Visual Speaker Verification System Depending on Arabic Syllables

Khadidja SADEDDINE, Fatma Zohra CHELALI, Rachida DJERADI, Amar DJERADI
Speech communication and signal processing laboratory
Houari Boumedienne University of sciences and Technologies, USTHB
Box n°: 32 El Alia, 16111, Algiers
Algeria
sadeddine_khadidja@yahoo.com, chelali_zohra@yahoo.fr

ABSTRACT: *We develop in this work a speaker verification system depending on Arabic syllables by the study of the visual speech that contains the visual and acoustic modalities. The visual signal provides both additional information that is not present in the audio and also a visual representation of some of the information that is present in the audio. This is particularly evident in the perception of speech where the articulatory gestures of the speaker's lips and face can significantly improve the listener's detection. In order to analyze Arabic visual speech, we extract features such as pitch and LPC coefficients for the acoustic modality, and we use DCT coefficients for lip images for visual modality. Hierarchical ascendant classification (HAC) is applied for each modality. Simulation results show good recognition rate of speaker verification depending on phoneme for the two modalities.*

Keywords: Visual-speech Analysis, Audiovisual Speech Analysis, Lip Reading, Hierarchical Ascendat Clustering

Received: 2 February 2013, Revised 8 March 2013, Accepted 17 March 2013

1. Introduction

Perception of human speech is a multimodal process, which involves the analysis of the uttered acoustic signal and which includes higher level knowledge sources such as grammar, semantics and pragmatics. One information source which is mainly used in the presence of acoustic noise is lip reading or so called speech reading [1]. Scientists needed novel, nontraditional approaches that use other sources of information to the acoustic input that not only significantly improve the performance in severely degraded conditions, but also are independent to the type of noise and reverberation [2]. Visual speech is one such source, obviously not perturbed by the acoustic environment and noise. It is well known that humans have the ability to lip read: We combine audio and visual information in deciding what has been spoken, especially in noisy environments [2].

Both human speech production and perception are bimodal in nature [3]. The visual modality such as seeing a speaker's lips move, can facilitate auditory speech perception and benefit to speech intelligibility in noise [3].

Bimodal integration of audio and visual stimuli in perceiving speech has been demonstrated by the McGurk effect because what we hear is influenced by visual information of articulatory movements, which is demonstrated by the McGurk effect [4, 5]. For example when a person '*hears*' the sound /ba/, but '*watches*' the sound /ga/, the person may not perceive either /ba/ or /ga/. Something close to a /da/ is usually perceived. The McGurk effect highlights the requirement for both acoustic and visual cues in the perception of speech. This effect has been shown to occur across different languages and in infants. In addition, visual speech is of particular importance to the hearing impaired: mouth movement is known to play an important role in both sign language and simultaneous communication between the deaf. The hearing impaired speech read well, and possibly better than the general population. There are many reasons why vision benefits human speech perception. One of those reasons that it provides complimentary information about the place of articulation which is due to the partial visibility of articulators, such as the tongue, teeth, and lips. In addition, jaw and lower face muscle movement is correlated to the produced acoustics and its visibility has been demonstrated to enhance human speech perception [2].

These facts have motivated significant interest in automatic recognition of visual speech, formally known as automatic lip reading [3]. Researches in this field aim to improve audio visual speech recognition AVSR by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, leading to audiovisual systems.

Many works have been realized to exploit the bimodality of visual speech and the correlation that exists between audio and visual speech features. The first automatic speech reading system was reported in 1984 by Petajan [2]. Given the video of the speaker's face, and by using simple image thresholding, he was able to extract binary mouth images, and subsequently, mouth height, width, perimeter, and area, as visual-speech features. He then developed a visual-only recognizer based on dynamic time warping to restore the best two choices of the output of the baseline audio-only system. His method improved AVSR for a single-speaker, isolated word recognition task on a 100-word vocabulary [2] [3]. Jiang and Guoyun examine and track head and extract lip's contour through the Bayesian tangent shape model and two types of acoustic features; MFCC features and PLP features for the recognition of audio visual speech [6]. Potamianos investigates the use of Fisher-Rao linear discriminant analysis (LDA) as a means of visual feature extraction for hidden Markov model based automatic speech reading [7]. Almajai and Milner examine the correlation between audio and visual speech features where two audio features (MFCCs and formants) and three visual features (active appearance model, 2-D DCT and cross-DCT) are considered. Nefian and Liang describe in their work the use of two statistical models for audio-visual integration, the coupled HMM (CHMM) and the factorial HMM (FHMM), and compare the performance of these models with the existing models used in speaker dependent audio-visual isolated word recognition [9].

Integration of visual features in speech recognition system aims to solve situations where audition is insufficient to insure speech comprehension, when speech is exposed to noise, bandwidth limitations, hearing limitations or other disturbances [10]; or in particular the tasks of isolated word speech and text-dependent speaker recognition [11]. Thus, audiovisual recognition of speech improves traditional classifiers performance especially when extracting mouth movements are added to the audio signal and many works in the literature have proved that movements of the mouth can be used as one of the speech recognition channels.

In speech analysis (audio modality), wide variety of features have been used, including time domain measurements such as energy, zero crossings, bandpass filter outputs; frequency domain measurements such as spectral coefficients, cepstral coefficients, spectral derivative; or linear predictive coding (LPC) [12]. Mel frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) are too the two most common feature extraction techniques in speaker identification. MFCC is generally used because of its robustness in speaker identification. Since the elements of feature vectors are generally correlated , a large number of mixtures with full covariance matrix are necessary to provide good approximation. The GMM with diagonal covariance matrix is used for both speaker identification and verification because of its computational simplicity [13].

Most speech coding algorithms make use of the source-filter model of human speech production, where speech is modeled as the response of a time varying linear synthesis filter to an input signal called the excitation. Examples of speech coders based on this model include the numerous variations of the linear predictive coding LPC vocoder, the large family of linear prediction based analysis by-synthesis (LPAS) coders including code-excited linear prediction (CELP) as well as some harmonic coding algorithms of current interest. The synthesis filter determines the short-term spectral envelope of the synthesized speech and is characterized by the linear prediction (LP) coefficients obtained from LP analysis on the input speech. These coefficients are commonly called LPC coefficients, which may refer generically to any of several different but equivalent parameter sets that specify the synthesis filter [14].

In the present paper, we consider the speaker's lip movements as visual feature by extracting lips from face image of talking speaker and then analyzing the sound produced in the purpose to verify the speaker depending on Arabic syllable uttered.

We'll use LPC parameters with other acoustic feature, the fundamental frequency called pitch (F_0) in the recognition process to significantly increase the performance of automatic speech recognition system.

A number of pitch detection algorithms have been reported by using time domain and frequency domain methods with varying degrees of accuracy [15]. In this paper, we'll employ the autocorrelation method for pitch detection. (We'll use the terms F_0 and pitch interchangeably, although technically, pitch is a perceptual attribute, whereas F_0 is an acoustic property, generally considered to be the primary cue for pitch [15]).

