Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

The Hmm Based Amazigh Digits Audiovisual Speech Recognition System

Ilham Addarrazi [1*], Ouissam Zealouk [1], Hassan Satori [1], Khalid Satori[1]

[1] Laboratory Computer Science, Image processing and Numerical Analysis, Faculty of Sciences Dhar Mahraz (FSDM), Sidi Mohammed Ben Abbdallah University,B.P. 1796, Fez, Moroccoe-mail

* ilham.addarrazi@usmba.ac.ma

Article Info
Page Number: 2261-2278
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract
In this paper, we present an Amazigh audio-visual speech recognition system that combines the information coming from the audio and visual modalities. The proposed system is considered, as far as we know, the first audio-visual system that uses Amazigh language. We develop each subsystem in different platforms. In order to building a visual subsystem, we extract the features from the region of the mouth using DCT to be modeled using Hidden Markov Models (HMM).Whereas, the audio subsystem is based on the Carnegie Mellon University Sphinx tools based on HMM. The two sub-systems use the AmDigit_AVSR (Amazigh Digit _ Audio-visual Speech Recognition System) database. The combined system obtained best performances of 93,99 % using "OR" based-rules. Our experiments show that the combination of the visual and acoustic information improves the performance of speech
Keywords: Speech recognition system, lip-reading, multimodal integration, late integration.

Introduction

Automatic Speech Recognition (ASR) system is a technology that allows humans to use their voices to communicate with a computer interface. This system aims to recognize the spoken words by transforming the acoustic cues into a sequence of words. A powerlessness of most ASR systems is their failure to perceive well when the signal is corrupted by noise. Hence, one of the biggest problems that remain in the ASR is noise robustness.

Naturally, the speech perception of human is a bimodal process. The vision of the speaker's mouth region yields extensive information to speech recognition in order to determine what has been spoken especially when noise is present. Thus, audio-visual speech recognition (AVSR) system aims to assume the bimodal nature of human speech perception by combining the audio information with the visual cues available from talker's mouth region (lip-reading). The purpose of this kind of system is to improve the performance of speech recognition in noisy environment.

The integration of lips shape information of the speaker seems to be helpful for speech recognition issues in noisy environment (McGurk effect) [1]. The information from visual modality is considered as one of the speech recognition channels. Thus, the audio-visual speech recognition (AVSR) is a field that associates the disciplines of image /speech

processing and integration of multimodal information. One of the most difficult and challenging issues facing the AVSR is how to combine the acoustic and the visual modalities. Thus, three strategies are proposed in literature to the integration problems: Early, intermediate and late integration [2].

Therefore, in the present study, we propose an alternative solution to recognize the Amazigh speech by combing the speech recognition system based on the open source Sphinx-4, and a visual model which is implemented in OpenCV using late integration strategy. Thus, the characteristics of the Amazigh language are studied to extract the best parameters adapted to it. For this purpose, we have used an in house Amazigh audio-visual database named AmDigit_AVSR [2] which is, as far as we know, the first audio-visual database uses the Amazigh language.

Our contributions are summarized as follows:

- Proposed a stochastic model that allows visual recognition of a digit through the selection of HMM parameters calculated from visual cue.
- Proposed a stochastic model that allows speech recognition of a digit through the selection of HMM parameters calculated from acoustic cue.
- Proposed a strategy to combine the decisions of the visual and speech recognition system.
- Built and created, manually, an audio-visual database to train and test our model.
- Proposed an open source platform to test the performance of our model.

This paper is organized as follows: Section 2 describes the review of some related works. The overview of the proposed system is presented in section 3. The Section 4 demonstrates the evaluation and discussion of the performance of the proposed. The conclusions and future works are presented in the final section.

Related works

The AVSR systems have appear in recent years as an active filed. A number of researches concentrated on designing the audio-visual speech recognition system to improve the performance of speech research community. A number of systems are developed so far in literature.

The first audio-visual recognition system was proposed in 1984 by Petajan [4]. The authors use the geometric information (such as width, height, area and perimeter) of the mouth for construction a lip-reading system to improve the performance of speech recognition system.

The AVSR system prposed by [5] used Zernike moments for computing visual features and and mel frequency cepstral coefficients for audio features. They applied thoses techniques on visual vocabulary of independent standard words dataset which contains collection of isolated set of city names of ten speakers. The performance of recognition of isolated words based on visual only and audio only features results in 63.88 and 100 % respectively.

Makhlouf et al. [6] combine the visual modality based on DCT features and the acoustic modality based on RASTA-PLP features. The authors present an algorithem based on Hidden Markov Model (HMM) hybridized with the Genetic Algorithm (GA) to modelling the multimodal data. They use two databases, the CUAVE database and their AVARB database. The experiment show that the results of the system that trained using the GA/HMM give a higher rate recognition compared to the HMM trained by the Baum–Welch algorithm.

