, tongue tip constriction degree (TTCD) tongue body constriction degree (TBCD) and velum (VEL) (Ghosh and Narayanan, 2011).

2.1.3. TISV system front end processing

After energy based voice activity detection (VAD), non-speech frames were eliminated and cepstral features were extracted. Real and estimated articulatory signals were also truncated based on the VAD results and then re-sampled at 100 Hz. A 25 ms Hamming window with 10 ms shifts was adopted for MFCC extraction. Each utterance was converted into a sequence of 36-dimensional acoustic feature vectors, each consisting of 18 MFCC coefficients and their first derivatives. Additionally, since different dimensions of those articulatory features are correlated, we applied PCA for dimension reduction and whitening.

Cepstral mean and variance normalization (MVN) normalization were performed to normalize the MFCC and real articulatory features to zero mean and unit variance. For global MVN, the mean and variance are calculated using the entire training database, while MVN operated on a per utterance basis is denoted as utterance MVN.

As shown in Fig. 4, after MVN or PCA, MFCCs are concatenated with real or estimated articulatory features to generate the MFCC-real-articulation and MFCC-estimated-articulation enhanced feature sets.

2.1.4. TISV system GMM baseline modeling

A UBM in conjunction with a maximum a posteriori (MAP) model adaptation approach (Reynolds et al., 2000) was used to model different speakers in a supervised manner. All the data in the background set were used to train a 512-component UBM, and MAP adaptation was performed using the training set data for each speaker. A relevance factor of 16 was used for the MAP adaptation. We performed AT-norm to calibrate the scores. Every testing utterance is scored on every target sample to generate the trials. The reason to use the GMM baseline here rather than the state-of-the-art i-vector PLDA method is that the XRMB data set is too small to train a large scale factor analysis model.

We do use i-vector representation with PLDA modeling (in Section 2.2) for the TDSV task since the scale of the RSR2015 database is larger.

2.2. TDSV system

In the TDSV task, the acoustic-to-articulatory inversion as well as the feature front end processing is the same as in the TISV task. The only difference is the modeling part. Since there are enough data for training, we adopted the state-of-the-art i-vector PLDA method here (Dehak et al., 2011a; Prince, 2007; Garcia-Romero and Espy-Wilson, 2011). Simplified supervised i-vector modeling is used for efficiency (Li and Narayanan, 2014). The details of this system are presented in Li and Liu (2014).

Once we have the low dimensional i-vectors, PLDA is applied as the back-end classifier for the TDSV task. Specifically, each speaker with each lexicon content password is considered as a class and different phrases from the same speaker are labeled with separate classes in the PLDA model training (Larcher et al., 2014b). Since there are three target utterances for each enrollment, we used the multiple enrollment PLDA scoring approach (Rajan et al., 2014; Liu et al., 2014). Finally, we simply employed the equal weighted summation fusion approach at the score level to further enhance the performance.

3. Data

Since the training databases for the acoustic-to-articulatory inversion are described in Section 2.1.2, here we mainly introduce the evaluation databases and experimental protocols for our TISV and TDSV tasks.

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Table 1
The XRMB data set partition for the TISV experiments.

Sets	Data
Background	All sessions from JW11-40
Target	Session 11 from JW41-63
Test	Other sessions from JW41-63
Tnorm	Sessions 11,12,79,80,81 from JW11-40

3.1. X-ray Microbeam database for the TISV task

A key feature of the Wisconsin X-ray Microbeam database (XRMB) (Westbury et al., 1990) is that articulatory measurements with simultaneously recorded speech signal from multiple speakers are available. To evaluate our methods, we selected the read speech data subset (citation words, sentences and paragraphs) from sessions 1 to 101 for each speaker from JW11 to JW63 for experiments which yielded a total of 4034 utterances from 46 speakers with an average duration of 5.72 s per utterance. Note that we excluded speech sessions involving different speaking styles (such as fast or slow speech, emphasized speech, or stimuli that involved diadokinesis). We also omitted speaker sessions where a speaker had to repeat an utterance, as well as those which were found to contain severe pellet tracking errors, as detailed in the XRMB Manual (Westbury et al., 1990). We used this XRMB database for our TISV analysis and experiments.

Table 1 shows the "ALL" protocol that we adopted in the evaluation. We used all sessions from speaker JW11 to JW40 (26 speakers with 2295 utterances) as the background data and select session 11 (a paragraph session) of each speaker from JW41 to JW63 (20 speakers) as the target registration utterance. For the testing, protocol "ALL" selects all the sessions (excluding session 11 in the target set) from speaker JW41 to JW63 (a total of 20 speakers and 1719 utterances). In order to perform test segment score normalization (T-norm), we selected all other paragraph sentences (sessions 12,79,80,81, totally 95 utterances) in the background set as the T-norm set.

