INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speaker verification using short utterances with DNN-based estimation of subglottal acoustic features

Jinxi Guo[1], *Gary Yeung*[1], *Deepak Muralidharan*[1], *Harish Arsikere*[2], *Amber Afshan*[1], *Abeer Alwan*[1]

[1] Dept. of Electrical Engineering, University of California, Los Angeles, CA 90095, USA
[2]Data Analytics Lab, Xerox Research Center-India (XRCI), Bangalore, Karnataka, India

lennyguo@g.ucla.edu, garyyeung@g.ucla.edu, deepakm2308@ucla.edu, harish.asikere@xerox.com, amberafshan@ucla.edu, alwan@ee.ucla.edu,

Abstract

Speaker verification in real-world applications sometimes deals with limited duration of enrollment and/or test data. MFCC-based i-vector systems have defined the state-of-the-art for speaker verification, but it is well known that they are less effective with short utterances. To address this issue, we propose a method to leverage the speaker specificity and stationarity of subglottal acoustics. First, we present a deep neural network (DNN) based approach to estimate subglottal features from speech signals. The approach involves training a DNN-regression model that maps the log filter-bank coefficients of a given speech signal to those of its corresponding subglottal signal. Cross-validation experiments on the WashU-UCLA corpus (which contains parallel recordings of speech and subglottal acoustics) show the effectiveness of our DNN-based estimation algorithm. The average correlation coefficient between the actual and estimated subglottal filter-bank coefficients is 0.9. A score-level fusion of MFCC and subglottal-feature systems in the i-vector PLDA framework yields statistically-significant improvements over the MFCC-only baseline. On the NIST SRE 08 truncated 10sec-10sec and 5sec-5sec core evaluation tasks, the relative reduction in equal error rate ranges between 6 and 14% for the conditions tested with both microphone and telephone speech.

Index Terms: speaker verification, short utterances, subglottal acoustic features, deep neural networks.

1. Introduction

Factor analysis based i-vector framework has defined the state-of-the-art in speaker verification [1]. However, the performance degrades rapidly as the available amount of enrollment and/or testing speech decreases [2, 3]. The significant amount of speech data required for speaker enrollment and verification is a limitation for everyday applications. To address this issue, several approaches have been studied. In [4], the authors used a method to quantify the uncertainty associated with the i-vector extraction process and propagated it to a probabilistic linear discriminant analysis (PLDA) classifier. In [5], the effect of short utterance i-vectors was analyzed, and it was found that duration variability can be modeled as additive noise in the i-vector space. In [6], several techniques to attenuate the effect of the short utterance variance were proposed.

While the majority of the techniques have focused on i-vector compensation, not many studies have focused on the role of feature extraction. Previous research shows that significant variations of i-vectors can occur when the utterance lengths are varied because of changes in acoustic properties of various sounds and number of unique phonemes [5, 6]. Standard speech features like MFCCs have large variations for different phonemes, but relatively stationary features may help reduce the variation of the i-vectors. Our previous research indicates that subglottal acoustics (capturing the acoustics of the tracheo-bronchial airways) are speaker specific and the spectral characteristics are much less variable than the spectral characteristics of the speech waveform [7, 8].

To record this subglottal acoustic data, a noninvasive accelerometer is generally used [9]. In the past, we studied the properties and applications of subglottal acoustic features, including automatic estimation of the subglottal resonances (SGRs) for speaker height estimation and adaptation [10, 11, 12] and estimating subglottal features from MFCCs for GMM-based speaker recognition [13]. However, no research has been done to find a good representation of subglottal features for the state-of-the-art i-vector/PLDA framework and to show their effects on the standard speaker verification datasets.

Motivated by this, we investigate the utility of the subglottal acoustic features for i-vector speaker verification using short utterances. The focus is specifically on the estimation of subglottal acoustic features using a DNN-based spectral feature mapping model, and their ability to discriminate between speakers and improve the performance for standard short utterance speaker verification tasks.

In Section 2, we describe the proposed system. Section 3 presents the estimation algorithm of the subglottal acoustic features and evaluation results. Section 4 describes the speaker verification experimental setup and results, and Section 5 concludes the paper.

