Robust Multi-modal Person Identification with Tolerance of Facial Expression*

Niall A. Fox
Dept. of Electronic and Electrical Engineering
University College Dublin
Belfield, Dublin 4, Ireland
niall.fox@ee.ucd.ie

Richard B. Reilly
Dept. of Electronic and Electrical Engineering
University College Dublin
Belfield, Dublin 4, Ireland
richard.reilly@ucd.ie

Abstract -*The research presented in this paper describes audio-visual speaker identification experiments carried out on a large data set of 251 subjects. Both the audio and visual modeling is carried out using hidden Markov models. The visual modality uses the speaker's lip information. The audio and visual modalities are both degraded to emulate a train\test mismatch. The fusion method employed adapts automatically by using classifier score reliability estimates of both modalities to give improved audio-visual accuracies at all tested levels of audio and visual degradation, compared to the individual audio or visual modality accuracies. A maximum visual identification accuracy of 86% was achieved. This result is comparable to the performance of systems using the entire face, and suggests the hypothesis that the system described would be tolerant to varying facial expression, since only the information around the speaker's lips is employed.*

Keywords: Multi-modal, fusion, audio-visual, person identification, classifier combination, lip modality.

1 Introduction

Biometrics is a field of technology devoted to verification or identification of individuals using biological traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas identification, a multi class problem, involves identifying a user from a set of subjects. Due to this fact, speaker identification is inherently a more difficult task, particularly when the number of registered subjects is large. Speaker identification systems based on the analysis of audio signals achieve high performance when the signal to noise ratio (SNR) of the audio signal is high. However the performance degrades quickly as the test set SNR decreases [11], which we refer to as an audio train\test mismatch.

The area of audio-visual signal processing has received much attention over the past ten years. Recent state of the art reviews indicate that much of the research carried out focuses on audio-visual speech recognition [8], [6]. The most important issues are, how to account for the reliabilities of the two modalities and at what level to carry out the fusion. The benefits of audio-visual fusion for the purpose of speaker identification have been shown in [11]. However, the fusion method employed used modality weightings found by exhaustive search to optimize the fusion scores. While this highlights the potential of audio-visual fusion, it is not useful in a practical real world scenario. Other audio-visual speaker identification approaches that use more automated fusion techniques [4] do not address the issue of an audio train\test mismatch. In [22], audio visual speaker verification experiments are carried out on 36 subjects, however, only an audio train\test mismatch was tested, whereas a visual train\test mismatch was not considered. In [7], robust audio-visual classifier fusion under both audio and visual train\test mismatch conditions is described. The adaptive fusion results were encouraging, with improved audio-visual scores better than either modality alone. However, the experiments were carried out on small database of just eight subjects. The work presented in this paper describes audio-visual speaker identification experiments on a large data set of 251 subjects from the XM2VTS audio-visual database [16]. In the context of this paper, the visual modality refers to a sequence of mouth images extracted from a video utterance. Both, the audio and visual modalities are degraded to emulate train\test mismatches.

This paper is organized as follows. In Section 2, the audio-visual corpus employed is described. Sections 3 and 4 describe how we performed the audio and visual identification respectively. Section 5 investigates audio-visual fusion techniques and describes how the fusion is carried out in this study. In Section 6 we present our results and finally in Section 7, the paper is summarized and some conclusions are offered.

2 Audio-Visual Corpus

The XM2VTS audio-visual database [16] was used for the experiments described in this paper. The database consists of video data recorded from 295 subjects in four sessions, spaced monthly. The first recording per session of

* 0-7803-8566-7/04/$20.00 © 2004 IEEE.

the third sentence (*"Joe took fathers green shoe bench out"*) was used for this research. This sentence was used because it was phonetically balanced. Some sentence recordings had the start of the word *Joe* clipped and in some cases it was totally missing. Due to this and other errors in the sentences, only 251 out of a possible 295 subjects were used for our experiments.