To evaluate the performance of the acoustic-only baseline as well as the acoustic-estimated-articulatory system, we followed the protocol (as shown in Table 1) exactly. For the speech-real-articulation system, a subset of data were removed from the train, target and testing sets due to the missing data in some articulatory channels (Westbury et al., 1990). We name this modified "ALL" protocol as the "ALL-small" protocol. In the "ALL-small" protocol, each utterance is shorter and there are 1849, 18, and 1389 utterances in train, target and testing sets, respectively.

3.2. RSR2015 database for the TDSV task

In the RSR2015 database (Larcher et al., 2014b), the number of speakers in the background, development and evaluation sets are 47, 47 and 49, respectively. The part I background data consists of parallel recordings of 30 TIMIT phrases uttered by 47 female speakers, each of whom participated in 9 recording sessions on 3 different recording devices. This same part I background data set is used for UBM, i-vector and PLDA model training. The Part I female portion of the RSR2015 database is adopted as our TDSV evaluation dataset. We adopted the same development and evaluation data in Larcher et al. (2014b) to demonstrate the system performance and we did not use the development data for training.

The number of trials for each of the four text dependent speaker verification scenarios on the Part I of the RSR 2015 database is shown in Table 2. We can see that only the target speaker uttering the correct lexical content is considered as the true trial, the other cases are all non-target trials. In order to show the results for all three types of non-target trials, we evaluate the system performance separately for each type of trials the same way as in Larcher et al. (2014b).

4. Experimental results and discussion

We evaluate the proposed methods on both TISV and TDSV tasks.

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Table 2
Number of trials for each of the four text dependent speaker verification scenarios on the Part I of the RSR 2015 database.

Speaker	Lexical content	Trial type	Female	
			Development	Evaluation
Target	correct	true	8419	8631
Target	wrong	false	244123	250229
Imposter	correct	false	387230	414249
Imposter	wrong	false	5612176	6006596

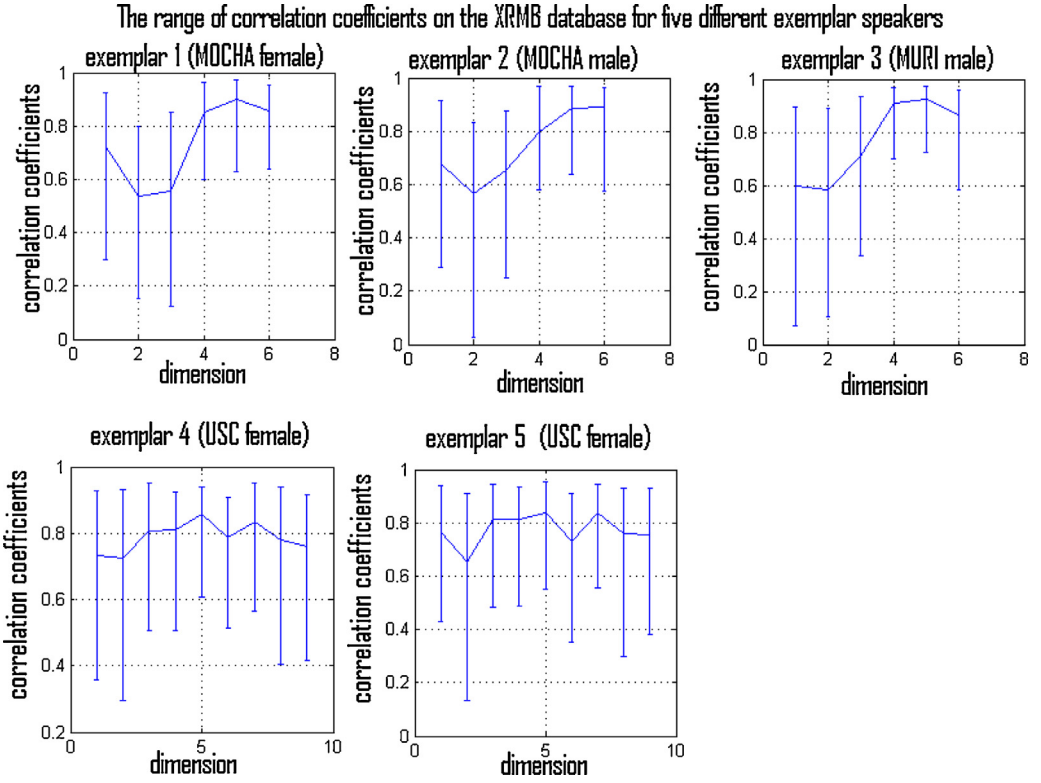

Fig. 5. The range of pair-wise correlation coefficients between the estimated articulatory features (after DTW) of session one from all 46 speakers in the XRMB database (all speak the same word sequence, totally 1035 pair-wise DTW and correlation). The nine dimensions of the estimated articulation for exemplar 4 and 5 are LA, PRO, JAW_OPEN, TTCD, TBCD, TDCD, TTCL, TBCL, and TDCL, respectively. The six dimensions of the estimated articulation for exemplar 1, 2 and 3 are LA, PRO, JAW_OPEN, TTCD, TBCD, and VEL, respectively.