2. System description

We first train a DNN model to map speech spectral features to their corresponding subglottal features and then use the trained DNN model as the feature extractor for a speaker verification experiment. A score-level method is used to fuse the information provided by MFCCs and estimated subglottal features under the i-vector/PLDA framework. Figure 1 shows the system overview and the implementation details are provided in Section 3 and 4.

The i-vector system aims at modeling the overall variability of the training data and compressing the speaker information to a low-dimensional vector. The main idea is that the speaker- and channel-dependent GMM supervector s can be modeled as:

$$s = m + Tw \quad (1)$$

where m is the UBM GMM mean supervector, T is a low-rank matrix representing the total variability space, and w is a random vector following a standard normal distribution. The i-

Work supported in part by NSF Grant No. 0905381.

http://dx.doi.org/10.21437/Interspeech.2016-282

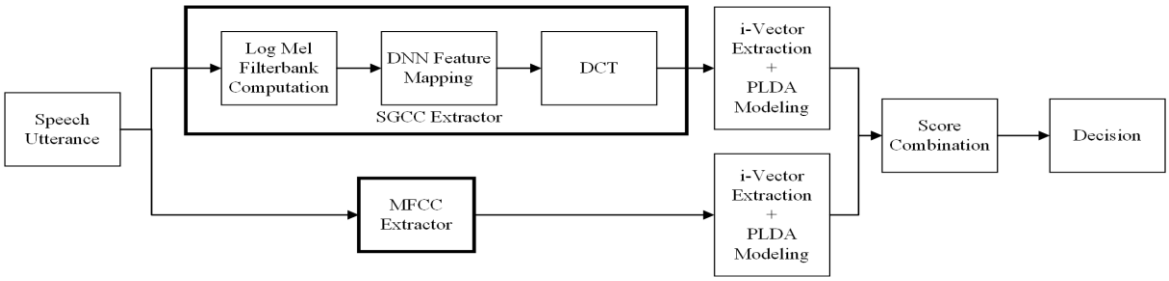

Figure 1: Block diagram of the proposed framework.

vector is the MAP point estimate of w, and T is referred to as the i-vector extractor. Using the PLDA model, the log-likelihood ratio can be computed for the hypothesis test on whether the two i-vectors are generated by the same speaker.

3. Estimating subglottal acoustic features

DNNs have been shown to be effective for feature mapping of speech signals [14, 15]. We adopt the approach here for subglottal feature estimation and evaluate it on the WashU-UCLA corpus [8] (which contains time-synchronized recordings of speech and subglottal acoustics).

We train a DNN regression model to learn the spectral feature mapping from speech to subglottal acoustics. The objective function for optimization is based on the mean square error. Eq. (2) is the cost function for each training sample:

$$\mathcal{L}(y, x; \mathrm{w}) = \sum_{k=1}^{K} \left(y_k - f_k(x)\right)^2 \qquad (2)$$

where y_k and $f_k(\cdot)$ are the desired and the actual output of the kth neuron in the output layer, respectively, and w denotes the weights to be learned during training.

The trained DNN regression model provides a mapping from a more variable speech spectral domain to the less-variable subglottal spectral domain (viewed in some sense as a many-to-one mapping).

3.1. Implementation details and evaluation setup

The WashU-UCLA corpus consists of 35 monosyllables (14 "hVd" and 21 "CVd" words, where V includes all the AE monophthongs and diphthongs) in a phonetically neutral carrier phrase ("I said a __ again"), with 10 repetitions of each word by each speaker. The corpus has simultaneous microphone and (subglottal) accelerometer recordings of 25 adult male and 25 adult female speakers of American English, and in total 17500 individual microphone(and accelerometer) waveforms. To avoid redundancy and keep the phonetic balance in the data that is used to train the DNN regression model, only the vowel segments of the monosyllables are isolated and used. Another reason why we only extract the vowel segments is that the accelerometer signals show little information for consonants. Since we only have the DNN mapping for vowels, we need a way to deal with non-vowel segments while estimating subglottal acoustic features for the speaker verification experiment. Section 4 explains the specific method used.

