Amazigh Audiovisual Speech Recognition System Design

Ilham ADDARRAZI
Department of Mathematics and Computer Science
FSDM, USMBA
Fez, Morocco
ilham.adrz@gmail.com

Hassan SATORI
Department of Mathematics and Computer Science
FPN, UMP
Nador, Morocco
hssatori@gmail.com

Khalid SATORI
Department of Mathematics and Computer Science
FSDM, USMBA
Fez, Morocco
khalidsatori@gmail.com

***Abstract*— It is well known that speech recognition is a multimodal process which uses information not only from audio but also from vision. This paper describes our experience to design an audio visual speech recognition system, which relates the acoustic and the visual information in order to improve noise robustness of automatic speech recognition. The accuracy rate for face and mouth detection using Viola-Jones approach was satisfactory (reaches to 99% and 96.6% for face and mouth detection respectively).**

Keywords— Audio-visual recognition; Automatic Speech Recognition; lip reading; HMM

I. INTRODUCTION

Automatic Speech Recognition (ASR) is the technology that permits human to communicate with a computer interface using their voices. One of the biggest problems that stay in the ASR is noise robustness [1, 2]. Several methods have been developed in order to improve the speech recognition in a noisy environment. To overcome this limitation many works focus on the bimodal nature of speech perception in human, combining both the acoustic signal information and visual information. Indeed, this combination has been helpfully confirmed by the McGurk effect [3], and therefore the integration of visual information in the recognition system is a solution. In this sense, the movement of the mouth is considered as one of the speech recognition channels.

The first Audio-visual recognition system (AVRS) was reported in 1984 by Petajan et al. [4]. The authors exploited the width, height, area and perimeter of the mouth for building a successfully lip-reading system to aid their speech recognition system. Yau et al. [5] proposed an approach to adopt a visual speech model based on the viseme representation in a Moving Picture Experts Group 4 (MPEG-4) standard.

Too many AVRS systems are dedicated to English languages [6, 7, 8], French [9], Arabic [10] and Korean [11]. This is an exciting progress in AVSR.

The big challenge in AVSR is the combination of the acoustic and visual information; recent studies have proposed an algorithm for modeling the multi-modal data, as Hidden Markov Model (HMM) hybridized with the Genetic Algorithm (GA) [10] this algorithm is composed by the Baum–Welch algorithm, which gives a useful approximation of the probabilities of the HMM. Other approaches could be based on support vector machines [12, 13].

Due to the importance of visual cues, we integrate it, in this work, with our already designed Amazigh Speech Recognition System [14]. This fusion of the audio and visual streams makes an AVRS highly configured able to augment the level of speech Amazigh perception.

This paper describes our experience to design an AVSR system, combining an Amazigh speech recognition system based on the open source Sphinx-4 [15] and a visual model which is implemented in OpenCV [16]. Our system as far as we know, is the first audio-visual system uses the less-resourced Moroccan Amazigh language able to perform better in noisy environments.

The rest of this paper is organized as follows: section II describes the architecture of proposed audiovisual system. The face and mouth detection experimental results is presented in section III. The conclusion is drawn in section IV.

II. THE ARCHITECTURE OF PROPOSED AUDIOVISUAL SYSTEM

AVRS represent an essential branch in the human computer interaction domain. It is a new addition to speech recognition and has attracted much attention in the last few years. This system is recognition of speech using acoustic and visual features which are considered as one of the most hopeful solutions for reliable speech perception, principally when the audio is corrupted by noise [17]. It presents novel tasks compared to ordinary ASR.

The overall structure of the proposed system is depicted in Figure 1. Three main modules comprise the AVRS system:

- Visual recognition module, extracted from video input.
- Auditory recognition module, extracted from the audio input.
- Fusion of visual and acoustic modules.

