CUHK NIST SRE 2010 system description

Weiwu Jiang
May 2010

The CUHK NIST SRE 2010 system consists of four different classifiers using two types of cepstral features. The first classifier is based on the generative GMM-UBM approach, while the second and third classifiers are based on discriminative SVM techniques and the last one is based on fishervoice framework [6]. The combinations of feature types and classifiers are listed in Table I.

Table I Combinations of cepstral features and classifier techniques (both generative and discriminative) that form the CUHK NIST SRE 2010 *speaker* recognition system.

Classifier	Feature
JFA (s=m+Vy)	MFCC
JFA-SVM (linear kernel)	MFCC
JFA-SVM (cosine kernel withspeaker factor y)	PLP
Fishervoice (speaker factor y)	PLP

1. Feature Extraction

The first stage of the feature extraction process is the voice activity detector (VAD). ETSI Adaptive Multi-Rate (AMR) GSM VAD [1] was applied to remove silence frames and to retain only the high quality speech. Then input speech utterances were converted to sequences of feature vectors by HTK [2]. Finally, the feature vectors were processed by mean-variance-normalization (MVN) and feature warping. We used two different types of cepstral features, namely, MFCC and PLP, as shown in Table I. For the MFCC parameters, frames are composed of 16MFCC+energy, its first derivatives and second derivatives. For the PLP parameters, frames are composed of 12PLP+energy, its first derivatives, second derivatives and third derivatives.

2. Training, Development and Score Normalization Data

Although not all subsystems used the same training and development data, a general data division was made as follows: The NIST SRE 2008 serves for subsystem development testing, calibration and fusion.

The gender-dependent 2048 Gaussian UBMs is a combination of 1024 mixtures trained on NIST SRE 2004-2006 telephone data and 1024 mixtures trained on NIST SRE 2005-2006 microphone data.

The gender-dependent eigenvoice matrix *V* is trained using LDC releases of Switchboard II Phase 2, Phase 3, Switchboard Cellular Parts 2, NIST SRE 2004, NIST SRE 2005 and NIST SRE 2006, including 893 male speakers with 11204 utterances and 1365 female speakers with 16556 utterances. The rank of the speaker space is set to 300.

For channel space training, we used telephone data of NIST SRE 2004, 2005 and 2006 to train a telephone channel space (100 channel factors), microphone data of NIST SRE 2005 and 2006 to train a microphone channel space (75 channel factors) and the

interview data of NIST SRE 2008 to train interview channel space (75 channel factors). We combined these three sub-spaces to get a full channel space of 250 channel factors.
For the SVM training, a special background data set was constructed on NIST SRE 2004-2006 training data, Switchboard Cellular Parts 2 and part of NIST SRE 2008 training data.
For the Fishervoice training, gender dependent fisher discriminative project matrix was constructed on NIST SRE 2004-2006 telephone data, including 400 male speakers and 400 female speakers, which each speaker contains 8 different utterances.
The outputs of the classifiers JFA and Fishervoice are normalized with TZnorm, while classifiers based on SVM are normalized with Tnorm. The NIST SRE2004-2006 training data is used for training cohort models in Tnorm. The Switchboard II Phase 2, Phase 3, Switchboard Cellular Parts 2 training data is used for training cohort models in Znorm.

3. Individual System Descriptions

3.1 JFA

The training procedure of JFA system closely follows Patrick Kenny's paper [3], the only difference that CUHK system is composed by speaker factors and channel factors without diagonal matrix *D*. The model is trained on MFCC. Furthermore, in the verification stage, we apply log-likelihood ratio (LLR) based scoring which is similar to [4].The implementation of this approach is to subtract the estimated noise in the feature level

3.2 JFA-SVM on linear kernel

This system uses JFA supervectors ($s = m + Vy$) to construct kernels of support vector machines by using LibSVM [5]. Given a speaker's speech data, a GMM is estimated by using JFA with MFCC parameters. The means of mixture components in the GMM are concatenated to a GMM supervector, which is used as SVM linear kernels.

3.3 JFA-SVM on cosine kernel with speaker factor *y*

This system uses JFA speaker factor y to construct kernels of support vector machines (SVM). Given a speaker's speech data, a JFA based GMM is estimated by using PLP parameters. The 300 dimension speaker factor *y* is calculated with SVM cosine kernel.

3.4 Fishervoice

This system uses JFA speaker factor *y* as input vector to create a fisher discriminative project matrix []. Each target's speaker factor is projected by this matrix and regarded as target model. Similar to training stage, each test utterance is first projected via fisher discriminative project matrix. Then direct cosine distance is calculated as each trial score.

4. Processing Speed

Performances of all subsystems were measured separately on only one core of an Intel Core 2 Quad CPU 2.53GHz. The breakdown of the runtime factor is summarized in Table II for each of the classifiers.

Table II The runtime factors comparison of four classifiers used in the system.

Classifier	Run Time (training)	Run Time (test)
JFA (s=m+Vy)	0.24	0.26

JFA-SVM (linear kernel)	0.28	0.24
JFA-SVM (cosine kernel withspeaker factor y)	0.24	0.24
Fishervoice (speaker factor y)	0.24	0.24

5. Reference

[1] GSM 06.94, “Digital cellular telecommunication system (Phase 2+); Voice Activity Detector VAD for AdaptiveMulti Rate (AMR) speech traffic channels; General description,” Tech. Rep. V.7.0.0, ETSI, February 1999.

[2] http://htk.eng.cam.ac.uk/download.shtml

[3] P. Kenny, et al., “Improvements in factor analysis based speaker verification,” ICASSP 2006.

[4] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, “A straightforward and efficient implementation of the factor analysis model for speaker verification,” in Proc. Interspeech’2007, Antwerp, Belgium, 2007, pp. 1242–1245.

[5] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[6] Z. Li, W. Jiang and H.Meng “FISHERVIOCE: A DISCRIMINANT SUBSPACE FRAMEWORK FOR SPEAKER RECOGNITION,” ICASSP 2010