EXPLORING UNIVERSAL SPEECH ATTRIBUTES FOR SPEAKER VERIFICATION

Sheng Zhang[1]*, Wu Guo*[1]*, Guoping Hu*[2]

[1]National Engineering Laboratory of Speech and Language Information Processing,
University of Science and Technology of China, Hefei, P.R.China
[2]Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, Hefei, China
zs1234@mail.ustc.edu.cn, guowu@ustc.edu.cn, gphu@iflytek.com

ABSTRACT

The universal speech attributes for speaker verification (SV) are addressed in this paper. The aim of this work is to exploit fundamental characteristics across different speakers within the deep neural network (DNN)/i-vector framework. The manner and place of articulation form the fundamental speech attribute unit inventory, and new attribute units for acoustic modelling are generated by a two-step automatic clustering method in this paper. The DNN based on universal attribute units is used to generate posterior probability in total variability modelling and i-vector extracting for the speaker recognition procedure. Furthermore, Gaussian mixture models (GMMs) are used to fit the distribution of the features associated with a given context-dependent attribute unit to improve performance. The experiments are carried out on the core test from the NIST SRE 2008 corpus; the proposed system can obtain better performance than all other state-of-the-art systems.

Index Terms— Speaker verification, DNN, universal speech attributes

1. INTRODUCTION

In recent years, i-vector [1] based speaker verification systems have become very popular because of their good performance and ability to compensate for channel variations. The i-vector algorithm provides a method to map a speech utterance to a low dimensional vector while retaining the speaker identity. Within this i-vector space, variability compensation methods such as linear discriminant analysis (LDA) [2] and within-class covariance normalization (WCCN) [3] are performed to reduce channel variability. Until now, the best performance has been obtained by modelling i-vector distributions through a generative model known as probabilistic linear discriminant analysis (PLDA) [4, 5, 6], which is adopted as a backend classifier in this paper.

DNNs have clearly shown their superiority over GMMs for automatic speech recognition (ASR) [7, 8]. Methods combining recent advances in DNNs with speaker verification have attracted researchers' attention [9, 10, 11, 12]. In [13, 14], a generalized i-vector framework is proposed in which the decision tree senones (tied triphone states) of a DNN model in the ASR system are employed to generate posterior probabilities rather than the unsupervised GMM-universal background model (UBM). In [15], S. Cumani *et al.* analysed the benefits of using different settings for the number of the DNN output states and for the number of Gaussians per DNN state, and they achieved obvious improvement over [14]. We adopted similar techniques in this paper.

There is growing interest in exploiting the discriminative properties of universal speech attributes in speech processing [16, 17, 18, 19, 20, 21], especially in language recognition. In paper [22], we first investigated universal speech attributes for speaker recognition. The DNN/i-vector framework is adopted in speaker modelling. The difference between phoneme-based systems and the proposed system in paper [22] is that universal attribute units are used to replace phonemes in DNN acoustic modelling. Universal attribute units are generated by referring to the English phoneme set. Comparative results with the phoneme-based DNN/i-vector system have been obtained in [22].

One limitation of the phoneme-based framework is lack of universal characterization across different speakers. In this paper, we improve the speaker recognition system in two aspects: the more elaborate generation of universal attribute units and the use of GMMs to fit the distribution of features associated with a given context-dependent attribute unit in total variability modelling. New attribute units are generated by a two-step automatic clustering method. The first step is the same as in paper [22]. The manner and place of articulation are first combined to generate enough universal speech attribute units, and triple attribute states are tied in content-dependent acoustic modelling. In the second step, different tied triple states are merged to new speech attribute units by means of automatic clustering in accordance with a likelihood calculation. In total variability modelling, we adopted the above hybrid DNN/GMM system, balancing attributive and acoustic precision to verify the

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

effectiveness of universal speech attributes for speaker verification [15].

The remainder of this paper is organized as follows. In section 2, we summarize the DNN/GMM approach. In section 3, we describe how to obtain effective universal speech attribute units. In section 4, we present results using attribute-based DNN/GMM systems on the NIST SRE 2008 corpus. Finally, we conclude our paper in section 5.

