INTEGRATING ACOUSTIC AND LABIAL INFORMATION FOR SPEAKER IDENTIFICATION AND VERIFICATION

Pierre Jourlin [1,2], *Juergen Luettin* [1], *Dominique Genoud* [1], *Hubert Wassner* [1]
[1] IDIAP, rue du Simplon 4, CP 592, CH-1920 Martigny, Switzerland
[2] LIA, 339 chemin des Meinajariès, BP 1228, 84911 Avignon Cedex 9, France
jourlin@univ-avignon.fr, (luettin, genoud)@idiap.ch, wassner@ensta.fr

ABSTRACT

This paper describes a multimodal approach for speaker verification. The system consists of two classifiers, one using visual features and the other using acoustic features. A lip tracker is used to extract visual information from the speaking face which provides shape and intensity features. We describe an approach for normalizing and mapping different modalities onto a common confidence interval. We also describe a novel method for integrating the scores of multiple classifiers. Verification experiments are reported for the individual modalities and for the combined classifier. The performance of the integrated system outperformed each sub-system and reduced the false acceptance rate of the acoustic sub-system from 2.3% to 0.5%.

1 INTRODUCTION

Automatic verification of a person's identity is a difficult problem and has received considerable attention over the last decade. The ability of such a system to reject impostors, who claim a false identity, becomes a critical issue in security applications. The use of multiple modalities like face, profile, motion or speech is likely to decrease the possibility of false acceptance and to lead to higher robustness and performance. Brunelli et al. [1] have previously described a bimodal approach for person identification. The system was based on visual features of the static face image and on acoustic features of the speech signal. The performance of the integrated subsystem was shown to be superior to that of each subsystem.

The cognitive aspect of lip movements in speech perception has been studied extensively and the complementary nature of the visual signal has been successfully exploited in bimodal speech recognition systems. The fact that temporal lip information not only contains speech information but also characteristic information about a person's identity has largely been ignored, until recently, where Luettin and al. [9] have proposed a new modality for person recognition based on spatio-temporal lip features.

In this paper, we extend this approach and address the combination of the acoustic and visual speech modality for a speaker verification system. We describe the normalization and mapping of different modalities and the determination of a threshold for rejecting impostors. A scheme for combining the evidence of both modalities is described and we show that the performance of the multimodal system outperforms both unimodal subsystems.

2 DATABASE FEATURES

The M2VTS audio-visual database has been collected at UCL (Catholic University of Louvain) [13]. It contains 37 speakers (male and female) pronouncing in French the digits from zero to nine. One recording is a sequence of the ten digits pronounced continuously. Five recordings have been taken of each speaker, at one week intervals to account for minor face changes like beards and hairstyle. The images contain the whole head and are sampled at 25 Hz. We have divided the database into 3 sets : the first three shots were used as training set, the 4th shot as validation set and the 5th shot as test set. The 5th shot represents the most difficult recordings to recognize. This shot differs from the others in face variation (head tilted, unshaved), voice variation (poor voice SNR), or shot imperfections (poor focus, different zoom factor).

3 LIP FEATURE EXTRACTION

We are interested in facial changes due to speech production and therefore analyse the mouth region only. Common approaches in face recognition are often based on geometric features or intensity features, either of the whole face or of parts of the face [2]. We combine both approaches, assuming that much information about the identity of a speaker is contained in the lip contours and the grey-level distribution around the mouth area. During speech production the lip contours deform and the intensities in the mouth area change due to lip deformation, protrusion and visibility of teeth and tongue. These features contain information specific to the speech articulators of a person and to the way that person speaks. We aim to extract this information during speech production and to build spatio-temporal models for a speaking person.

3.1 Lip Model

Our lip model is based on active shape models [3] and has been described in detail in [11]. It is used to locate, track and parameterize the lips over an image sequence of a speaking person. Features are recovered from tracking results. They describe the shape of the inner and outer lip contours and the intensity at the mouth area. The shape features and the intensity features are both based on principal component analysis which was performed on a training set. The intensity model deforms with the lip contours and therefore represents shape independent intensity information. This is an important property of the model. We obtain detailed shape information from the shape parameters and therefore would like the intensity model to describe intensity information which is independent of the lip shape and lip movements [10].

