语音及语言信息处理国家工程实验室
National Engineering Laboratory for Speech and Language Information Processing

The 10th International Symposium on Chinese Spoken Language Processing

First Investigation of Universal Speech Attributes for Speaker Verification

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

1. University of Science and Technology of China
2. Ministry of Public Security

Abstract

- One limitation of the phoneme based DNN/i-vector framework is lack of universal acoustic characterization. Universal speech attributes are more related to the pronunciation habits of a person than the speech content.
- Two methods to generate the attribute units are proposed in this paper: one is that the manner and place of articulation are directly combined to generate more robust universal speech attribute units, and the other is that the different context-dependent speech attribute units are merged to a new speech attribute unit set by means of automatic clustering in accordance with likelihood calculation.
- The novel generated attribute units based system can achieve a better performance than that of the phoneme based system. Furthermore, the attribute based system has demonstrated a good complementarity with the GMM-UBM/i-vector and phoneme based DNN/i-vector systems.

CPM units

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
CPM	mid_vowel, alveo-palatal_affricate, alveolar_voice-stop, low_diphthong, palatal_glide, mid_diphthong, velar_unvoiced-stop, high_vowel, velar_voice-stop, alveo-palatal_sibilant, low_vowel, alveolar_unvoiced-stop, dental_fricative, labio-dental_fricative, alveolar_sibilant, high_diphthong, bilabial_voice-stop, bilabial_glide, alveolar_liquid, alveolar_nasal, bilabial_nasal, bilabial_nasal, velar_nasal, bilabial_unvoiced-stop, glottal_fricative

Universal speech attributes

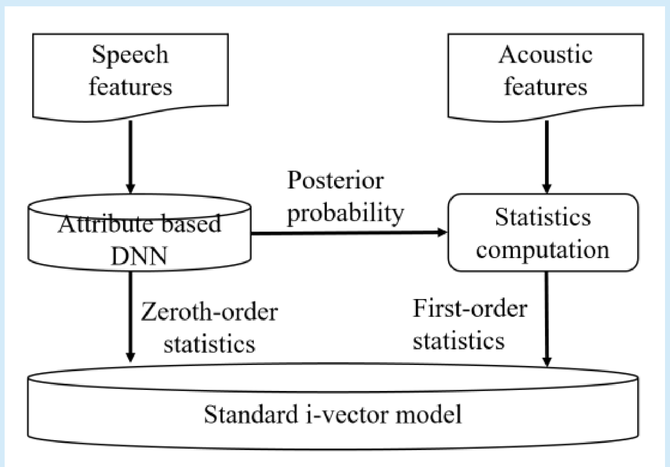
- The set of universal speech attributes is listed in the first two rows, which consists of the place and manner of articulation.
- The numbers of manner and the place of articulation are 11 and 10, respectively, which are much fewer than the phoneme set (approximately 40 in English ASR) in conventional LVCSR system.
- Accordingly, it is unwise to separately use the place and manner of articulation in acoustic model for SV systems.

The procedure of combining place and manner of articulation

- 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.
- 23 universal speech attribute units are obtained by combining the place and manner of articulation. For convenience, we call these units CPM.

