Loquendo - Politecnico di Torino
System Description
NIST 2006 Speaker Recognition Evaluation

Claudio Vair, Daniele Colibro*, Alba Fiorio*, Fabio Castaldo^, Emanuele Dalmasso^, Pietro Laface^*

Loquendo, Torino , Italy*
{Claudio.Vair, Daniele.Colibro, Alba.Fiorio}@loquendo.com
Politecnico di Torino, Italy^
{Fabio.Castaldo,Emanuele.Dalmasso,Pietro.Laface}@polito.it

1. Systems Overview

The primary system for SRE06 is the linear fusion of two independent GMM systems: a Phonetic GMM (PGMM) and a classical GMM. The results of the primary system, designed as LPT_1, were supplied for all the test conditions. A second and a third set of full results were supplied, related to the PGMM and GMM systems alone, designed as LPT_2 and LPT_3 respectively.

Phonetic GMM system

The system used for SRE06 shares the same architecture and scoring strategy of the system used in SRE05 but it includes a new feature-level compensation technique [1] based on the intersession variability compensation approach described in [2]-[4].

The system decodes the speaker utterance, both in enrolment and in verification, producing phonetic labeled segments. The decoder is a hybrid HMM-ANN model trained to recognize 11 language independent phone classes. Each phone class is modeled by a three state left-to-right automaton with self-loops. The ANN is a Multilayer Perceptron that estimates the posterior probability of each phone class state, given an acoustic feature vector. The ANN has been trained using 20 hours of speech of 10 different languages using corpora not specifically collected for speaker recognition evaluations.

The UBM [1] and the voiceprints consist of a set of phonetic GMMs: each phone class state has associated its GMM. The maximum number of (diagonal covariance) Gaussians per mixture per state is 64, for a total of approximately 2000 Gaussians. A gender – and nearly language – independent UBM has been trained on the same data that were used for training the ANN model.

In enrolment, the labels and the boundaries of the phonetic segments are used for MAP adaptation of the parameters of the phone class-dependent GMMs. In recognition, the phonetically labeled audio segments are scored against their corresponding GMMs. Thus, the likelihood of a given observation vector is computed selecting the GMM corresponding to the phone class decoded at that time frame. The PGMM classifier is currently the core technology of the Loquendo Free Speech Identification (LFSI) system.

The system uses 19 Mel Frequency Cepstral Coefficients (MFCC). Feature warping to a Gaussian distribution is then performed, for each static parameter stream, on a 3 sec sliding window excluding silence frames [6]. 36 parameters per frame are obtained discarding the C0 cepstral parameter, and computing the usual delta parameters on a symmetric 5 frame window.

Concerning the new feature-level compensation technique, an intersession variability vector, defined in a constrained low-dimensional subspace, is computed using the PSA technique [7]. The low rank matrix, defining the constrained intersession variability subspace, has been trained on SRE04 and SRE05 data. The intersession variability vector is then projected to the feature space and subtracted from the original observation vectors. More details about the feature domain compensation are given in [1]. The same compensation technique is applied in both enrollment and verification. The dimension of the subspace used in the evaluation was 40.

GMM system

The GMM can be considered a special case of the Phonetic GMM system where the number of classes is just one. The procedures for the adaptation of the speaker models and for scoring the test utterances are the same. The GMM system is characterized by a reduced set of mixtures (512), and features (the first 13 cepstral parameters and their deltas, excluding C0). The gender independent UBM has been trained using data from the NIST 2000, the OGI National Cellular, and HTIMIT databases.

Moreover, feature mapping [7] is performed before applying the same feature-level intersession compensation described for the PGMM system. Gender and channel dependent models have been used for feature mapping, with the channels labeled as Carbon, Electret, GSM, Analog, and Digital (a total of 10 models).

Fast selection of Gaussian is achieved by means of a “road-map” based approach [9].

In the 10sec4w-10sec4w test, an eigenvoice adaptation approach [10] has been used for training the voiceprints.

2. Multi-speaker conversations trials

For the multi-speaker conversations trials we use unsupervised speech segmentation to detect speaker clusters, followed by voiceprint creation and scoring. Our procedure for speaker clustering is described in [9]. For the two wire tests, speaker pre-segmentation is performed, and each putative speaker cluster is scored against the speaker models in the index list. For each model, we select the speaker cluster that gives the best score.

3. Score normalization

Similar to the 2005 evaluation, the evaluation has been carried out with score normalization. First the raw score are speaker-normalized by means of Z-norm. The Z-norm parameters for each speaker model have been evaluated using a subset of speaker samples included in the NIST SRE04 database. Separate statistics have been collected for the female and male speakers, using 2 audio samples of 80 speakers for each gender.