4.1. The inverted articulatory features

Fig. 5 shows the range of pair-wise correlation coefficients between the estimated articulatory features of the XRMB database session one (including all 46 speakers) after temporal alignment on the utterance pairs. All spoke the same word sequence in this case, allowing us to compare the inter-speaker variations by this method. Dynamic time warping (DTW) (applied on the estimated articulation) was used to remove possible speaking-rate confounds for this correlation study. Fig. 5 shows that tongue constriction degree features (dim 4, 5 and 6) have less inter-speaker variations than other dimensions. Jaw opening and lip protrusion (dimension 1 and 2) shows relatively large correlation range, implying that their inter-speaker variations are larger than the other tract variables.

Fig. 6 shows the estimated articulatory features (LA after DTW) on the XRMB database session one from the two-speaker pairs. Within all the speaker pairs, the pair of speaker JW15 and 32 has relatively high correlation, while the pair of speaker JW 46 and JW 59 has relatively lower correlation. This correlation difference matches with the

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

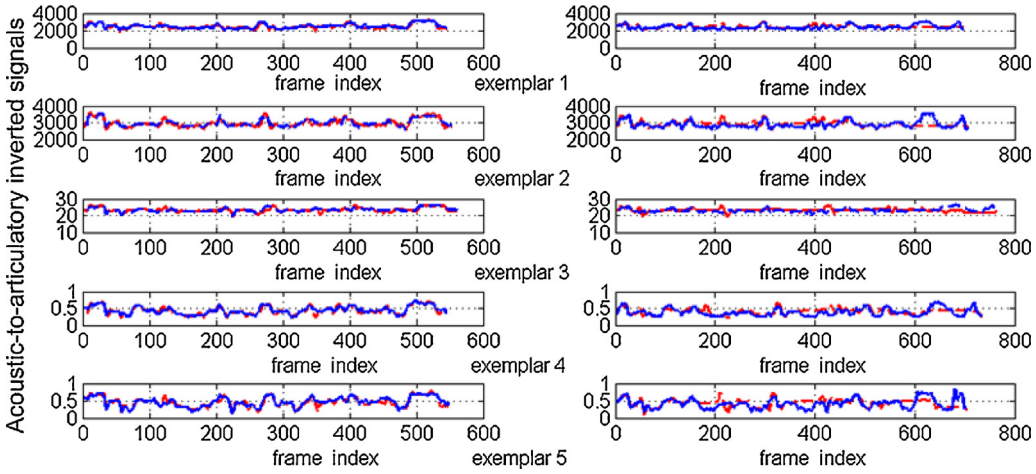

Fig. 6. Estimated articulatory features (LA after DTW) of the XRMB database session one from two-speaker pairs. Each row represent each exemplar. The left column plots are for the pair of speaker JW 15 and 32, and the right column plots are for the pair of speaker JW 46 and 59. The first speaker pair shows high correlation, while the other pair shows low correlation.

Table 3
The performance of 46 class speaker identification (closed set) systems based on mean and variance features derived from the estimated articulatory data (GSC method) on the XRMB data with different number of exemplar speakers.

Exemplar	Speaker and systems	1	2	3	4	5
3	MURI male	√	√	√	√	√
4	USC female 1		√	√	√	√
5	USC female 2			√	√	√
1	MOCHA female				√	√
2	MOCHA male					√
	Feature dimension	12	30	48	60	72
	Accuracy	22%	42%	53%	55%	68%

speech only speaker verification system score difference (JW15, JW32 and JW 59 are male, JW46 is female), therefore the inverted articulatory features carry a certain amount of speaker identity information.

4.2. TISV

In order to test our hypothesis that the mean and variance of inverted articulatory features carry inter-speaker variability information, we performed a simple multi-class SVM experiment before the main TISV experiment. Table 3 shows the performance of speaker identification with estimated articulatory trajectories using the GSC method on the XRMB data with different number of exemplars. The number of speaker classes is 46. Sessions 12, 79, 80 and 81 of all 46 speakers in the background data set was used for the train set (167 utterances), and session 11 was used as testing data. Table 3 shows the performance of 5 systems based on different numbers of exemplar speakers. We used the Liblinear toolkit (Fan et al., 2008) as the linear kernel SVM implementation and the cost value is set to 3. By using only mean and variance, system 5 achieves around 68% accuracy in a 46 class closed set identification task, indicating that they do carry valuable information regarding inter-speaker variations. This result may also suggest to normalize mean and variance of estimated articulatory parameters for minimizing speaker-dependent information for ASR applications.

Now we follow the protocol defined in Section 3.1 to evaluate the system performance in terms of speaker verification equal error rate (EER) and 2008 norm OptDCF cost value (NIST, 2010) for the TISV task.