2. DNN/I-VECTOR

2.1. DNN/i-vector Framework

In the i-vector framework, each utterance is represented by its zeroth- and first-order Baum-Welch statistics extracted with UBM. In paper [14], Y. Lei *et al.* made an important modification to estimate the statistics. They adopted an ASR DNN model to generate the zeroth-order statistics of feature vector o_t. In the DNN/i-vector framework, UBM can be trained in a supervised fashion. In this paper, we use a DNN/i-vector framework similar to [14]. The only difference is our replacement of phoneme-based DNN with the proposed attribute-based DNN. The flowchart of the attribute-based DNN/i-vector system is shown in Fig. 1.

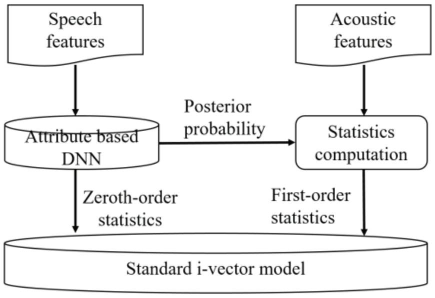

Figure 1. The flow diagram of DNN/i-vector framework attributes

2.2. HYBRID DNN/GMM

As noted in paper [15], DNNs are trained discriminatively, and the separation surface among its states is more complex than that provided by a single Gaussian. For these reasons, the effects of the granularity of the attributive DNN model and of the precision of the corresponding GMM models should be taken into consideration. To increase the accuracy of the distribution of the acoustic, S. Cumani *et al.* proposed to augment the number of Gaussians per merged state [15]. By combining the supervised and unsupervised methods, the balance of the attributive and acoustic precision is achieved. GMMs are also used to fit the distribution of features associated with a given context-dependent attribute unit in total variability modelling in this paper.

3. SPEECH ATTRIBUTES-BASED SV SYSTEMS

3.1. Universal speech attributes

The set of universal speech attributes is listed in Table 1 and include the place and manner of articulation [17]. The number of manners and places of articulation are 11 and 10, respectively, which are much fewer than the phoneme set (approximately 46 in English ASR) in the conventional phonetic LVCSR system. In LVCSR acoustic model training, context-dependent (CD) models are always adopted to improve recognition accuracy. Even when context-dependent models are used, the number of attribute units is not sufficient for good recognition performance. Accordingly, it is unwise to separately use place and manner of articulation in acoustic models for speaker verification systems. To increase the accuracy of the modelling units, we propose to generate attribute units with the following two steps: firstly combining place and manner of articulation directly, and secondly generating speech attribute units by automatic clustering.

Table 1. Universal Speech Attributes list for manner and place of articulation

manner	affricate, fricative, nasal, vowel, voice-stop, unvoiced-stop, glide, liquid, diphthong, sibilant
place	alveolar, alveo-palatal, dental, glottal, high, bilabial, labio-dental, low, mid, palatal, velar

3.1.1. Combine Place and Manner of articulation directly

The place and manner of articulation are combined to increase the number of attribute units and to take advantage of both in representing the pronunciation habits of speakers. Because there is a direct mapping between phonemes and attribute units, we can use phonemes to generate attribute units. We look up the corresponding place and manner of articulation of a phoneme. If they are different from those of other phonemes, we define a new attribute unit. For example, the manner and place of phoneme /ah/ are /vowel/ and /mid/, respectively, so we define a new attributes unit /mid_vowel/. The English phone set is used in our experiments, and 23 universal speech attribute units are obtained by combining place and manner of articulation. For convenience, we call these units CPMs.

3.1.2. Generate New Speech Attribute units by Automatic Clustering

CPMs can be used as acoustic modelling units in DNN training, but speaker recognition performance of the DNN/i-vector system based on CPMs is not satisfactory [22]. We aim to find better attribute modelling units, under which the acoustic model can be more accurate and speaker recognition accuracy can be improved. We use an automatic clustering approach to re-generate speech attribute units on the basis of CPMs. The detailed procedure is as follows.

