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

ABSTRACT

✓ The universal speech attributes for speaker verification (SV) are addressed in this paper.

✓ 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.

✓ The hybrid DNN/GMM framework is used to improve performance.

SYSTEM FRAMEWORK

✓ DNN/i-vector framework

In the DNN/i-vector framework, The only difference is our replacement of phoneme-based DNN with the proposed attribute-based DNN.

✓ DNN/GMM framework

By clustering DNN output states and augmenting the number of Gaussians per merged state, the balance of the attributive and acoustic precision is achieved.

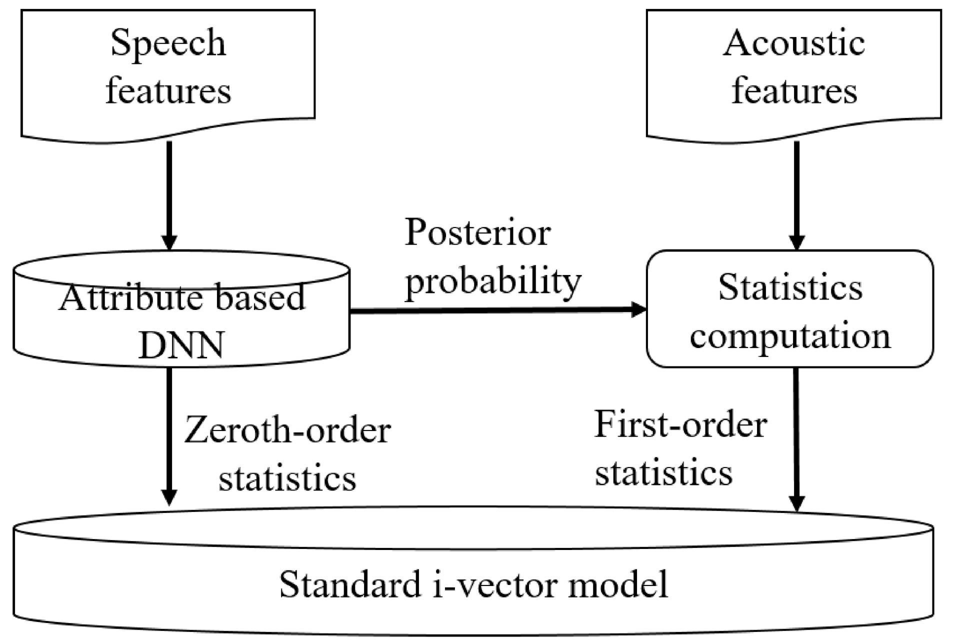

Figure 1. The flow diagram of DNN/i-vector framework attributes

SPEECH ATTRIBUTES-BASED SV SYSTEMS

✓ universal speech attributes

The set of universal speech attributes is listed in Table 1 and include the place and manner. Because the number of attribute units is not sufficient to train precise acoustic model, we propose to generate new attribute units in the following two steps.

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

✓ Combine Place and Manner of articulation directly(CPM)

The place and manner of articulation are combined to increase the number of attribute units. There is a direct mapping between phonemes and 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, we define a new unit /mid_vowel/. For English, 23 universal speech attribute units are obtained.

✓ New Speech Attribute units by Automatic Clustering(CAC)

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 first trained. In phoneme-based systems, there exist three or more HMM states for each phoneme. In our experiment, each tri-CPM is modelled by only one HMM state. Our purpose is to cluster these tri-CPMs into new universal attribute units.
2. The K-means algorithm is then used to cluster the large number of states into a pre-set number of clusters. The mean of each Gaussian is used as the input feature of K-means. 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. If n_k is the number of the observation in k-th cluster, the log-likelihood of the cluster would be $L_k = -\frac{1}{2} n_k[\log\left((2\pi)^d \|\Sigma\|\right) + 1]$, where d is the dimension of the feature vectors. Two clusters j and k are merged if $L_{j+k} - (L_j + L_k)$ is minimum for all clusters. This pairwise merging process is repeated until we have I clusters.

✓ CAC units-based acoustic model

As we lack linguistic knowledge for CAC units, the method of automatic clustering and generation of contextual questions is used. The training procedure of CAC units-based acoustic model is identical to that of the conventional phoneme based systems.

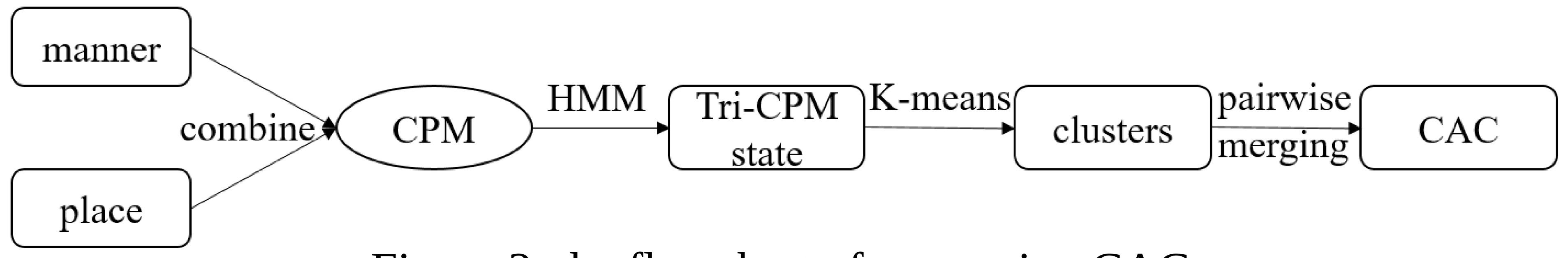

Figure 2. the flowchart of generating CAC

EXPERIMENTS

✓ Experimental setup

The experiments are carried out on common conditions 6, 7 and 8 of the NIST SRE 2008 database.

✓ DNN/i-vector

- The experiment is conducted to compare the unsupervised GMM-UBM/i-vector system with the supervised DNN/i-vector systems. The unsupervised GMM system achieves better performance on the multilingual condition (i.e., C6). Furthermore, compared to the phoneme-based system, a slight improvement is obtained by the CAC-based system on some conditions.
- In addition, we adopt different clustering sizes of CAC units for comparison.

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

✓ DNN/GMM

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. the product of number of DNN output states and the number of Gaussians per DNN state are set to 2048. The results have proved our previous hypothesis that universal speech attributes are more fundamental across different speakers than phonemes for speaker verification.

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