The LIA Speaker Verification system for NIST SRE 2006

Note: All the systems are based on the ALIZE toolkit [3], except the acoustic parameterization which is performed using the SPRO software. SVM-based systems are using the SVMLight Toolkit

I. Description of the LIA Submission

Three different systems have been submitted LIA_1, LIA_2, LIA_3

Here are the primary systems for the following task conditions:
* Train: 1conv4w - Test: 1conv4w => LIA_1
* Train: 1conv4w - Test: 10sec4w => LIA_2
* Train: 3conv4w - Test: 1conv4w => LIA_2

II. Commonalities

Background Data and Tnorm Speakers

The development corpora is the NIST SRE 2005 campaign.
For background data as well as for Tnorm speakers, the Fisher corpus Part 1 has been used.
More precisely, only speakers having a unique conversation has been used, resulting in 1464 speakers.
This data has been used to train the different background models, as well as to produce the negative examples for SVM-based systems.
Concerning Tnorm speakers, two gender dependent cohort of 180 speakers has been used, with aa balance in respect to the channel type of the file, precisely 60 cellular, 60 landline, 60 cordless for each gender.

III. Front End Processing

Parameterization

The signal is characterized by 50 coefficients including 19 linear frequency cepstral coefficients (LFCC), their first derivative, their 11 first second derivatives and the delta-energy.
They are obtained as follows. 24 filter bank coefficients are first computed over 20ms Hamming windowed frames at a 10ms frame rate. Bandwidth is limited to the 300-3400Hz range.

The frame removal process is the same as last year for 2005 data. The energy component is used to train a three component GMM, resulting in about 30% of the frames being selected.

Once the speech segments of a signal are selected, a final process is applied in order to refine the speech segmentation:

- overlapped speech segments between both the sides of a conversation are removed,
- morphological rules are applied on speech segments to avoid too short ones, adding or removing some speech frames.

Finally, the parameter vectors are normalized to fit a 0-mean and 1-variance distribution. The mean and variance estimators used for the normalization are computed file by file on all the frames kept after applying the frame removal processing.

A gender-dependent feature mapping (Reynolds 03) process is applied to all the data for channel robustness. Three channel conditions have been used (landline, cellular, cordless). All channel dependent models are derived using MAP adaptation (mean and variance).

IV. LIA_SpkDet: GMM/UBM system

World model
The world modeling relies on three steps: initialization, training and warping. Resulting world models are 2048 gender dependent Gaussian Mixture Models with diagonal covariance matrices.

For a better separation of initial classes, frames are selected among the entire learning signal via a probability followed by an iteration of the EM algorithm, to estimate the GMM parameters.

During the estimation of the world model parameters, instead of using all the learning signals in their temporal order, 10% of frames is selected randomly at each new iteration. For the two last iterations, the entire signal is classically used in its temporal order. Learning is also driven by different streams, which corresponds to different channel conditions. An *apriori* weight for each condition can be hence given to drive the learning process. During all the process, a variance flooring is applied so that no variance value is less than 0.5. The warping process described last year is applied, so that the global mean and variance of GMMs are 0 and 1.

Client model
Client models are derived by a Bayesian Adaptation (1 iteration of the Maximum A Posteriori method) of the world model. Only means of each gaussian are adapted. The amount of adaptation for each mean is related to the amount of data available, this is achieved via a relevance factor of 14. The warping process described last year is applied, so that the global mean and variance of GMMs are 0 and 1.

Test, normalization, and decision
Speaker detection test relies on log-likelihood ratio, computed on the 10 best gaussian components. Classical Tnorm normalization technique is then applied on each test likelihood ratio. Finally, the Tnormed log likelihood scores are compared with a threshold to make the decision. This threshold is gender dependent and set on the best DCF point estimated on SRE'05 development set.

V. SVM/UBM system

The SVM/UBM system uses a novel approach to perform a speaker verification task using the UBM model. Indeed, contrary to feature-based kernel (aka GLDS), this system propose an extension of the TFLLR kernel by using the UBM Gaussian indexes as tokens.

The resulting kernel formulation is close to the theory of score-space (a generalization Fisher Kernel) but is computationally efficient and can easily be applied in large-scale evaluation.
The paper [1] presented in Odyssey'06 goes deeper in the explanation of this system.
The UBM used is the same as the one in the GMM/UBM system. The SVMLight toolkit was used for the experiment. 732 impostors examples have been used for each gender.

Two different SVM/UBM systems have been submitted, which differs by the type of input vectors given to the SVM. The first one comes with the "raw" features as in the traditional system *(SVMUBM raw)*, the other one comes with all features being rank normalized *(SVMUBM rank)*.
The fusion of both is especially interesting.

VI. AES: The LIA Acoustic Event System

The AES system follows the paradigm of Speaker Detection using Acoustic Event Sequences.
This system uses a novel approach [2] to perform speaker detection using acoustic events produced by a Gaussian Mixture Model and by an analysis of their sequences. It basically assumes that a GMM with a maximum of training data, such as background models, contains enough information to extract a base structure on which a sequence analysis is speaker specific.
This base structure is referred to as acoustic events.