978-1-5090-4062-9/17/$31.00 ©2017 IEEE

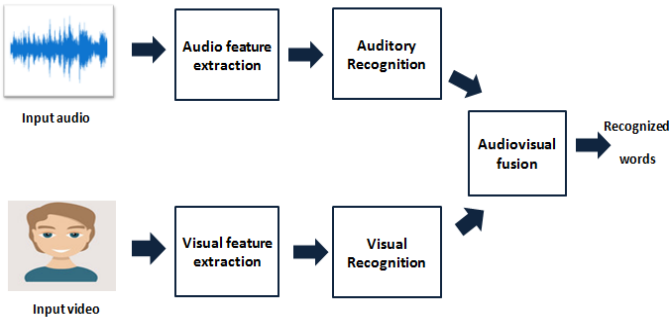

Fig. 1. Overview of the proposed audio-visual recognition system.

The system is usually takes as input the video, which is divided into audio file and image files. The first step is the extraction of the acoustic and visual features and then the fuse of these features.

A. *Visual feature*

The first necessary step in AVSR is the visual features extraction. These features are extracted from the mouth, which is considered as a region of interest (ROI) for our proposed system. The Figure 2 shows the procedure of visual feature extraction:

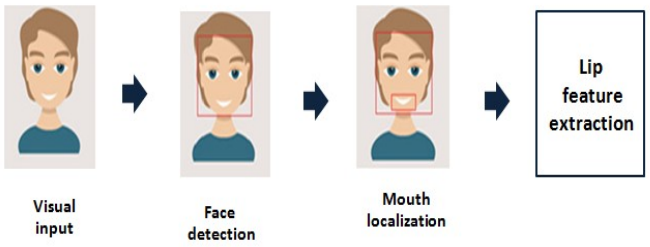

Fig. 2. Visual extraction processes

Face detection and mouth area localization are tracked from the video sequence. These elements are used to extract the visual features.

1) *Face detection*

Feature points like eye, nose, eyebrows and mouth, are used for many applications and they are detected within the face. To select our region of interest, which is the mouth, in each frame in a video sequence, the face is selected first.

Face detection is used in many applications that identify human faces in video sequence. For this purpose, many algorithms are presented in literature. The Viola-Jones algorithm [18] is used to realize robust detection. This approach was proposed in the year 2001 which is implemented in Opencv and based on 4 concepts:

- Integral image.
- Haar-like features.
- AdaBoost method.
- Cascade classifier.

This method allows analyze an image, without the need to study each their pixels. Indeed, the idea of the integral image is used as a quick method of calculating the sum of pixel values. This representation allows the Pseudo-Haar feature used by the detector to be computed and selected very quickly by boosting. The weak classifiers trained by AdaBoost are cascaded to a final strong classifier which is used to achieve the detection.

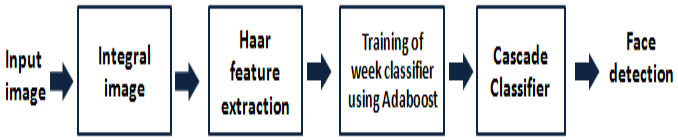

Fig. 3. The principle of the Viola and Jones algorithm

2) *Mouth localization*

Once the face is detected it is necessary to locate the mouth. The moth is the facial feature to have attracted the biggest attention, typically because it contains the most visual speech information in AVSR. This part is an important step for any speech recognition system.

The first work on the visual front-end is to extract a normalized ROI, in the form of a rectangle, revolved around the mouth of the speaker. The mouth region detection is carried out also using Viola-Jones approach.

This stage returns a rectangle around the detected mouth area. A simple output of face and mouth detection for an image from a speaker is shown in Figure 4.

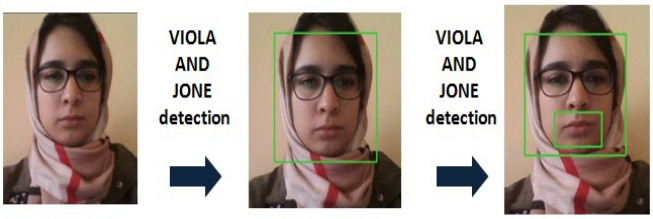

Fig. 4. Example result of Viola and Jones face and mouth detection

3) *Lip feature extraction*

In AVRS the concentration will be on the lips because they constitute a large part of our research. The feature extraction of lip contour is challenging to track due to its elastic shape and non-rigid motion.