3 Audio Identification

Audio based speaker identification is a mature topic [5], [20]. Standard audio methods have been employed in this paper. The audio signal was first pre-emphasized to increase the acoustic power at higher frequencies using the filter $H(z) = 1/(1-0.97z^{-1})$. The pre-emphasized signal was divided into frames using a Hamming window of length 20 ms, with an overlap of 10 ms to give an audio frame rate of 100 Hz. Mel-frequency cepstral coefficients (MFCC's) [9] of dimension 16 were extracted from each frame. The energy [23] of each frame was also calculated and used as a 17^{th} static feature. Static features refer to features extracted from individual audio frames that do not depend on other frames. Seventeen first order derivatives or *delta* features were calculated using W_D adjacent static frames, where W_D is the delta window size. The *delta* frames were appended to the static audio features to give an audio feature vector of dimension 34. These were calculated using the available HTK functions [23] employing a W_D value of five frames. *Cepstral mean normalization* [23] was performed on the audio feature vectors (to each audio utterance) in order to compensate for long term spectral effects of the channel.

A text dependent speaker identification methodology was tested. For text dependent modeling [5], the same utterance is spoken by the subject for both training and testing. It was employed, as opposed to text independent modeling [20], due to its suitability to the database used in this study.

The N subject classes S_i, $i = 1,2,...,N$, are represented by N speaker hidden Markov models (HMMs) [19], where $N = 251$ here. There was one background or global HMM. The first three sessions were used for training and the last session was used for testing. The background HMM was trained using three of the sessions for all N subjects. The background model was initialized using a prototype model. A prototype HMM consists of the initial HMM parameters. The background model captures the audio speech variation over the entire database. Since there were only three training utterances per subject, there was insufficient training data to train a speaker HMM directly from a prototype model. For this reason, the background model was used to initialize the training of the speaker models. Since HMM classifiers are employed, the classifier output scores are in *log-likelihood* form, denoted by $ll(O|S_i)$. The classification task is the calculation of maximum likelihood class. The audio speaker models were trained on the "clean" audio speech, which was the original audio data. Additive white Gaussian noise was applied to the clean audio at SNR levels ranging from 48 dB to 21 dB in decrements of 3 dB. All models were trained using clean speech and tested using the various SNR levels. This provides for a mismatch between the testing and training audio conditions. HMM training and testing was carried out using the HTK toolkit, version 3.1 [23].

4 Visual Identification

Visual speech feature analysis has also received much attention recently [15], [18]. Transform based features were used to represent the visual information based on the Discrete Cosine Transform (DCT), which was employed because of its high energy compaction [17]. The visual mouth features were extracted from the mouth region of interested (ROI), which consists of a *49x49* colour pixel block; see Fig. 1. The position of the mouth ROI was determined by manually labelling the left and right labial corners and taking the centre point. Frames were manually labelled for every 10^{th} frame only; the ROI positions for the other frames were interpolated. The ROI blocks were converted to gray scale values. The gray scale ROI was then histogram equalized and the mean pixel value was subtracted. This image pre-processing was carried out to account for varying illumination conditions across sessions. Carrying out both histogram equalization and mean subtraction on the images was found to improve the performance of the visual system. The DCT was applied the gray scale pixel blocks. Considering that most of the information of an image is contained in the lower spatial frequencies [17], the first 15 DCT coefficients were used, taken in a zigzag pattern to form the visual frame observation feature vector.

For our study, in order to account for practical video conditions, the video frame images were compressed using JPEG compression [21]. In our experiments, ten values of JPEG quality factor (QF) were examined, $QF \in \{50, 25, 18, 14, 10, 8, 6, 4, 3, 2\}$, where a QF of 100 represents the original uncompressed image. The compression was carried out using Matlab, version 6.5 [13], and applied to each individual video frame image. The mouth ROI was then extracted from the compressed images. The mouth ROI was then extracted from the compressed images. The manually labelled mouth coordinates were employed, so that any drop in visual performance would be due to mismatched testing conditions rather than to poorer mouth tracking. The variation of the mouth ROI with respect to JPEG QF is shown in Fig. 1. The JPEG blocking artifacts are evident at the lower QF levels.

The visual sentences were modeled using the same HMM methodology as described for the audio sentences. The number of states employed was adjusted to achieve the best visual accuracy. In all cases the visual speaker HMMs were trained on the "clean" visual images and tested on the degraded visual images. This provides for a mismatch between the testing and training visual conditions.