In other side, the image carries large amount of data, thus it requires efficient techniques to reduce the total number of bits necessary to represent the image without information loss or, more practically, with an acceptable and controlled level for information loss (image distortion). The operation of reduction of the amount of data is called data compression. Generally, transform coding for image data compression achieves better results [16].

Many unitary transforms have been proposed and applied (e.g. the Discrete Fourier Transform, the Hadamard Transform, the Haar Transform, etc.), each approximating to a different degree the behaviour of the optimal Karhunen-Loeve Transform KLT. It has been also shown that the Discrete Cosine Transform (DCT) closely behaves like the KLT for highly correlated images [16]. The DCT, owing to its important properties has found applications in image and signal processing. Recently, it has become the industry standard in image coding [17].The visual features in this study DCT coefficients of lip images.

In our work, we develop a speaker recognition/verification system depending on Arabic syllables (consonant-vowel CV) or phonemes. For that, we use the LPC coefficients and the pitch for speaker verification as audio features. Whereas, we extract DCT coefficients from still lip images for visual modality.

The paper is organized as follows. Section B, describes the Arabic phoneme database acquisition used in the experiments. Section C details the extraction of audio-visual features. In section *D* a general survey on hierarchical ascendant classification is presented. Section *E* discusses the results obtained. Finally, we conclude our work in section *F*.

2. System Description

2.1 Phoneme data base

The database was recorded in speech and communication laboratory, where four Algerian subjects (male and female) are speaking the 28 phonemes of the Arabic language.

The recordings were made on camera canon video with 25 frames of size 576 * 720 pixels per second. Then, the data were transferred into computer through IEEE1394 card. The corpus consists of 20 repetitions of every syllabus (phoneme with short vowel) produced by each speaker, 20 still images in format bmp and 20 repetitions for audio file in format wav.

The database includes 2240 face images for four different subjects. The speech signals are acquired during different sessions with a sampling frequency of 22 KHz.

3. Audio-visual Features Extraction

Our approach is composed of three steps: First, we proceed by extracting pertinent visual features (DCT coefficients) from the speaker's lip image; the extracted features are used for classification and recognition. Secondly, pitch and LPC parameters are also extracted from the corresponding audio file. The visual speech recognition system proposed for Arabic language and the extraction of its features are shown in figure 1.

3.1 Visual features extration

We begin by localizing lips from speaker's face images and convert the corresponding lip images.

3.1.1 Discrete cosine transform (DCT)

DCT is a well-known signal analysis tool used in compression due to its compact representation capability. It has an excellent energy compaction property for highly correlated data. This helps in reduction of feature dimension.

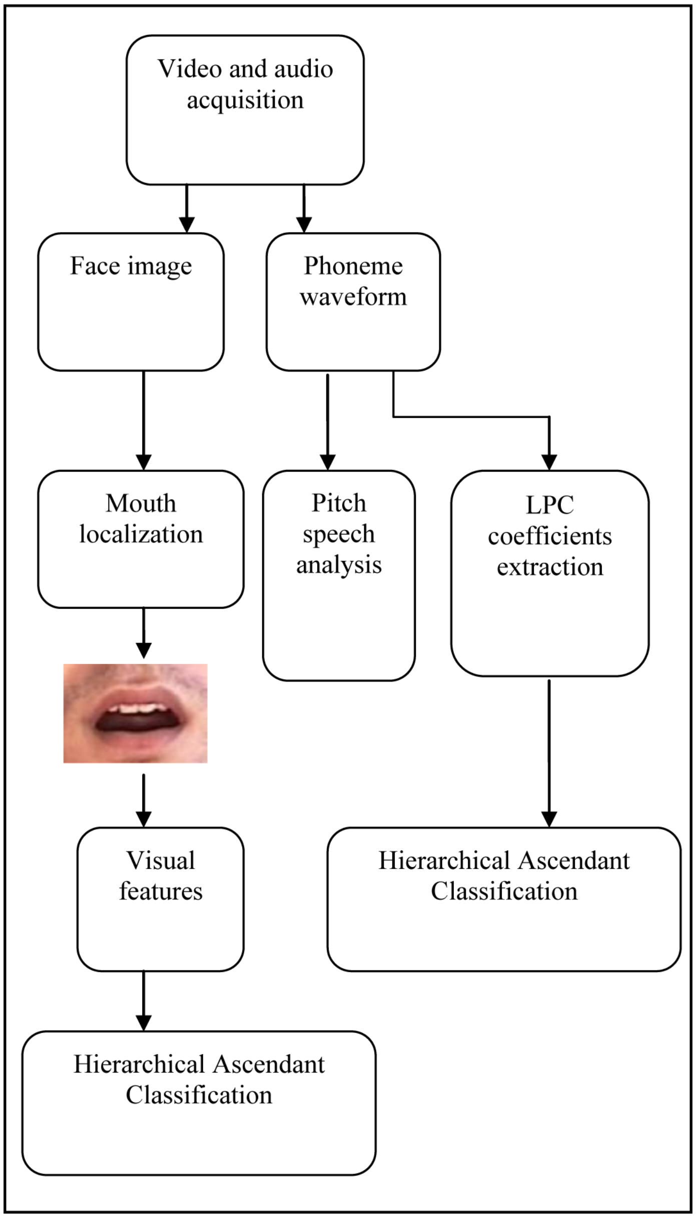
Video and audio acquisition
Face image
Phoneme waveform
Mouth localization
Pitch speech analysis
LPC coefficients extraction
Visual features
Hierarchical Ascendant Classification
Hierarchical Ascendant Classification

Figure 1. General corpus of phoneme acquisition

DCT transforms an image from the spatial domain to the frequency domain, where the image is decomposed into the combination of various frequency components. As a result, DCT is capable of extracting the different features in the frequency domain to encode the different facial information that is not directly accessible by the methods working in the spatial domain [19].

Let image $f(x, y)$ is represented as $f(m, n)$ of size $M \times N$. The 2-D DCT of an image $f(m, n)$ is given as [20]:

$$B(p, q) = \alpha(p)\,\alpha(q) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m, n) \cos \frac{\pi(2m+1)p}{2M} \times \cos \frac{\pi(2n+1)q}{2N} \qquad (1)$$

$p = 0 ... M - 1$and $q = 0 ... N - 1$.

Where p and q denote the frequencies.

