BRIEF DESCRIPTION OF THE SYSTEM

Konstantin Biatov

NetMedia Center
Fraunhofer Institute for Media Communication, Sankt Augustin, Germany
konstantin.biatov@imk.fraunhofer.de

1. Feature extraction

For speaker detection task the feature vectors consist of 15 mel-cepstral coefficients extended with delta energy. For MFCC calculation a 30 ms window and 10 ms step size are used.

2. Feature normalization

For feature normalization feature warping technique was used. For each of 15 cepstral dimensions the feature warping over sliding window with the size 300 ms and with the one frame step was applied.

3. Background model training

The gender independent universal background model (UBM) was trained using data from NIST 2005 speaker detection task. For training data 40 sec from each of 275 male speakers and 30 sec from each of 373 female speakers were extracted. Then this data was normalized using feature warping. In order to get initial approximation of 2048 mixtures k-means clustering algorithm was used before the training. Then the 2048 GMM with the diagonal covariance matrix using 5 iteration of the EM algorithm was trained.

4. Cross likelihood ratio calculation and normalization

In the speaker detection task for distance between train and test segments cross likelihood ratio (CLR) was used. The CLR distance measure between train segments c_i and test segment c_i is defined as:

$$CLR\ (c_i, c_i) = log\ \frac{L(x_i|M_j)}{L(x_i|M_{ubm})} + log \frac{L(x_j|M_i)}{L(x_j|M_{ubm})}\ ,$$

where $L(x_i \mid M_j)$ is the average likelihood per frame of data x_i from the segment c_i given by model M_j and $L(x_i \mid M_{ubm})$ is the average likelihood per frame of data x_i from the segment c_i given by model M_{ubm}. The model M_j is MAP adapted model and M_{ubm} is universal background model (UBM). The Gaussian means and weights of the target models are adapted using MAP adaptation. The final target model had 32 Gaussians. The factor τ was 26.

5. Other implemented techniques

Other well known techniques for speaker verification such as z-norm, t-norm, zt-norm were implemented and tested on the core test 2006. Unfortunately it was not enough time to evaluate the complete test using mentioned implemented techniques and to deliver the obtained results.

6. Speaker detection process description

For the system one trial unit, UBM, the list of the 10 imposter's models (for t-norm) and the list of imposter's segments (for z-norm) were used as the parameters. MFCC extraction, features warping and adaptation were done in the flight.

7. Threshold calculation

The threshold was calculated using keys for NIST 2005 evaluation.

8. Processing time

For speaker detection task 2006 computer cluster was used. This cluster includes 16 nodes each with the Pentium dual processor. The processing time for each train-test combination was approximately 2 minutes depending on the test and train data size. All calculations with exception of universal background model were done in the flight. The core task 1conv4w-1conv4w with the 53967 train-test combination takes 4 days.