UNIFR-INT System Description

NIST SRE 2006

Asmaa El Hannani[1,2] and Dijana Petrovska-Delacrétaz[2]

[1]DIVA group, University of Fribourg, Switzerland;
[2]EPH Dept., Institut National des Télécommunication, Evry, France
asmaa.elhannani@unifr.ch, dijana.petrovska@int-evry.fr;

1 Submitted systems

The University of Fribourg (UNIFR) in collaboration with the Institut National des Télécommunications d'Evry (INT) submitted two systems for the **1conv4w-1conv4w** and **8conv4w-1conv4w** tasks.

- The primary system is the fusion of the GMM, the ALISP N-gram and the ALISP LM systems.
- The secondary system correspond to the GMM system.

2 Systems Description

2.1 GMM system

The speech parameterisation for the Gaussian Mixture Model (GMM) system is done with Linear Frequency Cepstral Coefficients (LFCC), calculated on 20 ms windows, with a 10 ms shift. For each frame a 16-element cepstral vector is computed and appended with first order deltas and the delta-energy. Bandwidth is limited to the 300-3400Hz range. The parameter vectors are normalized to fit a zero mean and a unit variance distribution. The mean and variance used for the normalization are computed file by file on all the frames kept after applying the frame removal processing. The parameterisation is carried out using SPRO tools.

The feature vectors are modeled by a 2048 gender-dependent GMM. The background GMMs are created using data from the Fisher collection and the 2003'NIST SRE data. For each target speaker, a specific GMM with diagonal covariance matrices is trained via a maximum a posteriori adaptation. The verification is performed using the 10-best Gaussian components. This system is based on ALIZE-LIA-SpkDet tools.

2.2 ALISP systems

2.2.1 Training of the ALISP recognizer

The speech parameterization for the ALISP recognizer is done with Mel Frequency Cepstral Coefficients (MFCCs), using HTK [1]. Mel frequency bands are computed in the 300-3400 Hz range. Cepstral mean substraction is applied to the 15 static coefficients, estimating the mean on the speech-detected parts of the signal. The energy and delta components are appended, leading to 32 coefficients in each feature vector.

The systems described bellow use in the first stage a data-driven segmentation Automatic Language Independent Speech Processing (ALISP) tools [2]. This technique is based on units acquired during a data-driven segmentation, where no phonetic transcription of the corpus is needed. In this work we use 65 classes. The modelling of the set of data-driven speech units, denoted as ALISP units, is achieved through the following stages. After the pre-processing step for the speech data, first Temporal Decomposition is used, followed by Vector Quantization providing a symbolic transcription of the data in an unsupervised manner. Hidden Markov Modeling is further applied for a better coherence of the initial ALISP units. Each ALISP unit is modeled by a left-to-right HMM having three emitting states and containing up to 8 Gaussians each. The number of Gaussians is determined through a dynamic splitting procedure. The gender dependent ALISP HMMs are trained on data from (1999, 2001 and 2003) NIST SRE data sets.

2.2.2 ALISP N-gram System

The focus here is to capture high-level information about the speaking style of each speaker. Speaker specific information is captured by analyzing sequences of ALISP units produced by the data-driven ALISP recognizer. In this approach, only ALISP sequences are used to model speakers. For the scoring phase each ALISP-sequence is tested against a speaker specific model and a background model using a traditional likelihood ratio. The speaker specific ALISP-sequence models is generated using a simple n-gram (1-gram, 2-gram, 3-gram) frequency count [3]. The background models are trained using data from the Fisher and the 2003'NIST SRE data.

2.2.3 ALISP LM System

In this system [4], the label sequences produced by the ALISP recognizer are used to train ALISP trigrams using the HTK LM tools. The trigrams construction is a tow stage process. Firstly, the training text is scanned and the trigrams are counted and stored in a database of gram files. Secondly the resulting gram files are used to compute trigram probabilities which are stored in the language models file. The trigram language models is used to predict each symbol in the sequence given its tow predecessors. During the testing phase, the label sequences of a previously unseen test text is scored against the language model (see 14^{th} chapter of the HTK book [1] for more details). The speaker specific language models are adapted from a background models which are trained on Fisher and 2003'NIST data.

2.3 Systems fusion

The scores from the different systems are fused using the LIBSVM software. The SVM combiner uses a RBF kernel and is trained on NIST 2004 SRE trials.

3 Memory requirement and processing times

A cluster of 13 nodes, each 2x AMD Operon with 4 Gb RAM and 3Gb swap space, was used to perform the training and the testing of our systems.

The time spent to build models from the training data and to process the test segments is the following:

1. 1conv4w-1conv4w
   - ALISP systems
     - train : 17h
     - test : 10h
   - GMM system
     - train : 13h
     - test : 20h
2. 8conv4w-1conv4w
   - ALISP systems
     - train : 24h
     - test : 7h
   - GMM system
     - train : 22h
     - test : 17h

References

1. S. Young. Hidden markov model toolkit (HTK). [Online]. Available: http://htk.eng.cam.ac.uk/
2. G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot, "Towards ALISP: a proposal for Automatic Language Independent Speech Processing," *In Keith Ponting, editor, NATO ASI: Computational models of speech pattern processing Springer Verlag*, 1999.
3. A. E. Hannani and D. Petrovska-Delacrétaz, "Exploiting high-level information provided by alisp in speaker recognition," *Non Linear Speech Processing Workshop (NOLISP 05)*, 19-22 April 2005.
4. A. El-Hannani, D. T. Toledano, D. Petrovska-Delacrétaz, A. Montero-Asenjo, and J. Hennebert, "Using data-driven and phonetic units for speaker verification," *to appear in proc. of ODYSSEY06, The Speaker and Language Recognition Workshop.*