Audio-Visual Gender Recognition in Uncontrolled Environment Using Variability Modeling Techniques

Laurent El Shafey, Elie Khoury, Sébastien Marcel
Idiap Research Institute
Martigny, Switzerland
{laurent.el-shafey, elie.khoury, sebastien.marcel}@idiap.ch

Abstract

The problem of gender recognition using visual and acoustic cues has recently received significant attention. This paper explores the use of Total Variability (i-vectors) and Inter-Session Variability (ISV) modeling techniques for both unimodal and bimodal gender recognition, and compares them to several state-of-the-art algorithms. The experimental evaluation is conducted on the FERET and LFW databases for face-based gender recognition, on the NIST-SRE database for audio-based gender recognition, and on the MOBIO database for audio-visual gender recognition. Results on LFW show that the i-vectors technique outperforms state-of-the-art algorithms, which are based on Support Vector Machines (SVM) applied either on raw pixels, on Local Binary Patterns (LBP) or on Gabor filters, with an accuracy rate of about 95%*. Results on NIST-SRE show that the i-vectors system is also superior to state-of-the-art GMM-based gender recognition systems, with a relative gain of about* 11%*. Finally, results on MOBIO show that i-vectors and ISV also take advantage of combining visual and acoustic cues using logistic regression. The resulting bimodal systems achieve accuracy rates of about* 98%*.*

1. Introduction

Information about gender, age, ethnicity, and emotional state are important ingredients that lead to rich behavioral informatics. Such information can be extracted from visual or audio modalities. In this work, we focus on the problem of gender recognition using both visual and audio cues.

Automatic gender recognition is crucial for a number of applications of human-computer or human-robot interaction. It serves to (1) enrich the metadata of visual and audio documents in an indexing and retrieval system, (2) improve the efficiency and the accuracy of people (both face and speaker) recognition, diarization and surveillance systems by reducing the search space to subjects from the same gender, and by building gender-dependent models which are often better than gender-independent models, (3) enhance human-machine interaction by suggesting user-friendly interface (e.g. gaming, social networks) and personalized advertisements (e.g. interactive voice response system, in-store cameras), (4) increase the intelligibility of human-robot interaction, and (5) collect passive demographic data.

Due to these various applications, the problem of automatic gender recognition has recently received significant attention. Researchers have often addressed this problem with a unimodal aspect. For image-based gender recognition, readers can refer to [23, 32, 19, 20, 1]. For audio-based gender recognition, one can cite the work of [14, 15, 10, 5, 31, 17]. In contrast, only few works (e.g. [18, 26]) have taken into account both visual and acoustic cues to solve the problem of gender recognition. They have shown that audio-visual fusion can improve the accuracy of gender classification system especially under degraded conditions and temporal unavailability of one of the modalities. However, their evaluations were conducted on in-house or small databases, using simplistic unimodal systems.

In this work, we explore the recently proposed *total variability* (**TV**, also known as i-vectors) [9] and *inter-session variability* (**ISV**) [34] techniques for both audio-based and face-based gender recognition problems. We apply logistic regression [25] to combine the two modalities at the decision level. The proposed systems are compared to several unimodal and bimodal state-of-the-art algorithms. The experimental evaluation is conducted on the FERET and LFW databases for face-based gender recognition, on the NIST-SRE database for audio-based gender recognition, and on the MOBIO dataset for audio-visual gender recognition.

The remainder of this paper is structured as follows: In Section 2, we review existing work on unimodal- and multimodal-based gender recognition. Section 3 presents the proposed unimodal gender recognition systems, relying on the **TV** and **ISV** techniques. Section 4 describes the evaluation metrics, the databases, the experimental setup and the results. Section 5 concludes the paper.

2. Related work

2.1. Visual gender recognition

Gender recognition from face images has received significant attention recently, and several approaches were explored. In [23], authors show that *support vector machines* (SVMs) are superior to *linear discriminant analysis* (LDA), *nearest-neighbor*, and *radial basis function networks*. They conducted their experiments on images selected from the FERET database [24]. However, the experimental protocol suffers from lack of information that prevents the reproducibility of the results.

Authors in [19] made an effort towards making available the details about the protocol used on FERET, and thus helping to benchmark the different approaches. One of the findings of their work is that SVMs (with face pixels) are slightly superior to *neural networks* (with face pixels) and *AdaBoost* (with Haar-like features). In [1], authors proposed an approach that combines several SVM classifiers applied on intensity, shape and texture features gathered at different scales. This approach obtains a better accuracy than the techniques presented in [19]. The drawback of the FERET database is that the images are acquired in controlled conditions, and the dataset corresponding to the available protocol [19] is small (only 411 face images).

Gallagher and Chen [12] used contextual features to recognize people's gender in images of groups of people (family portraits, wedding photos, etc.). Their images were collected from Flickr (uncontrolled conditions) and made available for researchers [11]. However, they did not provide a standard evaluation protocol.