We extract the 40 dimensional log Mel-filterbank coefficients for both speech and accelerometer segments, and use the filterbank feature vectors of the speech segments as input and their corresponding subglottal filterbank feature vectors as output for the DNN model. The input and output features are normalized using the L2 norm of the feature vector. The activation functions of both the hidden layers and the output layer are the *tanh* functions. Three hidden layers are used and each hidden layer has 256 neurons. We use backpropagation with mini-batch stochastic gradient descent to train the DNN model, and the optimization technique uses adaptive gradient descent along with a momentum term. The THEANO DNN toolkit is used for DNN training [16].

The DNN regression model is evaluated using 5 rounds of cross-validation. All available vowel segment pairs (17500 in total) are split into a training set and a validation set. The training set has roughly 80% of the data and the rest is for validation. All signals are down sampled to 8 kHz (from the original sampling rate of 48 kHz), which is consistent with the NIST SRE dataset. The log Mel-filterbank coefficients for both speech and subglottal acoustic signals are extracted at 10ms intervals using a 20 ms Hamming window.

3.2. Results

To evaluate the performance of the DNN-based estimation model, we use two methods: (1) computing the correlation between actual and estimated log Mel-filterbank coefficients for each frame of subglottal recordings, and (2) comparing the actual and estimated subglottal features with regards to their ability to discriminate between speakers.

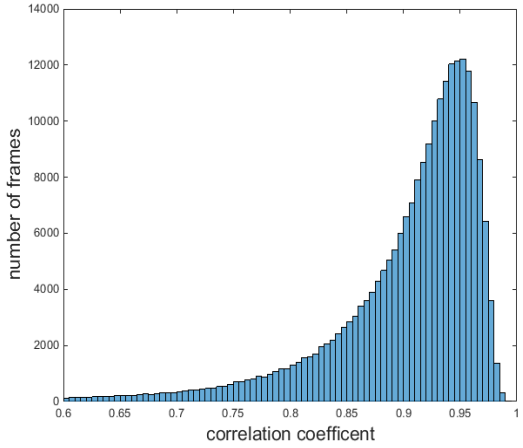

Figure 2: Histogram of the correlation coefficient of the actual and estimated subglottal Mel-filterbank coefficients for each frame in the validation dataset.

Figure 2 shows the histogram of the correlation coefficients for all frames in the validation dataset. The average value of the correlation coefficients for all 5 rounds of

cross-validation is 0.9, which indicates the sufficiency of DNN-based estimation model.

To compare the actual and estimated subglottal filterbank features in terms of speaker discriminability, the J-Ratio [17], which measures class separation, is used. Before calculating the J-Ratio, we compute DCT of the log Mel-filterbank coefficients, since it will decorrelate the filterbank features and be consistent with MFCC features. We refer to the subglottal features after taking DCT on the log Mel-filterbank coefficients as subglottal cepstral coefficients, which is denoted as SGCCs. The zeroth cepstral coefficient is discarded and the first 20 coefficients are used for both MFCCs and SGCCs. Given feature vectors for N speakers, the J-Ratio can be computed using Eqs. (3), (4) and (5):

$$S_w = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{3}$$

$$S_b = \frac{1}{N}\sum_{i=1}^{N} (M_i - M_o)(M_i - M_o)^T \tag{4}$$

$$J = \text{trace}\{(S_b + S_w)^{-1} S_b\} \tag{5}$$

where S_w is the within-class scatter matrix, S_b is the between-class scatter matrix, M_i is the mean vector for the ith speaker, M_o is the mean of all M_is, and R_i is the covariance matrix for the i th speaker (note that a higher J-Ratio means better separation).

Feature sets	J-Ratio
MFCCs(x_1-x_{20})	4.92
Actual SGCCs(x_1-x_{20})	5.48
Estimated SGCCs(x_1-x_{20})	5.47

Table 1: J-ratio, a measure of class separation for different features. Features were extracted from isolated vowel recordings of speech and subglottal acoustics, for all the 50 speakers in the WashU-UCLA corpus.

Table 1 shows the J-Ratio value for different feature sets. The results show that: (1) SGCCs offer better separation compared with MFCCs, which is partly attributable to the stationarity of subglottal acoustics and the low within-class variance that results from it, and (2) the estimated SGCCs are similar in performance to actual SGCCs, which is due to the great effectiveness of the DNN-based feature mapping model.

4. Speaker verification experiments

4.1. Task description and experimental settings

We evaluate our features and proposed system on the NIST SRE dataset under the state-of-the-art i-vector/PLDA framework.