The authors in [7] present an active appearance model (AAM) based multiple-camera AVSR system. The visual information is extracted from jaw and lip region to enhance the performance in car environments. They use a Four cameras in car audio-visual corpus is used to perform the experiments. The four visual streams are fused using a synchronous hidden Markov model to be combined with a single stream acoustic HMM to build five-stream AVSR.

Rahmani et al. [8] introduce an AVSR that combine the acoustic MFCC features with the visual DBNF features using the DNN and HMM based on early integration startegy. Their experiment give a 97,9% in Clean Audio and a 83.3% using 0 dB SNR.

Amazigh language

Before the creation of the proposed system, a good knowledge of the language is recommended. It's evident that the Berber language (Amazigh) is interpreted as one of the first languages of worlds. Presently, this language is used by the North Africa population. Among 50 % of Morroco's population speaks this language which is separted into three main regional dialects: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South West and the High Atlas. Thus, in 2003, the Royal Institute of the Amazigh Culture (IRCAM) present an official system named Tifinaghe-IRCAM, for writing Amazigh.

Each spokean language has thier own phonemes and visemes. These units are the most important composants that the AVSR system is based. The number of phonemes and visemes varies from a language to another (e.g For English there are 40 to 45 different phonemes and 11 to 14 visemes. In Dutch there are 42 different phonemes and 16 visemes).

The Tifinaghe-IRCAM system involve: [9]

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

The approved syllables in Amazigh language are: V, CV, VC, CVC, C, CC and CCC where V indicates a vowel while C indicates a consonant. Table 1 presents the ten Amazigh digits used in this proposed system, with their English, Arabic and Amazigh transcription scripts, and the type of syllable.

Table 1. The English, Arabic and Tifinagh transcription for 10 Amazigh digits

English transcription	Arabic transcription	Digits	Tifinaghe transcription	Syllables
AMYA	اميا	0	ⴰⵎⵢⴰ	VC-CV
YEN	يان	1	ⵢⴰⵏ	CVC
SIN	سين	2	ⵙⵉⵏ	CVC
KRAD	كراض	3	ⴽⵔⴰⴷ	VC-CVC
KOZ	كوز	4	ⴽⴽⵓⵣ	CVC
SMMUS	سموس	5	ⵙⵎⵎⵓⵙ	CCV-VC
SDES	سضيس	6	ⵙⴷⵉⵙ	CCVC
SA	سا	7	ⵙⴰ	CV
TAM	تام	8	ⵜⴰⵎ	CVC
TZA	تزا	9	ⵜⵣⴰ	CC-CV

The proposed system architecture

Audio-visual speech recognition (AVSR) system is a technique that interprets the speech using the audio stream with the visual information. It has emerged in recent years as an active filed, used by a vastly wider range of researchers in diverse applications. As illustrated in fig.1, the proposed AVSR consist of three modules: The visual recognition module (Lip-reading system), the acoustic recognition module (speech recognition system) and the modalities fusion.

However, as a first step in acoustic module, the features are extracted from the audio input to be recognized using Cmu-Sphinx. As for the visual module, the features are extracted from the speaker's mouth region, to be recognized using HMM. In the final module, the decisions of two previous modules are integrated to make a final decision for recognize the spoken words.

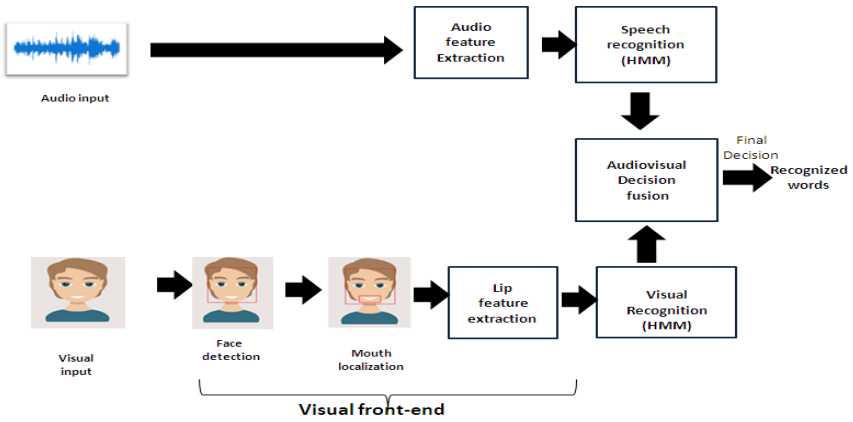

Figure 1. Overview of the proposed audiovisual recognition system

Visual speech recognition sub-system

The first principal issue in AVSR systems is the design of visual front-end. As illustrated in fig.1, the visual-front end consists of three major steps: Face detection, mouth region localization and mouth features extraction:

