LIA human-based system description for NIST HASR 2010
Nicolas Audibert, Anthony Larcher, Christophe Lévy, Juliette Kahn, Solange Rossato, Driss Matrouf, Jean-François Bonastre

To cite this version:
Nicolas Audibert, Anthony Larcher, Christophe Lévy, Juliette Kahn, Solange Rossato, et al.. LIA human-based system description for NIST HASR 2010. 2010. hal-01317695

HAL Id: hal-01317695
https://hal.archives-ouvertes.fr/hal-01317695
Submitted on 19 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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.

This paper describes in section 3, the automatic system used for scores mapping is presented. In section 2 the algorithms used for listening stimuli generation and the protocol for samples listening and rating. 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. 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

In order to perform extracts selection, recordings are pre-processed by using Linear-Frequency Cepstral Coefficients (LFCC). 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;

2. Morphological rules are applied on speech segments to discard segments too short to be used in a listening task, by adding or removing speech frames.

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. Extracts selection was guided by 2 criteria:

- Selected extracts should include a proportion as large as possible of speech frames;
- For both model and test segments, the extracts should include only speech frames corresponding to the interviewee, excluding speech turns of the interviewer audible in the channel of interest.

This selection is achieved by means of tools implemented in the MISTRAL/ALIZE [1] toolkit.

Once the signal parameterization and speech frames detection described supra is performed, 6 seconds-long extracts are extracted from the channel of interest of the model and test segments. Interventions of the interviewer are excluded thanks to the speech frames redundancy across channels. For each original file (model and test segment), the initial threshold is set to a minimum of 70% speech frames not appearing in the other channel and iteratively decreased when necessary to end up with a minimum number of selected segments of 7 (i.e. a minimum total duration of 42 seconds for each file of a model/test pair). Although this method generally succeeds in selecting extracts that mainly include speech
Table 1: Mapping of human decision and SVM-based automatic system scores, for each target speaker gender and each inter-listener level of agreement.

<table>
<thead>
<tr>
<th>Listeners decision</th>
<th>System decision</th>
<th>Confidence score calculation</th>
<th>Male score (N=36)</th>
<th>Female score (N=114)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 false</td>
<td>Certain false</td>
<td>Average impostor score - $2 \sigma_{imp}$</td>
<td>-2.45 (N=12)</td>
<td>-2.13 (N=33)</td>
</tr>
<tr>
<td>2 false, 1 true</td>
<td>Uncertain false</td>
<td>Average impostor score</td>
<td>0.62 (N=8)</td>
<td>0.69 (N=20)</td>
</tr>
<tr>
<td>1 false, 2 true</td>
<td>Uncertain true</td>
<td>Average client score</td>
<td>6.51 (N=7)</td>
<td>5.44 (N=39)</td>
</tr>
<tr>
<td>3 true</td>
<td>Certain true</td>
<td>Average client score + $2 \sigma_{target}$</td>
<td>12.19 (N=9)</td>
<td>10.85 (N=22)</td>
</tr>
</tbody>
</table>
3.1. Front-end processing
Parameters extracted from speech signal (using the open source SPro toolkit [5]) are based on a filter-bank analysis (linear filter). Feature vectors are composed of 19 Linear-Frequency Cepstral Coefficients (20ms window, 10ms shift), its derivatives, the first 11 second derivatives and the delta energy. The frequency window is restricted to 300-3400 Hz. An energy labelling is performed on the signal and only the frames deemed to be speech are processed by the speaker recognition engine. Then simple feature normalisation 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 [6] and microphone recordings from the NIST-SRE 2005 database.

3.3. Speaker model using Factor Analysis
According to the Latent Factor Analysis (LFA) modelling [3], speaker models are formed of three different components: a speaker and session independent background model, a speaker dependent and a session dependent components [7], [4]. The resulting model can be written as:

\[ m_{(h,s)} = m + Dy_s + Ux_{(h,s)} \]  (1)

where \( m_{(h,s)} \) 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 \( \tau \) 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 normalised with respect to the session variability. A distance between GMMs is computed by using a probabilistic kernel \( K \) [8]. 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 \( s' \).

\[ K(\mathcal{X}_s, \mathcal{X}_s') = \sum_{g=1}^{M} (\sqrt{\alpha_g \Sigma^{-\frac{1}{2}} g m_s^g})^t (\sqrt{\alpha_g \Sigma^{-\frac{1}{2}} g m_s'^g}) \]  (2)

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

The LIA SpkDet toolkit benefits from the LIB-SVM library [9] 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 behaviour). The negative labelled examples are speakers form the normalisation 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

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

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

<table>
<thead>
<tr>
<th>Automatic processing step</th>
<th>Mean time (seconds)</th>
<th>σ (seconds)</th>
<th>Time (xRT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal parametrization</td>
<td>3.70</td>
<td>0.08</td>
<td>0.008 xRT</td>
</tr>
<tr>
<td>Listening stimulus building</td>
<td>0.65</td>
<td>0.14</td>
<td>0.001 xRT</td>
</tr>
<tr>
<td>Automatic speaker verification</td>
<td>26.88</td>
<td>0.94</td>
<td>0.056 xRT</td>
</tr>
</tbody>
</table>

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.

5. Conclusions

The results submitted for the HASR part of the HASR 2010 evaluation were based on majority voting by 3 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