3.2 Lip Tracking

Experiments were performed on all 5 shots of the M2VTS database. The database consists of colour images which were converted to grey-level images for our experiments. Several subjects have a beard or did not shave between different recordings. We used examples from the training set to build the lip model. The model was then used to track the lips over all image sequences of all three sets. This consisted of analysing over 27 000 images which we believe is the largest experiment reported so far for lip tracking. It is important to evaluate the performance of the tracking algorithm and we have previously attempted to do this by visually inspecting tracking results [11]. However this task is very laborious and subjective. Here we omit direct performance evaluation of the tracking algorithm. Instead we try to evaluate the combined performance of the feature extraction and the recognition process by evaluating the person recognition performance only. Person recognition errors might therefore be due to inaccurate tracking results or due to classification errors.

4 SPEAKER VERIFICATION

4.1 Test Protocol

We use the sequences of the training set (first 3 shots) of the 36 customers for training the speaker models. The validation set serves for computing the normalization and mapping function for the rejection threshold and the test set is used for the verification tests. Subject 37 is only used as impostor, claiming the identity of all 36 customers. Each customer is also used as an impostor of the 35 other customers. The verification mode is text-dependent and based on the whole sequence of ten digits. For the verification task, we make use of a *world* model, which represents the average model of a large number of subjects (500 speakers for the acoustic model and 36 for the labial one) For each digit we compute the corresponding customer likelihood and the world likelihood. We can so obtain a customer and a world likelihood for all speech data. The difference between the ratio of the two scores and the threshold is then mapped to the interval [0, 1] using a sigmoïd function [5]. Then, the acceptation threshold for a final score is 0.5. Several methods have been proposed in order to find an a priori decision threshold, according to various criteria, e.g. Equal Error Rate and Furui method [4]. Due to the small amount of speech data for each speaker we calculated a customer independent threshold, based on a dichotomic method. The use of this method implies that the function of verification errors is convex. This function is computed on the *validation* set and the value, for which the number of false acceptance and false rejection errors is minimum, is used as threshold value.

4.2 Acoustic Sub-system

Since the word sequences are known in our experiments, we use a HMM based speech recognition system to segment the sentences into digits. The recognizer uses the known sequence of digit word models, which were trained on the *Polyphone* database of IDIAP [8], to find the word boundaries. Each digit HMM has been trained with 110 to 200 examples of 835 speakers. The segmentation is performed on all three sets. The segmented training set is used to train one model for each digit and speaker. These models are called *customer* models. The acoustic parameters are Linear Prediction Cepstral Coefficients with first and second order derivatives (39 components). We used left-right HMMs with between 2 and 7 emitting states, depending on the digit length. Each state is modelled with one single Gaussian mixture with diagonal covariance matrix. The same configuration is used for the world model. The world model is trained on the *Polyphone* database using 300 examples from 500 speakers for each digit. When an access test is performed, the speech is first segmented into digits. The test protocol described above is applied, where the customer and world likelihoods are obtained by the product of all digit likelihoods, using the customer and world models, respectively. The mapping function, obtained from the validation set, is used in the test set to map the score into the confidence interval. The results are shown in Figure 1 and Table 1.

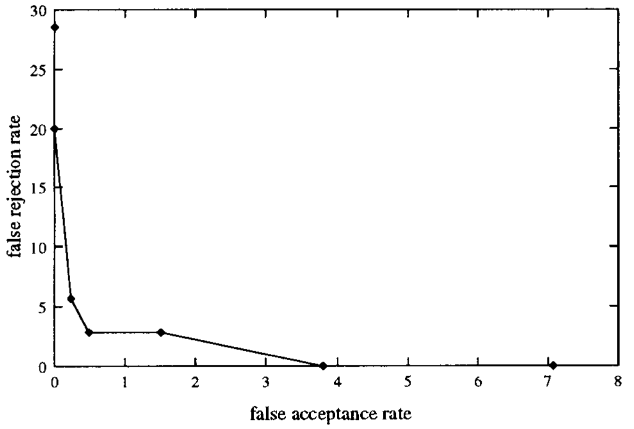

Figure 1: Receiver Operating Characteristic (ROC) curve for the acoustic sub-system (validation set)