Arabic Alphabet	Phonetic Transcription (API)	Symbol
ء	[?]	**A**
ب	[b]	**Ba**
ت	[t]	**Ta**
ث	[θ]	**Tha**
ج	[z]	**Dja**
ح	[ħ]	**H**
خ	[x]	**Kha**
د	[d]	**Da**
ذ	[δ]	**Dha**
ر	[r]	**Ra**
ز	[Z]	**Za**
س	[s]	**Saa**
ش	[š]	**Cha**
ص	[ś]	**Ssa**
ض	[ð]	**Daa**
ط	[ť]	**Taa**
ظ	[δ]	**Dhaa**
ع	[ς]	**Aa**
غ	[γ]	**Gha**
ف	[F]	**Fa**
ق	[q]	**Qua**
ك	[k]	**Ka**
ل	[l]	**La**
م	[m]	**Ma**
ن	[n]	**Na**
ه	[h]	**Ha**
و	[w]	**Wa**
ي	[j]	**Ya**

Table 1. The Phonetic Transcription and Symbol Used for Phoneme [18]

$$\alpha(p) = \begin{cases} \frac{1}{\sqrt{M}} & ; p = 0 \\ \sqrt{\frac{2}{M}} & ; p = 1 \dots M - 1 \end{cases} \tag{2}$$

$$\alpha(q) = \begin{cases} \frac{1}{\sqrt{N}} & ; p = 0 \\ \sqrt{\frac{2}{N}} & ; p = 1 \dots N - 1 \end{cases} \tag{3}$$

Where *M* and *N* denote the row and column size of *f* (*m*, *n*) respectively.

The transform coefficient *B* (0, 0) represents the average value of the input sequence and is denoted by DC coefficient, while all other transform coefficients are denoted by AC coefficients [21].

A general transform coding scheme involves subdividing an $N \times N$ image into smaller non overlapping $n \times n$ sub-image blocks and performing a unitary transform on each block. It aims to decorrelate the original data and compact a large fraction of the signal energy into a relatively small set of transform coefficients. In this way, many coefficients can be discarded after quantization and prior to encoding

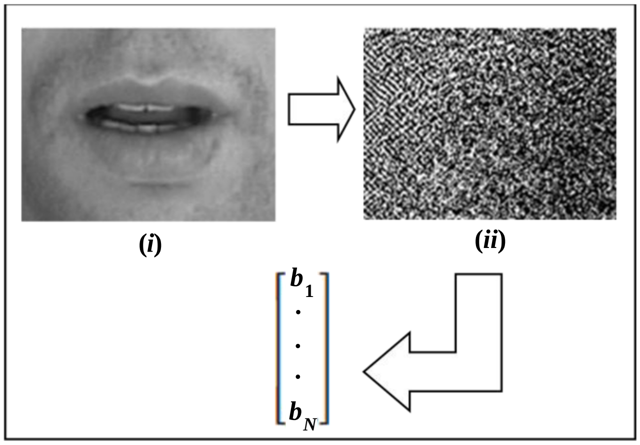

Figure 2. (i). Lip image in gray scale; (ii). Its 2-D DCT transform

After converting the lip images obtained (acquired in RGB space) into gray scale, we extract DCT coefficients (visual features) from the speaker's lip image. Those features are used for hierarchical classification HAC.

3.2 Acoustic features extraction

Acoustic speech is a naturally varying continuous signal whose statistics change with time. A problem results into how to break it up into observation feature vectors suitable for processing. Fortunately, speech varies slow enough to assume its statistics are quasi-stationary over segments up to 100 ms in duration [2] [22]. A hamming window is employed to segment the speech signal. For speech and speaker recognition applications a window size of 25ms is employed to segment the speech signal into 10ms blocks. For the experiments conducted in this study the window takes the form [2]:

$$w(n) = 0.54 - 46 \cos\left(\frac{2\pi\,(n-1)}{N-1}\right), 1 \leq n \leq N \tag{4}$$

Where n is the n^{th} sample and N is the number of samples.

After windowing, the segmented speech observation is then passed through a pre-emphasis filter. This filtering is performed to flatten the frequency characteristics of the speech signal, which typically has most of its energy situated in the low frequency range [22]. The pre-emphasis filter takes the form of:

$$E(Z) = 1 - az^{-1} \tag{5}$$

Where $0.99 < a < 1$ [22].

After pre-processing, we used two acoustic features such as the pitch and the LPC parameters.

3.2.1 Pitch extraction

Pitch is one of the most important perceptual features of sound. It conveys prosody and speaker identity in speech and melody in music, and it is one of the most important cues for segregating sounds from different sources in the environment [22].

The time from opening time of vocal cords to the next opening time of it is called fundamental period T_0, the vibrating speed of vocal cords is called fundamental Frequency F_0. As the term pitch is often being used as the same meaning with fundamental frequency, there are subtle differences between the two, but in general the pitch is often used to mean the word of the fundamental frequency [23].

Thus for the voiced speech, F_0 is usually defined as the rate of vibration of the vocal folds. Periodic vibration at the glottis may produce speech that is less perfectly periodic because of movements of the vocal tract that filters the glottal source waveform. However, glottal vibration itself may also show periodicities, such as changes in amplitude, rate or glottal waveform shape, or intervals where the vibration seems to reflect several superimposed periodicities, or where glottal pulses occur without an obvious regularity in time or amplitude [3].These factors conspire to make the task of obtaining a useful estimate of speech F_0 rather difficult [3].

The resonant frequencies of the speech signal are formed in the vocal tract and are known as formants. Typically, pitch ranges between 80 Hz and 160 Hz for male speakers, and between 160 Hz and 400 Hz for female speakers. Formant frequencies are usually greater than the pitch frequency and can lie in the kilo-Hertz range. Estimation of pitch and formants finds extensive use in speech encoding, synthesis, and recognition. Some well-known pitch detection methods employ cepstrum, adaptive prediction, and data reduction [24]. Because determination of pitch is very important for many speech processing algorithms, a commonly used method to estimate pitch is based on detecting the highest value of the autocorrelation function in the region of interest. Our perception of pitch is strongly related to periodicity in the waveform in the time domain [25].

In practice, we need to obtain an estimate the autocorrelation from knowledge of only N samples. The empirical autocorrelation function is given by [25]:

$$\hat{R}\,[m] = \frac{1}{N}\sum_{n=0}^{N-1-|m|}(w\,[n]\,x\,[n]\,w[n+|m|]\,x[n+|m|]\,) \tag{6}$$

Where $w\,[n]$ is, a window function of length N.

To detect the pitch, we take a window of the signal, with a length at least twice as long as the longest period that we might detect. Using this section of signal, we generate the autocorrelation function $\hat{R}\,[m]$ defined as the sum of the point wise absolute difference between the two signals over some interval [26].

3.2.2 LPC features extration

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate. It provides extremely accurate estimates of speech parameters, and is relatively efficient for computation. LPC starts with the assumption that the speech signal is produced by a buzzer at the end of a tube. The glottis (the space between the vocal cords) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). The vocal tract (the throat and mouth) forms the tube, which is characterized by its resonances, which are the formants [27].

Linear Prediction analysis provides an efficient means of representing speech.

Linear prediction (LP) analysis, attempts to model the windowed speech segment via an all pole filter of the form [2]:

$$H(z) = \frac{G}{1-\sum_{i=1}^{p} a_i z^{-i}} \tag{7}$$

Where G is the gain factor used to control the intensity of the excitation with the p linear predictive coefficients (LPCs) a_i describing the autoregressive model. The use of LPC can be extended to speech recognition since the FIR coefficients are the condensed information of a speech signal of typically 20-30ms [28].