More recently an evaluation protocol[1] on *Labeled Faces in the Wild* (LFW) database was proposed as one of the BeFIT (*Benchmarking Facial Image Analysis Technologies*) challenges. As for Gallagher's database, LFW images are acquired in realistic scenarios under large variability in illumination, facial expressions and head pose. Authors in [7] used this protocol to evaluate state-of-the-art systems. They found that *Gabor jets* and *local binary patterns* (LBPs) obtain similar accuracy rates, and perform generally better than pixels. They also found that SVM classifier works slightly better than LDA.

2.2. Acoustic gender recognition

Several approaches were proposed to cope with the problem of acoustic gender recognition. In [14], *Mel frequency spectral coefficients* (MFSC) with *neural networks* were used. Their database was collected from French and English radio stations. However, the details needed to replicate the experiments were not provided. Authors in [15] proposed a two-stage classifier where pitch thresholding is applied in the first stage, and *Mel frequency cepstral coefficients* (MFCC) extraction followed by GMM-based classification is done in the second stage. In [10], authors described an unsupervised system that jointly uses MFCC-based and Pitch-based classifiers, and any disagreement between the two classifiers is resolved by using a pitch-shifting mechanism. In both [15] and [10], the experiments were conducted on clean data (TIDIGITS in [15] and TIMIT in [10]). This explains the high accuracy rates reported in their work.

In 2010, a challenge on gender and age detection was conducted [31] on the aGender database [5], which contains recordings from German telephone speech. Several gender recognition algorithms were explored on this database such as GMM, SVM, MLP, GMM-Mean-SVM, GMM-MLLR-SVM, using both prosodic and acoustic features. Readers can refer to the work in [17] where seven sub-systems based on SVM and GMM were evaluated and combined. aGender contains one group of children speakers. However, its evaluation protocol does not distinguish between female and male children. This makes it difficult to be used independently for gender recognition.

In all these works, none of the recently proposed *inter-session variability* (**ISV**) [34] and *total variability* (**TV**) [9] modeling techniques were explored.

2.3. Audio-visual gender recognition

Contrarily to unimodal gender recognition, the audio-visual gender recognition has not been well explored in the literature. To the best of our knowledge, the first attempt of recognizing the gender using acoustic and visual cues was done in [35]. In this work, authors found that SVMs are better than *nearest-neighbor* and *k-nearest neighbors*. The main drawback of their work is that they used two separate unimodal databases to compare their audio and visual systems. This prevents them from making an objective and fair comparison between the two unimodal systems, and furthermore, it hinders them from combining both modalities to improve the performance of their system.

This issue was partially solved in the work of Liu *et al.* [18] where an audio-visual database was used. In their work, the audio gender classifier is based on GMM, whereas the visual classifier is based on SVM. The acoustic features used are the MFCC coefficients and their first derivatives, whereas the visual features used are the intensities of the pixels. At the fusion level, they combine the non-compatible scores (posterior probability for GMM and distances for SVM) from the two classifiers using a naive linear combination that was tuned directly on the test set. They reported gender classification accuracy rates of 85%, 84.75% and 91.25% on audio-only, visual-only and audio-visual cues, respectively. The main drawback of this work is the use of a private database without giving the full details about the conditions in which the data was collected.

[1] http://face.cs.kit.edu/431.php

3. Proposed audio-visual gender recognition

In this work, we propose to address the task of gender recognition by modeling the feature distribution using *Gaussian mixture models* (GMMs). Several classification and extraction techniques can be applied on top of this modeling for both visual and audio modalities. One possibility is to rely on the generative probabilistic framework for classification based on GMMs, introduced for speaker recognition in [27, 28] and then successfully applied for audio-based gender recognition [17]. Furthermore, to cope with the problem of high intra-class variability, we additionally investigate two recent session variability modeling techniques derived from GMMs: *inter-session variability* (**ISV**) [34] and *total variability* (**TV**) [9]. To the best of our knowledge, none of these two methods were used for gender recognition.

3.1. Feature distribution modeling using GMM

Two separate feature extraction processes are employed for face image and audio data. Considering visual data, a decomposition in the spatial domain is performed, leading to the extraction of parts-based features, as originally proposed for the task of face recognition in [29]. For audio data, the signal is decomposed in the time domain by extracting MFCC at equally-spaced time instants using a sliding window approach, as commonly performed in the field speaker recognition. For both modalities, this means that a set $\boldsymbol{O}$ of K feature vectors ($\boldsymbol{O} = \{\boldsymbol{o}^1, \boldsymbol{o}^2, \cdots, \boldsymbol{o}^K\}$) is extracted from each sample $\boldsymbol{\mathcal{O}}$, where each feature vector is of dimensionality M.

