Audiovisual Speaker Identification Based on Lip and Speech Modalities

Fatma Chelali and Amar Djeradi
Faculty of Electronics Engineering and Computer Science, University of Science and Technology Houari Boumedienne, Algiers

Abstract: *In this article, we present a bimodal speaker identification method, which integrates both acoustic and visual features and where the two audiovisual stream modalities are processed in parallel. We also propose a fusion technique that combines the two modalities to make the final recognition decision. Experiments are conducted on an audiovisual dataset containing the 28 Arabic syllables pronounced by ten speakers. Results show the importance of the visual information that is provided by Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) in addition to the audio information corresponding to the Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Predictive (PLP). Furthermore, some artificial neural networks such as Multilayer Perceptron (MLP) and Radial Basis Function (RBF) were investigated and tested successfully in this dataset by presenting good recognition performances with serial concatenation for the acoustic and visual vectors.*

Keywords: *Audiovisual speaker recognition, DCT, DWT, PLP, MFCC.*

Received April 13, 2014; accepted December 16, 2014

1. Introduction

Human speech is bimodal in nature: Audio and visual. While the audio speech signal refers to the acoustic waveform produced by the speaker, the visual speech signal refers to the movements of the lips, tongue, and other facial muscles of the speaker. Such bimodality has two aspects, the production and the perception. Speech is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs, including the nasal cavity, tongue, teeth, velum, and lips [1].

A lot of studies have been developed to demonstrate the importance of the use of visual information in addition to the audio information in different domains like: Speaker identification/verification, speech analysis and synthesis, speech production and speech perception.

The human speech perception is also bimodal: Humans combine audio and visual information in deciding what has been spoken, especially in noisy environments. The visual modality benefit to speech intelligibility in noise has been quantified as far back as in Sumby and Pollack [20].

Audio-only speaker/speech recognition systems are far from being perfect especially under noisy conditions. Furthermore, it is a known fact that the content of speech can be revealed partially through lip-reading. Performance problems are also observed in video-only speaker/speech recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions may have detrimental effects [2, 3]. Hence, robust solutions for both speaker and speech recognition should employ multiple modalities, such as audio, lip texture and lip motion in a unified scheme [4].

A number of experiments on humans have demonstrated the role of vision in speech identification, particularly in acoustical noise. The effect of additive noise in Automatic Speaker Recognition system (ASR), that is a crucial task in real life applications, has been studied by a lot of researchers [5]. Sheela and al demonstrate the robustness of the ASR system; by using a robust acoustic feature for representation of speaker and an efficient modelling scheme to yield good recognition accuracy [6].

Multimodal information fusion is a challenging problem, with the primary difficulties lie in the identification of the inherent relationship between different modalities, and the design of a fusion strategy which can effectively utilize and integrate the complementary information provided in different channels [7].

The fusion of multimodal information is usually performed at three different levels: Data/feature level, score level, and decision level. In general, information fusion at score or decision levels are more effective given that the associated multimodality data are uncorrelated, i.e., the semantics derived from individual modalities are statistically independent [7].

In our work, we will demonstrate that the fusion of bimodal information at feature level, known as earlier fusion, using the cepstral analysis Mel Frequency Cepstra Coefficients (MFCC) and Perceptual Linear

Predictive (PLP) combined with the frequency analysis Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) can improve the Recognition Rate (RR) of the speaker identification system depending on the Arabic syllables. We develop a speaker identification system using the two modalities and we analyze the score obtained for each modality and for the bimodal system using the feature level fusion.

2. Description of the System

Figure 1 presents the Audio Visual speaker Recognition System depending on Arabic Syllables (AVRS-AS) that consists of a face recognition module based on the lip image, a speaker recognition module and a fusion module. The two recognition modules represent unimodal systems, an audio-video camera, and a database of facial and speaker features assigned for each known individual. A bimodal speaker recognition system will be studied to recognize a person using the lip image and the audio file. The speech signals concern the Arabic syllables.