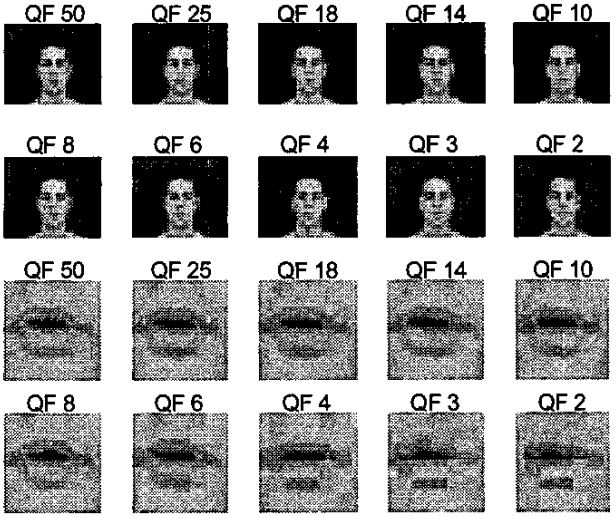

Fig. 1: Ten levels of JPEG compression and ROI images.

5 Audio-Visual Fusion

The fusion of classifiers is a mature research topic, which predates work on audio-visual speech fusion [14]. The main difference between audio-visual speech fusion and general classifier fusion is the similarity of the audio and visual modalities. For example, in this paper, both modalities are modelled using a similar HMM methodology. This enables fusion to be carried out at not only the classification level but at pre-classification levels also, which results in a large variety of possible fusion approaches. The two audio-visual fusion approaches most commonly investigated are feature fusion (early integration) and score fusion (late integration). Feature fusion, while being very simple to implement via feature vector concatenation of the two modalities, has several disadvantages. The audio-visual feature vector has a larger dimension, and due to the "curse of dimensionality", this results in making the training of parametric models, such as HMMs, less practical because the models will be under trained unless a large amount of training data is available, which is rarely the case for audio-visual speaker modelling. In addition to this, feature fusion does not take the reliability of either modality into account; if one modality is of a very poor reliability, the combined audio-visual feature vector will be compromised and catastrophic fusion may occur; where the combined audio-visual accuracy is poorer than either of the single modalities; this has been demonstrated in [11]. This issue is especially important for the visual modality where a tracking failure can occur. Another issue with feature fusion is that the frame rates of the audio and visual features are usually very different. Audio features are usually extracted at a frame rate in the region of *100Hz*, whereas visual features are limited to a frame rate equal to the video capturing speed, which is usually in the region of *25/30Hz* (or frames per second). Hence, for feature fusion, the visual features need to be up-sampled (usually via interpolation) in order to synchronise the two feature frame rates.

Score fusion consists of using the audio and visual classifier outputs to provide an audio-visual classification. The benefit of this method is that the audio and visual classifier outputs can be weighted in such a way that takes the reliability of both modalities into account. Most automatic audio-visual fusion techniques only use an audio reliability measure [12] and the visual signal is assumed to be of a constant quality. Moreover, it is assumed that the visual modality is equally distinguishable for all speakers, i.e. it only performs poorly if there is a train\test mismatch. In a practical scenario, neither of these assumptions is correct. For any given modality a particular speaker may not be very distinguishable. If the reliability parameter is determined before separate modality classification takes place, e.g. by measuring the audio noise levels or the mouth tracking integrity, then, for a given speaker, an indistinguishable modality could not receive a lower weighting. Taking these points into account, it is better to calculate a reliability measure based on the classifier score distribution, as this can quantify both the train\test mismatch and the ability to distinguish a speaker for a given speaker utterance. For these reasons, the fusion method employed here uses a reliability measure based on the classifier output score distribution and takes the reliability of both modalities into account.

The log-likelihood scores of the two modalities are normalized and integrated on a per utterance basis. No prior statistics of the log-likelihood score distributions are employed. The scores were normalization by scaling the top M scores into the range $[0,1]$. Using a low value of M reduces the potential of audio-visual fusion, with the limit of $M=1$ amounting to fusion by method of voting [14]. Fusion by voting is usually only carried out when many classifiers are used; the chosen class is the class that received the highest rank by the most classifiers. Hence no reliability information is considered. In the case of a high value of M, the worst scores (outliers) can unfairly skew the distribution of the normalized scores. Tests showed that the system performance degraded for $M<50$ and $M>100$. A value for M of 75 was chosen. Since the top M normalized scores are not log-likelihoods; instead of using the log-likelihood $ll(O|S_j)$, we use the likelihood $l(O|S_j)$.

