Classical vs. Biometric Features in the 2013 Speaker Recognition Evaluation in Mobile Environments

Luis Miguel Mazaira-Fernández, Agustín Álvarez Marquina, Pedro Gómez Vilda, Rafael Martínez-Olalla, Cristina Muñoz-Mulas

Grupo de Informática Aplicada al Procesado de Señal e Imagen (GIAPSI) Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid – SPAIN.

e-mail: luismiguel.mazaira@upm.es

Abstract. MFCC coefficients extracted from the power spectral density of speech as a whole, seems to have become the de facto standard in the area of speaker recognition, as demonstrated by its use in almost all systems submitted to the 2013 Speaker Recognition Evaluation (SRE) in Mobile Environment [1], thus relegating to background this component of the recognition systems. However, in this article we will show that selecting the adequate speaker characterization system is as important as the selection of the classifier. To accomplish this we will compare the recognition rates achieved by different recognition systems that relies on the same classifier (GMM-UBM) but connected with different feature extraction systems (based on both classical and biometric parameters). As a result we will show that a gender dependent biometric parameterization with a simple recognition system based on GMM-UBM paradigm provides very competitive or even better recognition rates when compared to more complex classification systems based on classical features.

Keywords: Speaker Characterization, Speaker Recognition, GMM-UBM, Source-Tract separation, MOBIO database.

1 Introduction

Although been an area that has been active at least since early 70s [2], Automatic Speaker Recognition (ASR) keeps on been an active research area. As a result of this interest, some institutions have provided a common framework to test the advances in this area. For instance, the National Institute of Standards and Technology (NIST) have been organizing the NIST SRE on a regular basis since 1996. The main objective pursued in these evaluations is to measure the state-of-the-art in text-independent speaker recognition subject to high variability covering different aspects, such as: varying length recordings (both in train and test), vocal effort variability, different

1 www.nist.gov

Disclaimer: This is a draft copy (restricted to non-commercial personal use) of the paper published in the Proceedings of the JVHC 2013, I Jornadas Multidisciplinares de Usuarios de la Voz, el Habla y el Canto, Las Palmas de Gran Canaria - 27-28 de junio 2013, pp. 122-131, ISBN: 84-695-8101-5 © held by Authors
communication channels, etc. A more novel evaluation, (the second one taking place on 2013) is the one proposed by the Biometric Group at the Idiap Research Institute, which is based on the MOBIO database [3]. The aim of this competition is mainly the same as the one of NIST SRE, i.e. determine whether a specified target speaker is present or not in a given segment of speech. The main difference with respect to NIST SRE is that this competition is restricted to speech recorded in mobile environments, and more important the average speech duration of MOBIO recordings are significantly lower (around 8s ) than those in NIST databases.

To meet the challenges of these evaluations different classification methods [4] (UBM-GMM, total variability spaces, Gaussian Supervectors, etc.) have been proposed combined with normalization post-processing steps such as LDA, WCCN, or NAP. However, regarding the front-end that feeds those classification systems, gender-independent MFCC coefficients extracted from the power spectral density of speech as a whole, seems to be the only choice, as demonstrated by the fact that all systems participating on the 2013 SRE on Mobile Environments used them [1].

Given that none of the presented evaluations performs cross-gender trials and that as is well known male and female voices present different characteristics (the clearest case is for example in pitch values), the present work defends the idea that a gender dependent parameterization that takes into account spectral characteristics of vocal tract (acoustic-phonetic) and glottal (phonation-gesture) estimates of voicing speech, as well as classical MFCC, will provide a better characterization of speakers and thus help to increase recognition rates.

The paper is organized as follows: section 2 briefly describes both the MOBIO database as well as the 2013 SRE in Mobile Environments. Section 3 presents both the biometric gender-dependent front-end as well as the whole recognition system build to participate in the SRE. Section 4 evaluates the performance of the presented system and compares it with the other systems. Finally, some conclusions will be exposed in section 5.

2 The MOBIO database and the 2013 SRE in Mobile Environments

The MOBIO database is a bi-modal database that consists of faces and voice information, captured on mobile devices (actually using a NOKIA 93i mobile phone with a sampling rate of 16 kHz). The database contains a total of 152 speakers (100 male and 52 female) that have been recorded in 2 different phases, with 6 different sessions in each phase, in 6 different sites. Regarding the audio data, all data was collected in English, though English may not always be the native language of the speakers. Additionally, as it was captured using a mobile device (not placed in a fixed position), it contains high variability in terms of quality and acquisition environments (which means real noise background). A deep description of the MOBIO database can be found in [3].
Regarding the competition, the main objective, as in NIST SRE, is to determine whether a specified target speaker is present or not in a given segment of speech. In order to provide a common framework for all systems taking part in the evaluation, the speakers (and thus its recordings) included in the MOBIO database, are split into three different subsets as follows:

- **Background training set.** This subset is only used to learn the background parameters of the algorithm (UBM, subspaces, etc.) or for normalization purposes.