This system has been described in the LIA 2005 system.
Certain changes can however be precise:

- The dictionary size this year has been fixed to 128
- The analysis of Ngram have changed from a fixed length to a variable length strategy. Hence, this year all Ngrams of length 2 to 4 have been used for the analysis
- A cost factor of 300 is applied in the SVM learning to compensate the imbalance between target and impostors.
- A Tnorm has been applied

VII. C-AES: The LIA class-dependent AES

The C-AES follows the main principles of the AES system. It can be seen has a multi-resolution framework for the AES system.

Figure 1 shows the general idea of a C-AES system. Two different set of acoustic events are generated:

- Feature Events, the same as in an AES system
- Class Events, of the same type as Feature Events but at a much lower resolution.

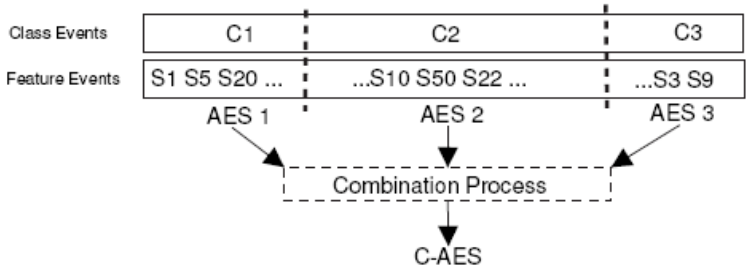


Figure 1: *Combination of multiclass information: an AES system is applied on each Class Event on their own Feature Events. A combination process is performed before the concatenation of all information.*

The C-AES system can hence perform multiple sequential analysis on different classes, focusing on inter-classes information.
Information coming from the Class Event has to be combined into a single vector for SVM classification. In order to perform this task, an extension of the TFLLR kernel is necessary. Indeed, apriori information on the Class Events is estimated using a MAP estimation.
The dimensionality of the Feature Event is 128, and 8 for the Class Event.
The SVM learning and score computation strategy follows the AES one.

VIII. System Combination

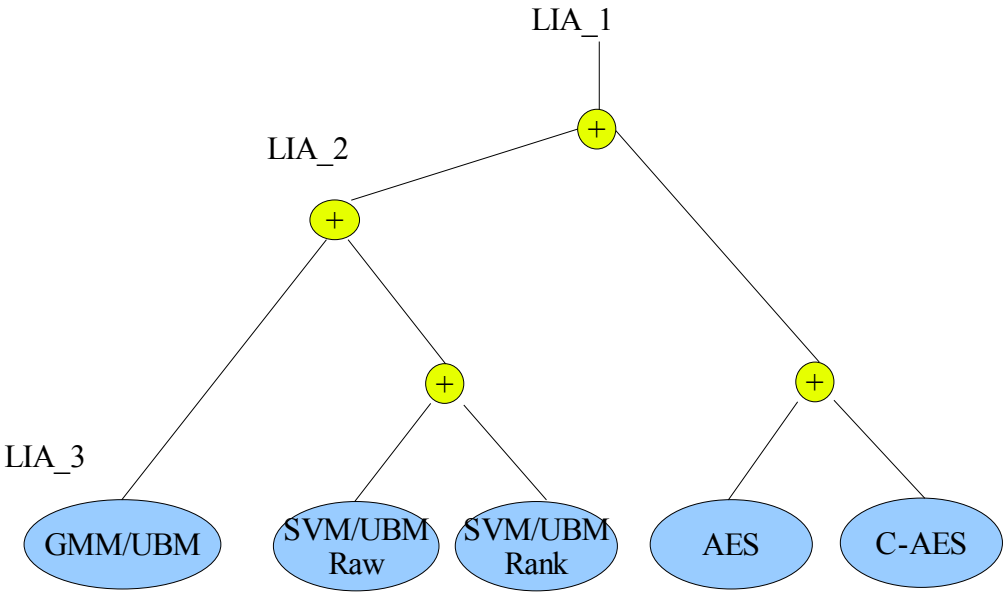
The different LIA systems resides in the combination process:

- LIA_3 is the GMM/UBM system,
- LIA_2 is the GMM/UBM + SVM/UBM combined together,
- LIA_1 is the combination of all the systems presented.

The combination process consists in:

- Performing a Tnormalization on each system,
- Tune fusion parameter on SRE'05,
- Apply these parameters on SRE'06.

The fusion this year is an arithmetic mean of the different systems in an hierarchical way. The figure below illustrates the process.



[1]. *"UBM driven discriminative approach for Speaker Verification"*, Nicolas SCHEFFER, Jean-François BONASTRE, Odyssey 2006.
[2]. "*Speaker Verfication using Acoustic Event Sequences"*, Nicolas SCHEFFER, Jean-François BONASTRE, EUROSPEECH 2005
[3]. "*ALIZE Toolkit",* http://www.lia.univ-avignon.fr/heberges/ALIZE/