Table 4 shows the performance of the "ALL" protocol on the XRMB data with exemplar 1 and 4 using the GSC inversion against the baseline. By combining the inverted articulatory features with MFCCs, we can find out that PCA achieves the best performance against the global and utterance level MVN. Specifically, we believe that utterance MVN

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Table 4
Performance of the "ALL" protocol on the XRMB data for TISV task with exemplar 1 and 4 using the GSC inversion method.

System ID	Methods	Exemplar	Articulatory post-processing	Dimension	Cost08	EER (%)
1	GMM-MAP	None	None	36	0.38	7.56
2	GMM-MAP	1	Global MVN	42	0.42	7.45
3	GMM-MAP	1	Delta+global MVN	48	0.54	9.37
4	GMM-MAP	1	PCA	40	0.38	7.33
5	GMM-MAP	4	Global MVN	45	0.44	8.44
6	GMM-MAP	4	Utterance MVN	45	0.5	12.91
7	GMM-MAP	4	PCA	40	0.41	7.50

Table 5
Performance of the "ALL" protocol on the XRMB data for TISV task with all 5 exemplar speakers using the GSC inversion method.

System ID	Methods	Exemplar	Articulatory post-processing	Dimension	Cost08	EER
1	GMM-MAP	none	none	36	0.38	7.56%
10	GMM-MAP	1	PCA	40	0.38	7.33%
11	GMM-MAP	2	PCA	40	0.37	7.10%
12	GMM-MAP	3	PCA	40	0.38	7.39%
13	GMM-MAP	4	PCA	40	0.41	7.50%
14	GMM-MAP	5	PCA	40	0.39	7.90%
15	Equal weight score level fusion 12 + 13				0.39	6.86
16	Equal weight score level fusion 12 + 13 + 14				0.37	6.69
17	Equal weight score level fusion 10 + 12 + 13 + 14				0.37	6.40
18	Equal weight score level fusion 10 + 11 + 12 + 13 + 14				0.36	6.46
19	Equal weight score level fusion 1 + 10				0.36	6.75
20	Equal weight score level fusion 1 + 18				0.36	6.28
21	GMM-MAP	1 + 2 + 3 + 4 + 5	Feature level fusion + PCA	44	0.39	7.10%

Table 6
Performance of the "ALL" protocol on the XRMB data for TISV task with all 5 exemplar speakers using the DNN inversion method.

System ID	Methods	Exemplar	Articulatory post-processing	Dimension	Cost08	EER
1	GMM-MAP	None	none	36	0.38	7.56%
22	GMM-MAP	1	PCA	40	0.37	7.10%
23	GMM-MAP	2	PCA	40	0.39	7.60%
24	GMM-MAP	3	PCA	40	0.37	7.00%
25	GMM-MAP	4	PCA	40	0.40	7.85%
26	GMM-MAP	5	PCA	40	0.39	7.90%
27	Equal weight score level fusion 22 + 23 + 24 + 25 + 26				0.34	6.57
28	Equal weight score level fusion 1 + 22				0.34	6.46

on the inverted articulatory data may lose the speaker information in terms of mean and variance which are informative for inter-speaker variations as illustrated in Table 3.

Table 5 shows the results for both feature level and score level fusion approaches. From system 10–14, we can see that systems based on different exemplar speakers perform differently. For exemplar 1–3, the enhanced feature sets achieve better performance in terms of EER. Systems based on exemplars 4 and 5 generate comparable results against the baseline. However, significant improvement (7.56% $\rightarrow$6.28 % absolute, 17% relatively, p-value = 0.0218 by a right-tailed two-sample t-test on the scores) is achieved after score level fusion. The results of systems 15–18 match with the findings in Table 3 that combining information from multiple exemplar speakers helps. It is worth noting that the feature level fusion using multiple sets of inverted articulatory features only achieves moderate gain. This might be because different sets of inverted articulatory features are correlated and the dimensionality of the feature level multiple exemplar speakers concatenated articulatory features is too high.

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Table 7
Performance of the "ALL-small" protocol on the XRMB data for TISV task with real articulatory measurements.

ID	Methods	Measurement	Articulatory post-processing	Dimension	Cost08	EER (%)
1	GMM-MAP	None	None	36	0.38	7.56
22	GMM-MAP	Real	PCA	48	0.1	1.3
23	GMM-MAP	Real	PCA + utterance MVN	48	0.1	1.58
24	GMM-MAP	Real	Utterance MVN	52	0.19	3.96

Table 8
Performance of the proposed systems on the development set of RSR 2015 Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker text	Type	Target		Imposter		System1: MFCC baseline	System2: MFCC + exemplar1 (GSC)	System3: MFCC + exemplar3 (GSC)
		Correct	Wrong	Correct	Wrong			
Trials	1	tar	non	–	–	0.77%/0.05/0.30	0.62%/0.04/0.2	0.68%/0.04/0.28
	2	tar	–	non	–	6.26%/0.32/0.8	7.15%/0.36/0.9	7.06%/0.36/0.87
	3	tar	–	–	non	0.1%/0.01/0	0.08%/0.01/0	0.1%/0/0

Similar performances were observed when using the DNN based inversion method as shown in Table 6. The results are not sensitive to the different acoustic-to-articulatory inversion methods. By fusing the GMM-MAP baseline system (ID 1) with the enhanced feature system (ID 22) at the score level, overall system EER was reduced from 7.56% to 6.46%.