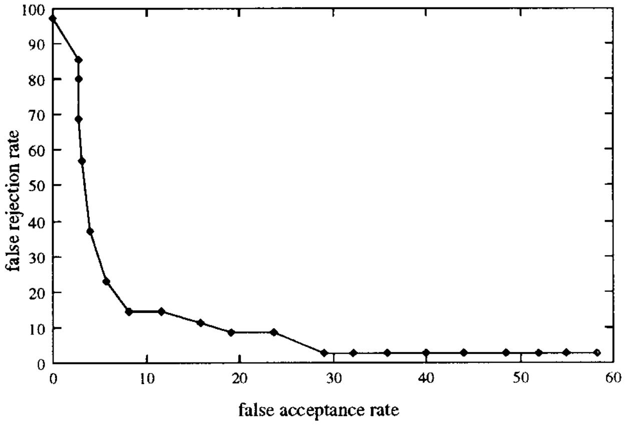

Figure 2: Receiver Operating Characteristic (ROC) curve for the labial sub-system (validation set)

4.3 Labial Sub-system

For segmenting the labial data we use the previous acoustic segmentation. Lip features can improve speech recognition results [7], but they do not provide enough information to segment speech into phonemic units. Lip movements may be useful in addition to the acoustic signal for segmenting speech, especially in a noisy acoustic environment (see [12] and [6]). We did not use visual information for segmentation since our acoustic models were trained on a very large database and are therefore more reliable than our labial models. We used the same scoring method for labial verification as for acoustic verification, except the world model, which was trained on the 36 customers from the M2VTS database. Labial data has a four times lower sampling frequency than acoustic data. The number of emitting states was therefore chosen to be 1 or 2, depending on the digit length. !la The parameter vectors consisted of 25 components: 14 shape and 10 intensity parameters and the scale. The same test protocol, which was used for acoustic experiments, was now used for labial verification. The results are shown in Figure 2 and Table 1.

4.4 Acoustic-Labial System

The acoustic-labial score is computed as the weighted sum of the acoustic and the labial scores. Both scores have been normalized as described in the previous sections. The process uses individual threshold values for each modality and maps the scores into a common confidence interval. The normalization process is a critical point in the design of an integration scheme and is necessary to ensure that different modalities are mapped into the same interval and share a common threshold value. The different modalities are now normalized but they provide different levels of confidence. We therefore need to weight the contribution of each modality according to their confidence level. The weight is α for the acoustic score and $1-\alpha$ for the labial one. The same dichotomic algorithm, used to compute the thresholds, is now used to find the optimal weight α. The function of verification results on the validation data is used for the dichotomic search, for the same reasons as described for threshold search.

The following results were obtained on the test set : using a weight of 0.86, we obtain a false acceptance rate of 0.5%, a false rejection rate of 2.8% and a correct identification rate of 100.0%. The absolute gain is a 1.8% reduction in cumulated verification errors (FA+FR) and an increase of 2.8% in the identification rate (ID). Fig 3 shows the effects of weighting on acoustic-labial results, when the acceptance threshold is optimaly fixed for each modality. Table 1 sums up the results.

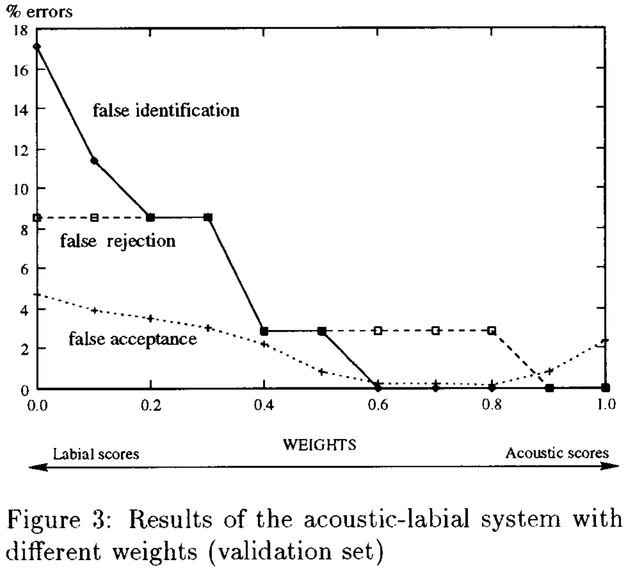

Figure 3: Results of the acoustic-labial system with different weights (validation set)

We have followed a data-driven approach for fusion, where data fusion is present at different levels. At the first stage, learning and decoding of the labial models use the segmentation obtained from the acoustic models.

