INTERSPEECH 2009 BRIGHTON

Multimodal Speaker Verification Using Ancillary Known Speaker Characteristics Such as Gender or Age

Girija Chetty and Michael Wagner

National Centre for Biometric Studies, University of Canberra, Australia

girija.chetty@canberra.edu.au, michael.wagner@canberra.edu.au

Abstract

Multimodal speaker verification based on easy-to-obtain biometric traits such as face and voice is rapidly gaining acceptance as the preferred technology for many applications. In many such practical applications, other characteristics of the speaker such as gender or age are known and may be exploited for enhanced verification accuracy. In this paper we present a parallel approach determining gender as an ancillary speaker characteristic, which is incorporated in the decision of a face-voice speaker verification system. Preliminary experiments with the DaFEx multimodal audio-video database show that fusing the results of gender recognition and identity verification improves the performance of multimodal speaker verification.

Index Terms: multimodal, face-voice, speaker verification, speaker characterisation

1. Introduction

In recent years, automatic identity verification systems have been deployed increasingly in forensic, government and commercial applications [1, 2]. Predominantly these are still unimodal systems, based on the voice biometric, although there has been significant research on benefits of multimodal approaches for improving the performance and robustness of speaker verification systems [3, 4], and a variety of multimodal speaker verification approaches have been developed using the face and voice biometrics. Multimodal systems can overcome certain problems such as a noisy signal in one modality, e.g. audio in surroundings affected by traffic-noise or video in unfavourable lighting conditions, which may be compensated by a good-quality signal in another modality. Also, an authentication system that uses the face and voice biometrics, such as speaking-face video sequences, is better able to ascertain the liveness of the biometrics by tracking the synchrony of the facial movements and the audio [5, 6].

A second area of research endeavour, which has emerged relatively recently is that of speaker characterisation, which seeks to identify speaker characteristics such as age, gender, ethnicity, eye colour, height or weight [7, 8, 9]. However, up to now, the two paradigms of speaker verification and speaker characterisation have almost always been considered in separation.

In this paper, we suggest that in most practical speaker verification systems, ancillary information about such speaker characteristics is collected during the enrolment phase and is therefore known to the system. We propose to utilise this information in a paradigm that verifies both the claimed speaker identity and the known speaker characteristics that are associated with that claimed identity. Specifically in this paper, we test the hypothesis that the combination of speaker verification and gender verification will improve the performance of a speaker verification system over that of a system that does not utilise the ancillary information about the speaker's gender. This argument could then be extended to include other ancillary speaker information such as age or dialect, in order to further improve the accuracy of practical speaker verification systems.

The rest of the paper is organised as follows. Section 2 presents the technique we have used for automatic extraction of gender information using face and voice features. In Section 3 we propose a parallel strategy with audio-visual gender classification and audio-visual speaker verification. The experimental results are presented in Section 4, and the conclusions and further research plans follow in Section 5.

2. Automatic gender extraction

In order to utilise ancillary speaker characteristics such as gender, age and ethnicity, there must be a mechanism to automatically extract these features from the user during the recognition phase. As users interact with the identity verification system through their primary biometric traits, the system should be able to automatically measure the ancillary speaker characteristics in a nonobtrusive manner, i.e. without additional interaction with the speaker. For a face-voice based speaker verification system this happens automatically, as the face and voice features are used for both verification and characterisation. The observed speaker characteristics are then used to supplement the identity information.

Abundant literature exists on methods to identify the gender, age, accent, ethnicity, and pose of the users, but most are based on unimodal biometric traits, either face images [9, 10, 11] or voice signals [12, 13, 14]. The use of voice information for automatic estimation of age using speaker modelling techniques is proposed in [12]. Another similar technique for recognising elderly users is proposed in [13]

For the research reported in this paper, we propose a multimodal approach to gender classification, using both visual and speech cues.

2.1 Visual features for gender classification

For the visual features, a hybrid face coding method by fusing appearance features and geometry features was used as shown in Figure 1. We used Haar wavelets with AdaBoost to represent the appearance features, while for the geometric features we used a representation based on Active Appearance Models (AAM).

