LIA human-based system description for NIST HASR 2010

Nicolas Audibert[1], *Anthony Larcher*[1], *Christophe Lévy*[1], *Juliette Kahn*[1,2]
Solange Rossato[2], *Driss Matrouf*[1], *Jean-Francois Bonastre*[1]

[1]University of Avignon - CERI/LIA - France
[2]University of Grenoble - LIG - France
{firstname.lastname}@univ-avignon.fr, solange.rossato@imag.fr

1. Introduction

This paper describes the participation of the LIA laboratory to the Human Assisted Speaker Recognition (HASR) evaluation, which is part of the NIST-SRE 2010 campaign. The submission of the LIA for this task is based on a human decision. Samples were rated by three listeners, system decision being based on majority voting. Confidence scores were defined by mapping human decision to scores distribution of a SVM-based automatic system.

The algorithms used for listening stimuli generation and the protocol for samples listening and rating are first described in section 2. Subsections 2.1 and 2.2 describe the algorithms used for automatic extracts selection from each model or test segment, and for extracts normalisation and concatenation. Subsection 2.3 describes the listeners involved and the listening protocol. Subsection 2.4 presents the calculations made on human decisions to obtain scores submitted to NIST. In section 3, the automatic system used for scores mapping is presented. The computing times required by different steps of the automatic processing are listed in section 3. Finally, the characteristics of the submitted system are summarized and perspectives for future work are presented in section 5.

2. Human evaluation protocol

2.1. Extracts selection

For each trial, 6 seconds-long extracts are automatically selected from the model and test segments and concatenated to build the audio stimulus presented to listeners. This selection is achieved by means of tools implemented in the MISTRAL/ALIZE [3] toolkit.
In order to perform extracts selection, recordings are preprocessed by using Linear-Frequency Cepstral Coefficients (LFCC). The extraction of these parameters is described in Section 3.1. Once the LFCC parameters are computed, the energy coefficients are first normalised using a mean removal and variance normalisation in order to fit a 0-mean and 1-variance distribution. The energy component is then used to train a three component GMM, which aims at selecting informative frames. The frames carrying the highest level of energy are selected through the GMM and labeled *speech*. Once the *speech* segments of a signal are selected, a final process is applied in order to refine the speech segmentation:

1. Overlapping speech segments between both sides of a conversation are removed, in order to avoid selecting speech turns of the interviewer that can be heard in the channel of interest;

2. Morphological rules are applied on speech segments by adding or removing speech frames, to obtain 6 seconds-long segments with a proportion as large as possible of speech frames.

For each model or test segment, the minimum number of 6 seconds-long selected segments is set to 7 (i.e. a minimum total duration of 42 seconds for each file of a model/test pair). This selection is achieved by applying strict morphological rules as a first step, i.e. selecting only segments with a large proportion of speech frames, and iteratively decreasing the selection threshold when necessary until the minimum number of selected segments is reached. Although this method generally succeeds in selecting extracts that mainly include speech frames corresponding to the interviewee speech turns, for 3 model/test pairs out of 150 (2 male target speakers, 1 female) it was unable to find in either the model or test segment extracts including enough useful information to make human decision possible. As a result, the selection of appropriate 6 seconds-long extracts was performed manually by one of the listeners for these 3 files.

2.2. Rules for the generation of stimuli

Selected extracts are then combined in the audio stimulus, generated using the Praat software [9]. Extracts from the model segment in the one hand and from the test segment in the other hand are chosen alternatively. A 1000 Hz, 50 milliseconds-long beep surrounded by two 75 milliseconds-long silent parts is inserted between consecutive segments to signal inter-extract switching. All

extracts included in the generated stimulus are normalized to the same acoustic intensity. This level of normalized intensity was set to 70 dB on HASR data, and lowered when necessary (down to 66 dB for 4 model/test pairs) to avoid clipping.