	ID	FA	FR
Validation set			
Acoustic	100.0	2.5	0.0
Labial	82.3	4.9	8.8
Bimodal	100.0	0.6	0.0
Test set			
Acoustic	97.2	2.3	2.8
Labial	72.2	3.0	27.8
Bimodal	100.0	0.5	2.8
Number of tests	36	1332	36

ID	:	Correct identification rate
FA	:	False acceptance rate
FR	:	False rejection rate

Table 1: Validation and test set results

The first score normalization is performed by normalizing the scores with respect to a world model for each modality. The final normalization is obtained by finding an optimal mapping in the interval [0, 1] for each modality. At this stage, the two scores are normalized, but we know that each modality has different levels of reliability. So, the last level of the fusion process is to find the optimal weight for the two sources of information.

5 CONCLUSION

The bimodal speech processing is a very new domain. Moreover, lip feature extraction is quite new in the computer vision field and is well known to be a difficult problem. Despite of these difficult conditions, the results we have obtained are very promising. The speech part of the M2VTS database seems to be quite small compared with other acoustic databases, but it is maybe the largest existing database for audio-visual speaker verification. The number of tests is quite small compared to other acoustic speaker verification experiments. However, the reduction of the false acceptance rate for the multimodal system suggests that the supplemental use of lip information could improve the performance of an acoustic based speaker verification system.

References

[1] Brunelli, R. and D. Falavigna (1995). Person identification using multiple cues *IEEE Trans. Pattern Anal. Mach. Intell.* 17 (10), 955-966.

[2] Chellappa, R., C.L. Wilson and S. Sirohey (1995). Human and machine recognition of faces: A survey. *Proceedings IEEE* 83 (5), 705-740.

[3] Cootes, T.F., A. Hill, C.J. Talor and J. Haslam (1994). Use of active shape models for locating structures in medical images. *Image and Vision Computing* 12 (6), 355-365.

[4] Furui, S. (1994). An overview of speaker recognition technology. In: *ESCA Workshop on Automatic Speaker Recognition Identification Verification*, Martigny, Switzerland, 1-9.

[5] Genoud, D., F. Bimbot, G. Gravier and G. Chollet (1996). Combining methods to improve speaker verification decision. In: *Internat. Conf. Speech and Language Processing*, Philadelphia, PA, 1756-1759.

[6] Jourlin, P., M. El-Bèze and H. Méloni (1995). Bimodal speech recognition. In: *Internat. Workshop on Automatic Face and Gesture Recognition*, Zurich, 320-325.

[7] Jourlin, P. (1996). Handling disynchronization phenomena with HMM in connected speech. In: *European Signal Processing Conference*, Trieste, Italy, 133-136.

[8] Chollet, G., J-L. Cochard, A. Constantinescu and P. Langlais (1995). Swiss french polyphone and polyvar : telephone speech databases to study intra and inter speaker variability. Technical Report, IDIAP, Martigny, Switzerland.

[9] Luettin, J., N.A. Thacker and S.W. Beet (1996a). Speaker identification by lipreading. In : *Internat. Conf. Spoken Language Processing*, Philadelphia, PA, 62-65.

[10] Luettin, J., N.A. Thacker and S.W. Beet (1996b). Speechreading using shape and intensity information. In: *Internat. Conf. Spoken Language Processing*, Philadelphia, PA, 58-61.

[11] Luettin, J. and N.A. Thacker (1997). Speechreading using Probabilistic Models. *Computer Vision and Image Understanding* 65 (2), 163-178.

[12] Mak, M.W. and W.G. Allen (1994). Lip-Motion analysis for speech segmentation in noise. *Speech Communication* 14 (3), 279-296.

[13] Pigeon, S. and L. Vandendorpe (1997). The M2VTS Multimodal Face Database (Release 1.00). In: J. Bigün, G. Chollet and G. Borgefors, Ed.,*Audio- and Video-based Biometric Person Authentication*, Springer-Verlag, Berlin, 403-409.