After feature extraction, and separately for each modality, the distribution of resulting feature vectors can be modeled using a GMM. A GMM is a weighted sum of C multivariate Gaussian components:

$$p(\boldsymbol{o}|\boldsymbol{\Theta}_{\text{gmm}}) = \sum_{c=1}^{C} \omega_c \mathcal{N}\left(\boldsymbol{o}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right) \,, \tag{1}$$

where $\boldsymbol{\Theta}_{\text{gmm}} = \{\omega_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=\{1,\ldots,C\}}$ are the weights, the means and the covariances of the model. This GMM can be seen as a codebook that represents the feature distribution. In the following, GMMs have diagonal covariance matrices.

3.2. Gaussian mixture modeling

To use GMMs for gender recognition, we need to learn a GMM $\boldsymbol{\mathcal{G}}_i$ for each gender ($i \in \{\text{male}, \text{female}\}$) from a set of enrollment samples. There are different ways to learn GMMs. As in [27], we employ the *expectation-maximization* algorithm to seek a *maximum-likelihood* estimate. Once gender-specific models $\boldsymbol{\mathcal{G}}_i$ are enrolled, the probability that a test sample $\boldsymbol{\mathcal{O}}_t$ is from the class male is given by a *log-likelihood ratio* (LLR) score:

$$h_{\textbf{GMM}}\left(\boldsymbol{\mathcal{O}}_t\right) = \ln p\left(\boldsymbol{\mathcal{O}}_t|\boldsymbol{\mathcal{G}}_{\text{male}}\right) - \ln p\left(\boldsymbol{\mathcal{O}}_t|\boldsymbol{\mathcal{G}}_{\text{female}}\right) \tag{2}$$

3.3. Inter-session variability modeling

Another technique to learn GMMs consists of training a generic model called a *universal background model* (GMM UBM), and to adapt it to the enrollment samples of a specific class. The adaptation is commonly achieved by using *maximum a posteriori* (MAP) estimation [28], where only the means of the UBM are updated. A convenient and compact representation of mean-only MAP adaptation and other session variability modeling techniques is the GMM supervector notation [34]:

$$\boldsymbol{g}_i = \boldsymbol{m} + \boldsymbol{d}_i \,, \tag{3}$$

where $\boldsymbol{g}_i$ is the mean supervector of the GMM $\boldsymbol{\mathcal{G}}_i$, $\boldsymbol{m} = \left[\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_c^T, \cdots, \boldsymbol{\mu}_C^T\right]^T$ is the mean supervector of the UBM $\boldsymbol{\mathcal{M}}$, and $\boldsymbol{d}_i$ is a class-specific offset for $\boldsymbol{\mathcal{G}}_i$.

A powerful approach that relies on a GMM UBM is *inter-session variability* (**ISV**) modeling [34]. It aims to estimate and suppress the effects of within-class variations in order to create more discriminant gender models. **ISV** assumes that session variability results in an additive offset to the mean supervector $\boldsymbol{g}_i$ of the gender model. This offset can be added directly to the normal mean-only MAP adaptation representation. Given the j-th sample $\boldsymbol{\mathcal{O}}_{i,j}$ of gender i the mean supervector $\boldsymbol{\mu}_{i,j}$ of the GMM that best represents this sample is:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{i,j} + \boldsymbol{D}\boldsymbol{z}_i \,, \tag{4}$$

where $\boldsymbol{U}$ is a subspace that constrains the possible session effects, $\boldsymbol{x}_{i,j}$ is its associated latent session variable ($\boldsymbol{x}_{i,j} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right)$), while $\boldsymbol{D}\boldsymbol{z}_i$ represents the gender-specific offset.

Similarly to **GMM**, **ISV** scoring relies on a LLR, using compensated GMMs as follows:

$$h_{\textbf{ISV}}\left(\boldsymbol{\mathcal{O}}_t\right) = \ln \frac{p\left(\boldsymbol{\mathcal{O}}_t|\boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{\text{male},t} + \boldsymbol{D}\boldsymbol{z}_{\text{male}}\right)}{p\left(\boldsymbol{\mathcal{O}}_t|\boldsymbol{m} + \boldsymbol{U}\boldsymbol{x}_{\text{female},t} + \boldsymbol{D}\boldsymbol{z}_{\text{female}}\right)} \tag{5}$$

3.4. Total variability modeling

In [9] it was shown that session variability modeling techniques can fail to separate between-classes and within-class variations into two different subspaces. To address this issue, an alternative technique called *total variability* (**TV**, also known as i-vectors) modeling was developed for speaker recognition [9], and later applied to face recognition [36]. **TV** modeling aims to extract low-dimensional factors $\boldsymbol{w}_{i,j}$, so-called *i-vectors*, from samples $\boldsymbol{\mathcal{O}}_{i,j}$. More formally, **TV** can be described in the GMM mean supervector space by:

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w}_{i,j} \,, \tag{6}$$

where $\boldsymbol{T}$ is the low-dimensional total variability subspace and $\boldsymbol{w}_{i,j}$ the low-dimensional i-vector, which is assumed to follow a normal distribution $\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right)$. $\boldsymbol{T}$ is learned by maximizing the likelihood over a large training set.