2.3. Human evaluation participants and protocol

Three native French listeners (2 female aged 25 and 36, 1 male aged 31) with experience in phonetics and speech analysis, and without any known hearing impairment, evaluated independently the 150 stimuli generated. For each trial, they were requested to decide whether the extracts alternated in the stimulus had been uttered by the same speaker or not. Although this information was not directly used in submitted results, listeners were also requested to rate their confidence in this judgment in a 0-5 scale for further analyses.
Listeners evaluated the stimuli in a quiet environment using closed headphones. For stimuli with a considerable level of noise (especially low-frequency noise) in either the model or test segment, they were given the possibility to band-filter the signal using the Praat software [9] after visual inspection of the power spectrum, in order to reduce the perceptual heterogeneity caused by the difference of recording channels. When listeners considered a single listening as not sufficient for decision making, they were allowed to listen to the generated stimulus by selecting parts or as a whole as many times as necessary. Listeners took 12 to 180 seconds for decision making (mean: 66 seconds).

2.4. System scoring

For each trial, the decision of the human system submitted to NIST is defined by majority voting among the decisions taken by the three listeners. The confidence score submitted is defined from the level of agreement between listeners. In order to make comparisons between human performances and that of the SVM-based automatic system described in section 3 possible, the inter-listener level of agreement is mapped to the impostor and client scores distribution obtained with this system and SRE 2008 data. Prior to evaluating HASR data, an experiment was therefore performed by running the SVM-based system (cf. section 3) according to the NIST-SRE 08 short2-short3 protocol. In order to map the human decision on the automatic speaker recognition framework, mean and variance of both the client and impostor scores distributions were estimated.
For each HASR trial, the mapping values presented in table 1, defined according to these score distributions, are selected as a function of inter-listener agreement and target speaker gender. Table 1 also indicates the number of trials corresponding to each inter-listener level of agreement for each target speaker gender. Overall, all three listeners made the same decision on 51% of trials.

3. Description of the automatic system

The speaker recognition system chosen to determine the human decision scores is a classical GMM-SVM system (Gaussian Mixture Model - Support Vector Machine) using Latent Factor Analysis (LFA) [1], [2] for session variability modelling. This system is based on the open-source biometric platform MISTRAL/ALIZE [3].

3.1. Front-end processing

Parameters extracted from speech signals (using the open source SPro toolkit [4]) are based on a filter-bank analysis (linear filter). Feature vectors are composed of 19 Linear-Frequency Cepstral Coefficients (20ms window, 10ms shift), their derivatives, the first 11 second derivatives and the delta energy. The frequency window is restricted to 300-3400 Hz. An energy labeling is performed on the signal and only the frames deemed to be speech are processed by the speaker recognition engine. Then simple feature normalization is applied, so that the distribution of each cepstral coefficient is 0-mean and 1-variance for a given utterance.

3.2. World model

The UBM model size is set to 512 components (with diagonal covariance matrix). The UBM consists of a GMM trained on telephone conversations from the Fisher English database [5] and microphone recordings from the NIST-SRE 2005 database.

3.3. Speaker model using Factor Analysis

According to the Latent Factor Analysis (LFA) modelling [1], speaker models are formed of three different components: a speaker and session independent background model, a speaker dependent and a session dependent components [6], [2]. The resulting model can be written as:

$$\mathsf{m}_{(h,s)} = \mathsf{m} + Dy_s + Ux_{(h,s)} \quad (1)$$

where $\mathsf{m}_{(h,s)}$ is the session-speaker dependent mean super-vector, D is $S \times S$ diagonal matrix (S is the dimension of the supervector), y_s the speaker vector, U is the eigenchannel matrix of low rank R (a $S \times R$ matrix) and $x_{(h,s)}$ are the session factors. Both y_s and $x_{(h,s)}$ are normally distributed among $\mathcal{N}(0,1)$. D satisfies the following equation $I = \tau D^t \Sigma^{-1} D$ where τ is the relevance factor required in the standard MAP adaptation.

3.4. SVM modelling

According to Equation 1, the Factor Analysis model estimates speaker supervectors normalized with respect to

Listeners decision	System decision	Confidence score calculation	Male score (N=36)	Female score (N=114)
3 false	Certain false	Average impostor score - 2 σ_{imp}	-2.45 (N=12)	-2.13 (N=33)
2 false, 1 true	Uncertain false	Average impostor score	0.62 (N=8)	0.69 (N=20)
1 false, 2 true	Uncertain true	Average client score	6.51 (N=7)	5.44 (N=39)
3 true	Certain true	Average client score + 2 σ_{target}	12.19 (N=9)	10.85 (N=22)

Table 1: Mapping of human decision and SVM-based automatic system scores, for each target speaker gender and each inter-listener level of agreement .