Once the mouth region is found; there are many algorithms can be used to get lip contour feature which have their own strengths and weaknesses. The exiting techniques for feature extraction can be separated into three categories: (1) the visual shape which can be represented as a sequence of landmark points defined by some templates. (2) appearance-based visual features where features are directly obtained from pixels. (3) Hybrid approaches that are the combination of both (1) and (2).

a) Visual shape:

The shape based features can be divided into lip geometric features and lip model features methods:

- **Lip geometric features:** These uses the geometrical parameters such as height, width and area of lip contour [4].
- **Lip model features:** These uses a number of parametric models. Active shape models (ASM) [19], snakes [20], AAM [21] or deformable template [22] are well known. The result of these approaches is an ensemble of points (point distribution model-PDM) which defines lips form.

  The Figure 5 shows the lip tracking using ASM which is connected with a set of techniques used to extract the lip contour in an image/video. The basic of this method was present by Cootes & Tayor in 1995 and has been applied in numerous tasks [19].

Fig. 5. Lip contour tracking using ASM

b) Appearance-based visual features:

The appearance-based visual features include image transformation based feature extraction methods such as Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA). These algorithms are usually applied to dimensionality reduction and to eliminate redundant data. Therefore they are based on appearance and feature analysis to determine the Lip reading.

- **DCT features of mouth region:**

The extraction of the visual features is executed with the DCT. The DCT is similar to the discrete Fourier transform (DFT). It is an orthogonal transformation which transforms the image from the spatial domain to the frequency domain [23].

The 2D DCT of a given function F (u, v) of size M*N is defined by the following equation:

$$F(u,v) = a_u a_v \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cos\frac{(2x+1)u\pi}{2N} . \cos\frac{(2y+1)v\pi}{2N} \quad (1)$$

Where f(x, y) is the intensity of the pixel in row x and column y, and:

$$a_u, a_v = \begin{cases} \sqrt{\frac{1}{2}}, & u,v = 0 \\ 1, & u,v = 1,\dots N \end{cases} \quad (2)$$

DCT produces a matrix of features which has the same dimension of the input mouth image; this matrix is saved as lips features.

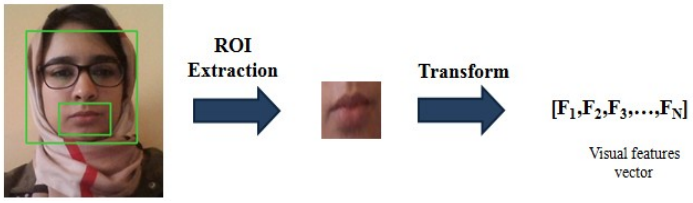

Fig. 6. The process of visual feature extraction

The result after feature extraction is based on classifier. The features from the DCT coefficients are classified using different classifiers as KNN, Neural network, SVM and HMM.

The classifiers consist of two phases: training phase and testing phase. In the training phase, features are extracted from the lip area by DCT and they are given as inputs to approximate the parameters of the classifier. Then, the word is recognized in the testing phase.

B. Audio feature

In our case, a Moroccan Amazigh speech is used to interact with the visual model.

1) AMAZIGH language

The Amazigh language is spoken by large populations in North Africa; particularly 28% of Moroccan population uses it partitioned into three main regional varieties, depending on the area and the communities: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South West and the High Atlas.