- **Development set.** The data assigned to this set is split into two subsets: enrolment and test. The first one is used to create a model of each of the target speakers included. The second one contains a list of audio samples that must be tested against all the target speakers. The data on this set is supposed to be used to tune meta-parameters of the algorithm (e.g. number of Gaussians, dimension of subspaces, etc.). The recognition rate, regarding EER, achieved with this development set is used to define a score threshold that will be used to evaluate the performance of the recognition systems.

- **Evaluation set.** The final evaluation performance is analysed using this set, which has a similar structure as the development set. A score must be provided for each trial, for instance in the form of log-likelihood, representing how accurately the test segment is classified as containing, or not, speech for the target speaker against which is confronted.

The performance of the systems is evaluated on terms of EER in two different ways. Using the development set, a score threshold, $\theta_{dev}$, is defined which provides the EER of that set. The threshold, $\theta_{dev}$, will provide the Half Total Error Rate (HTER) that can be defined as:

$$HTER = \frac{FAR_{eval}(\theta_{dev}) + FRR_{eval}(\theta_{dev})}{2}$$

where $FAR_{eval}$ is the False Acceptance Rate obtained using threshold $\theta_{dev}$, on the evaluation set, and $FRR_{eval}$ represents the False Rejection Rate achieved using threshold $\theta_{dev}$, on the evaluation set.

### 3 Gender-dependent biometric front-end and Recognition System Description.

A complete speaker recognition system has been design to participate in the 2013 SRE on Mobile Environments. This system can be divided into different blocks:

#### 3.1 Feature extraction.

In order to evaluate the influence of what we have called biometric parameters, an additional system (baseline front-end) was developed, also connected to the same modeling and scoring system. This baseline front-end performs a classical feature extraction, providing gender independent speaker features based on MFCC + $\Delta$. 
Alternatively to the feature set based on classical MFCC parameters where both male and female speakers are parameterized using the same set up, we propose an extended parameterization which integrates classical cepstral information with cepstral coefficients derived from vocal tract and glottal source estimates (biometric parameters) in a single feature vector. Moreover, as we have already pointed out, not only classical MFCC parameters are computed using a different set up depending on the gender, but also the MFCC parameters extracted from vocal tract and glottal source estimates are parameterized differently depending on the gender.

Fig. 2. Separation algorithm with lip radiation compensation using first order prediction lattice
The methodology used to separate vocal tract and glottal estimate of voice from continuous speech is based on the uncorrelation hypothesis between them and is carried out using Joint Process Estimation (JPE) algorithm. Fig. 2 shows the block diagram of the separation algorithm, while a comprehensive description of it can be found in [5].

Additionally to these new set of parameters, we have also investigated the used of other common parameters such as frame energy, delta energy, pitch and, formant 3 estimate.

The configuration finally sent to the SRE consisted of the following parameters:

- **Female Speakers (54-dim FV):**
  - 24MFCC+ΔMFCC (34mel-spaced filter bank)
  - 4MFCC (extracted from the glottal estimate)
  - F0 estimate
  - F3 estimate

- **Male Speakers (69-dim FV):**
  - 28MFCC+ΔMFCC (38mel-spaced filter bank)
  - 10MFCC (extracted from the glottal estimate)
  - ΔEnergy
  - F0 estimate
  - F3 estimate

### 3.2 Voice activity detection

An adaptive VAD algorithm based on energy detection has been implemented and computed over a 32ms-long window with 8ms overlap. Additionally, a built-in heuristic has been incorporated that removes or includes signal segments shorter than 56ms, depending on its relative location to longer voice segments. The algorithm has been tested on the MOBIO development dataset, providing accurate voice activity detection.

### 3.3 Feature Normalization

It is necessary to perform a noise reduction preprocessing step, as we are dealing with telephone conversations. In this case, a variation of the Ephraim-Malah spectral subtraction algorithm in a single channel is applied [6].

Once the set of MFCC feature vectors have been computed for the whole speech signal, the cepstral mean subtraction (CMS) [7] algorithm is applied. This algorithm mainly consists on computing the mean of each cepstral coefficient over the length of the current utterance, then the mean value is subtracted from the original cepstral coefficient, thus removing the channel induced effects as well as any other stationary speech component. Additionally, feature warping [8] has been applied to transform the original cepstral coefficients so that they follow a specific target distribution, for instance a normal distribution, over a window of speech frames, typically 3 seconds.
window. It provides a set of features that are supposed to be robust to channel mismatch, additive noise and nonlinear effects attributed to handset transducers. Finally, RASTA filtering [9] has been implemented in order to remove the spectral components that changes at different rate than the one present in speech, i.e. tries to remove convolutional and additive noise.

