USTC SSIP Laboratory
System Description
NIST SRE 2006

Yanlu Xie, Xi zhou, Zhiqiang Yao, Jixu Chen, Minghui Liu

1. Introduction

The USTC SSIP Laboratory submissions for the 2006 NIST Speaker Recognition Evaluation (SRE) is built upon five core sub-systems which are essentially GMM-UBM based systems. The main difference between the sub-systems is the feature used in them. Not all the five sub-systems are used in our submissions. The sub-systems are described in section2 and the submissions for different tasks are presented in section3.

2. Sub-systems

2.1 MFCC & GMM-UBM

As in previous SRE, the GMM-UBM based speaker verification system with MFCC parameters is our baseline system.

The long period silence is first removed before feature extraction. Band-limiting is then performed by only retaining the filter-bank outputs from the frequency range 300Hz-3400Hz. The speech is pre-emphasized by the filter $H(z)= 1-0.97z^{-1}$ and subsequently a 16-dimensional mel-cepstral vector is extracted from the speech signal every 10ms using a 20ms window. Delta cepstral are then computed and appended to the cepstral vector to produce a 32 dimensional feature vector. RASTA and CMS technology are also used in the feature space. Unlike the previous system, feature warping technology is also used to decrease the mismatch between training and test speech.

The UBM used here is trained with EM algorithm. Because the training data is different in each task, the number of mixtures is also different. Target models are derived by MAP estimation from UBM. Only a speaker-specific T-norm selection is used. The closest set of P cohort models are used to Tnorm during run time where P is chosen to be 50.

The GMM-UBM and T-norm selection technology is all used in the following three sub-systems.

2.2 LPCC & GMM-UBM

The difference between this system and the former system is the parameter used in them.

In this system, a 13-dimensional LPCC vector is extracted from the speech signal every 10ms using a 20ms window. Delta and delta-delta cepstral are then computed and appended to the cepstral vector to produce a 39 dimensional feature vector. RASTA, CMS and feature warping technology are also used in the feature space.

2.3 PITCH & GMM-UBM

We firstly split pitch and energy contours into segment with 7 frames length. 4 parameters related to pitch were extracted:

1. log (mean_F0) averaged over a segment
2. log (max_F0) of a segment
3. log (min_F0) of a segment
4. F0_slop of a segment

Another 4 parameters related to energy are extracted as above. Total 8 parameters of a segment comprise an 8-dimension vector. Pit1 sub-system is based on GMM-UBM scheme also.

2.4 WAVELET PITCH & GMM-UBM

First, we make wavelet analysis of the f0 and energy contour. Subsequently, the prosodic features are extracted only from the 3rd level approximation coefficients (cA3). Two kinds of parameter are extracted:

1. For the cA3 sequence of pitch contour, each four successive cA3 coefficients are combined as one 4-dimensional pitch feature.

2. For the cA3 sequence of log (energy) contour, one Energy Slope is extracted from each four successive cA3s.

The energy slope is combined with the corresponding pitch feature to form a new five dimensional feature vector. In this system we also use GMM-UBM scheme.

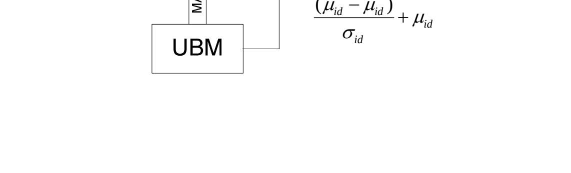
2.5 MFCC & SVM

Feature is extracted by adapted GMM. First, UBM is trained with EM algorithm. Then, for each train and test segment, a GMM is derived by MAP estimation from UBM. Therefore, feature can be extracted by adapted GMM just like this,

$$\frac{\mu_{id} - v_{id}}{\sigma_{id}} + v_{id}$$

Where μ_{id} and v_{id} are the means of the *d*th dimension feature in the *i*th Gaussian of GMM and UBM. $\sigma_{id}{}^2$ is the mutual variance.

Since the probabilistic statistical model GMM is the dominant approach for modeling in text-independent speaker recognition, we selected the Radial Basis Function (RBF) kernel which is more like GMM.


speech data
GMM
MAP
UBM
Feature Extraction
$\frac{(\hat{\mu}_{id} - \mu_{id})}{\sigma_{id}} + \mu_{id}$
SVM

2.6 FUSION

The scores from the sub-systems are fused with a perceptron classifier. The number of input nodes of the perceptron is the same as the number of sub-systems applied. There is no hidden layers and only one output node. The perceptron weights were trained using the entire development data.

3. submission systems

3.1 10seconds-10seconds submission systems

The primary systems both fuse 2.1, 2.2 and 2.3 sub-systems. We also present the second systems which fuse 2.1 2.2 and 2.3 sub-systems. The speech used to train UBM with 512 mixtures is from all of the training and test data of NIST 2004/2005 SRE. Therefore there are 2416 files (about 669 MB) for male and 3052 files (about 826 MB) for female to train UBM respectively. The T-norm set consisted of 246 male and 370 female speakers from previous NIST SRE database.

3.2 1conv-1conv submission system

The system fused 2.2, 2.3 and 2.5 sub-systems. The speech used to train UBM with 2048 mixtures is from some of the training data of NIST 2001 and 2004 SRE. Therefore there are 205 files (about 302 MB) for male and 222 files (about 343 MB) for female to train UBM respectively. The T-norm set consisted of 361 male and 469 female speakers from previous NIST SRE database.

3.3 1conv-10seconds, 3conv-10seconds and 8conv-10seconds submission systems

The three systems only use 2.2 sub-systems. The speech used in UBM and T-norm set is the same as in 1 1conv-1conv submission system.

3.4 3conv-1conv and 8conv-1conv submission systems

The systems fused 2.2 and 2.4 sub-systems. The speech used in UBM and T-norm set is the same as in 1 1conv-1conv submission system.

3.5 3 conversation summedchannel-1 conv and 3 conversation summedchannel-1 conv summedchan submission systems

The main work of the hierarchical agglomerative clustering partition the speech file into speaker homogeneous regions based on blind clustering for both training and test. After partition feature vectors into 100 frame segments (1s), these 100 frame segments form the initial set of clusters, each containing only one segment. Then the agglomerative clustering proceeds by computing the pair-wise distance between all clusters and merging the two clusters with the minimum distance. This is repeated till the number of clusters is reduced to 3. Since there are three training file this produces 9 clusters. The three closest clusters are merged into one final cluster, with the condition that only one cluster can come from each training speech file. Only 2.2 sub-system is used here.

4. Threshold setting

The threshold is selected to decide whether a test speech is according to the training speech. Here we use NIST 2005 SRE database to set the threshold. The same training and test is performed on these data. Threshold is tested when the minimal DCF is reached. The thresholds of male and female are decided separately.