For a given test utterance observation O, the audio and visual observation sequences are denoted with O_A and O_V respectively. The combined likelihood that O was produced by the i^{th} speaker S_i, is:

$$l_{AV}(O_A,O_V \mid S_i) = \alpha_A . l_A(O_A \mid S_i) + \alpha_V . l_V(O_V \mid S_i), \quad (1)$$

where α_A denotes the weight of the audio likelihood and α_V denotes the weight of the visual likelihood such that:

$$\alpha_A + \alpha_V = 1 \text{ and } \alpha_A, \alpha_V \in [0,1]. \quad (2)$$

Various reliability estimates have been used in the literature. Some examples include score dispersion [12], score entropy [12], score variance [22], cross classifier coherence coefficient [7] and the difference, ξ, of the top two best scores [2]. ξ is calculated as

$$\xi_m = l_m(O_m \mid S_\alpha) - l_m(O_m \mid S_\beta), \quad m \in \{A,V\}, \quad (3)$$

where S_α and S_β are the speakers achieving the best and second best scores respectively, and m denotes the modality. The difference of the top two best scores was employed for this study because, even though it is computational inexpensive, it performed well across all levels of audio and visual degradation. A high value of ξ indicates a confident score whereas a low value indicates a score of poor confidence since the separation of the highest speaker to the others is low.

A mapping between the reliability estimate and $\alpha_A \backslash \alpha_V$ is required. A sigmoidal mapping [10], can be used, but the parameters of the sigmoid curve require training. Another option is to form corresponding bins of reliability estimates and $\alpha_A \backslash \alpha_V$ values, effectively a lookup table, but again this requires extensive training. Considering the small amount of audio-visual training data available, just three speaker sessions, it was decided to use a non-learned approach to choosing the $\alpha_A \backslash \alpha_V$ values where α_V is varied from 0 to 1 in steps of 0.05 for a particular utterance. For each α_V value, the audio and visual scores are combined using (2). The top M audio-visual scores are then normalized, as above, and the audio-visual reliability estimate, ξ_{AV}, is calculated and maximized to give

$$\alpha_{V_{opt}} = \arg\max_{\alpha_V \in (0,1)} \{\xi_{AV} \mid \alpha_V\}, \quad (4)$$

where $\alpha_{V_{opt}}$ is the selected visual weighting that maximizes ξ_{AV}. Fig. 2 shows a sweep of α_V from 0 to 1 in steps of 0.05 for a particular utterance. The audio-visual reliability estimate reaches a maximum at an α_V value of 0.45.

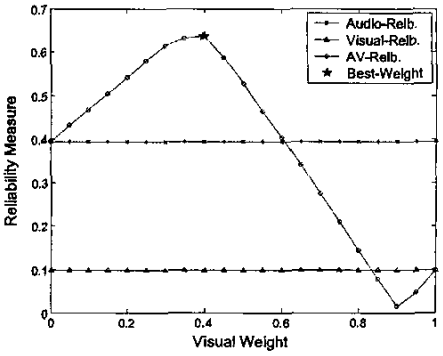

Fig. 2: Example of how the audio-visual reliability estimate varies with respect to α_V.

6 Results and Discussion

6.1 Audio Results

The number of audio HMM states that maximised the audio identification score was found to be eleven (two mixes per state). Fig. 3 shows how the audio performance with respect to the audio degradation. A maximum accuracy of 97.6% was achieved at 48dB. At 21dB the accuracy dropped to 37%. Further lowering of the SNR resulted in a random choice accuracy at -3dB, i.e. an accuracy of 0.398% or *100x(1/251)*. The audio experiments performed very well under near "clean" testing conditions, however the accuracy roll off with respect to SNR is very high.

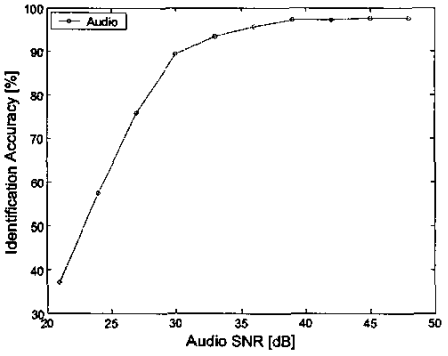

Fig. 3: Audio accuracy versus audio SNR.

6.2 Visual Results

We first tested the effect that the number of visual HMM states had on the visual identification accuracy. These tests were carried using matched training and testing data sets, i.e. images of QF 100. The number of states was increased from one, until a trend became apparent. The results of this are shown in Fig. 4. The visual features performed best with just one state and the performance decreased steadily with

increasing number of states. The fact that the visual features performed best with just one state indicates that HMMs may be not required to model visual speech when using static features, rather, a simpler Gaussian mixture model (GMM) approach [20] would be sufficient.