The predictor coefficients a_i are chosen to minimize the mean square filter error summed over the windowed speech segment [2] [28]. The minimization required to find the autoregressive coefficients a_i is accomplished through Levinson-Durbin recursion [29]. Upon solving of $H\,(z)$ the magnitude response represents the spectral envelope of the speech segment.

LPC estimates the current value $x\,(n)$ of a random sequence from p previous values. The estimate $\hat{x}\,(n)$ can be written as:

$$\hat{x}\,(n) = a_1 x(n-1) + a_2 x(n-2) + ... a_p x(n-p) \tag{8}$$

The error in the estimate is given by:

$$\varepsilon(n) = x(n) - \hat{x}(n) \qquad (9)$$

$$\varepsilon(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k) \qquad (10)$$

This is the output of a finite impulse response (FIR) filter [28]. In equation 10, the vector a_k $(1, a_1, a_2, ...a_p)$ is called linear prediction coefficients (where $a_0 = 1$).

$$\sigma_e^2 = E\{|\varepsilon(n)|^2\} = E\{(x(n) - \hat{x}(n))^2\} \qquad (11)$$

σ_e^2 is called the prediction error variance in linear prediction [28]. We minimize the error variance in order to find the optimal linear prediction coefficients a_k [28]. The resulting LPC coefficients vector $[a_0, a_1, a_2, ...a_p]$ for each phoneme is saved in matrix that represents the speaker and then all LPC coefficients are put in matrices containing the 28 Arabic phonemes. Hence, matrices are used for clustering system.

4. Hierarchical Ascendat Clustering (HAC)

Automatic clustering is to contribute to subsets the elements or data of set which means clustering is dividing this set into clusters where each cluster must be homogenous and in other side clusters must be different between each other. In addition, this sub setting is insufficient because we should look for a hierarchy of the formed groups (clusters) that constitute a binary tree called Dendrogram [18].

The definition of clustering leads directly to the definition of a single "*cluster*". Most definitions are based on loosely defined terms, such as "*similar*", "*alike*"....etc. The vectors are viewed as points in the l-dimensional space and the clusters are described as continuous regions of this space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points. We try to give some definitions for clustering [18].

Let X be our data set, that is,

$$X = \{x_1, x_2, ...x_N\}$$

We define as an m-clustering of X; R, the partition of X into m sets (clusters), $c_1 ... c_m$, so that the three following conditions are met [30]:

$$c_i \neq \varnothing,\ j = 1....m \qquad (12)$$

$$\cup_{i=1}^{m} c_i = X \qquad (13)$$

$$c_i \cap c_j \neq \varnothing,\ i \neq j, i, j = 1....m \qquad (14)$$

In addition, the vectors contained in a cluster c_i are "*more similar*" to each other and "*less similar*" to the features vectors of the other clusters [18].

However, not all clustering algorithms are based on proximity measures between vectors.

In the hierarchical clustering algorithms one has to compute distance between pairs of sets of vectors of X. In the sequel, we extend the preceding definitions in order to measure "*proximity*" between subsets of X. That is:

$$D_i \subset X\ i = 1....k \textit{ and } U = \{D_1 ... D_k\}$$

A proximity measure $\wp$ on U is a function:

$$\wp : U * U \rightarrow R$$

Usually, the proximity measures between two sets D_i and D_j are defined in terms of proximity measures between elements of D_i and D_j. Let be a set of observations, a hierarchy on this set is groups of observations collection. This technique defines similarities measure or distances measure between samples to be classified that leads to an aggregation criterion of clusters which is this measure. Aggregation criterion that we've chosen single linkage or k-nearest neighbour (*k*-nn) that uses a lot of metrics (City block, Mahalanobis, Euclidian …etc). The Euclidian distances L will be used as metric for hierarchical clustering where L is defined by [18]:

(14)

$$L = \| Y_i - Y_j \| = \sqrt{\sum_{k=1}^{N} (Y_{ik} - Y_{jk})^2} \qquad (15)$$

Y_i, Y_j represent vectors to classify and N is dimension of vectors. The hole set of clusters is ahierarchy of subsets or p-tree where the tree have a root and leaves. The root represents the hole set of samples which is always in the hierarchy and leaves are clusters those have only one element that is one image like shown in figure 3 and figure 4. The defined function DENDROGRAM (*Z*) generates a dendrogram plot of the hierarchical binary cluster tree *Z*. *Z* is an (*M* – 1)-by-3 matrix, generated by the LINKAGE function, where *M* is the number of objects in the original dataset [18].

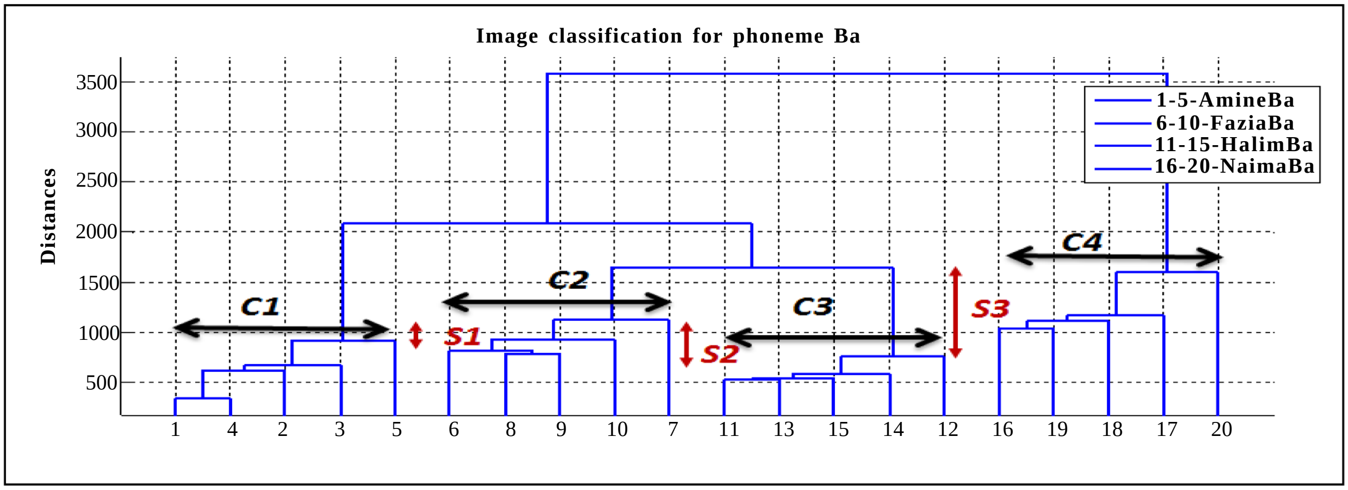

Figure 3. HAC clustering of lip images for phoneme “Ba / ب /”