Table 7 shows the performance on the MFCC only as well as the MFCC-real-articulation features systems with the "ALL-small" protocol. We can see that by augmenting the measured articulatory features with the MFCCs (even after utterance level mean and variance normalization), the enhanced feature set reduced the EER from 7.56% to 1.58%, a 70% relative EER reduction compared to the MFCC only system. Thus it is clear that, adding real articulation information enhances the speaker verification performance.

Although the gap between real articulatory measurements and estimated articulatory features through inversion is still big in terms of speaker verification performance, the performance improvements reported in Tables 5 and 6 do suggest the potential benefits that estimated articulatory features may provide in the TISV task. This is particularly important because in a real world TISV application, we have access only to the speech signal. In such scenarios it is only the estimated articulatory features that can provide information about speaker's articulatory characteristics. Experimental results in this work show that estimated articulatory features indeed provide production oriented information (complementary to the MFCCs) to discriminate different speakers. This is also shown by the Detection Error Trade-off (DET) curves in Fig. 7, which clearly demonstrate that adding estimated articulatory features improves the speaker verification performance.

4.3. TDSV

For the TDSV task, we evaluate the proposed methods on the female part of the RSR 2015 database as discussed in Section 3.2. The gender-dependent GMM UBM consists of 1024 mixture components. The sizes of i-vectors and the dimension of speaker-specific subspace in PLDA are 400 and 150, respectively.

Table 8 shows the performance of the proposed systems on the development set of Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF. We can see that the proposed systems with enhanced features (systems 2 and 3) outperform the MFCC baseline (system 1) for both type-1 and type-3 trials. This is because the introduction of the articulatory constraints introduced by the inverted features help the text dependent speaker verification system to reject wrong password trials. However, since more text dependent information (articulatory information conveys speech content information) is embedded in the enhanced feature, the

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

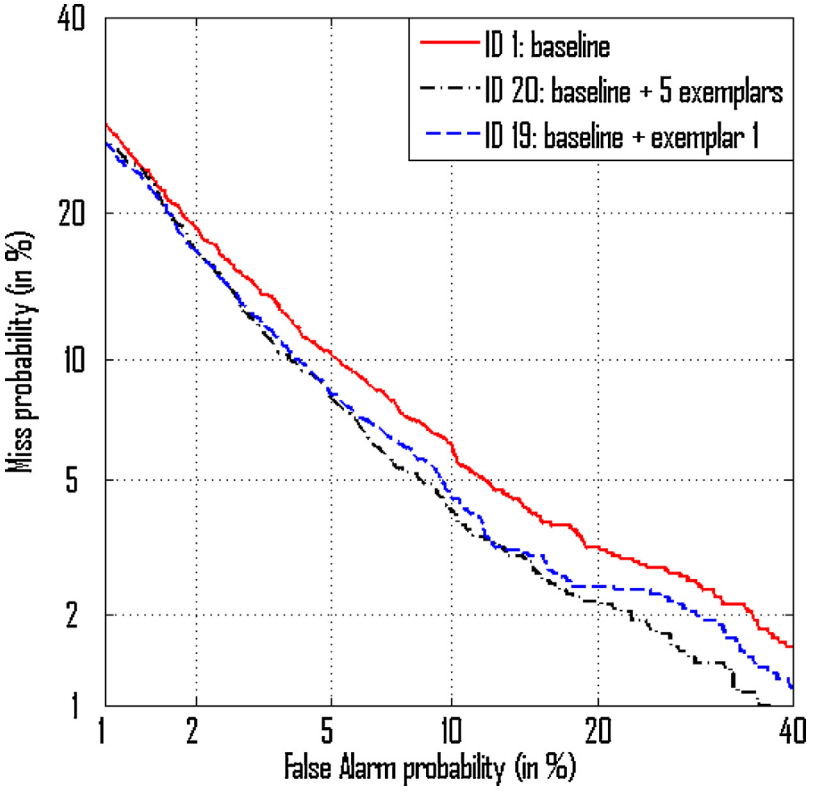

Fig. 7. The detection error trade-off (DET) curves of the systems with the "ALL" protocol on the XRMB data using the GSC inversion method.

system is less robust to the type-2 trials where the target and imposter speakers utter the same words. One possible solution is to classify the type of trial first and then apply different systems for different trials (Larcher et al., 2014a). In this work, we propose an alternative solution by simply fusing system 1 and 2 at the score level which not only maintains the error reduction for type-1 and type-3 trials, but also improves the performance on type-2 trials. From Table 9, we can see that the proposed fusion approach reduced EER from 5% to 19% relatively for different types of trials.

