Tsinghua-EE Telephone Speech Speaker Recognition System

Liu Jia

Department of Electronic Engineering, Tsinghua University, Beijing, China

Email: liuj@tsinghua.edu.cn

Our system contains 3 sub-systems. They are GMM-based system, SVM system and The Fusion System. The GMM-based System is our primary system.

1. GMM System

The system is an essential log-likelihood ratio detector with GMM-UBM models. Figure 1 shows the system architecture.

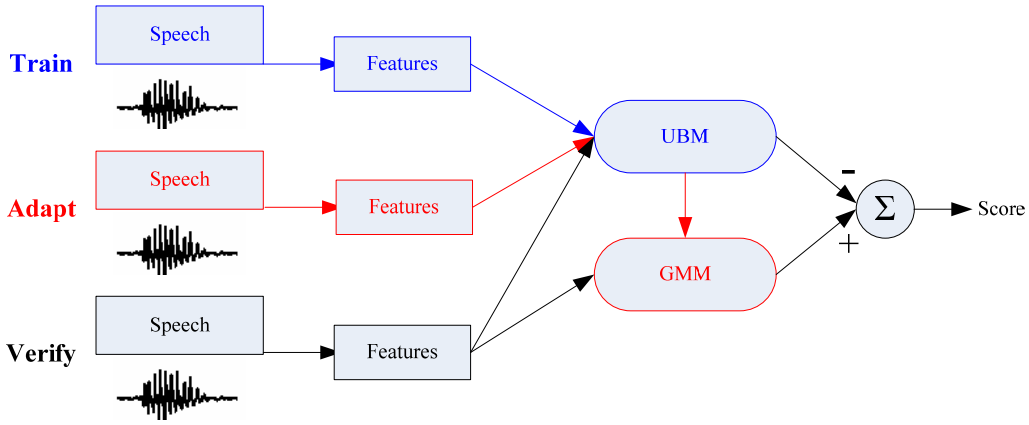

Figure 1 GMM system architecture

The whole system is processed by 3 steps. The first one is Train, the system will train a UBM model from long time speech; the second one is Adapt, the system will adapt the target speaker model from UBM with short-time speech; the third one is Verify, the system will get a match score from a segment of speech and a hypothesized speaker model, and give a decision with a fixed threshold.

The frond-end processing parts of the three steps are almost the same, the feature series will be extracted from the speech segment. Real speech is detected by two steps. The input speech is firstly processed by G.723.1 VAD (Voice Activity Detection) detector[1]; the detector is effective to remove long term silence. Then the active speech is processed frame-by-frame with a frame length of 20ms and the frame rate of 10ms. The bottom 20% frames sorted by energy are removed with a dynamic threshold. Then the 19 dimensions MFCC are extracted from each frame with a passband filter from 250 to 3450Hz, delta MFCC are also computed. Feature warping approach[2] is adopted to make the 38 dimensions vectors satisfied close to Gaussian distribution.

The feature series are used to train models in the Train and Adapt step, and to calculate the score with a hypothesized model in Verify step.

We use gender-dependent UBM models in the system. Each of them is trained from a 2-hours segment of speech, which is composed by the speech segments from the same amount speakers of each telephone handset type (Cordless, Cellular and Regular). The mixture number of the model is 2048.

In the Adapt step, the target speaker models are adapted from the UBM which has the same gender with them. EM algorithm is used to adapt just the means of the model[3]. Given a UBM and feature vectors, $X=\{x_1, \ldots, x_T\}$, we first determine the probabilistic alignment of the vectors into the UBM mixture components, for mixture i in the UBM, we compute

$$p(i/\boldsymbol{x}_t,\lambda)=\frac{C_i b_i(\boldsymbol{x}_t)}{\sum_{k=1}^{M} C_k b_k(\boldsymbol{x}_t)},$$

where C_i is the weight of the mixture, and $b_i(x)$ is the i-th Gaussian distribution. And then compute the sufficient statistics for the weight, mean, and variance parameters:

$$n_i=\sum_{t=1}^{T} p(i/\boldsymbol{x}_t),$$

$$E_i(\boldsymbol{X})=\frac{1}{n_i}\sum_{t=1}^{T} p(i/\boldsymbol{x}_t).$$

For the i-th mean, adapted parameter $\hat{\boldsymbol{\mu}}_i$ can be computed as follow:

$$\hat{\boldsymbol{\mu}}_i=\alpha_i^m E_i(\mathbf{X})+(1-\alpha_i^m)\boldsymbol{\mu}_i,$$

in the formulation,

$$\alpha_i^\rho=\frac{n_i}{n_i+\gamma^\rho}, \rho\in\{C,m,v\},$$

and $\alpha_i^C=\alpha_i^m=\alpha_i^v=n_i/(n_i+r)$, r=16.

In the Verify step, the LLR is computed with the UBM and a target model as follow:

$$S(\mathbf{X})=\frac{1}{T}\{\log p(\mathbf{X}/\lambda_{Tgt})-\log(p(\mathbf{X}/\lambda_{UBM})\},$$

The hypothesized speaker model was adapted from the UBM, so the fast scoring method[3] can be adopted.

2. SVM system

The SVM System is composed of the following parts. The system is implemented by using SVMTorch from IDIAP. The GLDS kernel[4] is implemented.

Feature Extraction is same as the part in GMM system. And in SVM model training, a typical 2 classes classifier is implemented. Both features from true speaker and background are assembled and sent to a wrapper. The wrapper then map the feature

sequence into a high dimension feature sequence using the traditional GLDS Kernel. The sequence is then sent to the SVM classifier and form a classification model. The model is composed of several support vectors and is then used for test task.
In the test, based on the SVM model and test feature sequence, the system calculates the corresponding scores of utterances. The scores show the likeliness of the test utterances. And final decision is then made according to the scores.

3. Fusion system

The linear fusion of the scores from GMM and SVM is used. The weight coefficients are calculated using Fisher Criteria.

4. Experiments on the development data

The results are shown in the Table 1, Table 2 and Table 3. The EER on the development sets is in Table 1. The fusion results are not very robust, because our SVM system is just finished, some parameters of system is not optimal. The absolute processing time and PC resources are in the Table 2 and Table 3.

Table 1 shows the experiments on the development data-NIST SRE'04 & 05 Corpus.

Corpus	GMM	SVM	Fusion
04Male	12.3	13.1	11.4
04Female	10.5	15.3	10.8
05Male	9.0	11.8	9.4
05Female	12.7	14.0	12.1
All Male	10.8	12.9	10.4
All Female	11.9	14.3	11.4

Note: In the Table 1 EER on the development data (%)

Table 2 shows the absolute processing time.

	Train	Detect
GMM System	5:44	86:58
SVM System	3:32	44:51
Fusion	9:16	131:49

Table 3 shows the PC condition.

CPU	Pentium4 2.8G
Memory	512M
Max Memorry Used	9.52M
HardDisk	250Gb/8M

Reference

[1] ITU. G.723.1 Annex A. Speech coders: Silence compression scheme. Geneva: ITU-T, Nov. 1996

[2] Pelecanos J, Sridharan S. Feature warping for robust speaker verification. Proceedings of Speaker Odyssey 2001 conference, June 2001

[3] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 2000, 10(1): 19-41

[4] Campbell W. Generalized linear discriminate sequence kernels for speaker recognition, Proceedings of ICASSP Proce edings Acoustics, Speech, and Signal Processing, 2002.: 161-164