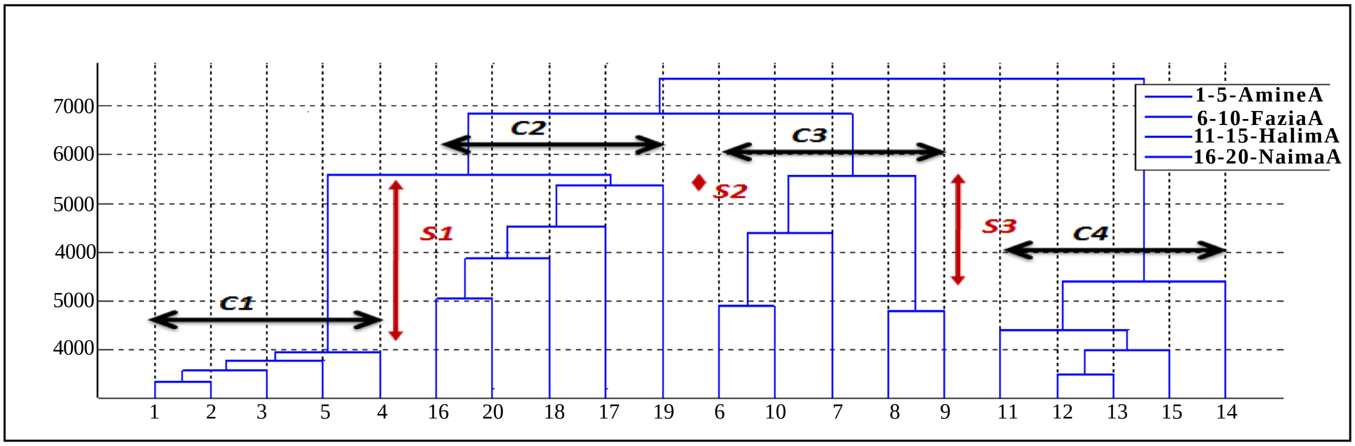

Figure 4. HAC clustering of lip images for phoneme “A /ء/”

5. Simulation Results and Discussion

5.1 Visual modality

In fact, DCT coefficients of lip images are the input of the classifier, by calculating distances; the k-nearest neighbor allows a hierarchical clustering for the input set where we have five lip images representing the phoneme for each speaker. Result is demonstrated in figure 3 that represents the HAC clustering for phoneme “*Ba* / ب /”. We noticed that the first, second, third, fourth and the fiveth images make one cluster and distances between them is very small.

C1, C2, C3 and C4 are clusters obtained where each cluster represents one speaker in our case. The inter-speaker variability of

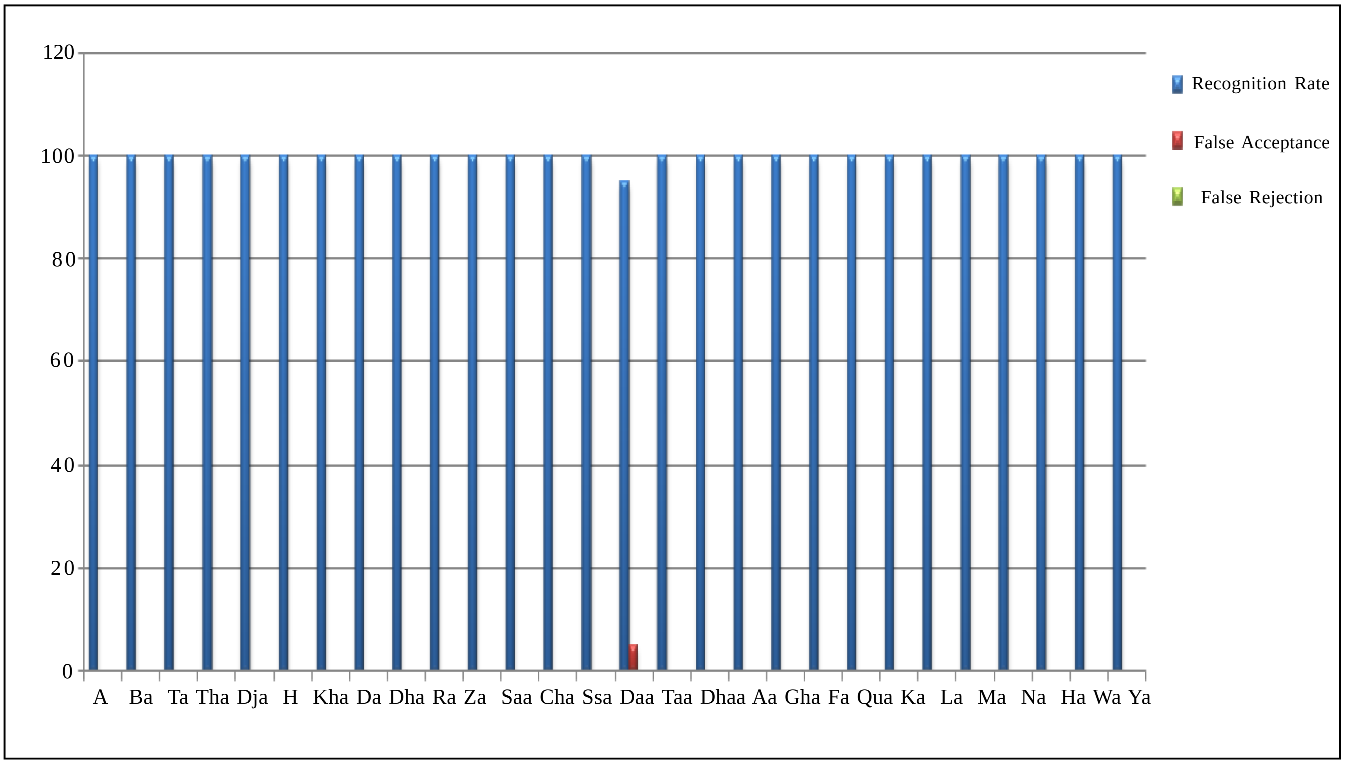

Figure 5. Recognition rate of HAC clustering of lip images for the 28 Arabic syllables