1) The context-dependent hidden Markov model (HMM) based on CPMs is trained. We use tri-CPMs in the same way as tri-phones in conventional phoneme-based HMM systems, and more than 10^4 tri-CPMs are obtained by counting training transcriptions. In phoneme-based systems, there exist three or more HMM states for each phoneme. Because CPMs are only intermediate units in our experiment, it is unnecessary to model CPMs in the traditional way. Each CPM is modelled by only one HMM state. Our purpose is to cluster these tri-CPMs into universal attribute units through data clustering methods.
2) After force alignment, the statistics of tri-CPMs can be modelled by a Gaussian distribution. As the number of tri-CPMs is large, it is a time-consuming process to cluster these nodes using the pairwise merging process in the following step. A K-means algorithm is first used to cluster the large number of nodes into a pre-set number of clusters. The mean of each Gaussian is used as the input feature in K-means clustering. After K-means clustering, the statistics of the same cluster are merged, and a Gaussian distribution is estimated for each merged cluster. After the K-means procedure, the number of clusters is reduced to 500 in our experiment.
3) After the K-means step, each cluster is modelled with a Gaussian distribution, where $\{\mu_k, \Sigma_k\}$ are the mean vector and covariance matrix of the *k*-th Gaussian component. If n_k is the number of the observation in *k*-th cluster, the log likelihood of the cluster would be

$$L_k = -\frac{1}{2} n_k \cdot \left[\log((2\pi)^d \cdot \| \Sigma \|) + 1 \right], \tag{1}$$

where *d* is the dimension of the feature vectors. Two clusters *j* and *k* are merged if

$$L_{j+k} - (L_j + L_k) \tag{2}$$

is minimum for all *j* and *k*, where L_{j+k} is the likelihood of the cluster formed by merging cluster *j* and cluster *k*. After this step, we now have *K*-1 clusters. This pairwise merging process is repeated until we have *I* clusters. We define these clusters generated by automatic clustering (CAC) as new attribute units.

3.2. CAC units-based acoustic model

The aforementioned CAC units are used in the following acoustic modelling exactly as phonemes are used in state-of-the-art ASR systems. Content-dependent modelling is adopted to improve performance. As we lack linguistic knowledge for CAC units, we cannot design a suitable question set for state tying. In this work, we generate a question set using the approach described in [23]. Tied triple states are obtained using decision trees. Standard HMM-GMM systems are used to generate the initial state alignments to train DNNs. Briefly, the training procedure of CAC units-based acoustic model is identical to that of the conventional phoneme based systems.

4. EXPERIMENTS AND RESULT ANALYSIS

4.1. Speaker verification system description

The experiments are carried out on common conditions 6, 7 and 8 of the NIST SRE 2008 database. The training and test conditions of these three common conditions are as follows.

- C6: All trials involving only telephone speech in training and test.
- C7: All trials involving only English language telephone speech in training and test.
- C8: All trials involving only English language telephone speech spoken by a native U.S. English speaker in training and test.

4.2. Acoustic model training based on CACs and phonemes

To obtain a fair comparison between CAC units-based DNN and conventional phoneme-based DNN, both HMM-GMM and HMM-DNN models are trained using approximately 300 hours of clean English telephone speech from Switchboard data sets. Except for the output layer, these two DNNs have identical architectures. The inputs of DNNs are 429-dimensional features, corresponding to 39-dimensional perceptual linear predictive (PLP) features within a context window of 11 (5+1+5) frames. There are 6 hidden layers with 2048 hidden units in each layer. To make the characteristics conform to a Gaussian distribution, the features are pre-processed with mean and variance normalization (MVN). The cross entropy criterion is used to train the DNN model.

4.3. System configurations for speaker recognition

The equal error rate (EER) and minimal detection cost function (DCF) are used to evaluate the performance of the systems.

PLP features are used in speaker recognition systems. Each speech signal is parameterized by the 13th order PLPs and their first and second derivatives. Further processing including relative spectral (RASTA) filtering, voice activity detector (VAD), cepstral mean subtraction (CMS) and Gaussianization are applied to all PLPs.