Similar results are shown in Tables 10 and 11 for the Part I evaluation set. Comparing with the state-of-the-art approaches (i-vector baseline in Larcher et al. (2014b)) in Table 9 and 11, our proposed method achieves significant performance improvement on type-1 and type-3 trials and comparable results on type-2 trials.

Table 9
Performance of the proposed fusion system and the reference systems on the development set of RSR 2015 Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF)

Speaker text	Target		Imposter		System 1+2	System 1+2+3	i-vector baseline (Larcher et al., 2014b)
	Correct	Wrong	Correct	Wrong			
Trials	tar	non	–	–	0.62%/0.04/0.3	0.64%/0.04/0.3	3.05%/0.17/–
	tar	–	non	–	6.03%/0.31/0.8	5.94%/0.30/0.8	7.87%/0.41/–
	tar	–	–	non	0.07%/0/0	0.08%/0.0/0	0.94%/0.05/–

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Table 10
Performance of the proposed systems on the evaluation set of RSR 2015 Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker text	Type	Target		Imposter		System1: MFCC baseline	System2: MFCC + exemplar1(GSC)	System3: MFCC + exemplar3(GSC)
		Correct	Wrong	Correct	Wrong			
Trials	1	tar	non	–	–	0.23%/0.01/0	0.16%/0.01/0	0.15%/0.01/0.05
	2	tar	–	non	–	3.85%/0.19/0.6	4.17%/0.22/0.7	4.17%/0.22/0.74
	3	tar	–	–	non	0.08%/0/0	0.07%/0/0	0.07%/0/0

Table 11
Performance of the proposed fusion system and the reference systems on the evaluation set of RSR 2015 Part I for different definitions of target and non-target trials in terms of EER, 08 norm min DCF and 10 norm min DCF (EER/08 norm min DCF/10 norm min DCF).

Speaker text	Target		Imposter		System 1+2	System 1+2+3	i-vector baseline (Larcher et al., 2014b)
	Correct	Wrong	Correct	Wrong			
Trials	tar	non	–	–	0.19%/0.01/0	0.18%/0.01/0	1.91%/0.11/–
	tar	–	non	–	3.45%/0.18/0.6	3.47%/0.18/0.6	6.61%/0.33/–
	tar	–	–	non	0.07%/0/0	0.06%/0/0	0.75%/0.04/–

5. Conclusions and future work

We propose a feature-level and score-level fusion approach by combining acoustic and (estimated) articulatory information for both text independent and text dependent speaker verification tasks. From a practical point of view, we study how to improve the speaker verification performance by combining the articulatory trajectories information characterizing the speech production. For the text independent speaker verification task, we find that concatenating articulatory features obtained from the measured speech production data with conventional MFCCs improves the performance dramatically. However, since access to the measured articulatory data is impractical for real world speaker verification applications, we also experimented with estimated articulatory features obtained using acoustic-to-articulatory inversion technique. We explore both the feature-level and score-level fusion methods and observe that the overall system performance is enhanced significantly. Since articulatory trajectories also contain speech content information, we also study the usage of inverted articulatory features in the text dependent speaker verification task. We demonstrate that the articulatory constraints introduced by the inverted articulatory features help to reject wrong password trials and improve the performance after score level fusion. Future work should include investigating the effects of different speaker independent acoustic-to-articulatory mapping methods in terms of speaker verification performance with training data from multiple exemplar speakers, especially in the direction of highlighting inter-speaker variations.

Acknowledgements

This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), USA NIH (DC007124), USA NSF, USA Department of Justice and CMU-SYSU Collaborative Innovation Research Center.

References

Arora, R., Livescu, K., 2013. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. In: Proc. ICASSP, IEEE, pp. 7135–7139.

Bharadwaj, S., Arora, R., Livescu, K., Hasegawa-Johnson, M., 2012. Multiview acoustic feature learning using articulatory measurements. In: Intl. Workshop on Stat. Machine Learning for Speech Recognition.

Brunner, J., Fuchs, S., Perrier, P., 2005. The influence of the palate shape on articulatory token-to-token variability. ZAS Pap. Linguist. 42, 43–67.

Brunner, J., Fuchs, S., Perrier, P., et al., 2009. On the relationship between palate shape and articulatory behavior. J. Acoust. Soc. Am. 125, 3936–3949.