In view of the importance of the Amazigh language in Morocco, the IRCAM has developed in 2003 an official graphic system which is called Tifinaghe-IRCAM, for writing Amazigh. This system contains: [24]

- 27 consonants including: the labials (ⵀ, ⵃ, ⴼ), the dentals (ⵜ, ⵟ, ⴷ, ⴹ, ⵏ, ⵍ, ⵔ, ⵕ), the alveolars (ⵙ, ⵚ, ⵣ, ⵛ), the palatals (ⵊ, ⵉ), the velar (ⴽ, ⵅ), the labiovelars (ⴽⵯ, ⵅⵯ), the uvulars (ⵇ, ⵅ, ⵖ), the pharyngeals (ⵄ, ⵃ) and the laryngeal (ⴰ);
- 2 semi-consonants: ⵢ and ⵡ;
- 4 vowels: three full vowels ⴰ, ⵉ, ⵓ and neutral vowel (or schwa) ⴻ which has a rather special status in Amazigh phonology.

2) AMAZIGH Speech Recognition:

The speech recognition system is constructed in our previous works [14]. The designed system is based on the CMU Sphinx tools and on hidden Markov model. It transfers the human speech from an audio signal to the text. The figure 7 shows the overall architecture of the ASR system which comprises of four components: [25]

- **Feature extraction**: extracts speech features which play an important role in ASR system performance.
- **Decoder**: joins the extracted features with data from the knowledge base, and performs a search to decide the sequences of words that could be represented by a series of features.

- **Acoustic models**: maps the observed features of phonemes provided by the front-end of the system and the HMMs.
- **Language models:** Contains a representation of the probability of occurrence of words.

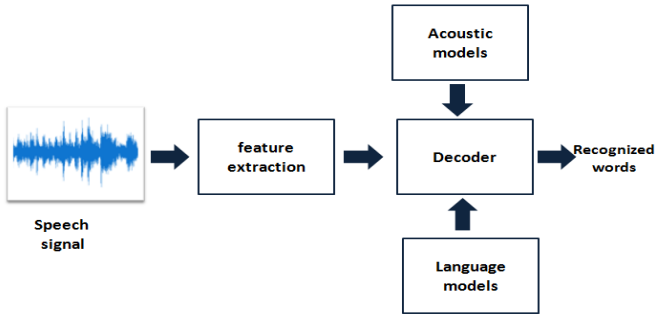

Fig. 7. ASR system architecture

The extraction of the features of acoustic cues is an important task to make enhanced recognition performance. These features, in our system, are extracted using the Mel-Frequency Cepstrum (MFCC) vectors.

The auditory parameters used in our designed speech recognition system (e.g. number of MFCC, number of Gaussian mixture, number of HMM, speech signal characteristics) were described in [14]. The acoustic model is prepared using the Amazigh phonetic properties of ten first digits, as shown in table 1, in order to have maximum performance.

TABLE I. TEN AMAZIGH FIRST DIGITS AND THEIR TRANSCRIPTION IN ENGLISH AND ARABIC

English transcription	Arabic transcription	Digits	Tifinaghe transcription	Syllables
AMYA	اميا	0	ⵄⵎⵢⴰ	VC-CV
YEN	يان	1	ⵢⴰⵏ	CVC
SIN	سين	2	ⵙⵉⵏ	CVC
KRAD	كراض	3	ⴽⵔⴰⴷ	VC-CVC
KOZ	كوز	4	ⴽ ⴽ ⵓ ⵣ	CVC
SMMUS	سموس	5	ⵙⵎ ⵎⵓⵙ	CCV-VC
SDES	سضيس	6	ⵙⴹ ⵉ ⵙ	CCVC
SA	سا	7	ⵙⴰ	CV
TAM	تام	8	ⵜ ⴰ ⵎ	CVC
TZA	تزا	9	ⵜ ⵣ ⴰ	CC-CV

C. *Audiovisual fusion*

The aim of the audio visual recognition system is the combination of audio and video cues in order to improve the performance of the speech perception when the noise is present. The integration of two cues is still an open problem.

The fusion can be classified into two different parts: feature fusion (or early integration) and decision fusion (or late integration) [26].

The features fusion, as shown in figure 8, is realized when the audio and vision feature vectors are concatenated before being presented to the classifier. This method combines the two modalities as a single classifier. Thus the features from acoustic and visual sources are joined in one feature vector. The recognition process is applied on the obtained and combined feature vector.