The audio and visual streams are processed in parallel by its corresponding module for feature extraction: MFCC and PLP coefficients for the audio modality and the DCT/DWT coefficients for the visual modality. The feature vectors can be concatenated into a single feature vector, which represent the person's identity, the fusion module combines the representative vectors with serial fusion to decide the corresponding speaker. A Multilayer Perceptron (MLP) and Radial Basis Function (RBF) classifier is used then for the training features for each modality and for the bimodal information.

The audio and visual streams are processed in parallel by its corresponding module for feature extraction, the cepstral analysis MFCC and PLP for the acoustic modality, DCT and DWT for the visual modality. The fusion module combines the two feature vectors extracted from the two modalities to make the final recognition decision. In addition, a recognition decision is made separately for each modality.

The corpus is a repetition of the 28 Arabic phonemes spoken by ten native Algerian speaking male and female subjects. A database of facial and speaker features is assigned for one of each known individuals. The recordings were made on camera canon video with 25 frames of size 576*720 pixels per second at the speech communication and signal processing laboratory in Algiers and the data were transferred into computer through IEEE 1394 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. For every phoneme, 20 different samples were recorded by ten subjects (Algerian university students: Fazia, Naima, Zohir, Amine, Halim, Nabil, khadidja, Dalila, Mohamed and Hanane). This gives 200 versions of each of 28 Arabic phonemes in our database. The database includes 5600 face images from ten different subjects. The speech signals are acquired during different sessions with a sampling frequency of 22 KHz.

Arabic phonemes contain two distinctive classes, which are named pharyngeal and emphatic phonemes. These two classes can be found only in Semitic languages. The allowed syllables in Arabic language are: CV, CVV, CVC, CVVC, CVCC and CVVCC where V indicates a (long or short) vowel while C indicates a consonant. Arabic utterances can only start with a consonant [17].

Arabic syllables can be classified as short or long. The CV type is a short one while all others are long.

The corpus consists of 20 repetitions of every phoneme (syllable: Phoneme with short vowel) produced by each speaker.

Our article presents a speaker identification system depending on Arabic syllables. A lot of studies were conducted for Arabic language; a state of the art was well presented in the work [30].

The facial image is then reduced to the lip area which will parameterize each individual. For each frame the lip area is manually located with a rectangle of size proportional to 120*160 and centred on the mouth, and converted to gray scale.

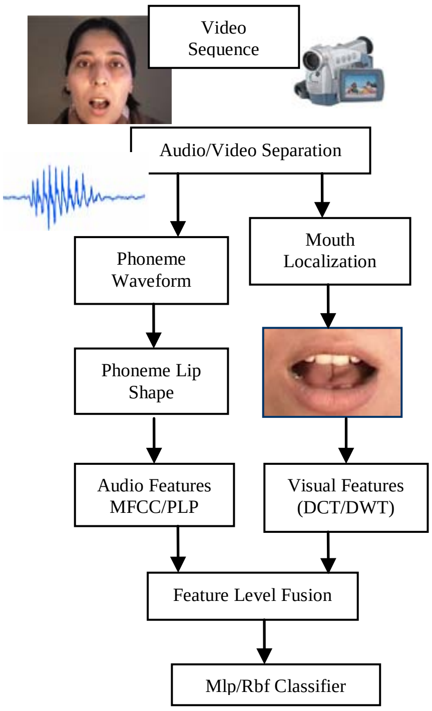


Figure 1. Overview of bimodal speech/speaker recognition.

3. Audio Modality Analysis

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), MFCC, PLP coefficients. Speech feature extraction is one of the important blocks of speaker recognition problem. We present in the following paragraph the theory of speech extraction based on the cepstral analysis MFCC and PLP.

3.1. MFCC Feature Extraction

In the first step, the continuous speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* (*M*<*N*). Typical values for *N* and *M* are *N*=256 and *M*=100 [18, 20]. Next processing step is windowing. If we define the window as *w*(*n*), 0≤*n*≤ *N*-1, where *N* is the number of samples, then the result of windowing is the signal [18, 20].