In contrast to **ISV**, **TV** does not explicitly perform session compensation and scoring. Hence, a set of preprocessing algorithms have been proposed to map i-vectors into a more adequate space [4, 13, 36]. For **i-vectors preprocessing**, *whitening* was proposed in [4] and shown to boost classification performance. Whitening consists of normalizing the **TV** space such that the covariance matrix of the i-vectors, of a training set, is turned into the identity matrix. Another efficient preprocessing technique is *length normalization* [13, 36], which aims at reducing the impact of a mismatch between training and test i-vectors. For **session compensation**, *within-class covariance normalization* (WCCN) is employed, which normalizes the within-class covariance matrix of a training set of i-vectors.

Once session compensation has been performed, any classification technique might be used. We investigate the simple and efficient *cosine similarity* measure [9, 36], as well as SVMs [33], leading to two systems **TV-Cosine** and **TV-SVM**, respectively.

4. Experimental evaluation

In this section, we evaluate the accuracy of unimodal and bimodal gender recognition systems on several databases. For both visual and audio modalities, and after feature extraction, four gender recognition systems are employed, relying on the modeling and classification techniques described in Section 3. We call them **GMM**, **ISV**, **TV-SVM** and **TV-Cosine**, respectively. GMMs are composed of 512 Gaussian components, and the rank of the subspaces are respectively set to 50 for **ISV** (matrix $\boldsymbol{U}$) and 400 for **TV** (matrix $\boldsymbol{T}$), respectively. Given the small size of the training set of FERET, the **TV** subspace has a rank of 200 on this database.

The development of the different systems has been performed relying on the open-source toolbox Bob [2][2]. Source code required to reproduce the experiments is available online[3].

Similarly to [19, 20, 7], the evaluation metrics used in our work are the *accuracy* (Acc), the *true positive rate* (TPR), the *true negative rate* (TNR) that are defined by:

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}}, \quad \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}}, \quad \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{N}}, \quad (7)$$

where TP is the number of samples correctly classified as positive (i.e. male), TN the number of samples correctly classified as negative (i.e. female), P the total number of positive samples and N the total number of negative samples. Furthermore, we used a variant of the *receiver operating characteristic* (ROC) curve, that plots the fraction of males classified correctly in terms of the fraction of females classified incorrectly [20].

4.1. Face-based gender recognition

The problem of face-based gender recognition has been tackled in [19, 1]. In their work, the experiments rely on a subset of the FERET database [24], for which an evaluation protocol is already established[4]. For the sake of comparison, we conducted a set of experiments on this small corpus (411 images), using the same annotations and the same protocol. Another drawback of using this database is the well controlled recording conditions of the images.

In contrast to FERET, images of the LFW database[5] [16] were acquired in an uncontrolled environment, leading to higher variability in term of pose, illumination and expression. In addition, the amount of samples is significantly larger ($13,233$ images). Experiments are conducted on this corpus using the BeFIT evaluation protocol (see Section 2.1).

We evaluated our proposed systems on both databases, using a very similar setup.

First, images are rotated, scaled and cropped to a fixed size, according to eye coordinate annotations and using a parametrization similar to the one in [19]. Next, visual features are extracted.

For the four proposed systems (**GMM**, **ISV**, **TV-SVM** and **TV-Cosine**), we rely on parts-based features that were initially proposed for the task of face recognition in [29]. The key idea is to decompose the face image into a set of overlapping blocks before extracting a feature vector from each of them. For this purpose, 12×12 blocks of pixel values are extracted from the preprocessed image using an exhaustive overlap. Pixel values of each block are normalized to zero mean and unit variance, prior to extracting the M ($M = 44$) lowest frequency *2D discrete cosine transform* (2D-DCT) coefficients [29] excluding the zero frequency coefficient. Finally, the 2D-DCT vectors are normalized to zero mean and unit variance.

We also evaluate other benchmarks, that apply SVM on raw pixels (**Raw-SVM**) or on LBP features (**LBP-SVM**) [19], as well as SVM on HOG [8] (**HOG-SVM**).

On FERET, we first evaluate all the systems at different image resolutions. Fig. 1 shows that the accuracy of the systems is stabilizing when image resolution is increasing. Therefore, we set the resolution of cropped images to the reasonable value of 80×80 in further experiments.

Additionally, Table 1 compares the accuracy of our systems to the results published in [19], using the same image resolution and cropping. At the largest resolution of 48×48, results suggest that the proposed **TV-SVM**, **ISV** and **GMM** systems outperfom the benchmarks. In particular, **ISV** reaches an accuracy rate of 90.7%, compared to 84.0% for the best system of [19] (**Raw-SVM**).

