Description of the INRIA-IRIT system for NIST'SRE-2010

Reda Jourani[1,3], Khalid Daoudi[2], Jérôme Farinas[1] and Régine André-Obrecht[1]

[1] SAMoVA Group, IRIT - UMR 5505 du CNRS
Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
Email: {jourani, Jerome.Farinas, obrecht}@irit.fr,

[2] INRIA Bordeaux-Sud Ouest
351, cours de la libération. 33405 Talence. France
Email: khalid.daoudi@inria.fr

[3] Laboratoire LRIT. Faculty of Sciences, Mohammed 5 Agdal University
4 Av. Ibn Battouta B.P. 1014 RP, Rabat, Morocco

This document briefly describes the speaker detection system for NIST'SRE-2010 developed jointly by INRIA (Bordeaux, France) and IRIT (Toulouse, France). This system is mainly based on the open source software ALIZE/Spkdet [1].

1. Feature extraction

The feature extraction is carried out by the filter-bank based cepstral analysis tool Spro [2]. Bandwidth is limited to the 300-3400Hz range. 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate and transformed into Linear Frequency Cepstral Coefficients (LFCC). Consequently, the feature vector is composed of 50 coefficients including 19 LFCCs, their first derivatives, their 11 first second derivatives and the delta-energy. The LFCCs are preprocessed by Cepstral Mean Subtraction and variance normalization.

We applied an energy-based voice activity detection (VAD) to remove silence frames and keep only the most informative frames. The energy coefficients are first normalized using zero mean and unit variance normalization, and then used to train a three components GMM. Finally the frames with lowest energy are discarded.

For the interview segments, the estimated intervals where the target speaker is speaking are determined based on the VAD. The VAD on the B channel of the interview segments determines the time intervals where the interviewer is speaking. Afterward, this estimated intervals are removed from the A channel speech segments, we thus process only the target speaker turns.

Once the speech segments of a signal are selected, a post-processing is applied to deal with the energized non-speech segments, the different distances of speakers from the microphones and possible time shifts, by removing some short speech segments. Based on tests done on the tarball of the NIST-SRE' 2010 development data and NIST-SRE' 2008 data, speech segments of interviews A channels are cleaned from the segments shorter than 20ms. Moreover, we purge telephone and microphone data from respectively, speech segments shorter than 40ms and 20ms.

Finally, the remaining parameter vectors are normalized to fit a zero mean and unit variance distribution.

2. Universal Background Models

Two gender-dependent Universal Background Models (UBMs) with 512 Gaussian components and diagonal covariance matrices are used. They were trained by the LIA laboratory (Laboratoire d'Informatique d'Avignon, France) using telephone data from the Fisher English Training Speech Part 1 (LDC:LDC2004S13), and microphone data from the NIST-SRE' 2005 data.

These UBMs were kindly given to us by J-F. Bonastre, A. Larcher and D. Matrouf of LIA.

3. Session variability modeling

A speaker model can be decomposed into three different components: a speaker-session-independent component, a speaker dependent component and a session dependent component. All the target speakers models are obtained by performing the Latent Factor Analysis modeling and by retaining only the speaker dependent components [3, 4].
We estimate two gender-dependent 40 rank U matrices (the session variability matrices) on NIST-SRE'2004 and 2005 data. The U matrices are trained using 194 male speakers and 134 female speakers with in average, 27 different sessions per speaker.

4. Score normalization

Gender-dependent T-norm is applied to the log-likelihood ratio scores. 200 male speakers and 200 female speakers from NIST-SRE' 2006 are used as background data. The half of the speakers models are trained on telephone date, while the remaining speakers are trained on microphone data.

5. Decision

Despite the different training and testing subsets proposed for the NIST-SRE' 2010 speaker recognition evaluation campaign, a unique gender-independent threshold is used. It is set on the EER point estimated on the male part of the NIST-SRE' 2008 primary task (short2-short3).

6. Speed and resources

Real-Time (RT) factors are estimated on a cluster, 8 x Intel XEON 64bits 3.16GHz, with 6MB of L2 cache per processor and 24GB of RAM.

The CPU execution time that was required to create client models from the training data is approximately 0.086xRT, using about 7.5 GB of memory.

The CPU execution time that was required to process the test segments is approximately 0.208xRT, using about 3.28 GB of memory.

Acknowledgments

We would like to thank the LIA laboratory, particularly J-F. Bonastre, A. Larcher and D. Matrouf, for providing the UBMs and for their continuous and valuable help during the whole campaign period.

References

[1] J.-F. B. et al., "Alize/spkdet: a state-of-the-art open source software for speaker verification," *in Speaker Odyssey*, 2008.

[2] Gravier, G., "SPro: Speech Signal Processing Toolkit," Online: http://www.gforge.inria.fr/projects/spro.

[3] 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*, Antewerp, Belgium, 2007.

[4] P. Kenny and P. Demouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.