$y(n) = x(n)\, w(n),\ 0 \le n \le N\text{-}1$

Among all the above, we used hamming window method most to serve our purpose for ease of mathematical computations, which is described as:

$$w(n) = \begin{cases} \alpha + (1-\alpha).\cos(2\pi n / N), & n \prec \dfrac{N}{2} \\ 0 & otherwise \end{cases} \tag{1}$$

If *α*= 0.54, we define then the Hamming window.

$$w(n) = W_{hamming} = 0.54 - 0.46 * cos(\frac{2\pi n}{N}) \quad for\ 0 \le n \le N - 1 \tag{2}$$

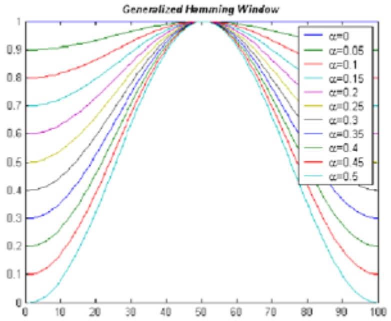


Figure 2. The generalized Hamming window.

The Fast Fourier Transform (FFT) is then applied, it converts a signal from the time domain into the frequency domain, which is defined on the set of *N* samples {*Xn*}, as follow:

$$X_n = \sum_{k=0}^{N-1} x_k . e^{-2\frac{j\pi kn}{N}} \quad n = 0, 1, ..., N - 1 \tag{3}$$

Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus, for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. Therefore, we can use the following approximate formula to compute the Mels for a given frequency *f* in Hz [18, 20].

$$F_{MEL} = 2595.log_{10}(1 + \frac{f_{HZ}}{700}) \tag{4}$$

One approach for simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale. That filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel frequency interval. The modified spectrum of S(ω) thus consists of the output power of these filters when S(ω) is the input. The number of Mel spectrum coefficients, K, is typically chosen as 20 [18, 20].

In the final step, the log Mel spectrum has to be converted back to time. The result is called the MFCCs. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. The MFCCs may be calculated using this equation [18, 20]:

$$\tilde{c}(p) = \sum_{m=1}^{M_l} log(\tilde{a}m) cos\left[p(m - \frac{1}{2})\frac{\pi}{M_l}\right] \tag{5}$$

Where *n*=1, 2, ..., K, *p*=1, 2, ..., M_l, K denotes the number of Mel cepstrum coefficients, is chosen as 20 (number of filters). This set of coefficients is called an acoustic vector. Note that we exclude the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal, which carried little speaker specific information [18, 20].

3.2. PLP Feature Extraction

PLP analysis was proposed by Hermansky [13]. PLP analysis is similar to LPC, except that the PLP technique also uses three concepts from the psychophysics of hearing. These three concepts are the critical-band spectral resolution, equal loudness curve, and intensity-loudness power law.

In the PLP technique, several well-known properties of hearing are simulated by practical engineering approximations, and the resulting auditory like spectrum of speech is approximated by an autoregressive all-pole model.

The speech segment is weighted by the Hamming window described by Equation 2. The typical length of the window is about 20ms.

The real and imaginary components of the short-term speech spectrum are squared and added to get the short term power spectrum [13].

$$P(w) = Re[s(w)]^2 + Im[s(w)]^2 \tag{6}$$

Critical-band spectral resolution, the spectrum *P*(*w*) is warped along its frequency axis *w* into the bark frequency *Ω* by:

$$\Omega(w) = 6ln\left\{w / 1200\pi + [(w / 1200\pi)^2 + 1]^{0.5}\right\} \tag{7}$$