The hybrid gender classification method consists of three main modules: the first normalises a given facial image after some geometric alignment and gray-level normalisation; the second extracts features from the normalised image to form a feature vector; and the third passes this vector to the gender classifier. This hybrid visual cue extraction technique allows

both global appearance features and local geometric features to contribute to the gender classification.

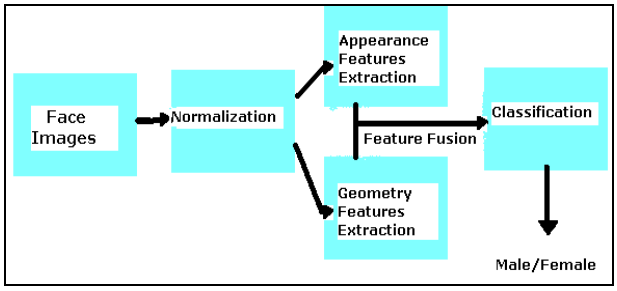

Figure 1: Gender classification schematic using hybrid visual cues.

2.1.1. Extraction of global features.

Many algorithms have been proposed for the extraction of global features [15, 16, 17]. In this paper, we chose AdaBoost to extract rectangular features from normalised facial images. Further, we chose Haar-like features introduced by [15], as shown in Figure 2a, and improved by [16], as shown in Figure 2b, to be the weak classifiers. For each box, the sum of the pixels in the white rectangles is subtracted from the sum of the pixels in the black rectangles. Within any image sub-window the total number of rectangular features is very large with the AdaBoost algorithm excluding "useless" features and retaining "useful" features, the latter containing more of the gender-related information.

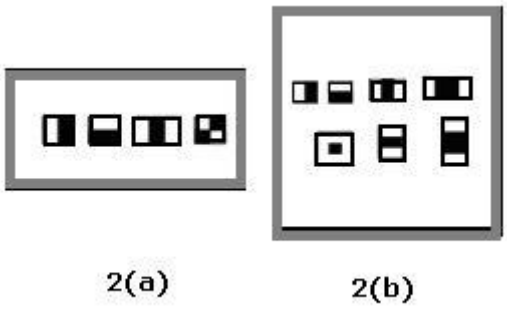

Figure 2: Haar-like features, (a) rectangles used by [15], (b) extended features used by [16]

2.1.2. Extraction of local features.

A number of sets of local features have been proposed for gender recognition, e.g. distance of eyebrow from eyes, eyebrow thickness and nose width [17], and a set of 406 geometric features in [18], of which 85% were shown to relate to significant differences in male and female features, but 18-20 sufficed to obtain a 96% correct gender classification. This conclusion lends strong support to the idea that geometry features can be regarded as a priori knowledge for gender classification. In our approach, we use AAM to get 83 landmarks from a face image and combine them with the 10 features in [18] where feature fusion with minimax normalisation is used to combine the two sets of global and local features. The performance of these visual features for gender classification is presented in Section 4.

2.2 Acoustic features for gender classification

Two different types of acoustic features, mel-frequency cepstra (MFC) and fundamental frequency (F_0) were extracted. The MFC were used for the identity verification subsystem and both MFC and F_0 were fused for gender classification. Speech exhibits significant variation from instance to instance for the same speaker and same text, but gender information in speech is to some extent time-invariant, phoneme-invariant, and speaker-independent for a given gender. Hence, cepstral features and fundamental frequency were both used for gender recognition. They were calculated for 16ms signal frames with 10ms overlap. The first 12 coefficients of the mel-cepstrum form one part of the acoustic feature vector. F_0 was calculated by an autocorrelation algorithm.

3. Two-part parallel fusion

In the parallel fusion framework proposed here, the system is divided into two subsystems. The first subsystem is the Speaker Verification Module, which uses the speaker-specific face-voice features, namely the hybrid visual features from the lip region and the acoustic features, discussed in Section 2, for verifying the identity of a client. The second subsystem is the Speaker Characteristics Module, which uses the same face-voice features to classify the gender of the client.

Figure 3 shows the block schematic of the proposed two-part parallel verification system, which fuses both the identity and the gender information.