[2] http://www.idiap.ch/software/bob
[3] https://pypi.python.org/pypi/xbob.gender.bimodal
[4] http://www.sis.uta.fi/~em55910/datasets/
[5] http://vis-www.cs.umass.edu/lfw/

Table 1. ACCURACY ON FERET. *This table reports the accuracy rate (in %) of the systems on FERET.*

Resolution	**TV-SVM**	**ISV**	**GMM**	**TV-Cosine**	**HOG-SVM**	**LBP-SVM** [19]	**Neural Network** [19]	**Raw-SVM** [19]	**AdaBoost** [19]
24 × 24	76.6	**91.6**	82.2	77.6	79.4	76.9	84.2	82.6	81.5
48 × 48	**88.8**	**88.8**	85.1	**88.8**	83.2	82.1	82.9	84.0	83.9

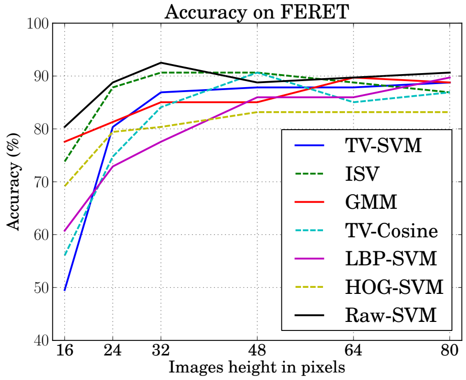


Figure 1. IMPACT OF IMAGE RESOLUTION ON FERET. *This figure shows the accuracy of all the systems on FERET by varying image resolution. The height and the width are set to identical values after cropping.*

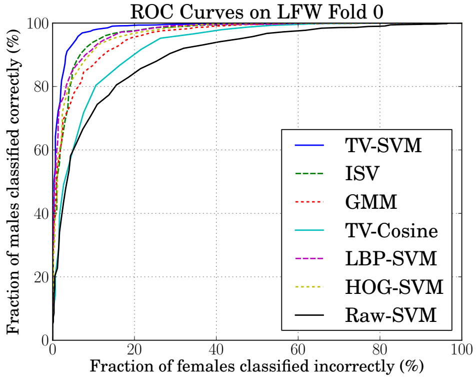


Figure 2. ACCURACY ON THE FIRST FOLD OF LFW. *This figure reports the accuracy of all the systems on the first fold of the LFW database.*

Experiments conducted on LFW show similar trends, the **ISV** system providing a good accuracy rate, as reported in Fig. 2. Nevertheless, the **TV-SVM** system significantly outperforms other systems and achieves state-of-the-art performances on this corpus (accuracy rate of 94.6% as shown in Table 2), compared to the best previously published results [7]. Looking at the errors made by the **TV-SVM** (examples are depicted in Fig. 3), results suggest that the high intra-class variability remains one of the main challenges. This variability is caused by recording conditions such as pose, illumination and expression on one side, and accessories, hair and make-up on the other side.

Table 2. ACCURACY ON LFW. *This table reports the accuracy, the true positive rate (TPR, for males) and the true negative rate (TNR, for females) on LFW after 5-fold cross-validation. The image resolution employed by each system is given in brackets.*

System	**Accuracy**	**TPR**	**TNR**
TV-SVM (80 × 80)	**94.6**	97.4	**85.0**
Gabor-PCA-SVM (120 × 105) [7]	94.0	**97.5**	82.2
LBP-PCA-SVM (120 × 105) [7]	93.8	97.0	83.0
Raw-PCA-SVM (120 × 105) [7]	89.2	95.4	68.1

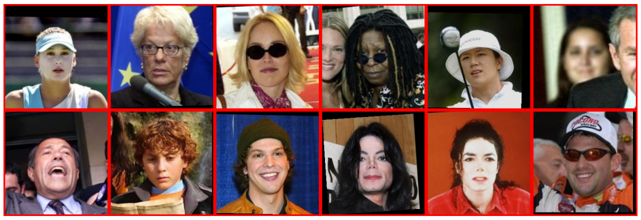


Figure 3. MISCLASSIFIED SAMPLES BY TV-SVM ON LFW, FOLD 0. *This figure shows misclassified samples (top row: females; bottom row: males) by the proposed **TV-SVM** gender recognition system. These are original images aligned with funneling from the LFW database, fold 0.*

Table 3. NIST-SRE PARTITIONNING. *This table reports the number of male and female speakers and the number of utterances on the training (TRAIN), development (DEV) and evaluation (EVAL) sets for NIST-SRE protocol.*

	TRAIN	**DEV**	**EVAL**
NIST-SRE series	2006	2010	2012
Number of Male speakers	481	235	763
Number of Female speakers	659	261	1,155
Number of utterances	14,735	22,848	73,106

4.2. Audio-based gender recognition

We evaluate our gender recognition systems on audio data from the MIXER corpus [6], which is provided by NIST since 2004 for the task of speaker recognition. The training (TRAIN) set uses data from NIST-SRE (*Speaker Recognition Evaluation*) 2006, whereas the development (DEV) and the evaluation (EVAL) sets use data from SRE 2010 and 2012, respectively. The recordings were collected in uncontrolled conditions (e.g. microphone, telephone, synthetic noise, real noise, duration variability, etc.). Statistics on the number of male and female speakers, and the number of utterances are reported in Table 3. To the best of our knowledge, this is the first large scale gender recognition experiment conducted on audio data[6].