The resulting warped power spectrum is then convolved with the power spectrum of the simulated critical-band masking curve *Ψ*(*Ω*). This step is similar to spectral processing in Mel cepstral analysis, except for the particular shape of the critical-band curve. In PLP technique, the critical-band curve is given by Equation 8:

$$\Psi(\Omega)=\begin{cases}0 & for\ \Omega < -1.3,\\ 10^{2.5(\Omega+0.5)} & for\ -1.3 \le \Omega \le -0.5,\\ 1 & for\ -0.5 \le \Omega \le 0.5,\\ 10^{-1.0(\Omega-0.5)} & for\ 0.5 \le \Omega \le 2.5,\\ 0 & for\ \Omega \succ 2.5.\end{cases} \tag{8}$$

The discrete convolution of $\Psi(\Omega)$ with (the even symmetric and periodic function) $P(w)$ yields samples of the critical-band power spectrum.

$$\theta(\Omega_i)=\sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{9}$$

The convolution with the relatively broad critical-band masking curves $\Psi(\Omega)$ significantly reduces the spectral resolution of $\theta(\Omega)$ in comparison with the original $P(w)$. This allows for the down-sampling of $\theta(\Omega)$ [13].

Equal-loudness preemphasis the sampled $\Theta[\Omega(w)]$ is preemphasized by the simulated equal-loudness curve:

$$\Xi\left[\Omega(w)\right]=E(w)\left[\Theta(w)\right] \tag{10}$$

The function $E(w)$ is an approximation to the non equal sensitivity of human hearing at different frequencies and simulates the sensitivity of hearing at about the 40-dB level [13].

The particular approximation is adopted from makhoul and Cosell (1976) and is given by:

$$E(w)=\left[(w^2+56.8*10^6)w^4\right]/\left[\begin{matrix}(w^2+6.3*10^6)^2 *\\ (w^2+0.38*10^9)\end{matrix}\right] \tag{11}$$

Finally, the values of the first (0bark) and the last (Nyquist frequency) samples (which are not well found) are made equal to the values of their nearest neighbors. Thus, $\Xi[\Omega(w)]$ begins and ends with two equal-valued samples [13].

Intensity-loudness power law the last operation prior to the all-pole modelling is the cubic-root amplitude compression.

$$\Phi(\Omega)=\Xi(\Omega)^{0.33} \tag{12}$$

This operation is an approximation to the power law of hearing (Stevens1957) and simulates the nonlinear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-loudness preemphasis, this operation also reduces the spectral amplitude variation of the critical band spectrum so that the following all-pole modelling can be done by a relatively low model order [13].

Autoregressive modelling in the final operation of PLP analysis, $\Phi(\Omega)$ is approximated by the spectrum of an all-pole model using the autocorrelation method of all-pole spectral modelling. We give here only a brief overview of its principle: The Inverse DFT (IDFT) is applied to $\Phi(\Omega)$ to yield the autocorrelation function dual to $\Psi(\Omega)$. The first M+1 autocorrelation values are used to solve the Yule-Walker equations for the autoregressive coefficients of the M^{th}-order all-pole model. The autoregressive coefficients could be further transformed into some other set of parameters of interest, such as cepstral coefficients of all-pole model [13].

Feature extraction (PLP coefficients) this operation is done for every individual and for all the phonemes used. For good speaker/speech recognition accuracy, nine PLP coefficients per frame were used.

Therefore, dimensionality reduction or speech parameterization is a very important step which will greatly improve the performance of the speaker recognition system.

The input matrix (the voice print matrix) has a dimension of M real values corresponding to 9 coefficients calculated for the total frames of each signal.

4. Visual Modality Analysis

Various sets of visual features for automatic speechreading or audio-visual speaker recognition have been proposed in the literature over the last 20 years. In general, they can be grouped into three categories: Video pixel (or, appearance) based ones; lip contour (or, shape) based features; and features that are a combination of both appearance and shape.

In our work, the Region Of Interest (ROI) containing the speaker’s mouth of the different syllabus is considered as the most relevant information used for speaker recognition dependant on syllables. The corresponding region was extracted; the result images of 120*160 pixels of the ROI are stored in a BMP format (original format). The DCT and the DWT have been employed for lip parametrization.