3.4 Classifier and model

Regarding the speaker modeling technique, we have applied the UBM-GMM (Universal Background Model – Gaussian Mixture Model) paradigm. The selection of this approach is twofold. First of all, GMM is a probabilistic model which has become the de facto reference method in text-independent speaker recognition. Second, the limited amount of data available on the MOBIO database (regarding both the number of speakers and channel variability) discourage the use of more complex recognition systems.

Specifically, we have used a standard mixture classifier with diagonal covariance matrix. Each speaker is represented by a GMM, \( \lambda_{\text{speaker}, k} \), which has been adapted from a gender-dependent UBM using the MAP algorithm [10] in which only the distribution means have been adapted. The UBM is also represented as a GMM, \( \lambda_{\text{UBM}} \), which has been trained on the MOBIO training set via the EM-algorithm. The number of Gaussians as well as the relevance factor used on the MAP-algorithm is in the case of female speakers 256 and 28 respectively; while for male speakers are 512 and 28.

Log-likelihood ratio (LLR) has been the score used to take a decision on whether a test audio segment is likely to be spoken by a specific speaker represented by a model \( \lambda_s \).

\[
\text{LLR} = \log P(X | \lambda_s) - \log P(X | \lambda_{\text{UBM}})
\]

3.5 Score Normalization.

Znorm [11], Tnorm [12] and ZTnorm have been tested using gender dependent cohorts, extracted from the MOBIO training set. However, only ZT norm score normalization, applied to male speakers, have provided some improvement on development test in terms of EER.

4 Performance Results

A battery of test has been conducted on the development set, in order to find the best configuration, in terms of number of MFCC, using the baseline front-end, which minimizes the EER for both genres. As expected, the results achieved, in terms of EER, with a gender independent configuration (GIC) are slightly worse than the ones obtained using a gender dependent characterization (GDC) even in the case of using just MFCC coefficients extracted from the power spectral density of speech as whole.
Additionally, we have introduce some alternative parameters (but also typically used in speaker recognition) that are added to the feature vector either alone or combined, to the GDC. Specifically, we have tested Energy, ΔEnergy, Pitch (F0) and formant F3. Finally, we have evaluated the performance of the gender-dependent biometric front-end. The approach that has been followed consists on incorporating the biometric coefficients (GDBP) into the best gender dependent configuration achieved so far (as presented in section 3.1).

The bests results achieved, on development set, are shown in table 1 (male) and table 2 (female), as well as the relative reduction in terms of EER if compare to the GIC. It must be noted that despite receiving the same name GDC and GDBP are different configurations depending on the gender. Additionally DET curves are presented in Fig.3.

Table 1. EER achieved for different configurations on development set, for male speakers, applying ZT Norm or no score normalization. Best results highlighted in bold.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>MALE</th>
<th></th>
<th>Relative Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZT Norm</td>
<td>No Norm</td>
<td>ZT Norm</td>
</tr>
<tr>
<td>GIC</td>
<td>10.31%</td>
<td>11.58%</td>
<td>-</td>
</tr>
<tr>
<td>GDC</td>
<td>9.70%</td>
<td>10.97%</td>
<td>5.92%</td>
</tr>
<tr>
<td>GDC+ ΔE+ F0 + F3</td>
<td>9.16%</td>
<td>10.47%</td>
<td>11.15%</td>
</tr>
<tr>
<td>GDBP</td>
<td>8.12%</td>
<td>8.91%</td>
<td>21.24%</td>
</tr>
</tbody>
</table>

Table 2. EER achieved for different configurations on development set, for female speakers, applying ZT Norm or no score normalization. Best results highlighted in bold.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>FEMALE</th>
<th></th>
<th>Relative Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ZT Norm</td>
<td>No Norm</td>
<td>ZT Norm</td>
</tr>
<tr>
<td>GIC</td>
<td>11.59%</td>
<td>12.16%</td>
<td>-</td>
</tr>
<tr>
<td>GDC</td>
<td>11.25%</td>
<td>11.64%</td>
<td>2.93%</td>
</tr>
<tr>
<td>GDC+ F0 + F3</td>
<td>11.57%</td>
<td>11.20%</td>
<td>0.17%</td>
</tr>
<tr>
<td>GDBP</td>
<td>10.79%</td>
<td>10.44%</td>
<td>6.90%</td>
</tr>
</tbody>
</table>

