The ALPineon System for the NIST SRE 2010

Boštjan Vesnicer <bostjan.vesnicer@alpineon.si>
Alpineon Research and Development, Ljubljana, Slovenia

I Introduction

We submitted only one system for the required core test. Although we produced scores for all conditions of the core test, there has been no development made on the interview and the microphone data. Therefor our focus was exclusively on the telephone part of the core test.

II System description

Our system is based on the state-of-the-art joint factor analysis (JFA) model of speaker and session variability, proposed by Kenny et al. [1]. There are three key differences between the system proposed in [2] and our system: a) we use unnormalized features, b) we do not use score normalization and c) our system is gender-independent. Despite those simplifications we managed to get promising results (0.027 DCF and 5.5 % EER) on the telephone part of the NIST SRE 2008 core test.

Configuration

For the development we used 16252 utterances from the NIST SRE 2004, 2005 and 2006 data. They were used for training the universal background model (UBM) model and also for estimating the JFA hyperparameters. Our gender-independent UBM consisted of 2048 Gaussian components and the dimensions of the speaker and channel subspaces were set to 300 and 100, respectively.

Acoustic features were 20-dimensional MFCCs, extracted every 10 ms from 25 ms-long windowed frame of the speech signal, using the HTK toolkit [3]. The static features were augmented with first and second order derivatives, which resulted in 60-dimensional feature vectors. As noted earlier, no feature normalization was performed. The silence removal was based on the available ASR-based transcriptions.

Scoring

We used two different decision criteria for calculating the verification scores. The first one was based on the likelihood ratio (LR) decision criterion and the second one on the

support vector machines[1] (SVMs). The final score was produced by summing together the two individual scores. The scores are not intended to be interpreted as log likelihood ratios.

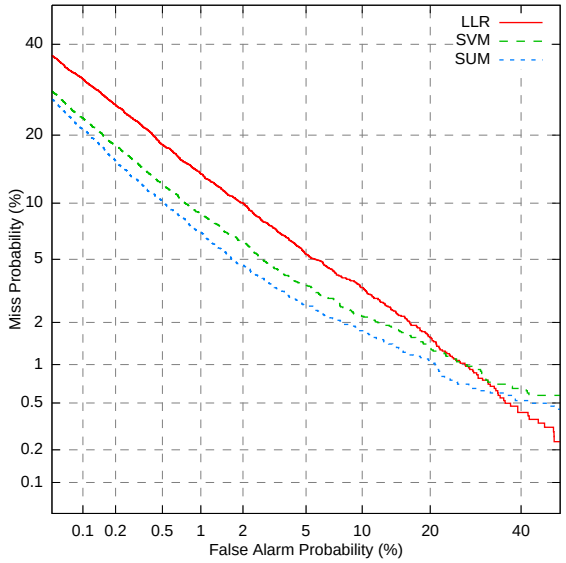
There exist various methods for calculating the likelihood (and its approximation) of the target speaker model given the test utterance. In our case we did not rely on the channel point estimate of the test utterance, but we instead integrated the likelihood function over the whole channel subspace.

For the SVM scoring, we first estimated the point estimates of the speaker supervectors for both the training and test utterances. These were subjected to rank normalization. The rank-normalized training supervector was used for estimating the linear SVM model. The verification score was calculated by measuring the distance between the rank-normalized test supervector and the hyperplane implied by the SVM model.

Parameter tuning

There was just one free parameter, namely the decision threshold, that we had to tune on the development data. Since this year NIST decided to change the DCF evaluation metric, we set up a large development test set consisting of all possible trial telephone-telephone pairs from the NIST SRE 2006 core test data.

III Results on the development data



	EER (in %)	DCF ($\times 10^4$)
LLR	5.2	6.5
SVM	4.0	5.3
SUM	3.4	5.1

[1] A custom-tailored code from LIBSVM [4] was used for that purpose.

IV References

1 Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, issue 4, pp. 1435–1447 (2007)

2 Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 16, issue 5, pp. 980–988 (2008)

3 Hidden Markov Model Toolkit (HTK). Available at `http://htk.eng.cam.ac.uk`

4 Chang, C.-C. and Lin C.-J. LIBSVM: a library for support vector machines (2001) Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`