The speaker verification module extracts speaker-specific face-voice feature vectors and uses a 3-mixture GMM classifier [6] to verify the identity of the test speaker by means of client and impostor likelihood scores. The gender classification module uses the gender-specific face-voice features and uses another 3 mixture GMM classifier to verify the gender of the speaker by means of male and female likelihood scores. The final accept-reject decision is arrived at by fusing the two scores with different weighting. The weight assigned to the speaker verification module is higher than that for the gender classification module. This ensures that the accept-reject decision is influenced more by identity information and to a lesser extent by the gender information.

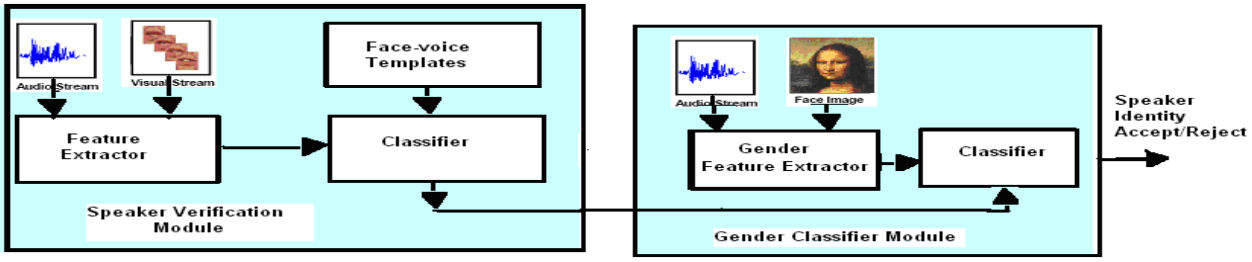

Figure 3: Block schematic for two-part parallel verification

Further, if we extend the system to include other speaker characteristics, some speaker characteristics may contain more information than the others. For example, the ethnicity of a speaker may give more information about the speaker, than gender. Therefore, we must introduce a weighting scheme for the speaker characteristics based on an index of distinctiveness and permanence, i.e. characteristics with smaller variability and larger distinguishing capability will be given more weight in the computation of the final score. A potential pitfall of the score weighting could be that an impostor could possibly spoof a client biometric more easily if the decision weights are biased in favour of speaker characteristics that are easier to modify than the primary identity characteristics. To avoid this problem, we assign smaller weights compared to those assigned to the primary biometric identifiers.

This differential weighting also has the implicit advantage of accounting for the expected error rate associated with a particular speaker characteristic such as age or gender. Even if a speaker characteristic is measured incorrectly (e.g., a male speaker is identified as a female), the effect on the speaker's likelihood score can be reduced by appropriate weight settings. The introduction of the weighting scheme (α, β) results in the following discriminant function LLR (log-likelihood ratio):

$$LLR(Ox, Oy) = \alpha \log\{p(O \mid \lambda_C\} + \beta \log\{p(O \mid \lambda_G\}$$

where $\alpha + \beta = 1$, and $\alpha > \beta$, Ox is the primary speaker-specific face-voice vector (concatenated hybrid visual vector from lip region and MFC), Oy is the secondary gender-specific face-voice vector (hybrid visual vector from the entire face region, MFC and F_0), and λ_C and λ_G are the client and gender GMMs, respectively. A decision threshold, θ is used to accept or reject the hypothesis H_0 that the joint face-voice test utterance (Ox, Oy) belongs to the claimed speaker model λ_C – if $LLR(Ox, Oy) > \theta$, H_0 is accepted, otherwise it is rejected.

Figure 4: DaFEx audiovisual corpus

The threshold θ can be adjusted to implement a verification system with the desired false-acceptance rate (FAR) or false-rejection rate (FRR) value. Since there is a trade-off involved, a system with lower FAR will suffer from a higher FRR and vice versa. As usual, the performance of the system is quantified by means of detection-error trade-off (DET) curves and resulting equal error rates (EER), for which the decision threshold is set such that FAR is equal to FRR.

4. Experimental results

Preliminary experimental results with an audio-visual emotion corpus DaFEx [19, 20] showed considerable improvement in speaker verification performance due to the utilisation of speaker characteristics, namely gender information.