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Brunner, J., Hoole, P., Perrier, P., 2007. Articulatory optimisation in perturbed vowel articulation. In: Proc. International Congress of Phonetic Sciences, pp. 497–500.
Campbell, W., Sturim, D., Reynolds, D., 2006. Support vector machines using gmm supervectors for speaker verification. IEEE Signal Process. Lett. 13, 308–311.
Cumani, S., Brummer, N., Burget, L., Laface, P., 2011. Fast discriminative speaker verification in the i-vector space. In: Proc. ICASSP, IEEE, pp. 4852–4855.
Dart, S., 1991. Articulatory and acoustic properties of apical and laminal articulations. In: Maddieson, I. (Ed.), UCLA Working Papers in Phonetics. , p. 79.
Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-end factor analysis for speaker verification. IEEE Trans Audio Speech Lang. Process. 19, 788–798.
Dehak, N., Torres-Carrasquillo, P., Reynolds, D., Dehak, R., 2011. Language recognition via i-vectors and dimensionality reduction. In: Proc. INTERSPEECH, pp. 857–860.
D'Haro, L.F., Cordoba, R., Salamea, C., Echeverry, J.D., 2014. Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition. In: Proc. ICASSP, IEEE, pp. 5379–5383.
Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: Proc. ICASSP, pp. 346–348.
Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C., 2008. Liblinear: a library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874.
Fant, G., 1960. Acoustic Theory of Speech Production. Mouton & Co., The Hague.
Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proc. INTERSPEECH, pp. 249–252.
Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.
Ghosh, P., Narayanan, S., 2011. A subject-independent acoustic-to-articulatory inversion. In: Proc. ICASSP, pp. 4624–4627.
Ghosh, P.K., Narayanan, S., 2011b. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 130, 251–257.
Ghosh, P.K., Narayanan, S.S., 2010. A generalized smoothness criterion for acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 128, 2162–2172.
Ghosh, P.K., Narayanan, S.S., 2013. On smoothing articulatory trajectories obtained from Gaussian mixture model based acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 134, 258–264.
Hatch, A., Kajarekar, S., Stolcke, A., 2006. Within-class covariance normalization for SVM-based speaker recognition. In: Proc. INTERSPEECH, pp. 1471–1474.
Hébert, M., 2008. Text-dependent speaker recognition. In: Springer Handbook of Speech Processing., pp. 743–762.
Hinton, G., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.
Honda, M., Fujino, A., Kaburagi, T., 2002. Compensatory responses of articulators to unexpected perturbation of the palate shape. J. Phon. 30, 281–302.
Kenny, P., Stafylakis, T., Ouellet, P., Alam, M.J., 2014. JFA-based front ends for speaker recognition. In: Proc. ICASSP, pp. 1724–1728.
Kim, C., Stern, R.M., 2012. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proc. ICASSP, IEEE, pp. 4101–4104.
Kim, J., Lammert, A., Ghosh, P.K., Narayanan, S.S., 2013. Spatial and temporal alignment of multimodal human speech production data: realtime imaging, flesh point tracking and audio. In: Proc. ICASSP, pp. 3637–3641.
King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am. 121, 723–742.
Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. 52, 12–40.
Lammert, A., Proctor, M., Katsamanis, A., Narayanan, S., 2011. Morphological variation in the adult vocal tract: a modeling study of its potential acoustic impact. In: Proc. INTERSPEECH, pp. 2813–2816.
Lammert, A., Procto, M., Narayanan, S., 2013a. Interspeaker variability in hard palate morphology and vowel production. J. Speech Lang. Hear. Res. 56, 1924–1933.
Lammert, A., Proctor, M., Narayanan, S., 2013b. Morphological variation in the adult hard palate and posterior pharyngeal wall. J. Speech Lang. Hear. Res., 521–530.
Larcher, A., Lee, K.A., Ma, B., Li, H., 2014a. Imposture classification for text-dependent speaker verification. In: Proc. ICASSP, pp. 739–743.
Larcher, A., Lee, K.A., Ma, B., Li, H., 2014b. Text-dependent speaker verification: classifiers, databases and rsr2015. Speech Commun. 60, 56–77.
Lee, L., Rose, R.C., 1996. Speaker normalization using efficient frequency warping procedures. In: Proc. ICASSP, pp. 353–356.
Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children's speech: developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. 105, 1455–1468.
Lei, Y., Scheffer, N., Ferrer, L., McLaren, M., 2014. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In: Proc. ICASSP, pp. 1695–1699.
Leung, K.Y., Mak, M.W., Kung, S.Y., 2004. Applying articulatory features to telephone-based speaker verification. In: Proc. ICASSP, IEEE, pp. 85–88.
Li, M., Kim, J., Ghosh, P.K., Ramanarayanan, V., Narayanan, S., 2013. Automatic classification of palatal and pharyngeal wall shape categories from speech acoustics and inverted articulatory signals. In: SPASR.
Li, M., Liu, W., 2014. Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenization and tandem features. In: Proc. INTERSPEECH.
Li, M., Narayanan, S., 2014. Simplified supervised i-vector modeling with application to robust and efficient language identification and speaker verification. Comput. Speech Lang. 28, 940–958.

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003

Liu, W., Yu, Z., Li, M., 2014. An iterative framework for unsupervised learning in the PLDA based speaker verification. In: Proc. ISCSLP, pp. 78–82.

Matejka, P., Glembek, O., Castaldo, F., Alam, M., Plchot, O., Kenny, P., Burget, L., Cernocky, J., 2011. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In: Proc. ICASSP, pp. 4828–4831.

Mooshammer, C., Perrier, P., Geng, C., Pape, D., 2004. An EMMA and EPG study on token-to-token variability. AIPUK 36, 47–63.