NIST SRE 2004, 2005, 2006 and switchboard corpora are used to train the UBMs. After the UBM is obtained, the conventional total variability matrix training and i-vector extraction procedures are performed. The total variability matrix with rank 400 is trained using NIST SRE databases before 2008. After extracting the i-vector, further processes including LDA, WCCN, whitening and length normalization are applied to improve performance. The PLDA algorithm is used as the backend classifier, where the sizes of speaker and channel matrices are 150 and 10, respectively.

4.4. DNN/i-vector

The first experiment is conducted to compare the unsupervised GMM-UBM/i-vector system with the supervised DNN/i-vector systems. The performance of the unsupervised GMM-UBM/i-vector system with 2048 components is shown in the first row of Table 2. The second row of Table 2 stands for the phoneme-based DNN/i-vector system, where the DNN model consists of 3992 states. The last two rows of Table 2 represent the CAC-based DNN/i-vector system, where the DNN model consists of 3979 states when $I = 50$ and 3996 states when $I = 80$.

Table 2. Experimental results for NIST SRE 2008 based on DNN/i-vector framework (EER% / minDCF08*1000)

Model	I	C6	C7	C8
acoustic GMM	--	**6.41/30.4**	2.87/15.8	2.64/14.3
phoneme DNN	--	6.50/31.9	2.04/**10.8**	1.81/10.0
attribute DNN	50	6.53/33.5	**1.90**/11.3	**1.67/9.89**
	80	6.65/34.2	2.01/11.8	**1.67**/9.97

Comparing the performance of these three systems, we can observe that the attributive- and phonetic-based DNN/i-vector systems achieve better performance than the acoustic GMM-based system on language matched conditions (i.e., C7 and C8). The unsupervised GMM system achieves better performance on the multilingual condition (i.e., C6). A reasonable explanation is that DNNs can provide more accurate posteriors than the unsupervised GMM on language matched conditions and vice versa. In addition, we adopt different clustering sizes of CAC units for comparison, such as $I = 50$, 80. Their performances are shown in Table 2. We can observe that the system with CAC number 50 is slightly better than the system with CAC number 80. Furthermore, compared to the phoneme-based system, a slight improvement is obtained by the CAC-based system on some conditions. This finding confirms that CAC is an effective presentation of universal speech attributes.

4.5. Hybrid DNN/GMM

To better analyse the contribution of the attributive information provided by the attributive DNN model with respect to the accuracy of acoustic GMMs, different settings for the number of the DNN output states and for the number of Gaussians per DNN state are compared.

In this section, automatic clustering in accordance with a likelihood calculation is adopted to reduce the number of states, which is analogous to [23]. For fair comparison, approximately 4000 DNN states (3996 CAC states and 3992 phoneme states) are merged to 256/128, while the number of Gaussians per state is set to 8/16, respectively. In the experiments, the product of number of DNN output states and the number of Gaussians per DNN state are set to 2048. Based on the experimental results in the previous section, the size of the CAC units is set to 50 in this section. The results of the CAC- and phoneme-based DNN/GMM systems are listed in Table 3.

Table 3. Experimental results for NIST SRE 2008 based on DNN/GMM framework. We reduce the number of DNN states by automatic clustering. (EER% / minDCF08*1000)

Model	DNN	GMM	C6	C7	C8
CAC	256	8	5.85/**26.9**	**2.00/10.4**	**1.47/8.59**
	128	16	**5.62**/27.3	2.09/10.9	1.64/8.66
phon-eme	256	8	5.98/28.6	2.06/11.1	1.66/8.84
	128	16	5.71/27.6	2.03/11.3	1.71/9.38

It is interesting to observe that the best performance in all conditions is obtained by CAC-based DNN/GMM systems. The results have proved our previous hypothesis that universal speech attributes are more fundamental across different speakers than phonemes for speaker verification. Compared to performance of DNN/i-vector systems in the previous section, obvious improvements are obtained. The reason is that the balance of attributive and acoustic precision can be more effective.