After these test carried out on the MOBIO database, and specifically in the development set, we can draw the following conclusions. First of all, a gender dependent characterization provides a clear improvement in the recognition rates. The use of additional parameters such as Pitch and formant F3 (and ΔEnergy in male speakers) also causes a decrease of the EER especially in the case of male speakers. Moreover, incorporating what we have called biometric parameters, results in a clear improvement of recognition rates. Finally, ZT Norm does not appear to provide a clear improvement in recognition rates for female speakers, unlike in the case of male speakers. This may be due, to the limited number of female speakers in the training set.
Finally, table 3 summarizes the results obtained in the SRE by different systems in both development and evaluation set. Systems marked with *, are actually fusion of different systems, while systems marked with + are those who used external/additional training data. It must be noted that, despite having developed a simple recognition system (based on the UBM-GMM paradigm), the fact of having achieved a better speaker’s characterization based on gender-dependent biometric parameters, allows us to get very competitive results. Moreover, the systems that improve the recognition rates of our system are only those that either performed a fusion of multiple systems or used additional data for training.
Table 3. EER % on the development (DEV) set and half total error rate (HTER %) on the evaluation (EVAL) set for the systems participating in 2013 SRE in Mobile Environments (from [1]).

<table>
<thead>
<tr>
<th>System</th>
<th>FEMALE DEV</th>
<th>FEMALE EVAL</th>
<th>MALE DEV</th>
<th>MALE EVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpineon*</td>
<td>7.982%</td>
<td>10.678%</td>
<td>5.040%</td>
<td>7.076%</td>
</tr>
<tr>
<td>ATVS+</td>
<td>16.836%</td>
<td>17.858%</td>
<td>14.881%</td>
<td>15.429%</td>
</tr>
<tr>
<td>CPqD*</td>
<td>14.348%</td>
<td>15.987%</td>
<td>11.824%</td>
<td>10.214%</td>
</tr>
<tr>
<td>CDTA</td>
<td>19.471%</td>
<td>22.640%</td>
<td>12.738%</td>
<td>19.404%</td>
</tr>
<tr>
<td>GIAPSI2</td>
<td>11.590%</td>
<td>12.813%</td>
<td>9.683%</td>
<td>8.865%</td>
</tr>
<tr>
<td>GIAPSI3</td>
<td>10.440%</td>
<td>13.115%</td>
<td>8.125%</td>
<td>8.854%</td>
</tr>
<tr>
<td>EHU</td>
<td>17.937%</td>
<td>19.511%</td>
<td>11.310%</td>
<td>10.058%</td>
</tr>
<tr>
<td>IDIAP</td>
<td>12.011%</td>
<td>14.269%</td>
<td>9.960%</td>
<td>10.032%</td>
</tr>
<tr>
<td>L2F*</td>
<td>13.484%</td>
<td>22.140%</td>
<td>10.599%</td>
<td>11.129%</td>
</tr>
<tr>
<td>L2F-EHU*</td>
<td>11.005%</td>
<td>17.266%</td>
<td>7.889%</td>
<td>8.191%</td>
</tr>
<tr>
<td>Mines-Telecom+</td>
<td>11.429%</td>
<td>11.633%</td>
<td>10.198%</td>
<td>10.779%</td>
</tr>
<tr>
<td>Phonexia+</td>
<td>8.364%</td>
<td>14.181%</td>
<td>9.601%</td>
<td>9.109%</td>
</tr>
<tr>
<td>RUN+</td>
<td>25.405%</td>
<td>23.112%</td>
<td>24.643%</td>
<td>22.524%</td>
</tr>
</tbody>
</table>

5 Conclusions

In order to test the importance of an accurate front-end to better characterize speakers, a complete system has been developed to participate in an international evaluation contest, 2013 SRE in Mobile Environments. It has been shown that the use of a gender-dependent biometric representation provides a more accurate description of the speakers than the one based on classical gender-independent MFCC. Thus confirming conclusions achieved in previous works [13][14] with different databases.

Additionally, as we are dealing with text-independent trials, it seems that including information, as MFCC parameters, from the vocal tract estimate provides no additional benefit. This may be mainly due to the fact that this component, in which voice can be split, is more related to the message carried out by voice rather than to the biometry of the speaker as is the glottal source estimate.

Although the developed system meets the state-of-art requirements is probably the simplest presented system, but at the same time achieves very competitive results thanks to incorporate enhanced description of speakers. Moreover, according to the published results [1] our system gets the best simple system performance on male. However, results seem to be still far away for the best results that can be achieved fusing all the presented systems (male eval. HEER=6.986%, female eval. HEER=4.767), providing still some room for improvement by using the presented gender-dependent biometric front-end combined with more promising classifiers such as GSV or i-vectors.

---

2 Original results.
3 Results achieved after post-processing.
6 Acknowledgement

This work is being funded by grant TEC2009-14123-C04-03 from Plan Nacional de I+D+i, Ministry of Science and Innovation; by project HESPERIA (http://www.proyecto-hesperia.org) from the Program CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain, and by project TEC2012-38630-C04-04.

References