Test dependent normalization is performed using T-norm [12]. A fixed set of impostor models have been selected among the voiceprints enrolled with data belonging to the SRE04 evaluation. The T-norm parameters for each test sample were estimated using the Z-normalized scores of the impostor voiceprints. We refer to the Z-Norm followed by T-Norm as ZT-Norm. Separate sets of impostor voiceprints have been enrolled according to the speaker gender and according to the training condition of the voiceprints used in each test trial: specifically 160 female and 160 male impostor voiceprints for each test condition.

The fusion of the PGMM and of the GMM systems is obtained by linear combination – with the same weights - of the normalized scores produced by the two systems.

4. Unsupervised adaptation

This year we submitted the results for the unsupervised adaptation 1conv4w-1conv4w test. The systems used for this test are the same LPT_1, LPT_2 and LPT_3 used for the other tests. We performed the tests using the primary, un-adapted and ZT-normalized score for selecting – with a rather conservative threshold set to 4.0 – the samples to be used for voiceprint adaptation. This threshold produced, on the SRE05 development data, a negligible number of false alarms and about 1/3 of false misses. The selected utterances are then used to update the adapted voiceprints used to produce the final ZT-normed scores.

5. Decision thresholds

The decision thresholds have been set according to past experience acquired performing the SRE05 evaluation. The decision thresholds have been estimated on subsets of the SRE 2004 and SRE 2005 data used as evaluation sets.

6. Execution time

PGMM runs on a dual Xeon 3.4GHz, SUSE Linux Server with the following average execution times:

- 10sec4w training: 0.6 sec/voiceprint
- 1conv4w training: 9.5 sec/voiceprint
- 3conv4w training: 26 sec/voiceprint
- 8conv4w training: 77 sec/voiceprint
- test/znorm/tnorm 1conv4w: 9 sec/audio (up to 400 voiceprint/audio)
- segmentation 1conv2w: 8sec/audio

GMM runs on a AMD Athlon 1.8GHz, Windows XP PC with the following average execution times:

- 10sec4w training: 1 sec/voiceprint
- 1conv4w training: 4.5 sec/voiceprint
- 3conv4w training: 14 sec/voiceprint
- 8conv4w training: 36.5 sec/voiceprint
- test/znorm/tnorm 1conv4w: 4 sec/audio (up to 400 voiceprint/audio)

References

[1] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, P. Laface, *Channel Factors Compensation in Model and Feature Domain for Speaker Recognition,* Odyssey 2006 Workshop on Speaker and Language Recognition.

[2] P. Kenny, P. Dumouchel, *Disentangling Speaker and Channel Effects in Speaker Verification*, Proc. ICASSP 2004, pp. I-37-40, 2004.

[3] N. Brümmer, “NIST SRE 2004 Evaluation Workshop”, Toledo, Spain, 2004.

[4] R. Vogt, B. Baker and S. Sridharan, “Modelling Session Variability in Text-independent Speaker Verification”, *Proc. INTERSPEECH-2005*, pp. 3117-3120, 2005.

[5] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*, Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[6] J. Pelecanos, S. Sridharan, *Feature warping for robust speaker verification*, Proc. 2001: a Speaker Odyssey, pp. 213-218, 2001.

[7] S. Lucey, T. Chen, *Improved Speaker Verification through Probabilistic Subspace Adaptation*, Proc. EUROSPEECH-2003, pp. 2021-2024, 2003.

[8] D. A. Reynolds, *Channel Robust Speaker Verification via Feature Mapping*, Proc. ICASSP 2003, vol. 2, pp. 53–56.

[9] Povey D. &Woodland P.C., *Frame Discrimination training of HMMs for Large Vocabulary Speech Recognition*, Proc. ICASSP'99, pp. 333-336, Phoenix.

[10] R. Kuhn J.C. Junqua, P. Nguyen, and N. Niedzielski, ”Rapid Speaker Adaptation in Eigenvoice Space”, IEEE Trans. on Speech and Audio Processing, Vol.8, No.6, Nov. 2000, pp. 695-707.

[11] E. Dalmasso, P. Laface, D. Colibro, C. Vair, *Unsupervised Segmentation and Verification of Multi-Speaker Conversational Speech*, Proc. INTERSPEECH-2005, pp. 1001-1004**.**

[12] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, *Score Normalization for Text-Independent Speaker Verification Systems,* Digital Signal Processing, 10 (2000), pp. 42-54.