DaFEx is an Italian audiovisual database composed of 1008 short videos from 8 professional actors (4 male and 4 female) primarily designed for emotion recognition (happiness, sadness, anger, fear, disgust, surprise and neutral). Both video and audio signals were recorded. Each actor recorded a sub-set of 126 videos, which includes all the emotions considered, at the three intensity levels and in the two different conditions.

Table 1: EER performance for face-voice speaker verification with speaker characteristics

Feature Set	EER (%)
Speaker Verification Module (no gender classifier)	10.5
Speaker Verification Module + gender classifier with acoustic features only	9.6
Speaker Verification Module + gender classifier with visual features only (appearance +geometric features)	7.5
Speaker Verification Module + gender classifier with both acoustic and visual features	6.8

The performance results in terms of DET curves and EERs in Table 1 and Figure 5 indicate that the inclusion of speaker characteristics, in this case gender, to aid speaker verification is quite promising. As can be seen from Table 1 and Figure 5, the EER for speaker verification system without gender classifier is 10.5% and the EER of the system including the gender classifier improves the speaker verification performance for all the gender classifiers we tested.

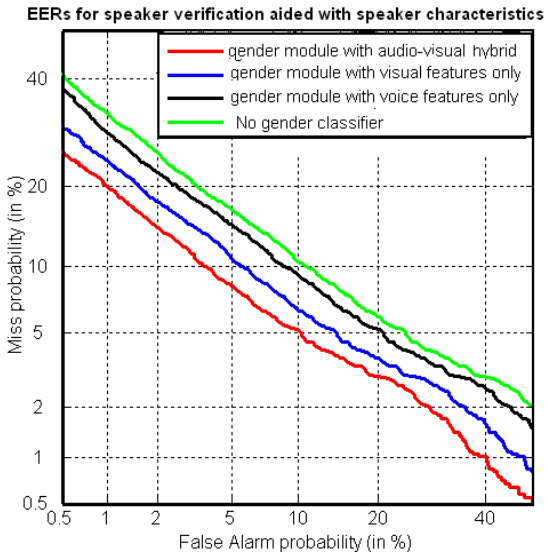

Figure 5: DET curve for face voice speaker verification aided by speaker gender. The 4 curves labelled red, blue, black and green in the top box, appear in order of increasing EER, the first (red) having the lowest EER and the last (green) having the highest EER.

With an audio-only gender classifier, the EER is 9.6% (a performance improvement of 8.6%), with hybrid visual features only, the EER is 7.5% (a performance improvement of 28.6%), and with both audio and hybrid visual features included, an EER of 6.8% (a performance improvement of 45.2%) is achieved.

It should be noted that the face-voice feature vectors for the gender classifier are fused in score-fusion mode, unlike the speaker verification module where the audiovisual features are fused in feature-fusion mode. This experimental investigation demonstrates the benefit of using known ancillary speaker characteristics along with the primary speaker identifiers for enhancing the performance of speaker verification systems.

5. Conclusions and further scope

In this paper we have presented a novel approach to utilise ancillary speaker characteristics like gender, height, weight, age, ethnicity, or colour of the eye/skin/hair, which are often known in practical authentication system, to complement verification that is solely based on identity modelling. We have shown that the ancillary speaker characteristic of age can be used in a separate parallel verification system to provide an additional verification score, which may be fused with the "traditional" audiovisual verification score to obtain improved speaker verification. Some ancillary speaker characteristics such as age may not be as permanent and reliable as the traditional biometric identity models, but they can provide important additional information about the identity of the speaker, which can lead to higher accuracy in establishing the speaker identity. In this paper, we have proposed a parallel fusion framework for integrating the speakers' gender information with the speakers' identity information based on face-voice feature vectors. Our initial experiments on a face-voice speaker verification system that uses gender as the additional information shows promising results and hence provides us with the motivation to include other speaker characteristics like dialect, ethnicity and age to be included in the speaker verification process. Our future work will also address techniques to determine the optimal set of weights for the ancillary speaker characteristics based on their distinctiveness and permanence.