4.1. The Discrete Cosine Transform

The DCT separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image’s visual quality). The DCT is similar to the discrete Fourier transform: It transforms a signal or image from the spatial domain to the frequency domain. The DCT coefficients are real. The general equation for a 2D (N by N image) DCT is defined by the following equation:

$$C(u,v)=\alpha(u)\alpha(v)\sum_{y=0}^{N-1}\sum_{x=0}^{N-1} f(x,y)\cos\left(\frac{\pi(2x+1)u}{2N}\right)\cos\left(\frac{\pi(2y+1)v}{2N}\right) \tag{13}$$

Where u, v=0, …, N-1.

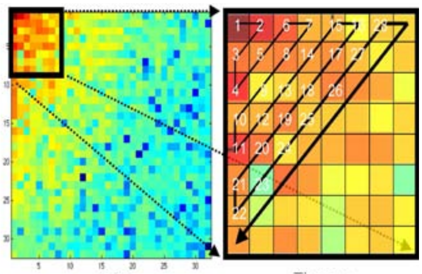
The corresponding inverse transform is:

$$f(x,y)=\sum_{v=0}^{N-1}\sum_{u=0}^{N-1}\alpha(u)\alpha(v)C(u,v)cos\left(\frac{\pi(2x+1)u}{2N}\right)cos\left(\frac{\pi(2y+1)v}{2N}\right)$$

$$\alpha(u)=\begin{cases}\sqrt{\frac{1}{2}} & pour\ u=0\\ 1 & pour\ u=1,\dots,N-1\end{cases} \tag{14}$$

Where $x, y=0, \ldots, N-1$.

For most images, much of the signal energy lies at low frequencies; these appear in the upper left corner of the DCT. Compression is achieved since the lower right values represent higher frequencies, and are often small enough to be neglected with little visible distortion as shown in Figure 3.



DCT Zig-zag

Figure 3. Extraction of DCT coefficients and the zigzag detection.

The role of the DCT is to reduce the dimensionality of the working space. If a face image can be considered as a m×n matrix, the DCT can be seen as a one-to one mapping for N-point vectors, or called it m×n feature coefficient matrix, between the time and the frequency domains and most of the energy is in the upper left corner. The coefficient, which is directly related to the average value of the time-domain block, is often called the DC coefficient, and the remaining coefficients of a block are called AC coefficients [14].

Joo *et al.* [16] investigate the illumination invariant property of the DCT by discarding its several low-frequency coefficients. It is well-known that the first DCT coefficient represents the DC component of an image which is solely related to the brightness of the image. Therefore, it becomes DC-free (i.e., zero mean) and invariant against uniform brightness change by simply removing the first DCT coefficient [8, 9, 16].

Some of the low-frequency components of DCT also account for the large area non uniform illumination variations. Consequently, the nonuniform illumination effect can be reduced by discarding several low-frequency DCT coefficients [16].

4.2. Discrete Wavelet Transform

DWT is a suitable tool for extracting image features because it allows the analysis of images on various levels of resolution. Typically, low-pass and high-pass filters are used for decomposing the original image.

The low-pass filter results in an approximation image and the high-pass filter generates a detail image. The approximation image can be further split into a deeper level of approximation and detail according to different applications [25].

Suppose that, the size of an input image is $N\times M$. At the first filtering in the horizontal direction of down-sampling, the size of images will be reduced to $N\times(M/2)$. After that, further filtering and down-sampling in the vertical direction, four subimages are obtained, each being of size $(N/2)\times(M/2)$. The outputs of these filters are given by Equations 15 and 16 [25].