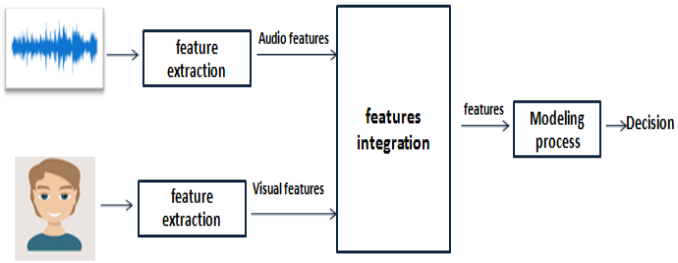

Fig. 8. Early integration

As for the decision fusion algorithms classify each modality (audio-only and visual-only) independently and then their fusion is achieved at the time of the decision.

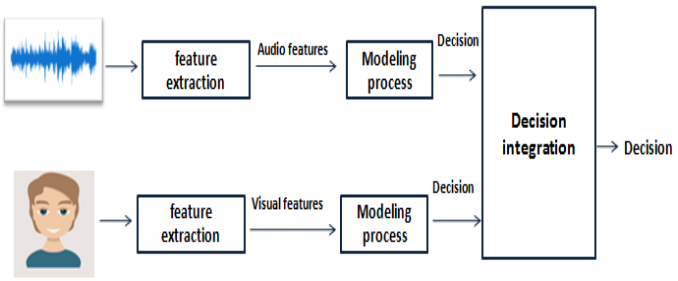

Fig. 9. Late integration

Due to the need of separate modeling for each modality, the decision fusion is more challenging compared to early integration. In our proposed system, late integration strategy is applied.

There are a number of techniques which have been used for the fusion steps, such as SVMs, graphical models(DBNs and HMMs), neural networks, and estimation algorithms, for example, Kalman filtering.

III. FACE AND MOUTH DETECTION EXPERIMENTAL RESULTS

In this part of the paper we present the experimental results of evaluating the face and mouth detection using an Amazigh audiovisual database.

A. *Audiovisual Database Description*

Every audio-visual database encloses audio and visual inputs that are obtained jointly while a speaker is talking.

The database used is an Amazigh audiovisual which includes ten Amazigh digits (see table II) for recognition. The speakers are 10 females and 10 males. Each of them pronounces 10 times each digit. All the videos are registered at 25 images per second with a resolution of 1280*720 and a sampling rate of 16 kHz for audio.

TABLE II. DATABASE PARAMETERS

	Female	Male
Number of speakers	10	10
Number of word repetition	10	10
Number of token	1000	1000

The frames are segmented manually before 0.2s from starting of each digit to produce isolated digit models.

The images of our database include the frontal face and the high part of the body with a relatively uniform background (see figure 4).

B. Experiment result

As mentioned above, the first stages in the visual front-end for our proposed system is face and mouth detection. To test the performance of face and mouth detections the Viola-Jones algorithm is used. Different images are selected from different speakers and their face and mouth detection rates are registered as presented in table III.

TABLE III. FACE AND MOUTH DETECTION

Visual evaluation	Detection rate (%)
Face detection	99
Mouth detection	96.6

IV. CONCLUSIONS

This paper presents the design of an Amazigh audiovisual speech recognition system that integrates the visual information with the speech recognition system in order to enhance the performance of the system in noisy environment. The description of our techniques to realize our system was presented. The accuracy rate for the face and moth detection was highly satisfactory.

Our next work will concern the integration of the modules of our system and test it in a normal environment using an Amazigh audio visual database.

REFERENCES

[1] B. H. Juang, " Speech recognition in adverse environments," Computer Speech Lang., vol. 5, pp. 275–294, 1991.

[2] J.Li, L. Deng, Y Gong and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745-777, 2014.

[3] H. McGurk and J. MacDonald, " Hearing lips and seeing voices," Nature, vol. 264, pp. 746-748, 1976.

[4] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in Proc, Global Telecomm, Conf, Atlanta, GA, pp. 265–272, 1984.