Acoustic features are extracted at equally-spaced time instants using a sliding window approach. First, *voice activ-*

[6] The protocol is made publicly available within the package.

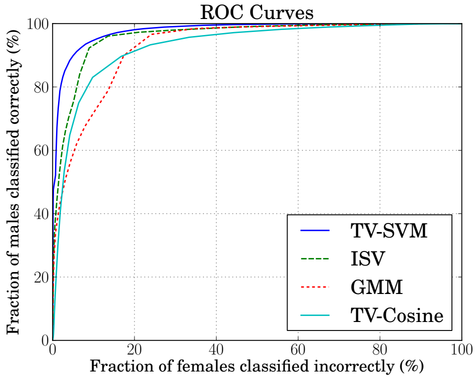


Figure 4. ACCURACY ON NIST-SRE. *This figures shows the ROC curves of the different algorithms on NIST-SRE dataset.*

ity detection (VAD) is performed jointly using the normalized log energy and the 4 Hz modulation energy [30]. Second, 19 MFCC and log energy features together with their first- and second-order derivatives are computed over 20 ms Hamming windowed frames every 10 ms. This results in acoustic feature vectors of dimensionality $M = 60$. Finally, *cepstral mean and variance normalization* (CMVN) is applied on the remaining feature vectors.

Results in Fig. 4 clearly show that **TV-SVM** and **ISV** systems outperform state-of-the-art **GMM** system by up to 11% of relative gain. It also shows that **TV-SVM** system is slightly superior than **ISV**. The accuracy rates of **TV-SVM**, **ISV**, **GMM** and **TV-Cosine** are 92.5%, 91.1%, 83.5% and 87.1%, respectively.

4.3. Bimodal gender recognition

We evaluated bimodal gender recognition on the MOBIO[7] database, which consists of 61 hours of audio-visual data of 150 people captured within twelve sessions. This corpus is challenging since the data is acquired on mobile devices with real noise. It has been used to evaluate several speaker, face and bimodal recognition systems [21]. The extracted images contain faces with uncontrolled illumination, facial expression, and occlusion, while the extracted speech segments are relatively short, partially even less than two seconds. A new protocol[8] for gender recognition is established, with separate training, development (DEV) and evaluation (EVAL) sets, each containing 50 identities.

For each modality, we employ the same features and parametrization as the ones introduced for the unimodal systems (cf. Sections 4.1 and 4.2). The combination of the two modalities is performed using *score fusion* (also called *high-level* fusion), which combines scores from several systems. For this purpose, we use the *linear logistic regression* approach, which has been successfully employed for combining heterogeneous speaker classifiers [25].

Table 4. ACCURACY ON MOBIO. *This table reports the accuracy, the true positive rate* (TPR, *for males) and the true negative rate* (TNR, *for females) of the systems on the evaluation set of MOBIO.*

		TV-SVM	ISV	GMM	TV-Cosine
Face	Accuracy	**94.5**	93.9	86.7	91.5
	TPR	**97.7**	97.0	91.9	93.0
	TNR	**88.4**	87.9	76.5	88.6
Audio	Accuracy	97.5	**97.8**	97.6	94.1
	TPR	96.9	**97.3**	97.0	92.1
	TNR	98.6	**98.7**	98.5	98.0

Let an audio-visual test sample $\mathcal{O}_t = (\mathcal{O}_t^{\text{a}}, \mathcal{O}_t^{\text{v}})$ be processed by both audio and visual systems. Each system produces an output score, $h_s^{\text{audio}}(\mathcal{O}_t^{\text{a}})$ and $h_s^{\text{visual}}(\mathcal{O}_t^{\text{v}})$ for audio and visual cues, respectively. The final fused score is expressed by the logistic function:

$$h_s^{\text{fusion}}(\mathcal{O}_t) = g\left(\beta_0 + \beta_1 h_s^{\text{audio}}(\mathcal{O}_t^{\text{a}}) + \beta_2 h_s^{\text{visual}}(\mathcal{O}_t^{\text{v}})\right), \tag{8}$$