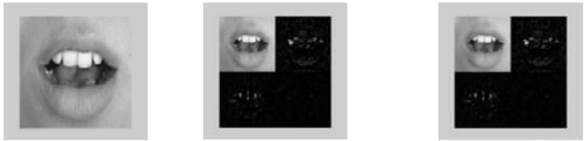
$$a_{j+1}[p]=\sum_{n=-\infty}^{n=+\infty} l[n-2p]a_j[n] \tag{15}$$

$$d_{j+1}[p]=\sum_{n=-\infty}^{n=+\infty} h[n-2p]a_j[n] \tag{16}$$

Where $l[n]$ and $h[n]$ are coefficients of low-pass and high-pass filters, respectively.

Accordingly, we can obtain four images denoted as LL, HL, LH and HH. The LL image is generated by two continuous low-pass filters; HL is filtered by a high-pass filter first and a low-pass filter later; LH is created using a low-pass filter followed by a high-pass filter; HH is generated by two successive high-pass filters as described in [25].

Every subimage can be decomposed further into smaller images by repeating the above procedure. The main feature of DWT is the multiscale representation of a function. By using the wavelets, a given image can be analyzed at various levels of resolution. Since the LL part contains most important information and discards the effect of noises and irrelevant parts, we adopt the LL part for further analysis in this paper. We extract features from the LL part of the second-level decomposition. The reasons are the LL part keeps the necessary information and the dimensionality of the image is reduced sufficiently for computation at the next stage.



a) Lip image. b) DWT (db) representation. c) DWT (Haar) representation.

Figure 4. First level decomposition by Haar and daubechies 4 wavelet.

One major advantage afforded by wavelets is the ability to perform local analysis that is, to analyze a localized area of a larger signal. A lot of tests have been carried; we adopt the Haar wavelet with second level decomposition. The Haar-DWT feature vector is assigned for each individual and for the whole database containing lip representation of the 28 Arabic syllables.

5. Unimodal Speaker Recognition System

The classifier used in our system is a MLP with only one hidden layer and a RBF neural network. First, our speaker recognition system is tested for the entire database. Comparison is done for each classifier. The characteristic vectors are obtained by the DCT and the DWT. The following paragraph presents a brief description of the MLP classifier.

5.1. Multilayer Perceptron for Face Recognition

MLP neural networks are feed-forward and use the back-propagation algorithm. We imply feed-forward networks and back-propagation algorithm (plus full connectivity). While inputs are fed to the ANN forwardly, the 'Back' in back-propagation algorithm refers to the direction to which the error is transmitted.

The multilayer perceptron is one of the most popular neural network models for solving pattern classification and image classification problems, it consists of several layer of perceptrons. Nodes in the *i* layer are connected to nodes in the (*i*+1) layer through suitable weights.

Learning process in backpropagation requires providing pairs of input and target vectors. The output vector y of each input vector is compared with target vector *d*. In case of difference the weights are adjusted to minimize the difference. Initially random weights and thresholds are assigned to the network [10].

These weights are updated every iteration in order to minimize the cost function or the mean square error between the output vector and the target vector.

The logistic function $f(x) = \frac{1}{1+exp(-x)}$ which maps the real numbers into the interval [-1 +1] and whose derivative, needed for learning, is easily computed {f' (x) = $f(x)$ [1- (x)]}. The reason for its popularity is the ease of computing its derivative [9].

Neural networks are adaptive statistical devices. This means that they can change iteratively the values of their parameters (i.e., the synaptic weights) as a function of their performance. These changes are made according to learning rules which can be characterized as supervised (when a desired output is known and used to compute an error signal) or unsupervised (when no such error signal is used).

Backpropagation consists of measuring the error term between target output $d(n)$ and the observed output $y(n)$ [19].

For better performance it can be useful to combine natural gradient learning with some standard super linear optimization algorithm. One such algorithm is the nonlinear Conjugate Gradient (CG) method. The CG method is a standard tool for solving high dimensional nonlinear optimization problems. During each iteration of the CG method, a new search direction is generated by conjugation of the residuals from previous iterations. With this choice the search directions form a Krylov subspace and only the previous search direction and the current gradient are required for the conjugation process, making the algorithm efficient in both time and space complexity [18].