a. Face and mouth detection

It is obvious that the face is the part of the body that contains the human communication elements (e.g the mouth provide speech, eye and nose produces information for face expressions, etc). The fact that the mouth includes the most visual speech information in AVSR, the concentrate will be on this facial part. İn order to select the mouth region called also region of interest (ROI), the face is selected first. To realize the face detection, the viola-jones algorithm [10] is used. The choice of this method is based on its simplicity and its quickness to give a high detection rate. This method provides a new representation called integral image to compute Haar-features locates the talker's face in the input image by generating a new representation named integral image. This representation aims to calculate speedily Haar features to train and select these features by Adaboost. To achieve the detection, the weak classifiers trained are cascaded to a strong classifier. As shown in fig.2, the result of Viola-Jone algorithm is a rectangle revolved around the region of detected face. Regarding the mouth region detection, the Viola-Jones algorithm is applied too. This methods return a rectangle fixed on the region of detected mouth in each face detected.

b. ROI features extraction

Once the region of the mouth is detected, it is necessary to extract the useful information. For this purpose, three main methods categories are proposed in literature [11]: (a) the appearance-based approaches where features are directly extracted from the image pixels; (b) the geometric-based where the features obtain the form of the mouth shape, height, width, and area; (c) Hybrid approaches that the mixture of (a) and (b). In this work, the Discrete Cosine Transform (DCT) technique is used as an appearance-based approach to extract features from the speaker's mouth region [12]. This method composes of two steps: (a) the extraction of DCT coefficients from the entire ROI image; (b) the DCT coefficient selection to obtain the

mouth region feature vectors. As a first step in DCT feature extraction, the coefficients are calculated as follows:

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

Where:

- M and N are the dimensions of the image.
- F(u, v) Are the DCT coefficients.
- f(x, y) is the intensity of the pixel in row x and column y.
- a_u and a_v are defined as follows:

$$a_u = \begin{cases} \sqrt{\frac{2}{M}}, 1 \le u \le M - 1^{\frac{1}{\sqrt{M}}, u=0} \end{cases} \quad (2)$$

$$a_v = \begin{cases} \sqrt{\frac{2}{N}}, 1 \le v \le N - 1^{\frac{1}{\sqrt{N}}, v=0} \end{cases} \quad (3)$$

DCT produces a coefficients matrix that has the same dimension of the mouth region input image. The upper left corner of this matrix contains the high coefficient and the bottom right contains the low coefficient. Then, those coefficients are extracted in a zigzag scanning pattern to be saved into a vector. The purpose of using this way of scanning is to place the high coefficients in top of vector. In order to minimize the calculations in the recognition step, Only 100 coefficients are used to characterize the mouth region image.

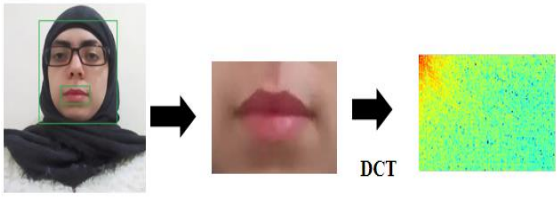

Figure 2. The reuslt of applying DCT on region mouth image

c. Visual modality recognition

- **The stochastic system**: Hidden Markov Model is a stochastic signal modeling system. It is considered as a continuation of Markov process. The HMM comprises of a set of a connected where, these states can be connected to each other by transitions annotated by probabilities. The HMM observations are presented as a set of the observed process output, where, each of them can be emitted by each state. Precisely, a HMM can be formulated as:

$$\lambda = (N, M, \Pi, A, B) \quad (4)$$

Where:

- N is a finite set of hidden states.
- M is a finite set of observed states.
- Π is probability of starting in hidden states, where:

$$\sum_{i=1}^{N} \Pi_i = 1 \quad (5)$$

- A: Matrix of transitions probabilities from one state to another, with A→ $\{a_{ij}\}$ N*N and a_{ij} represents the probability of transitioning from state i to j, where:

$$a_{ij} = p(X_t = j | X_{t-1} = i) \quad (6)$$

B: Matrix of probabilities of observations, with B →$\{b_{ij}\}$ N*M where b_{ij} represents the probability of state i in N emitting observation j in M.