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.C., Zhu, Y., Goldstein, L., Byrd, D., Bresch, E., Ghosh, P., Katsamanis, A., Proctor, M., 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research. J. Acoust. Soc. Am. 136, 130701311.

NIST, 2010. NIST 2010 Speaker Recognition Evaluation Plan, www.itl.nist.gov/iad/mig/tests/spk/2010/index.html.

Novoselov, S., Pekhovsky, T., Shulipa, A., Sholokhov, A., 2014. Text-dependent GMM-JFA system for password based speaker verification. In: Proc. ICASSP, pp. 729–733.

Ozbek, I., Hasegawa-Johnson, M., Demirekler, M., 2011. Estimation of articulatory trajectories based on Gaussian mixture model (GMM) with audio-visual information fusion and dynamic Kalman smoothing. IEEE Trans. Audio Speech Lang. Process. 19, 1180–1195.

Özbek, I.Y., Hasegawa-Johnson, M., Demirekler, M., 2009. Formant trajectories for acoustic-to-articulatory inversion. In: Proc. INTERSPEECH, pp. 2807–2810.

Palm, R.B., 2012. Prediction As a Candidate for Learning Deep Hierarchical Models of Data. Technical University of Denmark, Informatics, Lyngby, Denmark (Master's thesis).

Pechyony, D., Vapnik, V., 2010. On the theory of learning with privileged information. Adv. Neural Inf. Process. Syst. 23.

Perkell, J., 1997. Articulatory processes. In: Hardcastle, W., Laver, J. (Eds.), The Handbook of Phonetic Sciences. Blackwell, Oxford, pp. 333–370.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of vowels. J. Acoust. Soc. Am. 24, 175–184.

Prince, S., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: Proc. ICCV, pp. 1–8.

Rajan, P., Afanasyev, A., Hautamäki, V., Kinnunen, T., 2014. From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification. Digital Signal Process. 31, 93–101.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Process. 10, 19–41.

Richmond, K., 2011. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In: Interspeech, Florence, Italy, pp. 1505–1508.

Rudzicz, F., 2010. Adaptive kernel canonical correlation analysis for estimation of task dynamics from acoustics. In: Proc. ICASSP, pp. 4198–4201.

Shao, Y., Wang, D., 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In: Proc. ICASSP, pp. 1589–1592.

Silva, J., Rangarajan, V., Rozgic, V., Narayanan, S., 2007. Information theoretic analysis of direct articulatory measurements for phonetic discrimination. In: Proc. ICASSP, IEEE, pp. IV–457.

Siniscalchi, S.M., Lyu, D.C., Svendsen, T., Lee, C.H., 2012. Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data. IEEE Trans. Audio Speech Lang. Process. 20, 875–887.

Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.H., 2013. Universal attribute characterization of spoken languages for automatic spoken language recognition. Comput. Speech Lang. 27, 209–227.

Stevens, K., 1998. Acoustic Phonetics. MIT Press, Cambridge, MA.

Thibeault, M., Ménard, L., Baum, S., Richard, G., McFarland, D., 2011. Articulatory and acoustic adaptation to palatal perturbation. J. Acoust. Soc. Am. 129, 2112–2120.

Toutios, A., Margaritis, K., 2003. A rough guide to the acoustic-to-articulatory inversion of speech. In: 6th Hellenic European Conference of Computer Mathematics and its Applications, HERCMA-2003.

Uria, B., Renals, S., Richmond, K., Uria, B., Murray, I., Renals, S., Richmond, K., 2011. A deep neural network for acoustic-articulatory speech inversion. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. In: Proc. ICASSP, pp. 4080–4084.

Vinyals, O., Jia, Y., Deng, L., Darrell, T., 2012. Learning with recursive perceptual representations. Adv. Neural Inf. Process. Syst., 2834–2842.

Wang, H., Leung, C.C., Lee, T., Ma, B., Li, H., 2013. Shifted-delta MLP features for spoken language recognition. IEEE Signal Process. Lett. 20, 15–18.

Wang, J., Johnson, M.T., 2014. Physiologically-motivated feature extraction for speaker verification. In: Proc. ICASSP, IEEE, pp. 1709–1713.

Westbury, J., Milenkovic, P., Weismer, G., Kent, R., 1990. X-ray microbeam speech production database. J. Acoust. Soc. Am. 88, S56.

Wrench, A., 1999. MOCHA-TIMIT Speech Database. Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh.

Zhang, S.X., Mak, M.W., Meng, H.M., 2007. Speaker verification via high-level feature based phonetic-class pronunciation modeling. IEEE Trans. Comput. 56, 1189–1198.

Zhuang, X., Nam, H., Hasegawa-Johnson, M., Goldstein, L., Saltzman, E., 2009. Articulatory phonological code for word classification. In: Proc. INTERSPEECH, pp. 2763–2766.

Please cite this article in press as: Li, M., et al., Speaker verification based on the fusion of speech acoustics and inverted articulatory signals. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.05.003