INTEGRATING AUDIO AND VISION FOR ROBUST AUTOMATIC GENDER RECOGNITION

Marianna Pronobis [†‡] *and Mathew Magimai.-Doss* [†]

[†] Idiap Research Institute, Martigny, Switzerland
[‡] School of Electrical Engineering, Royal Institute of Technology (KTH), Sweden
{mprono,mathew}@idiap.ch

ABSTRACT

We propose a multi-modal Automatic Gender Recognition (AGR) system based on audio-visual cues and present its thorough evaluation in realistic scenarios. First, we analyze robustness of different audio and visual features under varying conditions and create two uni-modal AGR systems. Then, we build an integrated audio-visual system by fusing information from each modality at the classifier level. Our extensive studies on the BANCA corpus comprising datasets of varying complexity show that: (a) the audio-based system is more robust than the vision-based system; (b) integration of audio-visual cues yields a resilient system and improves performance in noisy conditions.

***Index Terms*— automatic gender recognition, audio-visual cue integration, feature selection, robustness under realistic scenarios**

1. INTRODUCTION

The ability to perform automatic recognition of human gender is crucial for a number of systems that process or exploit human-source information. Typical examples are information retrieval, human-computer or human-robot interaction. The outcome of an *Automatic Gender Recognition* (AGR) system can be used for generating meta-data information useful for annotating audio and video files. Moreover, gender is an important cue that can be exploited for improving intelligibility of man-machine interaction, or simply, for reducing the search space in speaker recognition or surveillance systems.

The problem of AGR was addressed in the past by several authors (see Section 2). In all these works, only one modality (audio or vision) was employed. The investigations were performed mainly under clean conditions and the robustness of AGR systems in real-world scenarios was seldom considered. In most typical applications, both audio and vision are available. Ideally, an AGR system should be able to exploit both modalities to improve robustness. Since each modality has different characteristics, audio-visual cues can provide a more comprehensive description of a subject than a single modality. Finally, integration of the cues may yield a AGR system that is resilient to the degradation of both, or even to temporal unavailability of one of the input signals.

In this paper, we investigate an audio-visual AGR (AV-AGR) system trained under clean conditions and, then tested under varying conditions. We also analyze different feature representations for each of the cue, and assess their robustness to varying conditions. The AV-AGR studied in this work is based on the high-level integration framework. In other words, first uni-modal AGR systems are trained, and then the cue integration is performed by fusing the evidences from the two systems. Through extensive studies under varying conditions (controlled, degraded, adverse) on the BANCA corpus, we show that: (a) the audio-based system is more robust than the vision-based system, and (b) integration of audio-visual cues yields a resilient system that preserves the performance of the best modality in clean conditions and, helps in improving the overall performance in noisy conditions. To the best knowledge of the authors, this is the first study on audio-visual integration for AGR.

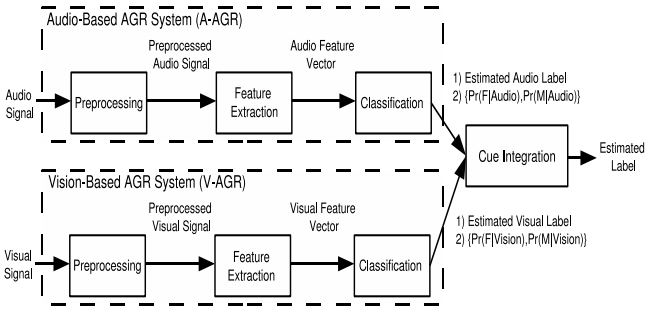

Fig. 1. Overview of the architecture of the AV-AGR system. The two modalities are processed separately, and then integrated at the classifier level.

2. RELATED WORK

The previously proposed solutions to the AGR problem were based on single modality, either on audio or vision. The first works on audio-based AGR aimed at identifying the most appropriate features of speech signal for the task. Comparison of voice source- (pitch frequency) and vocal tract-related features (first four formants with their respective frequency, amplitude and bandwidth) for ten vowels extracted from the clean-condition speech data of 52 speakers was presented in [1]. Further analysis of different parametric representations of speech signal (linear prediction, autocorrelation, reflection and cepstrum) was performed on the same database for vowels, voiced and unvoiced fricatives [2]. The evaluation of mel-cepstral features for different groups of phonemes like vowels, nasal, liquids etc. was conducted in [3]. Moreover, the last two studies analyzed an influence of different filter orders (from 8 to 20) and types of coefficients (static *vs.* delta) on the performance of an AGR system. More recently, the comparision of Support Vector Machines (SVMs) with nearest neighbor classifiers for the first 12 cepstral coefficients on high quality recordings of 150 speakers from the ISOLET corpus was presented with 100% AGR rate for SVMs [4].