Visual subsysem modeling: During the speaking, the duration of each mouth's shape changes. To model this variation, the HMM model is efficient. Once the features are extracted, a HMM model for each digit is built. Baum-Welch algorithm is used for training the construction of each HMM models. Hence, teen HMMs are trained to recognize the first teen digit. The recognition is based on observation sequence O obtained from the extracted feature vectors. Concerning the testing, the probability of an observed sequence of the digit is evaluated. In order to recognize an unknown digit 10 probability values are calculated using Viterbi algorithm to choose the HMM that has the elevated probability value to accept it as the recognition result (see fig.3)

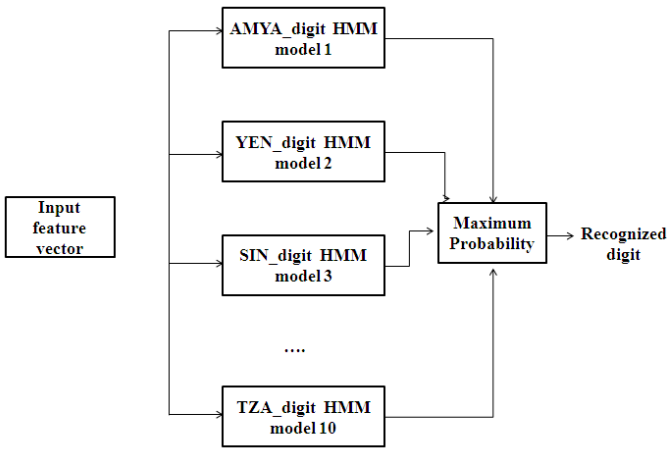

Figure 3. Ten HMM models used for digit visual speech recognition

Audio speech recognition sub-system

In this work, the creation of the speech recognition system is based on the open source sphinx-4, from the Carnegie Mellon University. It's an open source platform built completely in the Java TM programming language. The Sphinx-4 framework has been developed with a high level of flexibility based on hidden Markov models (HMMs).

CMU-Sphinx is a collection of the libraries used for development and realization of speech recognition applications. It consists of a set of packages helpful for different tasks and applications [13] [14] [15]:

- **Pocketsphinx**: Lightweight library of written recognition in C.
- **Sphinxbase**: support for libraries required by Pocketsphinx.
- **Sphinx4**: decoder for voice recognition search written in Java.
- **CMUclmtk**: Language model tools.
- **Sphinxtrain**: Acoustic model drive tool.
- **Sphinx3**: decoder for voice recognition search written in C

The fig.4 shows the overall ASR system architecture based on cmu-sphinx.

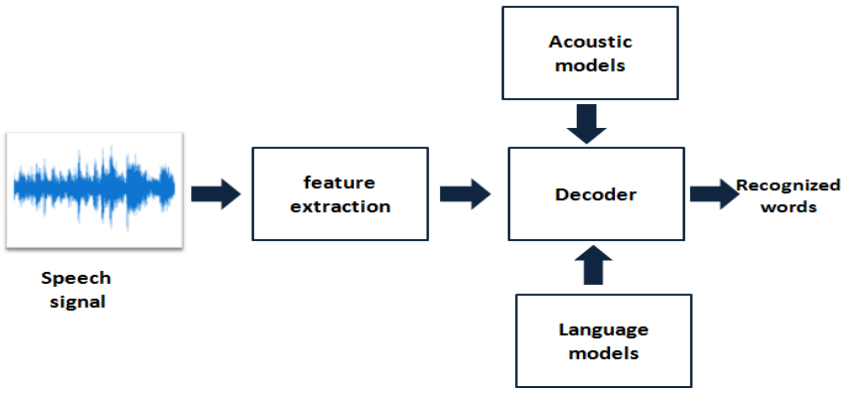

Figure 4. ASR system architecture

a. Features extraction

The first stage in any speech recognition system is to extract features. The aim of this module is to characterize a recorded signal (input) into a set of acoustic parameters. These features play an important part in speech decoding. The realization of this parameterization requires a set of communicating blocks called DataProcessors. Where, each of these blocks reads data from the predecessor as its input and executes it to carry out the output. The DataProcessors produce a data object collected of parameterized signals, called features, to be used by the decoder. The reason for features extraction creates a link between audio cues and statistical models in speech recognition system. Mel-frequency cepstral coefficients (MFCC) are the most widely used methods to extract acoustic features.

b. Acoustic model

Acoustic modeling describes the mapping between the observed features of phonemes (basic speech unit) produced by the front-end (features extraction component) of the system and the HMMs. This mapping considers the context and the position of the word. A HMM is a statistical model used for defining probability distributions over sequences of observations. A HMM encloses of two stochastic processes, the hidden states (invisible process) and observable events (visible process). HMMs models are trained with the forward-backward or Baum-Welch algorithm. In speech recognition, each phoneme has its own HMM. Then, the ASR system aims to identify the sequence of phonemes (represented by states) that correspond the sound pronounced (represented by observations). The build of the acoustic model is realized by the sphinx train. This model is based on a set of configuration files:

- Phone file: It contains the list of phones used in the dictionary with the phone SIL to represent a silence should have one phone.

