University of the Basque Country + Ikerlan System for NIST 2008 Speaker Recognition Evaluation

Maider Zamalloa[1,2], *Mikel Penagarikano*[1], *Luis Javier Rodriguez*[1], *German Bordel*[1], *Juan Pedro Uribe*[2]

(1) Department of Electricity and Electronics, University of the Basque Country, Spain
(2) Ikerlan - Technological Research Center, Spain
E-mail: maider.zamalloa@ehu.es

1. Introduction

This paper describes the speaker recognition system EHUIKER developed in the GTTS group of the University of the Basque Country in collaboration with IKERLAN Technological Research Center. The EHUIKER system submitted to the NIST 2008 Speaker Recognition Evaluationis is based on the standard Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) approach. In order to compensate for environment and/or channel mismatch, noise reduction and channel compensation techniques are applied to speech signals.[1]

2. Primary System: EHUIKER_1

The EHUIKER primary system is based on the standard Universal Background Model - GaussianMixture Model (UBM-GMM) approach and implements some recently developed noise reduction and channel compensation techniques [1]. The system was built by means of the Sautrela framework [2].

2.1. Preprocessing

The Qualcomm-ICSI-OGI (QIO) [3] noise reduction technique (based on Wiener filtering) was independently applied to the audio streams. The full audio stream was taken as input to estimate noise characteristics, thus avoiding the use of voice activity detectors on which most systems rely to contraint noise estimation to non-voice fragments.

2.2. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) were used as acoustic features, computed in frames of 25 ms at intervals of 10 ms. The MFCC set comprised 10 coefficients, not including the zero (energy) coefficient. Cepstral Mean Subtraction (CMS) and Feature Warping [4](3-second feature histograms warped into a normal distribution) were applied to cepstral coefficients. Finally, the feature vector was augmented with dynamic coefficients (first-order deltas), resulting in a 20-dimensional feature vector.

2.3. Channel Compensation

Channel compensation was performed by applying feature mapping [5]. Three gender dependent channel-specific models (cellular, cordless and landline) were MAP adapted from a gender dependent UBM using NIST SRE06 training data. Data used for channel adaptation were disjoint from those used to estimate the UBM. Due to a lack of gender dependent data for the new acoustic condition of SRE08, feature mapping was not applied to interview data.

2.4. UBM and Speaker Models

Three UBM were defined, for male telephone speech, female telephone speech and gender-independent microphone speech (interviews), each consisting of 512 Gaussian mixture components. The UBM for telephone speech were trained from two channel balanced male and female subsets of the NIST SRE06 training set. The UBM for microphone speechwas estimated from the Mixer5 development data provided by NIST for the evaluation.

Speaker models were derived from the UBM using *Relevance Maximum A Posteriori* (MAP) adaptation [6], with relevance factor $\tau = 16$. In the adaptation process, only the Gaussian means were adapted.

2.5. Development Corpus

The NIST SRE04 Evaluation core condition dataset (1 side train - 1 side test) was used for development purposes.

2.6. Scoring

Verification experiments were carried out by applying the standard top N [6] log-likelihood ratio scoring method (llr), with N=8. Scores were calibrated by means of a single scaling factor ($\alpha = 12$). The scaling factor was obtained from

[1] This work has been partially funded by the Basque Government, under program SAIOTEK, projects S-PE06UN48, S-PE07UN43, S-PE06IK01 and S-PE07IK03, and theUniversity of the Basque Country, under project EHU06/96.

the development corpus by maximizing Mutual Information which is equivalent to minimizing C_{llr}. Well-calibrated scores were used to obtain the verification decision through the minimum expected cost Bayes decision:

$$decision = \begin{cases} true, \quad if & llr > \ln\left(\frac{C_{fa}\cdot(1-P_{target})}{C_{miss}\cdot P_{target}}\right) \\ false, & otherwise \end{cases}$$

For the 10s training condition, no calibration was applied. In this case, the $true$ verification decision was output when $llr > 0$.

2.7. Processing Speed

Experiments were carried out on a dual AMD dual core 270 Opteron server (2.0 Ghz) with a 6GB RAM. The processing speed was measured by running one experiment in a single thread. The Java Virtual Machine memory usage was limited to 1GB. The resulting runtime factor was 0.71xRT for the UBM, 0.16xRT for the speaker models estimation and the test.

3. Contrastive System: EHUIKER_2

A second EHUIKER system was built for the NIST SRE08 evaluation, identical to the EHUIKER primary system but for the fact that it did not apply any channel compensation technique, and the EHUIKER telephone (gender dependent) UBM are estimated using the training data from the NIST SRE05 core condition (1 side train).

4. References

[1] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System", IEEE Transactions on Audio, Speech, and Language Processing, 15(7):1979-1986, Sept. 2007

[2] M. Penagarikano and G. Bordel, "Sautrela: A Highly Modular Open Source Speech Recognition Framework", in Proceedings of the ASRU Workshop, 2005.

[3] A. Adami et al., "Qualcomm-ICSI-OGI features for ASR", in Proc. ICSLP, 2002

[4] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification", in Proc. Speaker Odyssey, Crete, Greece, 2001, pp. 213-218.

[5] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping", in Proc. ICASSP, Apr. 2003, vol. II, pp. 53-56.

[6] D. A. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, vol. 10, pp. 19-41, January 2000.