5.2. Radial Basis Function Neural Network

An RBF neural network can be considered as a mapping: Rr→Rs.

Let $P \in R^r$ be the input vector and $C_i \in R^r (1 \le i \le u)$ be the prototype of the input vectors [16]. The out put of each RBF unit is as follows:

$$R_i(P) = R_i(\|P - C_i\|) \quad i = 1, \dots, u \tag{17}$$

Where $\|.\|$ indicates the Euclidean norm on the input space. Usually, the Gaussian function as shown in Figure 5 is preferred among all possible radial basis functions due to the fact that it is factorizable.

$$R_i(P) = exp\left[- \frac{\|P - C_i\|^2}{\sigma_i^2} \right] \tag{18}$$

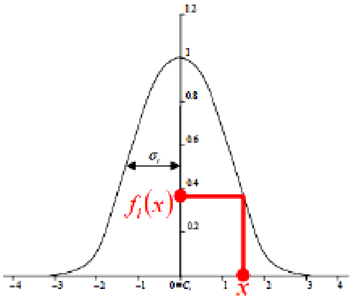


Figure 5. Radial basis function.

Where σ_i is the width of the ith RBF unit. The j^{th} output $y_i(P)$ of a neural network is:

$$y_i(P) = \sum_{i=1}^{u} R_i(P) * w(j,i) \tag{19}$$

Where R_0=1, $w(j, i)$ is the weight or strength of the i^{th} receptive field to the j^{th} output. In order to reduce the network complexity, the biais is not considered in the analysis.

We can see from Equations 18 and 19 that the outputs of an RBF neural classifier are characterized by a linear discriminant function. They generate linear decision boundaries (hyperplanes) in the output space. Consequently, the performance of an RBF neural classifier strongly depends on the separability of classes in the u-dimensional space generated by the nonlinear transformation carried out by the u RBF units [16].

According to cover's theorem on the separability of patterns where in a complex pattern classification problem cast in a high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space, the number of Gaussian nodes u>=r, where r is the dimension of input space. On the other hand, the increase of Gaussian units may result in poor generalization because of overfitting, especially, in the case of small training sets [16]. It is important to analyze the training patterns for the appropriate choice of RBF hidden nodes.

Geometrically, the key idea of an RBF neural network is to partition the input space into a number of subspaces which are in the form of hyperspheres. Accordingly, clustering algorithms (k-means clustering, fuzzy-means clustering and hierarchical clustering) which are widely used in RBF neural networks are a logical approach to solve the problems [16].

However, it should be noted that these clustering approaches are inherently unsupervised learning algorithms as no category information about patterns is used.

While considering the category information of training patterns, it should be emphasized that the class memberships are not only depended on the distance of patterns, but also depended on the Gaussian widths [16].

6. Speaker Recognition Results

6.1. Acoustical Speaker Recognition

This operation is realized for each syllable and for the entire database. Nine coefficients PLP and twenty MFCC per frame have been choice to characterized our acoustical signals. It has been proved that the use of cepstral vectors ameliorate significantly the performance of speaker recognition system. This operation is done for every individual and for the 28 syllables used.

A neural network is used to classify each frame as belonging to one of the ten specific speakers. The network has a three-layered architecture and is trained using the back-propagation algorithm. The number of the input nodes is equal to the size of the input vectors. The number of the output nodes is equal to the number of the registered to the system speakers. Finally, the number of the hidden nodes is chosen by the user.

The network will receive an input layer having a matrix of size (C*50); fifty corresponds to five training signals for the ten speakers. C corresponds to the total coefficients calculated for the total analyzed frames. The frame size depends on the syllable analyzed.

The features test matrix is defined with variables called target, the target matrix has the same dimension as the training matrix. The network is trained to output a 1 in the correct position of the output vector and to fill the rest of the output vector with 0's.

Twenty eight neural networks were constructed for each specified phoneme. All the NNs trained present fast convergence and the training process terminated within 300 or 500 epochs, with the Summed Squared Error (SSE) reaching the pre-specified goal (10-4).