The NIST SRE 2004, 2005, 2006 and Switchboard II datasets are used as development dataset. Gender dependent universal background models (UBM) with 2048 Gaussians are trained using a subset of the development dataset, which only has utterances from male speakers. The total variability subspace for i-vector extractor, channel compensation technique LDA and speaker variability subspace for PLDA are trained using all the male speakers from the development dataset. The Kaldi toolkit [18] is used to build the system.

MFCCs using first 20 coefficients (discarding the zeroth coefficient) with appended first and second order derivatives are extracted from the detected speech segments after voice activity detection. A 20 ms Hamming window, a 10 ms frame shift, and a 23-channel Mel-filterbank are used for baseline MFCC feature extraction. A total variability matrix T of 400 factors is used and the dimension is reduced to 200 using LDA before the PLDA modeling. Length normalization of the i-vectors is also used.

For SGCC feature extraction, non-vowel speech frames must be discarded since the DNN feature extractor is trained only on isolated vowels. A normalized autocorrelation peak value of 0.7 is used as a threshold to detect the strongly-voiced vowel frames. A 20 ms Hamming window and 10 ms frame shift are used to extract 40-channel Mel-filterbank coefficients from the voiced frames. Then, the filterbank coefficients are fed into the trained DNN feature extractor to extract the estimated subglottal features. The first 20 coefficients (excluding the zeroth coefficient) with appended first order derivatives are selected after taking the DCT on the estimated subglottal Mel-filterbank coefficients. A total variability subspace of 150 dimensions is used and the same number of latent components is adopted for PLDA modeling. Length normalization is also done here.

The NIST SRE 2008 core task [19], which has both microphone and telephone speech and channel matched and mismatched conditions, was used for the experiments. The enrollment and testing dataset are truncated to 10 seconds and 5 seconds for each utterance for our short-utterance speaker verification tasks. The core task contains 1993 female and 1270 male speakers. Only the male speakers with 39433 test trials are used here for evaluation.

Given an utterance, MFCCs and SGCCs are computed as described above. Each feature set will generate a set of scores for test trials. Scores from the two speaker verification systems were normalized to the range (0, 1) and are fused in a linear weighted fashion such that the weights sum up to 1. The fused scores are used to make decision.

4.2. Results and analysis

The fused MFCCs+SGCCs system gives improvement for almost all conditions for the truncated core task. The gains are higher and more significant for the conditions that better match the characteristics of the WashU-UCLA corpus used for DNN training. Therefore, we show the results for conditions C2 (interview speech from the same microphone types for both training and testing), C7 (English telephone speech), and C8 (English telephone speech spoken by native U.S. English speakers) in Table 2. Both equal error rate (EER) and minimum detection cost function (DCF) are used for evaluation.

The combined system yields the biggest improvement under matched microphone speech (C2), with a relative 11.5% EER reduction for the 10sec-10sec task, and 14.3% for the 5sec-5sec task. This may be due to the fact that the DNN feature extraction model is also trained under matched microphone speech. For English telephone speech, we can see that C8, which contains utterances spoken by Native American English speakers, gives relative better improvement compared with C7. This may also result from the fact that all the speakers in the WashU-UCLA dataset are native US speakers. The weights used for fusion are the same for both 10sec-10sec and 5sec-5sec tasks, which are 0.85 for MFCCs and 0.15 for SGCCs.

Conditions	Feature set	C2		C7		C8	
		EER	DCF	EER	DCF	EER	DCF
10sec-10sec	MFCCs	8.12	0.0255	19.51	0.0739	21.08	0.0783
	MFCCs+SGCCs	7.20	0.0240	18.24	0.0726	19.29	0.0759
	Relative improvement	11.5%		6.5%		8.5%	
5sec-5sec	MFCCs	14.11	0.0458	27.76	0.0955	27.83	0.0954
	MFCCs+SGCCs	12.10	0.0437	26.21	0.0945	26.07	0.0939
	Relative improvement	14.3%		5.6%		6.3%	

Table 2: EER and DCF for the MFCC baseline system and the proposed system on the NIST SRE 08 truncated 10sec-10sec and 5sec-5sec evaluation tasks. The relative improvement of EER is also shown.