6. References

[1] Jain, A.K., Ross, A., Prabhakar, S., "An Introduction to Biometric Recognition" in IEEE Transactions on Circuits and Systems for Video Technology, SpecialIssue on Image- and Video-Based Biometrics 14, 4–20 (2004)

[2] Wayman, J.L., Jain, A.K., Maltoni, D., Maio, D. (eds.), "Biometric Systems, Technology, Design and Performance Evaluation". Springer, Heidelberg (2005).

[3] Duc, B., Bigun, E.S., Bigün, J., Maitre, G., Fischer, S., "Fusion of Audio and Video Information for Multi Modal Person Authentication", Pattern Recognition Letters(18), No. 9, September 1997, pp. 835-843.

[4] Kittler, J.V., Matas, J., Jonsson, K., Sanchez, M.U.R., "Combining Evidence in Personal Identity Verification Systems", Pattern Recognition Letters(18), No. 9, September 1997, pp. 845-852.

[5] Bredin, H. and Chollet, G. 2007., "Audiovisual speech synchrony measure: application to biometrics". EURASIP J. Appl. Signal Process. 2007, 1 (Jan. 2007), 179-179.

[6] Chetty G., and Wagner M., "Robust face-voice based speaker identity verification using multilevel fusion", Image and Vision Computing, Volume 26, Issue 9, 1 September 2008, Pages 1249-1260.

[7] Lee J-E., Jain A. K., and Jin R.,Scars, "Marks and Tattoos (SMT): Soft Biometric for Suspect and Victim Identification", in Proceedings of IEEE Biometric Symposium, Biometric Consortium Conference, 2008.

[8] Golomb A., Lawrence D. T., and Sejnowski T. J. "SEXNET: A neural network identifies sex from human faces", Neural Information Processing Systems, 1991, pp: 572-577

[9] Cottrell G. W. and Metcalfe J.. EMPATH: "Face, emotion, and gender recognition using holons". Neural Information Processing Systems, 1991, pp: 564-571

[10] Costen, N., Brown, M., Akamatsu, S., "Sparse models for gender classification". Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FGR'04), 2004, pp: 201– 206..

[11] Shakhnarovich, G., Viola, P.A., Moghaddam, B., "A unified learning framework for real time face detection and classification", Proc. Internat. Conf. on Automatic Face and Gesture Recognition (FGR'02). IEEE, 2002, pp: 14–21.

[12] N. Minematsu, K. Yamauchi, and K. Hirose, "Automatic estimation of perceptual age using speaker modelling techniques," in Proceedings Interspeech 2003, Geneva, Switzerland, 2003, pp. 3005 – 3008.

[13] Mueller C., Wittig F., and Baus J., "Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs," in Proceedings Interspeech 2003, Geneva, Switzerland, 2003, pp. 1305 – 1308.

[14] Metze F., Ajmera J., Englert R., Bub U., Burkhardt F., Stegmann J., "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications," in ICASSP 2007 Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawai'i, USA, 2007, vol. 4, pp. 1089 – 1092.

[15] Viola P., Jones M. J., "Rapid Object Detection using a Boosted Cascade of Simple Features", Computer Vision and Pattern Recognition, 2001, Volume 1, pp: 8-14

[16] R. Lienhart, A. Kuranov, V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection", MRL(Microprocessor Research Lab, Intel Labs) Technical Report, 2002.

[17] R. Brunelli and T. Poggio. "Hyperbf networks for gender classification", DARPA Image Understanding Workshop, 1992, pp: 311–314.

[18] Samal, A., Subramani, V., Marx, D., "Analysis of sexual dimorphism in human face", J. Vis. Commun. Image R. 18 , 2007, pp: 453–463

[19] Battocchi, A.; Pianesi, F.. 2004. "DaFEx: Un Database di Espressioni Facciali Dinamiche". In Proceedings of the SLI-GSCP Workshop, Padova (Italy) 30 Novembre - 1 Dicembre 2004.

[20] Mana N., Cosi P., Tisato G., Cavicchio F., Magno E. and Pianesi F., "An Italian Database of Emotional Speech and Facial Expressions", In Proceedings of "Workshop on Emotion: Corpora for Research on Emotion and Affect", in association with "5th International Conference on Language, Resources and Evaluation (LREC2006), Genoa, Italy, 24-25-26 May 2006.