5. CONCLUSIONS

One limitation of the phoneme-based framework is lack of universal characterization across different speakers. This paper presents a method to generate CAC units that define universal speech attributes precisely using automatic clustering in accordance with K-means and likelihood calculations. In the hybrid DNN/GMM framework, the system balancing attributive and acoustic precision achieves better performance on the core conditions of NIST SRE 2008 than existing systems. Our experiment confirms that CACs are an effective representation of universal speech attributes and are more fundamental than phonemes. It benefits from the fact that universal speech attributes are more related to the pronunciation habits of a person than to speech content.

6. ACKNOWLEDGEMENT

This work was partially funded by The National Key Research and Development Program of China (Grant No.2016YFB100130300) and Natural Science Foundation of Anhui Province (Grant No. 1408085MKL78).

7. REFERENCES

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” IEEE Transactions on Audio Speech & Language Processing, vol. 19, no. 4, pp. 788 – 798, 2011.

[2] C. Bishop, “Pattern recognition and machine,” New York:Springer, 2006.

[3] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for svm-based speaker recognition.” in Interspeech, 2006.

[4] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel,”Plda for speaker verification with utterances of arbitrary duration,” in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7649 – 7653.

[5] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007, pp. 1 – 8.

[6] P. Kenny, “Bayesian speaker verification with heavy-tailed priors.” in Odyssey, 2010, p. 14.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” Signal Processing Magazine, IEEE, vol. 29, no. 6, pp. 82 – 97, 2012.

[8] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no. 1, pp. 30 – 42, 2012.

[9] M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak, and P. Dumouchel, “First attempt of boltzmann machines for speaker verification.” in Odyssey, 2012, pp. 117 – 121.

[10] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, “Preliminary investigation of boltzmann machine classifiers for speaker recognition.” in Odyssey, 2012, pp. 109 – 116.

[11] S. Yaman, J. W. Pelecanos, and R. Sarikaya, “Bottleneck features for speaker recognition.” in Odyssey, vol. 12, 2012, pp. 105 – 108.

[12] Vasilakakis V, Laface P, Cumani S. Speaker recognition by means of Deep Belief Networks [J]. Speaker Recognition, 2013.

[13] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” in Odyssey, 2014.

[14] Y. Lei, L. Ferrer, M. McLaren et al., “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 1695 – 1699.

[15] Cumani, Sandro, Pietro Laface, and Farzana Kulsoom. "Speaker recognition by means of acoustic and phonetically informed GMMs." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[16] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, “Exploiting context-dependency and acoustic resolution of universal speech attribute models in spoken language recognition,” in Eleventh Annual Conference of the International Speech Communication Association, 2010.

[17] ——, “Universal attribute characterization of spoken languages for automatic spoken language recognition,” Computer Speech & Language, vol. 27, no. 1, pp. 209 – 227, 2013.

[18] V. Hautamaki, S. M. Siniscalchi, H. Behravan, V. M. Salerno, and I. Kukanov, “Boosting universal speech attributes classification with deep neural network for foreign accent characterization,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[19] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, “Exploring universal attribute characterization of spoken languages for spoken language recognition.” in INTERSPEECH, 2009, pp. 168 – 171.

[20] Y. Wang, J. Du, L. Dai, and C.-H. Lee, “A fusion approach to spoken language identification based on combining multiple phone recognizers and speech attribute detectors,” in Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, 2014, pp. 158 – 162.

[21] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, “High-resolution acoustic modeling and compact language modeling of language universal speech attributes for spoken language identification,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[22] S. Zhang, W. Guo, G. Hu, “First Investigation of Universal Speech Attributes for Speaker Verification,” International Symposium on Chinese Spoken Language Processing. IEEE, 2016.

[23] R. Singh, B. Raj, and R. M. Stern, “Automatic clustering and generation of contextual questions for tied states in hidden markov models,” Proceedings of the Icassp Phonexi Az, vol. 1, pp. 117 – 120, 1999