- Training and testing fileid files: which are the text files include the list the recordings audio files paths (without audio file extension) used for training and testing.

- Training and testing transaction files: which are text files include the transcription for each audio file used for training and testing. Each line of the file starts with <s> and ends with </s> followed by an id in parentheses.

- Pronunciation dictionary (named also lexicon): it's a file text that contains the ten first Amazigh digits used for training following by their pronunciation. The fig.5 illustrates the dictionary file used in the training of our system:

AMYA	A M Y A
YEN	Y A N
YEN(1)	Y E N
SIN	S I N
KRAD	K R A D
KOZ	K O Z
SMMUS	S M U S
SDES	S D E S
SDES(1)	S D E S S
SA	S A
TAM	T A M
TAM(1)	T T A M
TZA	T Z A

Figure 5. Extract of the dictionary file used for speech recognition module of the proposed system

c. Language models

The Language Model is used to find the probability of each given word the observed word. This means attributing the likelihood of the nth word, using the n − 1 attributing words:

$$P(w_n|w_1,w_2,w_3...w_{n-1}) \quad (7)$$

Two type of langage model are known:Grammar Based Language Models (GBLM) and Probabilistic Language Models (PLM).

d. Decoder

The main import part in the speech recognition system is the decoder. It decides which sequence of words that could be represented the features by using the language models and acoustic model. This component observes the spoken sounds and generates the HMM correspond in the Acoustic Model.

The audiovisual integration

The integration of audio and video modalities is an open issue. These Two modalities treat different information coming from different source of the same scene. To assume the bimodal nature of the speech the integration of these two modalities is required. Three main integration strategies have been presented in the literature: early integration (or features fusion), intermediate integration (or classifier fusion) and late integration (or decision fusion). Due to the speech information diversify (audio and visual streams), the late integration strategy is used for fusion the audio and visual modalities. Moreover, the decision fusion model yields a procedure for modeling the reliability of audio and visual modalities. Several late integration fusion methods have been proposed [2]. In this present works, the decision fusion is made using "AND", "OR" rules.We presume that the output of visual recognition system is D_V and the output of speech recognition system is D_A. The values of D_V and D_A are:

$$D_A, D_v = \begin{cases} 1\ (\text{yes}), & \text{if the digit is recognized} \\ 0\ (\text{NO}), & \text{if the digit is not recognized} \end{cases} \quad (8)$$

Furthermore, we suppose that the output of the decision fusion is DAV. Consequently, the two fusion rules are explained as follows:

- **Using the "AND" rule**: the final decision will be passed only when all the two subsystem's outputs are equal to 1. Hence the "AND" rule can be illustrated as:

$$D_{AV} = \begin{cases} 1, & if\ \ D_A = 1\ \boldsymbol{AND}\ D_V = 1 \\ 0, & Otherwise \end{cases} \quad (9)$$

- **Using the "OR" rule:** the final decision will be passed only when one of the two system's output is equal to 1. The "OR" rule is stated as:

$$D_{AV} = \begin{cases} 1, & if\ \ D_A = 1\ \boldsymbol{OR}\ \ D_V = 1 \\ 0, & if\ \ \ D_A = 0\ \boldsymbol{AND}\ \ D_V = 0 \end{cases} \quad (10)$$

Results and Discussion

In order to test the performance of our proposed audiovisual speech recognition system, we use the AmDigit_AVSR (Amazigh Digit _ Audiovisual Speech Recognition System) database [3]. It is a first audiovisual corpus that uses Amazigh language. This database contains over 4000 video and audio files of the first ten Amazigh isolated digits uttered by 40 speakers (20 female and 20 male). All parameters of AmDigit_AVSR are presented in Table 2. All the videos of the used database are registered at 25 images per second with a resolution of 1280*720 and a sampling rate of 16 kHz for audio. Moreover, the images AmDigit_AVSR database comprise the frontal face and the high part of the speaker's body with a relatively simple background.

Table2. The AmDigit_AVSR database parameter

	Parameter	**Value**
Audio	Sampling rate	16 KHz
	Audio data file format	.wav
Video	Video data file format	.avi
	Resolution	1280*720
	images per second	25
Speakers	Speakers' age	18 and 45 years-old
	Speakers' gender	20 females, 20 males
Corpus	Number of repetitions per word	10
	Condition of environment	normal

The results

The proposed AVSR system for Amazigh language is a result on integration the two separated systems (visual recognition system and speech recognition system). Each of this sub-system is developed and created independently in different platforms using different techniques and algorithms, based on same database (AmDigit_AVSR corpus). The visual modality extracts the DCT features from each detected mouth image. Left-to-right HMM models with 3 states are used to model each viseme. Further, the construction of each model HMM for each digit is, simply, done by concatenating the appropriate viseme models. The components of this

modality are implemented using OpenCV libraries. In classification, result of each HMM is compared and maximum probability HMM digit is selected as result. Around 700 video files are used for training and 300 audio files for testing. Concerning the audio modality (speech recognition system), 700 audio files from AmDigit_AVSR database are used for training, where 300 audio files are used for testing. Based on the previous works [13] [14], the used ASR system uses 32 Gaussian mixture models (GMM) and 5 states per HMM to train each digit.