where

$$g(x) = \frac{1}{1 + \exp(-x)} \tag{9}$$

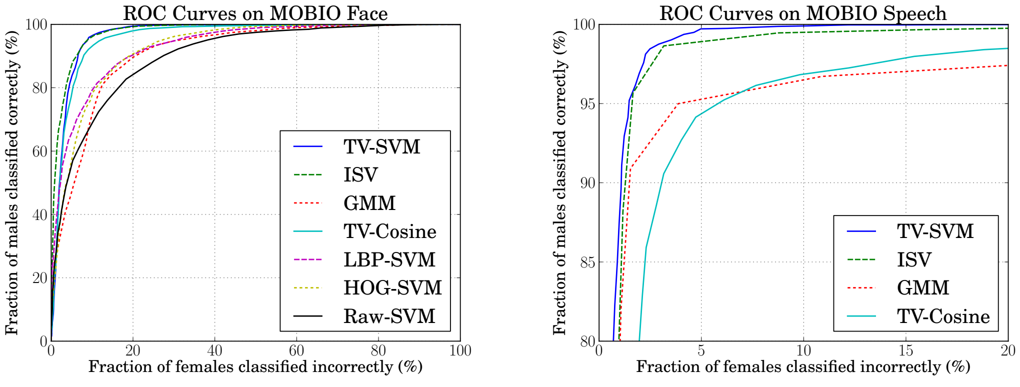
and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$ are the regression coefficients, that are computed by estimating the maximum likelihood of the logistic regression model on the scores of the development set. In this work, the optimization is done using the *conjugate-gradient* algorithm [22].

Performances of unimodal gender recognition systems on the MOBIO database are shown in Fig. 5. For the visual modality, **TV-SVM** , **ISV** and **TV-Cosine** outperform the **Raw-SVM** and **LBP-SVM** benchmarks. For the audio modality, **TV-SVM**, **ISV** and **GMM** achieve very high performances.

Interestingly, when comparing the two modalities (Table 4), a significantly higher accuracy rate (96.8%) is achieved with the audio modality, compared to the visual one (92.2%). Possible explanation is that the visual modality was subject to more variability (pose, illumination, expression) than the audio modality during the data collection. In addition, for audio-based gender recognition, the classification rates are comparable for the two classes male and female. In contrast, for face-based gender recognition, there is a large gap between male (TPR) and female (TNR) classification rates.

We investigate the fusion of several unimodal systems. Results depicted in Fig. 6 show that the fusion of the two modalities allows to drastically reduce the error rate, reaching an accuracy rate of about 98% for **TV-SVM** and **ISV** systems. Real classification examples of the **TV-SVM** systems (both unimodal and bimodal ones) are illustrated in Fig. 7. Sample 1 (first column) is classified correctly by all the unimodal and bimodal systems. In contrast, samples 2 to 5 are only classified correctly by one of the

[7] https://www.idiap.ch/dataset/mobio
[8] The protocol is made publicly available within the package.



(a) Visual

(b) Audio

Figure 5. ACCURACY ON MOBIO. *These figures show ROC curves for both visual and audio modalities on MOBIO. For the audio modality, a zoom is performed in the region of interest, as a high accuracy is achieved.*

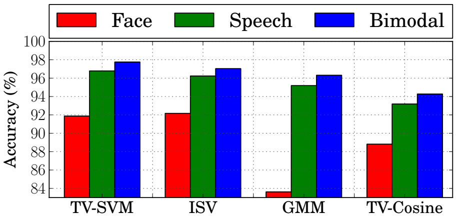


Figure 6. PERFORMANCES OF THE PROPOSED SYSTEMS ON MOBIO. *This figure reports the accuracy rate of several unimodal and bimodal systems on the evaluation set of MOBIO.*

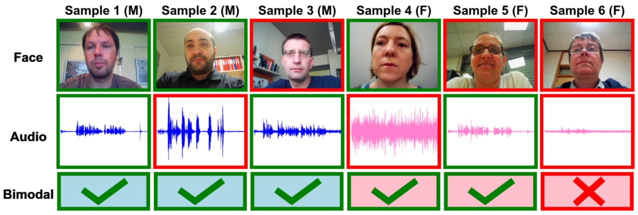


Figure 7. CLASSIFICATION EXAMPLES ON MOBIO. *This figure shows few classification examples of the bimodal* ***TV-SVM*** *system. Each column corresponds to a test sample, whereas each row corresponding to a modality (visual, audio and bimodal, respectively). A green bounding-box around a cell indicates that the sample has been classified correctly, whereas a red one indicates a misclassification.*

two unimodal systems, but the bimodal one is still able to take the right decision. This suggests that, when a modality is affected by challenging conditions (e.g. noise or accessories), the other one is available for the rescue. Sample 6 is very challenging since both modalities are subject to high deformations.

To quantify the significance of the improvement of the proposed systems over the baselines, we conducted an evaluation based on Eq. 15 and Fig. 2 of [3]. This evaluation shows that the improvement is statistically significant considering a 99% confidence interval for all databases except for the small FERET database.

5. Conclusions

This paper investigates the problem of audio, visual and bimodal gender recognition with two different variability modeling techniques: **ISV** and **TV**. For visual gender recognition, state-of-the-art performances are achieved on both FERET and LFW databases. For the audio modality, the large scale evaluation conducted on NIST-SRE shows that the **TV-SVM** system is achieving an accuracy rate of 92.5%. In addition, experiments were carried out on the bimodal MOBIO database. Results show that our proposed **TV-SVM** and **ISV** systems outperform state-of-the-algorithms on both modalities. Furthermore, additional improvements are obtained by combining them using linear logistic regression. The final accuracy rate of the bimodal system is about 98%.