Early research in vision-based gender recognition was focussed upon the use of artificial neural networks for feature extraction and classification on clean condition data [5, 6]. Latest research looked

This work was supported by the EU 6^{th} FWPISTIP AMIDA project (FP6-033812). The authors thank Sébastien Marcel and Jie Luo for help, discussions and comments.

into more complex lighting and pose variations, and for larger sets of subjects, such as in the FERET database [7]. The experimental studies suggested that for the AGR task based on visual cues, the SVMs with the RBF kernel are superior to the linear, quadratic, fisher linear discriminant, k-nearest neighbor classifiers as well as to more complex techniques such as large ensemble RBF networks [7, 4]. In [4], comparison of row data reprasentation with features obtained through principal component analysis (PCA), referred to as eigenfaces [8], was made on database consisting of 1640 frontal, unoccluded face images.

3. THE AUTOMATIC GENDER RECOGNITION SYSTEM

This section presents an architecture of our audio-visual AGR (AV-AGR) system.

3.1. System Overview

In designing the AV-AGR a two-fold approach was adapted. First, we studied the two cues separately by building audio-based and vision-based AGR systems (A-AGR and V-AGR). Second, these systems were integrated to provide the final decision based on both modalities. Overview of the system architecture is presented in Figure 1. In the proposed solution, the A-AGR system utilizes speech signal. Similarly, the V-AGR system exploits exclusively face images and no stature information is used. The A-AGR and V-AGR systems have similar architectures which consist of three parts performing the following functions: (a) data preprocessing, (b) feature extraction, and (c) classification. The role of the signal preprocessing block is extraction of useful fragments of the signal. The previous studies on audio suggested that voiced phonemes are more discriminative for gender than unvoiced phonemes [2, 3]. We use a voiced/unvoiced detection to obtain the most informative parts of the signal. In case of the V-AGR system, data preprocessing includes face detection, localization, and finally segmentation. The function of the second block is extraction of features from the preprocessed signal that allow for most accurate classification of the subject. The description of this block for the A-AGR and V-AGR system is given in Sections 3.2 and 3.3, respectively. Finally, classification of an instance to one of the two possible classes (female or male) is performed. The classification module employs the same algorithm in case of both A-AGR and V-AGR. The SVMs with the radial basis function (RBF) as a kernel, successfully applied in the previous studies [7, 4], were used as a classifier. In order to estimate the a posteriori probability of a particular class for each modality the Platt's method was used [9].

3.2. Audio Features

It is a commonly known phenomenon that females are characterized by a higher pitch frequency (F0) value than males. The discriminative role of F0 was experimentally confirmed in [1]. Thus, we first evaluate the effectiveness of F0 for the AGR problem under varying conditions. Furthermore, it was observed that female and male voices differ in the entire range of their spectral characteristics due to dissimilarities in the anatomical structure of the vocal tracts [10]. In consequence, different values of the formant characteristics (frequency, bandwidth, amplitude) are typical for females and males, and thus can be used to distinguish the gender. Another type of features that capture, *inter alia*, information encoded in the shape of the spectrum of speech signal is the parametric representation. We focused on the standard features used in the state-of-the-art Automatic Speech and Speaker Recognition systems, like Linear Prediction Cepstral Coefficients (LPCCs), Mel-frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) coefficients [11]. In order to compare different types of features, preliminary experiments[1] were performed for: (a) first four formants with their respective frequency and bandwidth, and (b) LPCCs, MFCCs and PLPs. The parametric features were constantly better than the formant related features (which is consistent with the previous results [2]), especially under noisy conditions. It may be due to the fact that format estimation is not robust. Moreover, all three parametric representations provided the same performance under clean conditions, however PLPs were slightly better than MFCCs and LPCCs under noisy conditions. As a result, in the rest of the paper, we report our studies with the F0 and PLP features.

3.3. Visual Features

The standard automatic face recognition systems use a low-dimensional representation of faces obtained by means of component analysis techniques. The suitability of the eigenface method for the V-AGR problem and clean-condition data was experimentally confirmed in [4]. However, in case of face recognition systems, this technique works efficiently only when constant face pose and lightning are preserved and tends to fail under varying conditions. To overcome this problem a technique that additionally uses linear discriminant analysis (LDA), referred to as the fisherface method, was introduced [12]. Both types of features, eigenfaces and fisherfaces were evaluated in this work.