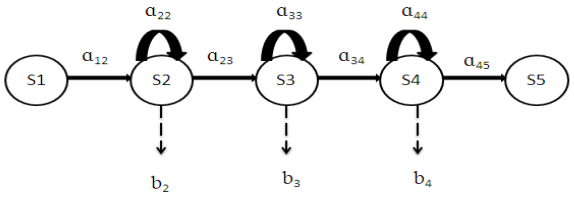

Figure 6. The 5-states HMM model for ASR system

The performance of the visual speech recognition system using HMM is presented in table 3, in the form of the confusion matrix. Moreover, the table 4 represents the performances of the speech recognition system. For the performance of the combined audiovisual speech recognition system using ‘’AND’’, “OR” rules-based is presented, respectively, in table 5 and table 6. The total recognition rate for each digit, in each system, is described in table 7.

Table 3. Confusion matrix of VSR subsystem

		Predicted Digits										
		AMYA	**YEN**	**SIN**	**KRAD**	**KOZ**	**SMMUS**	**SDES**	**SA**	**TAM**	**TZA**	**Omitted**
Actual Digits	**AMYA**	**27**								2		1
	YEN		**16**	7				5				2
	SIN		9	**17**				3				1
	KRAD				**26**						1	3
	KOZ					**23**	3					4
	SMMUS					4	**24**					2
	SDES		4	2				**21**				3
	SA	8							**13**	7		2
	TAM	2							6	**20**		2
	TZA					3			6		**18**	3

Table 4. Confusion matrix of ASR subsystem

	Predicted Digits										
	AMYA	**YEN**	**SIN**	**KRAD**	**KOZ**	**SMMUS**	**SDES**	**SA**	**TAM**	**TZA**	**Omitted**

Actual Digits	**AMYA**	**28**	1									1
	YEN		**27**	1								2
	SIN		2	**28**								1
	KRAD				**29**							1
	KOZ				1	**26**					1	2
	SMMUS						**27**	2				1
	SDES						2	**28**				0
	SA			1				2	**25**			2
	TAM	2								**27**		1
	TZA					1					**26**	3

Table 5. Confusion matrix of AVSR system using "AND" based-rule

		Predicted Digits										
		AMYA	**YEN**	**SIN**	**KRAD**	**KOZ**	**SMMUS**	**SDES**	**SA**	**TAM**	**TZA**	**Omitted**
Actual Digits	**AMYA**	**27**										3
	YEN		**14**									16
	SIN			**17**								13
	KRAD				**25**							5
	KOZ					**22**						8
	SMMUS						**23**					7
	SDES							**20**				10
	SA								**12**			18
	TAM									**18**		12
	TZA										**17**	13

Table 6. Confusion matrix of AVSR system using "OR" based-rule

		Predicted Digits										
		AMYA	**YEN**	**SIN**	**KRAD**	**KOZ**	**SMMUS**	**SDES**	**SA**	**TAM**	**TZA**	**Omitted**
	AMYA	**29**										1
	YEN		**28**									2
	SIN			**29**								1
	KRAD				**30**							0
	KOZ					**27**						3

Actual Digits	**SMMUS**						**28**					2
	SDES							**29**				1
	SA								**26**			4
	TAM									**29**		1
	TZA										**27**	3

Table 7. Recognition rate of each digit in visual-only, audio-only, audiovisual using "AND" rule and audiovisual using "OR" systems

Digits	**Rate recognition (%)**			
	visual-only	**Audio-only**	**Audiovisual using "AND"**	**Audiovisual using "OR"**
AMYA	90,00	93,33	90,00	96,66
YEN	53,33	90,00	46,66	93,33
SIN	56,66	93,33	56,66	96,66
KRAD	86,66	96,66	83,33	100
KOZ	76,66	86,66	73,33	90,00
SMMUS	80,00	90,00	76,66	93,33
SDESS	70,00	93,33	66,66	96,66
SA	43,33	83,33	70,00	86,66
TAM	66,66	90,00	60,00	96,66
TZA	60,00	86,66	56,66	90,00