6. Acknowledgment

The research leading to this work has received funding from the Swiss National Science Foundation (SNSF) under the LOBI project and from the European Community's Seventh Framework Programme (FP7) under grant agreement 284989 (BEAT) .

References

[1] L. Alexandre. Gender recognition: A multiscale decision fusion approach. *Pattern Recogn. Lett.*, 31(11):1422–1427, 2010. 1, 2, 4

[2] A. Anjos et al. Bob: a free signal processing and machine learning toolbox for researchers. In *ACM International Conference on Multimedia (ACMMM)*, 2012. 4

[3] S. Bengio and J. Mariéthoz. A statistical significance test for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004. 7

[4] L. Burget et al. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4832–4835, 2011. 4

[5] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann. A database of age and gender annotated telephone speech. In *International Conference on Language Resources and Evaluation (LREC'10)*, 2010. 1, 2

[6] C. Cieri, J.-P. Campbell, H. Nakasone, D. Miller, and K. Walker. The mixer corpus of multilingual, multichannel speaker recognition data. In *LREC*. European Language Resources Association, 2004. 5

[7] P. Dago-Casas, D. Gonzalez-Jimenez, L. Yu, and J. Alba-Castro. Single- and Cross- Database Benchmarks for Gender Classification under Unconstrained Settings. In *Intl. Conf. on Computer Vision Workshops*, 2011. 2, 4, 5

[8] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *Intl. Conf. on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, 2005. 4

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. on Audio, Speech, and Language Processing*, 19:788–798, 2011. 1, 2, 3, 4

[10] A. DeMarco and S. Cox. An accurate and robust gender identification algorithm. In *INTERSPEECH*, pages 2429–2432, 2011. 1, 2

[11] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[12] A. Gallagher and T. Chen. Understanding images of groups of people. In *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 256–263, 2009. 2

[13] D. Garcia-Romero and C. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252, 2011. 4

[14] H. Harb and L. Chen. Gender identification using a general audio classifier. In *International Conference on Multimedia and Expo (ICME)*, volume 2, pages 733–736, 2003. 1, 2

[15] Y. Hu, D. Wu, and A. Nucci. Pitch-based gender identification with two-stage classification. *Security and Communication Networks*, 5(2):211–225, 2012. 1, 2

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 4

[17] M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech Language*, 27(1):151–167, 2013. 1, 2, 3

[18] M. Liu, X. Xu, and T. S. Huang. Audio-visual gender recognition. In *International Symposium on Multispectral Image Processing and Pattern Recognition*, 2007. 1, 2

[19] E. Mäkinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008. 1, 2, 4, 5

[20] E. Mäkinen and R. Raisamo. An experimental comparison of gender classification methods. *Pattern Recogn. Lett.*, 29(10):1544–1556, July 2008. 1, 4

[21] C. McCool et al. Bi-modal person recognition on a mobile phone: using mobile phone data. In *IEEE Intl. Conf. on Multimedia and Expo (ICME), Workshop on Hot Topics in Mobile Multimedia*, 2012. 6

[22] T. P. Minka. Algorithms for maximum-likelihood logistic regression. Technical Report 758, CMU Statistics Department, 2001. 6

[23] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(5):707–711, 2002. 1, 2

[24] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transaction Pattern Analysis and Machine Intelligence (TPAMI)*, 22(10):1090–1104, 2000. 2, 4

[25] S. Pigeon, P. Druyts, and P. Verlinde. Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, 10(1–3):237–248, 2000. 1, 6

[26] M. Pronobis and M. Magimai-Doss. Integrating audio and vision for robust automatic gender recognition. Idiap-RR Idiap-RR-73-2008, Idiap, 11 2008. 1

[27] D. Reynolds and R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, 1995. 3

[28] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000. 3

[29] C. Sanderson and K. K. Paliwal. Fast features for face authentication under illumination direction changes. *Pattern Recognition Letters*, 24(14):2409–2419, 2003. 3, 4

[30] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1331–1334, 1997. 6

[31] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan. The interspeech 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, Sept. 2010. 1, 2

[32] N. Sun et al. Gender classification based on boosting local binary pattern. In *Advances in Neural Networks - ISNN*, volume 3972 of *Lecture Notes in Computer Science*, pages 194–201. Springer Berlin Heidelberg, 2006. 1

[33] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., 1995. 4

[34] R. Vogt and S. Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 22(1):17–38, 2008. 1, 2, 3

[35] L. Walawalkar, M. Yeasin, A. Narasimhamurthy, and R. Sharma. Support vector learning for gender classification using audio and visual cues: A comparison. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(03):417–439, 2003. 2

[36] R. Wallace and M. McLaren. Total variability modelling for face verification. *IET Biometrics*, 1(4):188–199, 2012. 3, 4