The 2006 France Telecom Research and Development Center (Beijing) Speaker recognition Systems

Xianyu Zhao, Yuan Dong, Hao Yang, Jian Zhao

France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China

1. INTRODUCTION

In NIST SRE'2006, France Telecom Research and Development Center (Beijing), abbreviated to be FTRDBJ in the following discussion, presented four systems for three different tasks, as the following table shows.

Table 1: Matrix of training and test segment conditions.

		Test Segment Condition			
		10sec 2-chan	1conv 2-chan	1conv summed-chan	1conv aux mic
Training Condition	10 seconds 2-channel	FTRDBJ_1 *FTRDBJ_4**	NA	NA	NA
	1 conversation 2-channel	NA	FTRDBJ_1 FTRDBJ_2 *FTRDBJ_3**	NA	NA
	3 conversation 2-channel	NA	NA	NA	NA
	8 conversation 2-channel	NA	NA	NA	NA
	3 conversation summed-channel	NA	FTRDBJ_1 *FTRDBJ_2**	NA	NA

Contents in each entry represent FTRDBJ systems that are presented for corresponding task. "NA" means no system is presented for the task. Primary system for each task is marked with this color, asterisk and *italic*.

2. BRIEF SYSTEM DESCRIPTION

2.1 FTRDBJ_1: the Baseline System

The baseline system is a standard GMM-UBM based speaker verification system. A universal background model GMM is used as the alternative hypothesis model and target models are derived from it through Maximum a Posteriori (MAP) adaptation. The logarithmic likelihood ratio between target and UBM models is compared with a predefined threshold to determine whether or not the target speaker matches the speaker in the test segment.

A 13-dimensional perceptual linear predication (PLP) vector is extracted from the speech signal every 10ms using a 25ms window. RASTA filtering is applied to the PLP feature sequence to mitigate linear channel bias effects. Delta, acceleration and triple-delta PLP coefficients are then calculated and appended to the original PLP vector to form a 52-dimensional feature vector. Feature mapping is then applied for channel normalization through a set of gender- and channel-dependent GMM models which are also derived from UBM with MAP adaptation. A Heteroscedastic Linear Discriminant Analysis (HLDA) transform is used to project the channel normalized 52-dimensional feature vector to 51-dimensional. After HLDA projection, Histogram Equalization (HEQ) is applied to each component of the feature vector to improve the channel and noise robustness of speaker verification system.

The UBM model is a gender independent GMM with 2048 mixture components trained using about 40 hours of data from the Switchboard and SRE'2004 evaluation database.

Target models are adapted from UBM using MAP estimation with the relevance factor set to be 16.

This baseline system is evaluated for three tasks: 10sec4w-10sec4w, 1conv4w-1conv4w and 3convs2w-1conv4w.

2.2 FTRDBJ_2: Baseline + TNorm

Instead of comparing directly the logarithmic likelihood ratio score of baseline with the threshold, this ratio score is normalized with TNorm. For each male trial, 354 imposter scores are computed against a set of male imposters; for each female trial, 468 imposter scores are calculated. The mean and standard deviation of these imposter scores are then used to adjust the target speaker score.

All the TNorm imposter speaker models are extracted from the Switchboard and SRE'2004 evaluation data and also adapted from the gender-independent 2048-mixture UBM used in the baseline system.

The FTRDBJ_2 system is evaluated for two tasks: 1conv4w-1conv4w and 3convs2w-1conv4w. And it is selected as the primary system for 3convs2w-1conv4w condition.

2.3 FTRDBJ_3: Baseline + ATNorm

In this system, a target specific imposter pool is selected and used to generate imposter scores for trials of the corresponding target speaker. The size of target-specific imposter pool is set to be 55. The 55 TNorm speakers which are nearest to the target are selected into the target-specific imposter pool. The distance used to define the neighborhood of two models, λ_i and λ_j, are calculated as,

$$d(\lambda_i, \lambda_j) = -\frac{1}{N_i}\log\left(\frac{p(x_i|\lambda_j)}{p(x_i|\lambda_{UBM})}\right) - \frac{1}{N_j}\log\left(\frac{p(x_j|\lambda_i)}{p(x_j|\lambda_{UBM})}\right)$$

where x_i and x_j are the training data for λ_i and λ_j respectively, N_i and N_j are the number of feature vectors of x_i and x_j.

The FTRDBJ_3 system is evaluated on the 1conv4w-1conv4w task, and is selected as this task's primary system.

2.4 FTRDBJ_4: Eigenvoice adaptation

For this system, the mean super vector of target model (built by concatenating all of the mixture component mean vectors) is represented as a linear combination of a set of eigenvoices. And, the target-specific combination coefficients are derived through Maximum a Posteriori Eigen-Decomposition (MAPED). Total 500 eigenvoices are used in our FTRDBJ_4 system.

This system is evaluated on the 10sec4w-10sec4w task, and is selected as this task's primary system.

2.5 Summed Channel Segmentation and Speaker Selection Process for 3convs2w Training Condition

The summed channel segmentation and speaker selection process for 3convs2w training condition are shown in Fig.1. This process uses generalized likelihood ratio (GLR) as means of change point detection, of cluster merging, and of speaker selection for multiple conversations.

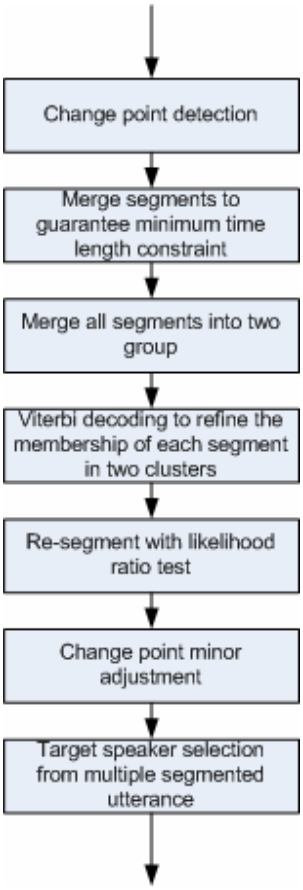

Fig.1 Summed channel segmentation and speaker selection

3. INTERPRETATION OF TRIAL SCORES

For all trials we have done, the output scores of our systems are the logarithmic likelihood ratio between target and UBM models. They are **NOT** intended to be interpreted as the ratio of the posterior odds and the prior odds defined by NIST.