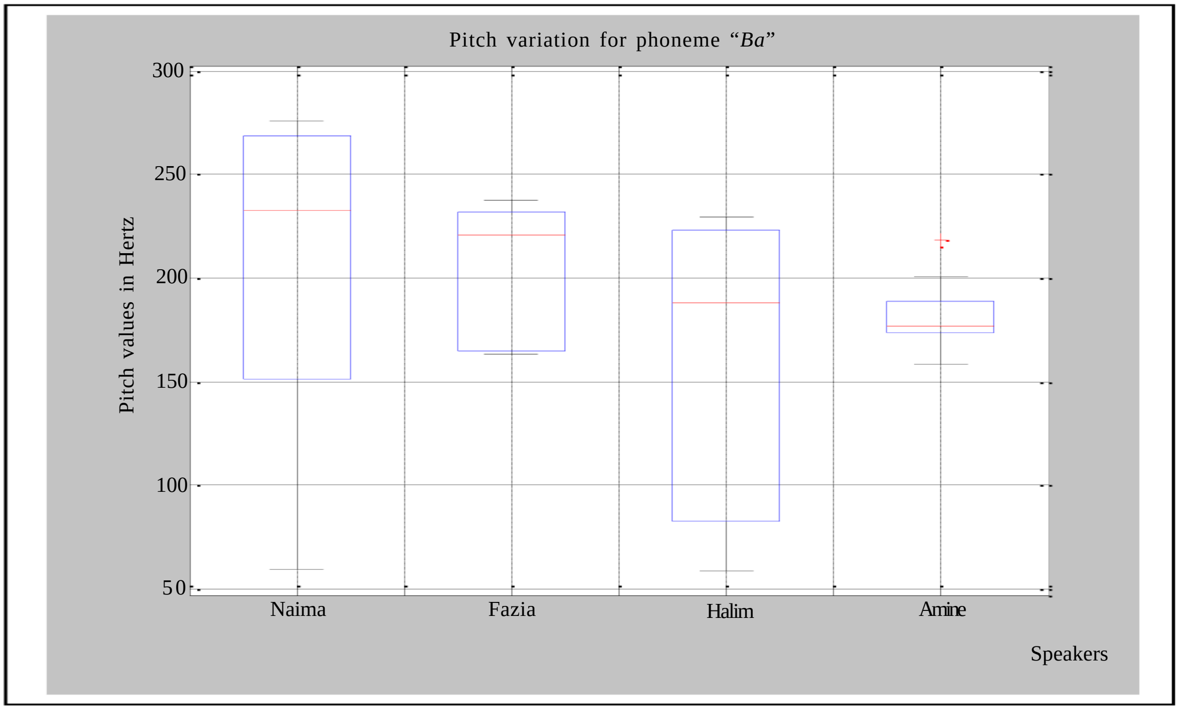

Figure 6. Pitch variation in Hertz for different speakers for phoneme "*Ba* / ب /"

distances S1, S2 and S3 is shown in this figure. Those distances are very small for this phoneme against to figure 4 where they

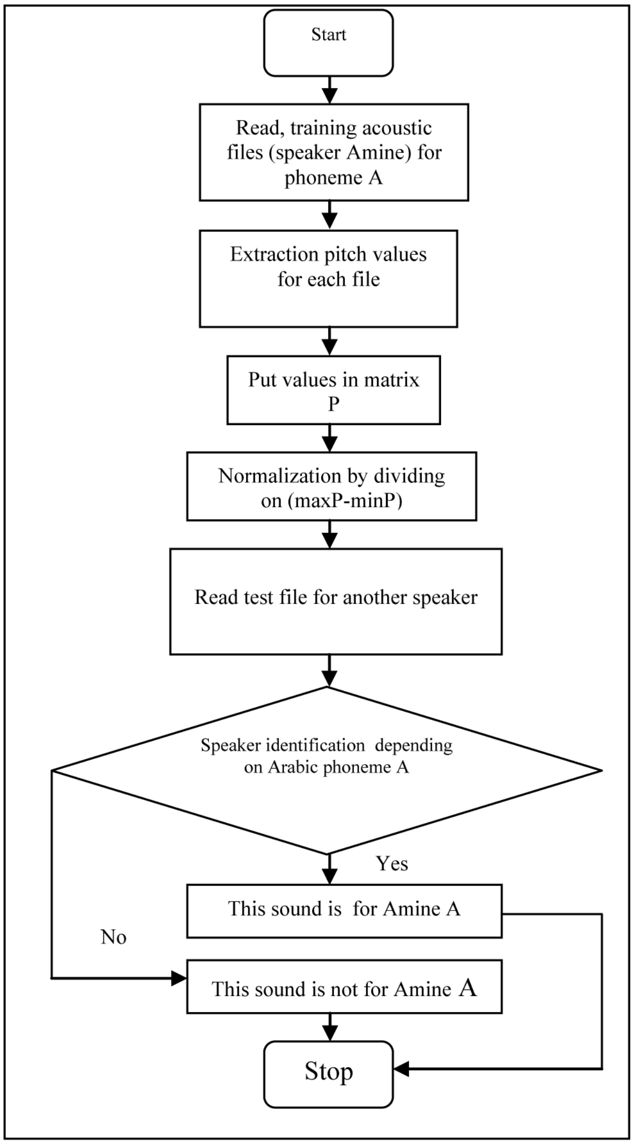
Start
Read, training acoustic files (speaker Amine) for phoneme A
Extraction pitch values for each file
Put values in matrix P
Normalization by dividing on (maxP-minP)
Read test file for another speaker
Speaker identification depending on Arabic phoneme A
Yes
No
This sound is for Amine A
This sound is not for Amine A
Stop

Figure 7. Speaker verification algorithm depending on Arabic phoneme “*A* / ء /” using pitch

are large for phoneme “*A* /ء /”. The existence of this inter-speaker variability of distances signifies that the speakers are visually discriminated so when those distances are large the speakers are well differentiated. There are other distances which are the intra-speaker distances that must be small to attribute phonemes to their speakers (clusters) and if those distance are large phonemes cannot be attributed to clusters.

We’ve applied the HAC classifier with the entire set of lip images for the 28 phonemes of Arabic language (figure 5). A good recognition rate has been obtained (100%) for all syllables except for phoneme “*Daa*/ض /” where it was 95%. We note that phoneme “*Daa* /ض /” is the only phoneme that distinguishes Arabic language from other languages.

5.2 Acoustic modality

5.2.1 Pitch feature

Figure 6 shows an important variability of pitch for the phoneme “*Ba*” spoken by four speakers (Naima, Fazia, Halim and Amine) which means that pitch is variable for the same phoneme pronounced from the same speaker.

The pitch for the entire set of phomemes varies (figure 6) for one speaker for many repetitions and for the same phoneme. Hence,