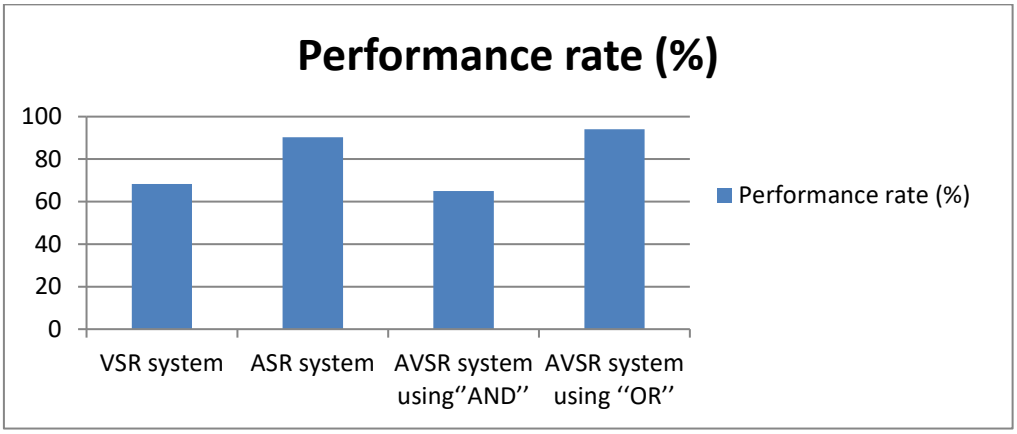

Figure 7. Comparison of recognition rates of audio-only, video-only and audiovisual ASR using "AND", "OR" based-rules

Our experimental results summarized in Table 7 shows that the decision fusion based on "OR" based-rule is more accurate than "AND" rule. As shown in table a high rate, of 93,99 %, is given by the combined audiovisual system using ''OR'' based-rule. The VSR system achieve, independently, a rate of 68,33 %. Whereas, the performance rate of ASR system is 90,33 %. Thus, the combined audiovisual system, using ''OR'' rule, achieve a high performance to either of the subsystems executing alone. This confirms our previous observation that the addition of visual information completes the acoustic information to decide what has been spoken imitating by the human speech perception. Based on ''OR'' rule, the spoken digit is recognized only when it's invested in one of the VSR system or the ASR system. By using the "OR" rule, the highest accuracy of 100% was obtained with digit "KRAD". This accuracy is due to this digit is always recognized in one of VSR system or in the ASR system which digit is recognized only when it's invested in one of the VSR system or the ASR system.

The linguistic discussion

The Amazigh language has adopted a number of consonantal phonetic from foreign languages, such as: Arabic, French and Spanish. In linguistics, the phonetics of a language are mostly attributing to the sound system of a language.

There are two types of articulations:

- The place of articulation (Velar, Palatal, Labial, Bilabial, Dental, Alveolar).
- The manner of articulation (Occlusive, Constrictive, Nasal, Fricative, Trill, Glide, Approximant).

The table 8 presents a list of phoneme used in Amazigh language. In our case, the phonetic analysis will help for getting the place and the manner of articulation for a digit that is more accurate. As shown in table 10, we can get out 3 types of the AVSR system average recognition rates: highly, medium and low.

In class highly the digit with a large number of phonemes (4 phonemes) gives a high rate (96,74%). Whereas in second category (Medium), the digits contain of 3 phonemes give a performance of 94,44%. A performance of 86% given by the digits that has 2 phonemes (Low category). Thus, the digit recognition is affected by its phonemes components.

Table 8. Phonological table of standard Amazigh consonant

	Bilabial	**Dental**	**Alveolar**	**Palatal**	**Velar**	**Uvular**	**Pharyngeal**
Occlusive	/b/	/t/, /d/			/k/, /g/	/q/	
Constrictive	/f/		/s/, /z/, /j/	/c/		/x/, /ɣ/	/h/, /ɛ/
Nasal	/m/	/n/					
Vibrant		/r/					
Lateral		/l/					
Semi-consonnes	/w/			/y/			

In category that achieve a highly performance, that is presented in table 9, we observe that the digit "KRAD'' contains a sequence of velar and dental phonetics. This phonetic categories combination gives the highest rate for the ASR system and the AVSR system. Therefore, in second category (Medium), the combination of the dental and alveolar phonetics attains a 93,33% in ASR system and 96,66% in AVSR system for the digit TAM. In Low category, the digit SA achieves a lowest rate in ASR, VSR and ASR system. This digit contains a dental phonetic category. Moreover, we observe that the pattern syllable of AMYA, KRAD and SMMUS are, respectively, VC-CV, VC-CVC and CCV-VC. Thus, these digits contain the largest number of syllables. Also, the syllable pattern of the digit SA (CV) demonstrates that this digit is the shortest in syllables number. Therefore, the digits that have largest number of syllable are most recognized with the system that diagnostic the smoking people [14]. Also, this type of digits is considered as the most noise-resistant digits [16]. Therefore, the number of syllables has an effect on Amazigh audiovisual speech recognition system.