[5] W. Yau, D. Kumar and A. Poosapadi , "Recognition of speech consonants using facial movement features," Integrated Computer-Aided Engineering, vol. 14, no. 1, pp. 49-61, 2007.

[6] G. Fanelli, J. Gall, , and L. J Van Gool, "Hough Transform-based Mouth Localization for Audio-visual Speech Recognition," In BMVC , pp. 1-11, September 2009.

[7] E. K. Patterson, S. Gurbuz, Z.Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference,Vol. 2, pp. II-2017, May 2002.

[8] S. Richie, C. Warburton and M. Carter, "Audiovisual database of spoken American English," Linguistic Data Consortium,2009.

[9] S. Foucher, F. Laliberte, G. Boulianne and L. Gagnon, " A dempster-shafer based fusion approach for audio-visual speech recognition with application to large vocabulary french speech," In IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1, pp. I-I, 2006.

[10] A. Makhlouf, L. Lazli and B. Bensaker, "Evolutionary structure of hidden Markov models for audio-visual Arabic speech recognition," Int. J. Signal and Imaging Systems Engineering, vol. 9, no. 1, pp. 55-66, 2016.

[11] J. Shin, J. Lee and D. Kim, "Real-time lip reading system for isolated Korean word recognition, Pattern Recognition", vol. 44, no 3, pp. 559-571, 2011.

[12] M. Gordan, C. Kotropoulos and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," EURASIP J, Appl, Signal, process, vol. 2002, no 1, pp. 1248-1259, 2002.

[13] S. Pachoud, S. Gong, and A. Cavallaro, "Space-time audio-visual speech recognition with multiple multi-class probabilistic Support Vector Machines," International Conference on Audio-Visual Speech Processing, pp. 155-160, 2009.

[14] H. Satori, F. ElHaoussi, "Investigation Amazigh speech recognition using CMU tools," International Journal of Speech Technology, vol. 17, no 3, p. 235-243, 2014.

[15] CMU Sphinx Open Source Speech Recognition Engines. (2016). Retrieved November 05, 2016, from http://cmusphinx.sourceforge.net/wiki/download.

[16] Opencv. (2015).Retrieved October 20, 2015, from http://opencv.org/downloads.html

[17] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE transactions on multimedia, vol. 2, no 3, pp. 141-151, 2000.

[18] Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features," In Computer Vision and Pattern Recognition, 2001. CVPR 2001, Proceedings of the 2001 IEEE Computer Society Conference on. IEEE, vol. 1, pp. I-511-I-518, 2001.

[19] T. F. Cootes, C. J Taylor, D. H. Cooper and J. Graham, "Active shape models-their training and application," Computer vision and image understanding, vol. 61, no 1, pp. 38-59, 1995.

[20] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models, International journal of computer vision," vol. 1, no 4, pp. 321-331, 1988.

[21] G. J. Edwards, C. J. Taylor and T. F. Cootes, "Interpreting Face Images Using Active Appearance Models," in Proceedings of the 3rd, International Conference on Face & Gesture Recognition, pp. 300-305, 1998.

[22] A. K. Jain, Y. Zhong and S. Lakshmanan, "Object matching using deformable templates," IEEE Transactions on pattern analysis and machine intelligence, vol. 18, no 3, pp. 267-278, 1996.

[23] N. Ahmed, T. Natarajan, And K.R. Rao, "Discrete cosine transform," IEEE transactions on Computer, vol. 100, no 1, pp. 90-93, 1974.

[24] A. A. Fadoua, And B. Siham, "Natural language processing for Amazigh language: Challenges and future directions," Language Technology for Normalisation of Less-Resourced Languages, vol. 19, 2012.

[25] H. Satori, M.Harti and N. Chenfour, "Introduction to Arabic speech recognition using CMUSphinx system". arXiv preprint arXiv:0704.2083, 2007.

[26] A. Adjoudani and C. Benoit, "on the integration of auditory and visual parameter s in an HMM-based ASR," In Speechreading by humans and machines, Springer Berlin Heidelberg, pp. 461-471, 1996.