3.4. Audio-Visual AGR System

The AV-AGR system is created by fusing evidences from the two modalities at the high level, after the single-cue classification is performed. The architecture of the AV-AGR system is presented in Figure 1. The a posteriori probabilities provided by the single-cue classifiers are combined using the sum or product rule to provide the final decision based on both modalities. Theoretical studies show that these two rules are most suitable for the two-class problem [13]. Additionally, we considered equal or unequal weighting of modalities during the experiments. The latter were performed in order to answer the question of different importance of the modalities in the correct classification.

4. EXPERIMENTAL SETUP

4.1. Database

The AGR system was evaluated on the BANCA database (English corpus) comprising three datasets of varying complexity [14]. Data acquisition was performed using 2 cameras and 2 microphones (poor-quality and good-quality) under three different types of conditions: (a) *controlled*: good-quality microphone and camera, uniform background and stable lighting; (b) *degraded*: poor-quality microphone and camera, non-uniform background; (c) *adverse*: good-quality microphone and camera, background noise, arbitrary conditions [15]. In every conditions, 4 sessions were scheduled during which 2 recordings from 52 subjects (26 females, 26 males) were collected. In our experiments, we used 1.3s of voiced speech signal and 5 images from each video file. The data were divided into three sets used for training, development and testing. In order to evaluate the system on the same number of known and unknown subjects, only a half of the subjects (26) were used for training. Then, all subjects (52) were divided into two groups consisting of 16 and 36 subjects which were used for development and testing. It is important to highlight the fact that data from different sessions were used

[1] Preliminary experiments were performed on the development set preserving the experimental setup presented in Section 4.

	Item	Train	Dev	Test
Common Setup	#Subj. Σ(F,M)	26(13,13)	16(8,8)	36(18,18)
	#Video Files	104	64	144
	#Images	520	320	720
	Audio Amount	136s	84s	188s
Protocol	Item	Train	Dev	Test
Controlled	Conditions	Con	Con	Con
	Session	s01,s02	s03,s04	s03,s04
Degraded	Conditions	Con	Deg	Deg
	Session	s01,s02	s07,s08	s07,s08
Adverse	Conditions	Con	Adv	Adv
	Session	s01,s02	s11,s12	s11,s12

Table 1. Experimental setup for different protocols. Abbreviations and symbols: 'Con'=Controlled, 'Deg'=Degraded, 'Adv'=Adverse, 'Subj.'=Subjects, 'Σ'=Total, 'F'=Females, 'M'=Males and 'sXX'=session number referring to the BANCA database [17].

for training, development and testing. The following three protocols were defined: *Controlled*, *Degraded* and *Adverse*. Both the common and specific parameters of these protocols are presented in the upper and lower part of Table 1. The system was always trained on controlled-condition data and tested under controlled, degraded and adverse conditions.

4.2. Analysis of Audio Data

The audio signal extracted form the video files, sampled at $16kHz$, was analyzed in frames of 25ms using a frame shift interval of 10ms. The ESPS pitch tracking method was used to estimate the F0 values [16]. This method, not only gives estimates of F0, but also provides binary information about frame voicing which was used to extract the 1.3s of voiced data from each file. We analyzed performance of the PLP features with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients included to the feature vector.

4.3. Analysis of Visual Data

In order to extract a face region from an image, first an automatic frontal face detector performing geometric normalization of the image in order to align eyes was applied. Then, each image was cropped to a size of 64x80. The PCA was performed on face images to compute the eigenfaces. The number of features was chosen to capture 99% of the data variations which, in this case, corresponds to the first 116 eigenvectors. Then, through LDA, the fisherface features were obtained. Due to the fact that the informative part of the LDA features is encoded in the first $n-1$ vectors, where n is the number of classes, each image was represented using only one feature.

4.4. System

We used the SVM classifier with the RBF kernel, in case of multi-dimensional feature vectors, and with the linear kernel, in case of one-dimensional feature vectors. The parameters of the SVMs (error penalty C) and the RBF kernel (variance γ) were estimated on the development set. While integration of the cues was performed using unequal weighting of modalities, weights were determined to maximize performance of the AV-AGR system for the development set. The classification accuracy was used as the measure of performance. We report the results for: (a) a single instance, like *frame accuracy* and *image accuracy*, and (b) a file, for 1.3s of voiced speech data or for 5 images (*file accuracy*).

5. RESULTS AND DISCUSSION

This section presents an evaluation of the AGR systems under varying conditions.