Conclusion

İn present work, we build an AVSR for Amazigh language based on two separated audio and visual recognition subsystems. Each proposed module is developed independently. For modeling the visual speech recognition an HMM with DCT features is used. Regarding the ASR subsystem is designed using the Cmu-sphinx based tools. In order to develop these two modalities, we use the AmDigit_AVSR database. The decision fusion strategy is applied to combine the decisions of the two sub-systems. The proposed system achieves a high accuracy. This accuracy demonstrate that the integration of visual and acoustic information provide better performance than given by audio-only and visual-only systems In future work, we will

test and perform the proposed system under different degree of noise. We are also planning to extend our system to the continuous speech.

Table 9. The performance categories

Number of phoneme	Digits	Rate recognition (%)			The occurrence of phonetic category in digit							average rate 0f AVSR	Performance
		ASR system	VSR system	AVSR system using ''OR''	bilabial	Dental	Alveolar	Palar	Velar	Uvelar	Pharyngeal		
4	AMYA	93,33	90	96,66	+							96,74	Highly
	SMMUS	90	80	93,66	+		++						
	SDESS	93,33	70	96,66		+	++						
	KRAD	96,66	86,66	100		++			+				
3	YEN	90	53,33	93,33		+						94,44	Medium
	SIN	93,33	56,66	96,66		+	+						
	TAM	90	66,66	96,66	+	+							
	TZA	86,66	60	90		+	+						
2	SA	83,33	43,33	86			+					86	Low

References

[1] H. McGurk & J. MacDonald, (1976). "Hearing lips and seeing voices". Nature. 264(5588), 746.

[2] I. Addarrazi, H. Satori, & K. Satori, (2020). "A Follow-Up Survey of Audiovisual Speech Integration Strategies". Embedded Systems and Artificial Intelligence. Springer, Singapore, pp. 635-643.

[3] I. Addarrazi, H. Satori & K. Satori, (2018, October). "Building A First Amazigh Database For Automatic Audiovisual Speech Recognition System". In Proceedings of the 2nd International Conference on Smart Digital Environment, pp. 94-99.

[4] E. Petajan, (1985). "Automatic lipreading to enhance speech recognition". Proc. CVPR'85.

[5] P. Borde, A. Varpe, R. Manza, & P. Yannawar, (2015). " Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition". International journal of speech technology, 18(2), pp. 167-175.

[6] A. Makhlouf, L. Lazli & B.Bensaker, (2016). "Evolutionary structure of hidden Markov models for audio-visual Arabic speech Recognition". International Journal of Signal and Imaging Systems Engineering, 9(1), pp 55-66.

[7] A. Biswas, P. K. Sahu & M. Chandra, (2016). "Multiple cameras audio visual speech recognition using active appearance model visual features in car environment". International Journal of Speech Technology, 19(1), pp. 159-171.

[8] M. H. Rahmani, F. Almasganj & S. A. Seyyedsalehi, (2018). “Audio-visual feature fusion via deep neural networks for automatic speech recognition”. Digital Signal Processing, pp. 82, 54-63.

[9] F. A. Allah, & S. Boulaknadel, (2010, July). “Amazigh Search Engine: Tifinaghe Character Based Approach”. In Proceeding of the International Conference on Information and Knowledge Engineering (pp. 255-259).

[10] M. Jones & P. Viola, (2003). “Fast multi-view face detection”. Mitsubishi Electric Research Lab TR-20003-96, 3(14), 2.

[11] I. Addarrazi, H. Satori & k Satori, (2020). "Lip Movement Modeling Based on DCT and HMM for Visual Speech Recognition System." Embedded Systems and Artificial Intelligence. Springer, Singapore,. 399-407.

[12] I. Addarrazi, H. Satori, & K. Satori, (2017, April). “Amazigh audiovisual speech recognition system design”. In Intelligent Systems and Computer Vision (ISCV), IEEE, pp. 1-5.

[13] H. Satori & F. Elhaoussi, 3 (2014). "Investigation Amazigh speech recognition using CMU tools." International Journal of Speech Technology, 17, pp. 235-243.

[14] O. Zealouk, H. Satori, M. Hamidi, N. Laaidi, & K. Satori, (2018). “Vocal parameters analysis of smoker using Amazigh language”. International Journal of Speech Technology, 21(1), pp. 85-91.

[15] M. Hamidi, H. Satori, O. Zealouk, & K. Satori, (2019). “ Speech coding effect on Amazigh alphabet speech recognition performance”. J. Adv. Res. Dyn. Control Syst, 11(2), pp.1392-1400.