DNN/i-vector Framework

The flow diagram of attribute based DNN/i-vector framework

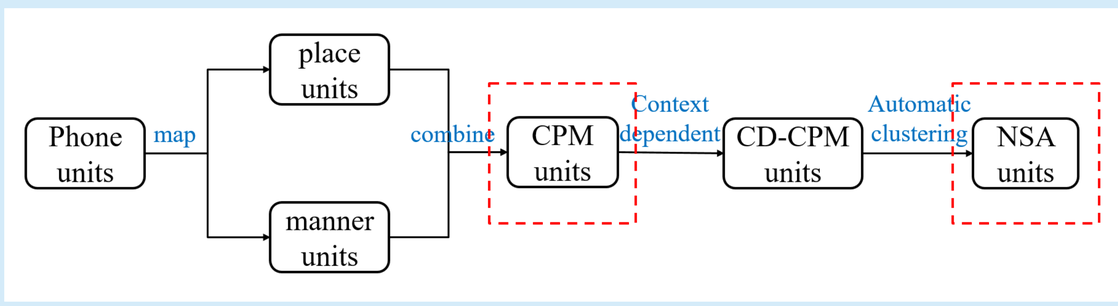


- In the GMM-UBM/i-vector framework, each utterance is represented by its zeroth- and first-order Baum-Welch statistics extracted with the unsupervised UBM
- The proposed attribute based DNN is used to replace the phoneme based DNN. The attribute based DNN is used to compute the posteriors for each frame.
- This new supervised UBM can replace the traditional unsupervised UBM to obtain the required statistics for the i-vector computation.

$$\gamma_{kt}^{(i)} \approx p(k|\boldsymbol{x}_t^{(i)})$$
$$N_k^{(i)} = \sum_t \gamma_{kt}^{(i)}$$
$$\boldsymbol{F}_k^{(i)} = \sum_t \gamma_{kt}^{(i)} \boldsymbol{x}_t^{(i)}$$

NSA units by Automatic Clustering

The flow diagram of generating NSA units



- the context-dependent hidden markov model (HMM) is trained firstly, where the speech arttribute units are set to single state.
- the number of new arttribute units genarated by automatic clustering based on likelihood calculation is set to 50 and 80 for comparison.

Experiments Results

Experimental results in NIST SRE (EER% / minDCF08*1000)

System description	Language/Unit number	female			male		
		C6	C7	C8	C6	C7	C8
S1: GMM-UBM/i-vector	–	**5.68/28.9**	2.54/12.8	2.91/13.3	**3.73/20.5**	1.65/9.04	1.09/5.47
S2: Phoneme based DNN	English / 42	6.33/33.6	**1.82**/10.5	1.93/10.5	5.09/22.1	**1.44**/8.22	0.64/3.82
S3: CPM based DNN	English / 23	6.79/34.2	1.93/11.1	2.16/10.9	5.12/22.6	2.03/7.47	0.64/4.26
S4: CPM based DNN	English+Mandarin / 23	6.38/33.7	1.91/10.6	1.93/11.0	5.11/21.2	1.55/8.10	0.87/3.61
S5: NSA based DNN	English / 50	6.28/32.7	1.84/11.1	1.95/10.7	4.99/22.6	1.88/**7.15**	0.62/3.83
S6: NSA based DNN	English / 80	6.33/34.1	1.93/10.5	2.17/10.8	4.71/22.3	1.69/7.58	0.78/**3.29**
S7: NSA based DNN	English+Mandarin / 50	6.32/33.7	1.84/10.5	1.87/11.1	5.11/22.3	1.88/8.47	0.83/6.12
S8: NSA based DNN	English+Mandarin / 80	6.21/33.6	1.90/**10.3**	**1.80/10.1**	5.21/21.8	2.11/7.76	**0.47**/4.93
Fusion:S1+S2+S3	–	5.16/27.9	1.71/9.02	1.82/**9.11**	3.50/19.1	1.35/7.64	0.57/3.49
Fusion:S1+S2+S4	–	5.21/27.5	1.73/9.21	1.79/9.17	3.50/**18.5**	**1.32**/7.26	0.57/2.85
Fusion:S1+S2+S5	–	**5.14**/27.8	1.76/9.31	1.80/9.70	3.53/18.9	1.48/7.51	0.57/2.85
Fusion:S1+S2+S6	–	5.35/27.7	1.76/9.37	1.87/9.24	3.49/19.1	1.42/**6.87**	**0.47**/2.85
Fusion:S1+S2+S7	–	5.37/**27.4**	1.65/**8.99**	1.76/9.28	**3.45**/18.8	1.40/7.23	0.62/3.60
Fusion:S1+S2+S8	–	5.31/27.5	**1.58/8.99**	**1.66/9.11**	3.48/18.7	1.43/6.89	0.57/**2.74**

- The CPM based system achieves comparable performance with the phoneme system
- The NSA based system achieves better performance on certain conditions
- the fusion of the attribute based system and other conventional systems can obtain an obvious improvement