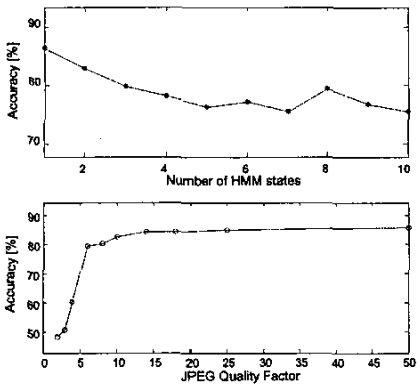

Fig. 4: Visual performance versus number of HMM states (top) and effects of JPEG compression (bottom).

The tests on the degraded visual data were carried using one HMM state. Fig. 4 shows how the visual features perform with respect to JPEG degradation. The visual features show a high level of robustness, with an accuracy of 48% at a QF of 2. This level of robustness conforms with the high level of speech recognition robustness to JPEG compression reported in [18]. In [3] face identification was carried out across five levels of JPEG compression. Again, a high level of robustness was reported, with no significant drop in performance, except for the lowest JPEG quality level. It should be noted for the experiments described in this paper, that if the mouth ROI was automatically segmented, rather than manually labelled, poorer robustness to visual degradations should be expected.

6.3 Audio-Visual Results

The results for the automatic fusion method described above are shown in Fig. 5 and Table 1. It is apparent that the automatic fusion accuracies are higher than either of the audio and visual modalities, at all levels of degradation. For example, the accuracies at the most severe levels of audio and visual modality train\test mismatch are 37% and 48% respectively, whereas the automatic fusion of the at these levels achieved an accuracy of 70%.

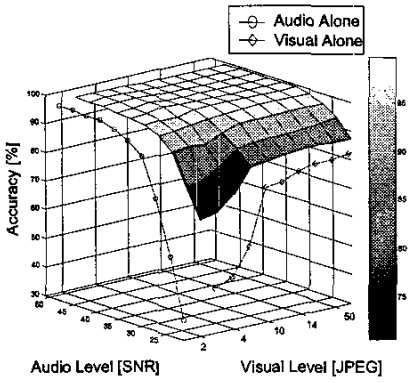

Fig. 5: Speaker identification accuracies for audio, visual and automatic audio-visual fusion.

Table 1: Automatic audio-visual fusion accuracies (%) for ten levels of audio (dB) and visual (QF) degradation.

QF \ dB	48	45	42	39	36	33	30	27	24	21
50	99.2	99.2	99.2	99.2	99.2	99.2	98.4	96.4	93.2	87.3
25	99.2	99.2	99.2	99.2	99.2	99.2	98.4	96.4	92.4	87.3
18	99.2	99.2	99.2	99.2	99.2	99.2	98.4	96.4	91.6	87.6
14	99.2	99.2	99.2	99.2	99.2	99.2	98.4	96.0	91.2	87.3
10	99.2	99.2	99.2	99.2	99.2	99.2	98.0	96.0	91.2	86.9
8	99.2	99.2	99.2	99.2	99.2	99.2	97.6	96.0	90.8	85.7
6	99.6	99.2	99.2	99.2	99.2	98.8	97.2	94.0	90.0	84.9
4	99.6	99.6	99.6	99.6	99.6	97.6	96.4	91.2	85.3	76.9
3	99.6	99.6	99.6	98.8	98.4	97.6	95.6	92.0	81.7	71.7
2	99.6	99.2	99.2	98.4	98.0	97.2	95.2	91.2	80.5	70.5

7 Conclusions

The XM2VTS database was recorded under extremely well controlled visual conditions; it does not represent practical real world scenarios. This is highlighted by the exceptionally high visual speaker identification accuracies achieved in this study, the best accuracy been 86%. The new BANCA database [1] is a large audio-visual database consisting of 208 subjects, that is recorded under controlled, degraded and adverse scenarios, which will provide data of a more practical nature for the testing of audio-visual fusion methodologies. Experimental results have being presented for audio-visual fusion with application to speaker identification. Ten levels of train\test mismatch of not only the audio modality but also the visual modality have been examined. The audio-visual fusion methodology uses reliability estimates of both the audio and visual modalities. Additional audio-visual training data is not required to tune the fusion process.