the session variability. A distance between GMMs is computed using a probabilistic kernel K [7]. This distance, well suited for a SVM classifier when only mean parameters of the GMM models are adapted, is given by Equation 2 for two sequences $\mathcal{X}_s$ and $\mathcal{X}'_s$ respectively spoken by speakers s and s'.

$$K(\mathcal{X}_s, \mathcal{X}'_s) = \sum_{g=1}^{M} (\sqrt{\alpha_g}\, \Sigma_g^{-\frac{1}{2}}\, \mathsf{m}_s^g)^t (\sqrt{\alpha_g}\, \Sigma_g^{-\frac{1}{2}}\, \mathsf{m}_{s'}^g) \quad (2)$$

where m_s is taken form the model in Equation 1 ($\mathsf{m}_s = \mathsf{m} + Dy_s$), and Σ_g is the covariance matrix of the component g shared by the two models.

The LIA SpkDet toolkit benefits from the LIB-SVM library [8] to estimate SVMs and classify instances. SVM are trained with an infinite (very large in practice) C parameter thus avoiding classification error on the training data (hard margin behavior). The negative labeled examples are speakers form the normalization cohort.

3.5. Automatic system performance

The system was developed on NIST SRE 2008 data. Performance of this system are reported in Table 2 for the 8 conditions of NIST-SRE 2008

4. Computation time

Each model segment is approximately 180 seconds-long, while each test segment is approximately 300 seconds-long. In addition to the time required by human processing presented in section 2.3, table 3 presents the computation time required for a trial by each step of the automatic processing of speech signals. The parameterization of speech signals, used both for listening stimuli generation and by the automatic system, is performed only once.

5. Conclusions

The results submitted for the HASR2 part of the HASR 2010 evaluation were based on majority voting by three listeners, after automatic selection of extracts of interests from the model and test segments and their concatenation in a listening stimulus for each of the 150 trials. Submitted confidence scores were obtained by mapping human decision to scores distribution obtained on SRE 2008 data with the SVM-based automatic system presented in section 3.

Comparison of human vs. automatic system performances will be presented at the NIST SRE workshop, together with an analysis of human performances. Moreover, human performance analysis will be extended by using individual confidence scores and by evaluating differences between the model and test segments of each trial according to numerous perceptual dimensions, including channel differences, specific phonetic and prosodic features, and speakers affective states.

6. References

[1] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, and Jean-Francois Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *International Conference on Speech Communication and Technology*, 2007.

[2] Patrick Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2005, vol. 1.

[3] Jean-François Bonastre, Nicolas Scheffer, Driss Matrouf, Corinne Fredouille, Anthony Larcher, Alexandre Preti, Gilles Pouchoulin, Nicholas Evans, Benoît Fauve, and John S.D. Mason, "ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2008, http://mistral.univ-avignon.fr/.

[4] G. Gravier, "SPro: speech signal processing toolkit," *Software available at http://gforge. inria. fr/projects/spro*.

[5] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Fourth International Conference on Language Resources and Evaluation*, 2004.

[6] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *European Conference on Speech Communication and Technology (Eurospeech)*. ISCA, 2005.

NIST-SRE08 test Condition		det1	det2	det3	det4	det5	det6	det7	det8
Male	**EER**	6.69	1.22	6.68	8.42	4.69	5.37	2.28	1.31
	DCFmin $\times 100$	3.23	0.40	3.25	2.82	2.07	3.35	1.26	0.74
Female	**EER**	10.03	2.10	9.88	10.81	8.55	8.59	3.55	3.95
	DCFmin $\times 100$	4.55	0.53	4.44	4.46	3.13	4.57	1.65	1.68

Table 2: Performance (% EER and DCFmin) of the SVM system used for HASR score mapping.

Automatic processing step	Mean time (seconds)	σ (seconds)	Time (xRT)
Signal parametrization	3.70	0.08	0.008 xRT
Listening stimulus building	0.65	0.14	0.001 xRT
Automatic speaker verification	26.88	0.94	0.056 xRT

Table 3: Computation time mean per trial and standard deviation, for each automatic processing step. Mean computation time is also indicated as a multiple of real time.

[7] W.M. Campbell, DE Sturim, and DA Reynolds, "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308, 2006.

[8] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," .

[9] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer (Version 5.0. 38)[Computer program]," *Retrieved November*, vol. 1, pp. 2008, 2008.