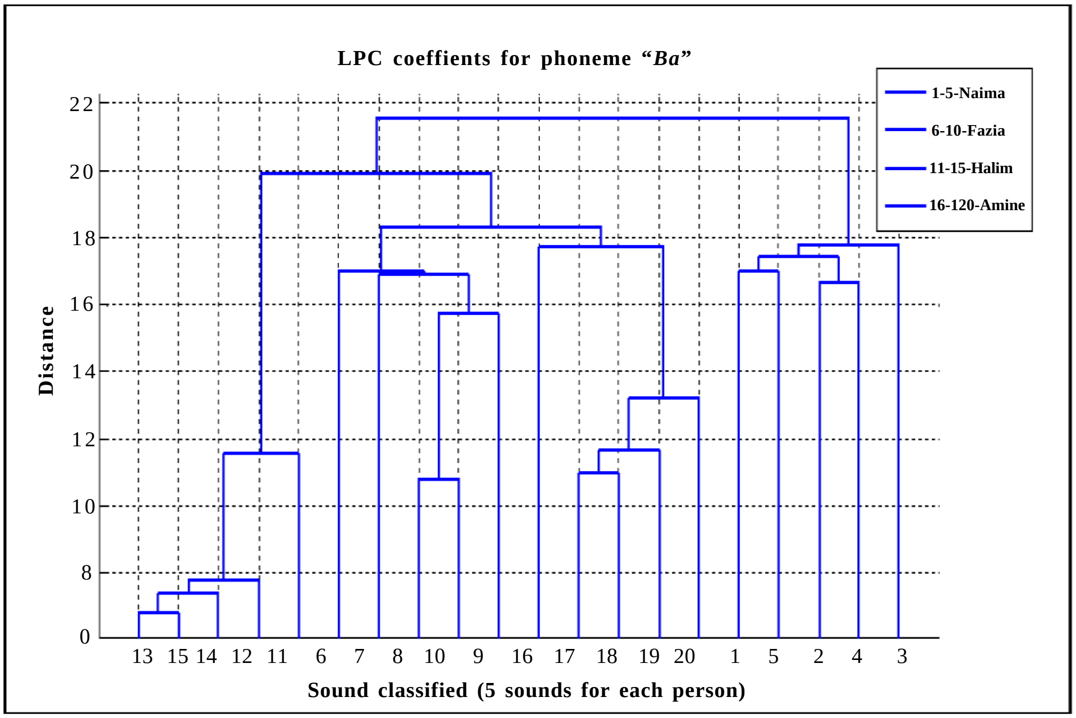

Figure 8.Clustering of LPC coeffients for phoneme "*Ba* /ب /"

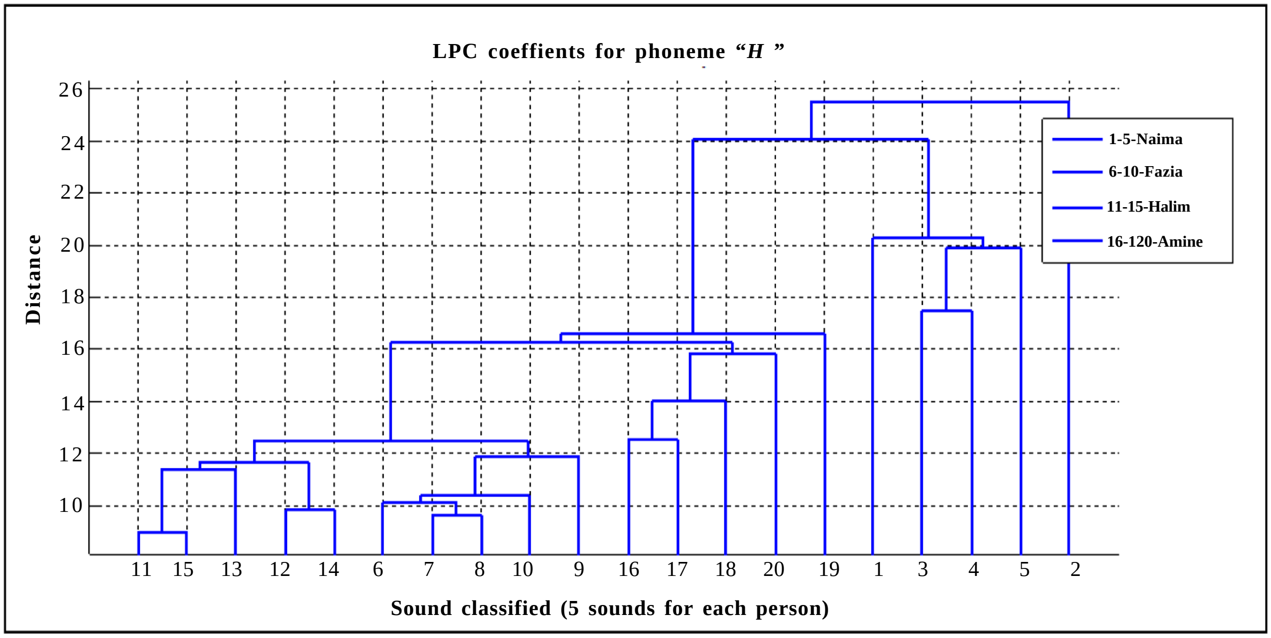

Figure 9. HAC clustering of LPC coeffients for phoneme "*H* / ح /"

the identification and the check of speaker uttering some phoneme using pitch becomes difficult.

In this work, we present an algorithm (figure 7) that alows the verification/idendification of speaker depending on phoneme.

The algorithm has been tried for the identification of the four speakers pronouncing different syllables of Arabic language. Figure 7 shows the synoptic scheme for speaker verification. We take the case when the speaker Amine saies "A /ء /" for example, this sound is the input of the algorithm. One of the two sentences below can be displayed : “*This sound is for Amine A*” or “*This souns is not for Amine A*”.

5.2.2 LPC coefficients feature

For each Arabic phoneme which is a syllable CV (table1) a speaker verification system depending on syllables is applied. Hence, LPC coefficients are determined for the 28 Arabic phonemes and then being the input of the classifier HAC by fixing order of prediction at 12. Figure 8 and figure 9 show the response for phonemes “Ba / ب /” and “H / ح /”.

The clustering of the prediction coefficients for each syllable permits us to say that a phoneme pronounced by one speaker is different from itself pronounced by another speaker because the intra-speaker distances have not shown important variations for majority phonemes against for the inter-speaker distances. These results allow us to say, when the speaker is pronouncing some phoneme; he is recognized for this phoneme using LPC coefficients.

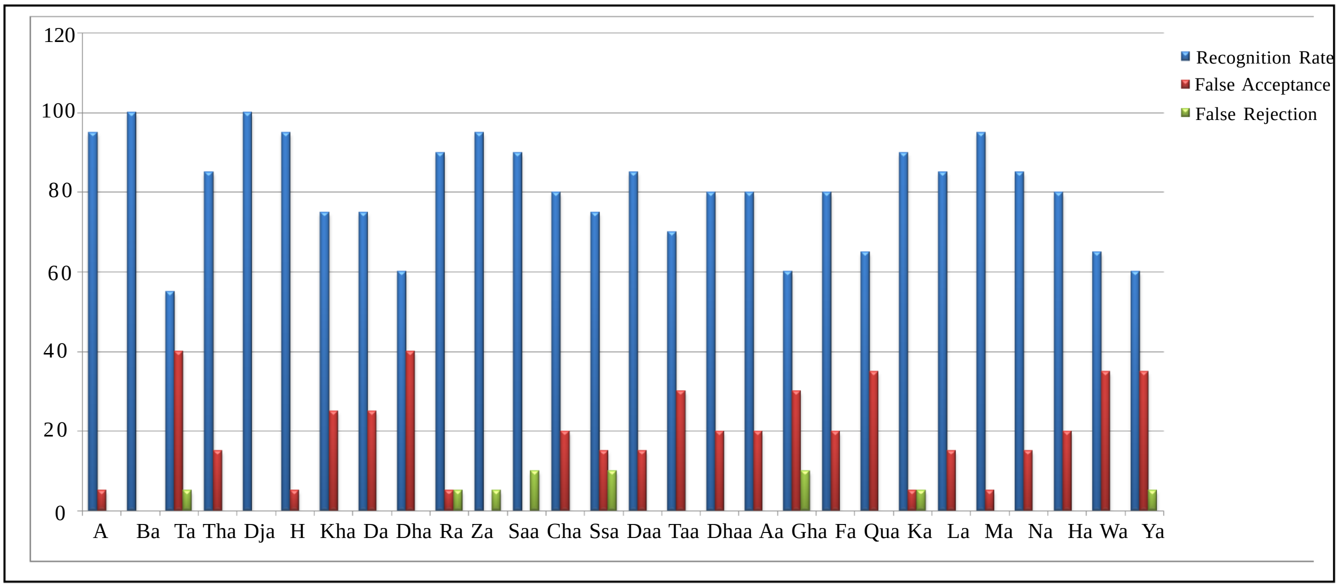

Figure 10. Recognition rate of HAC clustering of LPC coefficients for the 28 Arabic

Figure 10 demonstrates the variation of the recognition rate of LPC coefficients for the 28 Arabic syllables. This variation is between 55% and 100%. We notice that false acceptance varies from 5% and 40%.