The results are encouraging with the audio-visual accuracies exceeding both the audio and visual accuracies at all levels of audio and visual degradation, and some cases comparable to the accuracies achieved by the empirical fusion method. The fusion method is computationally inexpensive. The audio-visual system described has applications in practical scenarios, such as human computer interfaces and can also be extended to biometrical systems for robust person verification. Due to the audio-visual method employed, the audio articulation is being accounted for by the visual modality. For example, any outliers in the audio modality are compensated for, by the visual modality. A maximum visual identification accuracy of 86% was achieved. This result is comparable to the performance of systems using the entire face region, (93% facial identification reported in [10] on the same database) and suggests the hypothesis that the system described would be tolerant to varying facial expression, since only the information around the speaker's lips is employed. Ongoing work includes the recording of video data under controlled and degraded scenarios, which will supplement the data used in this study.

References

[1] E. B. Bailliere, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariethoz, J. Matas, K. Messer, V. Popovici, F. Poree, B. Ruiz, and J. P. Thiran, "The BANCA Database and Evaluation Protocol," *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA*, Guildford, UK, pp. 625-638, June 2003.

[2] S. Basu, H. S. M. Beigi, S. H. Maes, M. Ghislain, E. Benoit, C. Neti, and A. W. Senior, "Methods and Apparatus for Audio-visual Speaker Recognition and Utterance Verification." United States: Patent 6,219,640, 1999.

[3] D. M. Blackburn, M. Bone, and P. J. Phillips, "Facial Recognition Vendor Test 2000: Evaluation Report," 2000.

[4] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," *IEEE Transactions on Pattern Analysis and Machine Intelligience*, vol. 17, pp. 955-966, Oct. 1995.

[5] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, vol. 85, pp. 1437-1462, Sept. 1997.

[6] T. Chen, "Audiovisual Speech Processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 9-21, Jan. 2001.

[7] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "Adaptive Classifier Integration for Robust Pattern Recognition," *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, vol. 29, pp. 902-907, Dec. 1999.

[8] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-Based Bimodal Recognition," *IEEE Transactions on Multimedia*, vol. 4, pp. 23-35, Mar 2002.

[9] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357-366, Aug. 1980.

[10]N. A. Fox, R. Gross, P. de Chazal, J. F. Cohn, and R. B. Reilly, "Person Identification Using Automatic Integration of Speech, Lip, and Face Experts," *Proceedings of the ACM SIGMM Multimedia Biometrics Methods and Applications Workshop*, Berkeley, CA., pp. 25-32, Nov. 2003.

[11]N. A. Fox and R. B. Reilly, "Audio-Visual Speaker Identification Based on the Use of Dynamic Audio and Visual Features," *Proc. of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication*, Guildford, UK, pp. 743-751, June 2003.

[12]M. Heckmann, F.. Berthommier, and K. Kristian, "Noise Adaptive Stream Weigthing in Audio-Visual Speech Recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, pp. 1260-1273, Nov. 2002.

[13]http://www.mathworks.com/, "Matlab Ver. 6.5, R13."

[14]J. Kittler and F. M. Alkoot, "Sum versus Vote Fusion in Multiple Classifier Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 110-115, Jan. 2003.

[15]I. Matthews, T. F. Cootes, J. A. Bangham, J. A. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 198-213, Feb. 2002.

[16]K. Messer, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: "The Extended M2VTS Database"," *The Proceedings of the Second International Conference on Audio and Video-based Biometric Person Authentication (AVBPA'99)*, Washington D.C., pp. 72-77, March 1999.

[17]A. N. Netravali and B. G. Haskell, "*Digital Pictures*": Plenum Press, 1998.

[18]G. Potamianos, H. Graf, and E. Cosatto, "An Image Transform Approach for HMM Based Automatic Lipreading," *Proceedings of the IEEE International Conference on Image Processing, ICIP 98*, Chicago, vol. 3, pp. 173-177, Oct. 1998.

[19]L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, Feb. 1989.

[20]D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-83, Jan. 1995.

[21]G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, pp. 31-44, April 1991.

[22]T. J. Wark, S. Sridharan, and V. Chandran, "The use of Speech and Lip Modalities for Robust Speaker Verification under Adverse Conditions," *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, Florence, Italy, vol. 1, pp. 812-816, June 1999.

[23]S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "*The HTK Book (for HTK Version 3.1)*". Cambridge University Engineering Department: Microsoft Corporation, 2001.