4.3. Discussion

In order to show the relative stationarity property of subglottal acoustic features, we conduct an experiment to show performance degradation of the speaker verification experiments for both feature sets, when the length of the utterances for both enrollment and testing data were reduced from 30 to 10 seconds.

From Table 3, the average absolute EER increase for the selected conditions is 12.5% for MFCCs, and 3.5% for SGCCs, which indicates less variation of SGCCs and their robustness to short utterances.

	Absolute EER change from 30sec-10sec
MFCCs	12.5%
SGCCs	3.5%

Table 3: Absolute EER change between the 30-second condition and 10-second condition for MFCCs and SGCCs

While the J-Ratio analysis in Sec 3.2 shows that the estimated SGCCs can provide better speaker separation than MFCCs using the selected vowel segments, the SGCC-only system performs worse than the MFCC baseline on the NIST SRE dataset. This discrepancy could be due to (1) acoustic mismatch between the WashU-UCLA corpus and the speaker verification corpora, and (2) only the strongly-voiced vowel-like frames are selected for SGCC estimation.

5. Conclusion

In this paper, a DNN-based regression model is proposed to estimate the subglottal acoustic features from the speech signals. The results on the evaluation dataset show the efficacy of the estimation algorithm. The estimated features can provide improved speaker-verification performance when combined with conventional MFCC features at the score level, using the NIST SRE dataset with short utterances.

6. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, Mar. 2011.

[2] Vogt, R. J., Baker, B. J., & Sridharan, S.. "Factor analysis subspace estimation for speaker verification with short utterances." *Interspeech*, 2008.

[3] Kanagasundaram A, Vogt R, Dean D B, et al. "I-vector based speaker recognition on short utterances." in *Proc. of Interspeech* ,2011, pp. 2341-2344.

[4] Kenny P, Stafylakis T, Ouellet P, et al. "PLDA for speaker verification with utterances of arbitrary duration." *ICASSP*, 2013, pp.7649-7653.

[5] Hasan T, Saeidi R, Hansen J H L, et al. "Duration mismatch compensation for i-vector based speaker recognition systems" *ICASSP*, 2013, pp.7663-7667.

[6] Kanagasundaram A, Dean D, Sridharan S, et al. "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques." *Speech Communication*, 2014, 59: 69-82.

[7] S. M. Lulich, "Subglottal resonances and distinctive features," *Journal of Phonetics*, vol. 38, pp. 20-32, 2010.

[8] Lulich, S. M., Morton, J. R., Arsikere, H., Sommers, M. S., Leung, G. K., and Alwan, A., "Subglottal resonances of adult male and female native speakers of American English", *The Journal of the Acoustical Society of America*, 132(4): 2592-2602, 2012.

[9] Alwan, Abeer, Steven Lulich, and Mitchell Sommers. *The Subglottal Resonances Database LDC2015S03*. Hard Drive. Philadelphia: Linguistic Data Consortium, 2015

[10] Arsikere, H., Leung, G. K. F., Lulich, S. M., and Alwan, A. "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation", *Speech Commun*, 2013, 55(1): 51-70.

[11] H. Arsikere, S. M. Lulich, and A. Alwan. "Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency." *ICASSP*. 2013.

[12] Guo J., Paturi R., Yeung G., et al. "Age-dependent height estimation and speaker normalization for children's speech using the first three subglottal resonances." *Interspeech*. 2015.

[13] Harish Arsikere, H.A. Gupta and Abeer Alwan, "Speaker recognition via fusion of subglottal features and MFCCs," *Interspeech.* 2014, pp. 1106-1110.

[14] Han K., He Y., Bagchi D., et al. "Deep neural network based spectral feature mapping for robust speech recognition." *Interspeech*. 2015, pp. 2484-2488.

[15] Himawan, I., Motlicek, P., Imseng, D., Potard, B., Kim, N., & Lee, J. "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition." *ICASSP*. 2015, pp. 4540-4544.

[16] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley and Y. Bengio. "Theano: new features and speed improvements". *NIPS* 2012 deep learning workshop.

[17] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic Press, 1990

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The Kaldi speech recognition toolkit," in *Proc. of ASRU*, 2011, pp. 1–4.

[19] "The NIST year 2008 speaker recognition evaluation plan," tech. rep., NIST, April 2008.