5.1. Audio-Based AGR System

In case of the A-AGR system, we investigated two types of features: voice source (F0) and vocal tract related (PLPs). Results obtained under three types of conditions: controlled, degraded and adverse are presented in Figure 2. As expected, the performance of the AGR system decreases with increasing severity of conditions. The F0 is the best feature under controlled conditions attaining perfect recognition. However, its performance is highly affected by signal degradation and dropped to 93.0% (file accuracy) under adverse conditions. It is more due to the fact that F0 estimation is not robust to noisy conditions. We analyzed performance of the PLPs with respect to the number (9, 13, 19) and type (static vs. static+delta) of cepstral coefficients included to the feature vector. We observed that performance of the AGR system for the PLPs increased with the number of coefficients. In the preliminary experiments performed on the development set, the addition of dynamic coefficients slightly aids in performance under degraded conditions (1.6%). This is consistent with results obtained in [3]. In spite of the small improvement, we decided to use both static and delta coefficients in further experiments. For the test set, the following gains in accuracy were obtained 1.4%, 4.2% and 3.5% under controlled, degraded and adverse conditions, respectively. The PLPs yielded the better system than F0 under degraded and adverse conditions. This indicates that PLP features are more robust to signal degradation. The complementarity of F0 and PLPs was evaluated by combining these two features using low and high level integration (with sum rule). The latter method was superior to the former one under all conditions. The AGR system with the high level integration preserved performace of the best of the single feature systems under controlled and adverse conditions and, the accuracy of classification was improved by 0.7% under degraded conditions.

5.2. Vision-Based AGR System

In case of the V-AGR system, we compared two types of features: eigenfaces and fisherfaces. Results obtained under various conditions are presented in Figure 3. For both features performance of the V-AGR system decreases with degradation of the signal. As expected, the fisherfaces are superior to eigenfaces under controlled conditions. The difference in performance of 2.1% (file accuracy) was observed. However, eigenfaces are better than fisherfaces by 4.0% and 0.7% under degraded and adverse conditions, respectively. It may be a consequence of the mismatch between within-class variances that are significantly higher in the test than training set.

5.3. Audio-Visual AGR System

For the AV-AGR system, we evaluated different combinations of audio (F0, PLPs) and visual (eigenfaces, fisherfaces) features. Results obtained for the sum and product rule are presented in Figure 4. For all combinations of audio and visual features, the approach based on the summation was at least as good as the one based on the product rule. An advantage of the sum over product rule is especially visible when combining visual features with F0. Results for the combination approaches using equal or unequal weighting of modalities are provided in Figure 5. The unequal weighting of modalities outperforms the equal weighting of modalities for all combinations of features. While using unequal weighting, the audio cues obtained

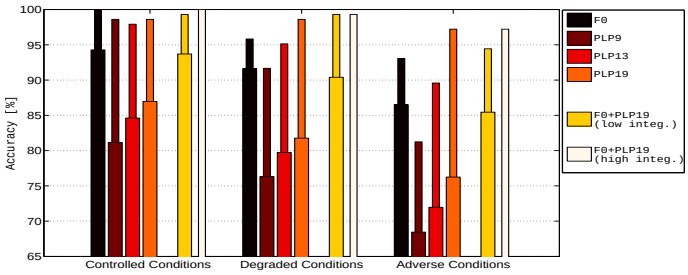

Fig. 2. Performance of the A-AGR system (wide bar - frame accuracy; narrow bar - file accuracy; for PLPs both static and dynamic features were used: 9, 13 and 19 refer to number of static features).

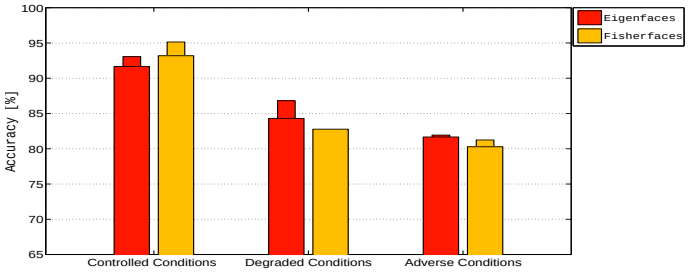

Fig. 3. Performance of the V-AGR system (wide bar - image accuracy; narrow bar - file accuracy).

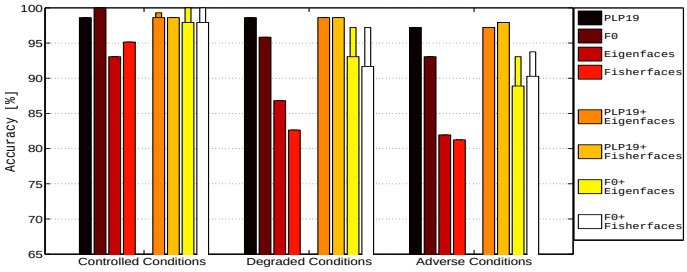

Fig. 4. Performance of AV-AGR systems using classifier combination based on sum or product rule is compared to results obtained for the single-cue systems (wide bar - product rule, narrow bar - sum rule; file accuracy, equal weighting of modalities).