For the acoustic modality (using both pitch and LPC coefficients), the speaker is verified depending on the spoken syllable.

In visual modality, lip images of the speakers (for the 28 syllables) are well associated for their corresponding clusters (speakers); so lip image of speaker is recognized depending on the syllable studied.

Simulation results show good recognition rate for both acoustic and visual modalities.

6. Conclusion

We’ve presented in this article a speaker verification system depending on Arabic syllables.

Two modalities of visual Arabic speech are studied in this paper, the acoustic modality and the visual one which is considered as secondary source of information for visual speech.

The acoustic modality shows good recognition rate basing on the LPC parameters comparing to the use of pitch.

The visual information was applied by calculating DCT of lip images and HAC classifier when the response was 100% for the majority of phonemes.

References

[1] Luettin, J., Dupont, S. Continuous Audio Visual Speech Recognition, IDIAP (Dalle Molle Institute for Perceptual Artificial Intelligence) and Faculté Polytechnique de Mons, Belgium.

[2] Lucey, S. (2002). Audio-visual Speech Processing. PhD Thesis, *School of Electrical & Electronic Systems Engineering*.

[3] Potamianos, G., Gravier, G. and Senior, A.W (2003). Recent Advances in the Automatic Recognition of Audiovisual Speech. *In*: Proc of the IEEE, 91 (9) 1306-1326.

[4] Hisanaga, S., Sekiyama, K., Igasaki,T., Murayama, N. (2009). Audiovisual speech perception in Japanese and English: Inter-language differences examined by event-related potentials. International Conference on Audio-Visual Speech Processing University of East Anglia, Norwich, UK.

[5] MacGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. Nature 264, p. 746–748.

[6] Jiang, D., Guoyun, L., Ravyse, I., Jiang, X., Zhang, Y., Sahli, H., Rongchun, Z. (2007). Audio Visual Speech Recognition and Segmentation Based on DBN Models, Robust Speech Recognition and Understanding, Michael Grimm and Kristian Kroschel (Ed.), *In*: Tech, Austria.

[7] Potamianos, G., Graf, H. P. (1998). LINEAR Discriminant Analysis For Speechreading. *In*: Proc of the IEEE Signal Processing Society Workshop on Multimedia Signal Processing, Los Angeles, CA, p. 221-226.

[8] Almajai, I., Milner, B., Darch, J. Analysis of Correlation between Audio and Visual Speech Features for Clean Audio Feature Prediction in Noise. *School of Computing Sciences University of East Anglia*, Norwich, UK.

[9] Nefian, A. V., Liang, L., Liu, X., Pi, Xi., Murphy, K. (2002). Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP Journal on Applied Signal Processing*, p. 1-15.

[10] Abbasi, A. R., Ahmad, N. Urdu Viseme Identification. p. 68-71.

[11] Lucey, S. (2002). Audio-visual Speech Processing. Phd thesis, *Queensland University of Technology*.

[12] Rabiner, R. (1978). On creationg refrence templates for speaker independant recognition of isolated words. *In*: the Proc of the IEEE 26 (1).

[13] Ajmera, P. K., Jadhav, D. V., Holambe, R. S. (2011). Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. *Elsevier Pattern Recognition*, 44, p. 2749-2759.

[14] Hagen, R., Paksoy, E., Gersho, A. (1999). Voicing-specific LPC quantization variable-rate speech coding. *IEEE Transl*, 5 (5).

[15] Zahoriana, S. A., Hongbing, H. (2008). A spectral/temporel method for robust fundamental frequency tracking. *Journal of Acoustical Society of America*, p. 4559-4571.

[16] Cappellini, V., Re, Ded, E. (1985). Image data compression by the discrete cosine transform. *Journal of Mathematics and Computers in Simulation*, 27, p. 599-608.

[17] Britanak, V. (1994). On the discrete cosine transform computation. *Elsevier Journal of Signal Processing*, 40, p. 183-194.

[18] Chelali, F. Z., Djeradi, A. (2011). Primary research on Arabic visemes, analysis in space and frequency domain. *Internationl Journal of Mobil Computing and Multimedia Communication*, 3 (4).

[19] Liu, Z., Liu, C. (2008). Fusion of the complementary Discrete Cosine Features in the YIQ color space for face recognition. *Elsevier Journal of Computer Vision and Image Understanding*, 111, p. 249-26.

[20] Jadhav, D. V., Holambe, R. S. (2009). Rotation, illumination invariant polynomial kernel Fisher discriminant analysis using Radon and discrete cosine transforms based features for face recognition. *Elsevier Journal of Pattern Recognition Letters*, 31, p. 1002-1009.

[21] Roma, N., Sousa, L. (2011). A tutorial overview on the properties of the discrete cosine transform for encoded image and video processing. *Elsevier Journal of Signal Processing*, 91, p. 2443-2464.

[22] Briley, P. M. , Breakey, C., Krumbholz, K. (2007). Evidence for Pitch Chroma Mapping in Human Auditory Cortex. *Cerebral Cortex Advance Access Oxford Journals*, p. 2387-2399.

[23] Park, H. W., Khil, A. R., Bae, M. J. (2012). Pitch Detection Based on Signal-to-Noise-Ratio Estimation and Compensation for Continuous Speech Signal. Springer-Verlag Berlin Heidelberg, p. 767-774.

[24] Zhao, W. W., Ogunfunmi, T. (1999). Formant and Pitch Detection Using Time-Frequency Distribution. *International Journal of Speech Technologie*, 3, 35-49.

[25] Chelali, F. Z., Djeradi, A. (2012). Visual speech analysis, Application to Arabic phonemes. SEDEXS'12, Settat, Morroco.

[26] Middleton, G. (2003). Pitch detection algorithms. Produced by The Connexions Project and licensed under the Creative Commons Attribution License, p. 1-6.

[27] Tubach, J. P. (1989). La parole et son traitement automatique. Calliope, Masson.

[28] Cui, Y., Takaya, K. (2005). Recognition of syllables in a continous stream of speech by parcor parameters of linear predictive vocodor. *IEEE, CCECE/CCGEI*, Saskatoon.

[29] Mammone, R. J., Zhang, X., Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. *IEEE Signal Processing Magazine*, 13, 58-70.