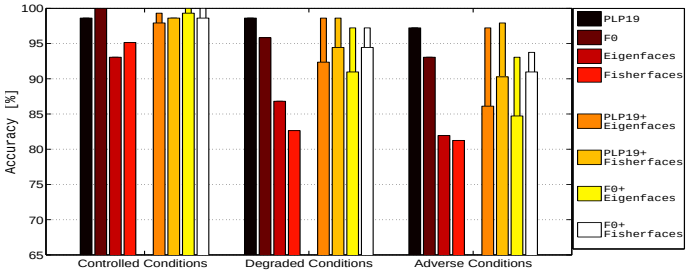

Fig. 5. Performance of AV-AGR systems using classifier combination and equal or unequal weighting of modalities is compared to results for the single-cue systems (wide bar - equal weighting, narrow bar - unequal weighting; file accuracy, sum rule).

higher weights than visual cues and their importance in correct classification increased with degradation of the data. Finally, the integration of audio-visual cues yielded a resilient system that preserved the performance of the best single AGR system in clean conditions and improved performance under degraded and adverse conditions by 0.7% and 16.7% for the combination of PLPs with fisherfaces while comparing to the A-AGR and V-AGR system, respectively. This is due to the fact that the V-AGR system was inferior to the A-AGR system. Future work will address the problem of improving the robustness of the V-AGR system.

6. CONCLUSION

In this paper, we proposed a multi-modal AGR system based on audio and visual cues and presented its thorough evaluation in realistic scenarios. The studies were conducted on the BANCA corpus comprising datasets of varying complexity (controlled, degraded and adverse). Our studies in the framework of single cue systems showed that the audio-based system is more robust than the vision-based system, and that the PLP and eigenfaces are less affected by changing conditions than pitch frequency and fisherfaces. The integration of audio-visual cues yielded a resilient system that preserved the performance of the best modality in clean conditions and, helped in improving performance in adverse conditions.

7. REFERENCES

[1] D. G. Childers and K. Wu, "Gender recognition from speech. part II: Fine analysis," *JASA*, vol. 90, pp. 1841–1856, 1991.

[2] K. Wu and D. G. Childers, "Gender recognition from speech. part I: Coarse analysis," *JASA*, vol. 90, pp. 1828–1840, 1991.

[3] J. W. Fussell, "Automatic sex identification from short segments of speech," in *Proc. ICASSP*, 1991, pp. 409–412.

[4] L. Walawalkar et al., "Support vector learning for gender classification using audio and visual cues: A comparison.," in *SVM*, 2002, vol. 2388 of *LNCS*, pp. 144–159.

[5] G. W. Cottrell and J. Metcalfe, "Empath: face, emotion, and gender recognition using holons," in *NIPS*, 1990, pp. 564–571.

[6] B. Golomb et al., "Sexnet: A neural network identifies sex from human faces," in *NIPS*, 1990, pp. 572–577.

[7] B. Moghaddam and M-H. Yang, "Gender classification with support vector machines," in *Proc. FG*, 2000, pp. 306–311.

[8] L. Sirovich et al., "Low-dimensional procedure for the characterization of human faces," *JOSA*, vol. 2, pp. 519–524, 1987.

[9] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[10] A. Bladon, "Acoustic phonetics, auditory phonetics, speaker sex and speech recognition: a thread," in *Computer Speech Proc.*, 1985, pp. 29–38.

[11] B. Gold and N. Morgan, *Speech and Audio Signal Processing*, John Wiley & Sons, 1999.

[12] P. Belhumeur et al., "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. PAMI*, vol. 19, no. 7, pp. 711–720, 1997.

[13] D.M. Tax et al., "Combining classifiers by averaging or multiplying?," *Pattern Recog.*, vol. 33, no. 9, pp. 1475–1485, 2000.

[14] E. Bailly-Baillire et al., "The BANCA database and evaluation protocol," *LNCS*, vol. 2688, pp. 1057–1072, 2003.

[15] S. Marcel et al., "On the recent use of local binary patterns for face authentication," *EURASIP JIVP*, 2007, Accepted.

[16] B. Secrest and G. Doddington, "An integrated pitch tracking algorithm for speech systems," in *Proc. ICASSP*, 1983, pp. 1352–1355.

[17] S. Bengio et al., "Experimental Protocol on the BANCA Database," Tech. Rep., Idiap, 2002.