We used log-sigmoid functions as a transfer function at all neurons (In hidden layer and output layer). Log-sigmoid is ideal for our system.

In order to show the importance of processing elements, we trained our MLP classifier with variable hidden unit from 5 to 45. For a small number of neurons (5 to 10) in the hidden layer we observed large MSE, so low accuracy. The MLP generalize poorly. After ~50 neurons, MSE came back to the levels of a system with only 5 neurons in the hidden layer. by adding more and more units in the hidden layer the training error can be made as small as desired but generally each additional unit will produce less and less benefit. When too many neurons, poor performance is a direct effect of over fitting. The system overfits the training data and does not perform well on novel patterns.

For the most syllables used in this work, 30 to 45 hidden neurons were used to accomplish this task and permit us to have good RR for the total speakers and for each syllable. Therefore, dimensionality reduction or speech parameterization is a very important step which will greatly improve the performance of the speaker recognition system.

The MLP and the RBF classifier were trained several times in order to find the optimal topology or architecture.

Our system achieves between 85% to 100% in identifying the correct speaker. Figure 6 shows the global RR obtained for each syllable .The choice of feature extraction is a very important step, and then the application of a non linear classifier such as MLP or RBF permits us to obtain good results for a developpement of a speaker identification process.

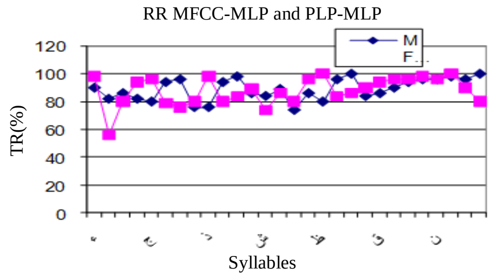


Figure 6. RR for MFCC-MLP and PLP-MLP.

It's clear that from the most syllables, RR obtained with the MFCC characterization is greater (superior) than the PLP characterization. We decide then to use the MFCC cepstral coefficients for the acoustical speaker recognition and also for the bimodal system.

By comparing the MFCC-MLP and the MFCC-RBF, results demonstrate a little variation between the score obtained. The RBF neural network has a faster convergence whereas, the MLP present a modified architecture in the training phase.

The MLP classifier gives better RR; the network was trained several times in order to find the optimal topology or architecture.

From our results, it's clear that recognition rate obtained for MLP classifier is appreciatively greater than the RBF classifier for the most syllables studied. We retain then the MLP classifier for the visual and the bimodal system.

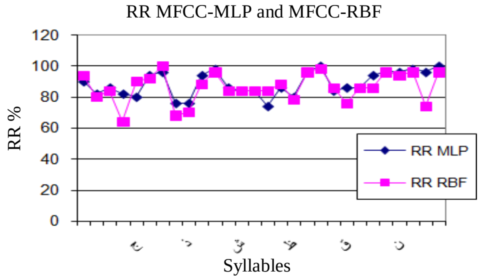


Figure 7. Comparison between MFCC-MLP and MFCC-RBF.

6.2. Visual Speaker Recognition

6.2.1. Lip Image Modality

An input image of 200x 180 pixels is converted in a matrix of size 140*130 by a manual detection in order to reduce the spatial representation of the image and detect only the face image. The DCT coefficients are then calculated and stored in a matrix.

After DCT, majority of transformation energy is concentrated on the coefficients of low frequency which reflect the main image and are bigger while ones of high frequency are smaller. After DCT, the part of low frequency is displaced in left-down angle and the part of high frequency is located in right-up angle among the matrix. The dimension of data that are transformed by DCT is very high. We keep only 100 features or DCT components in this paper. Figure 8 shows that the magnitudes of the first 100 coefficients are relevant to characterize lip image. Only 300 coefficients from 14400 corresponding to the total DCT calculated for the whole lip image converted to 120*120 pixels.

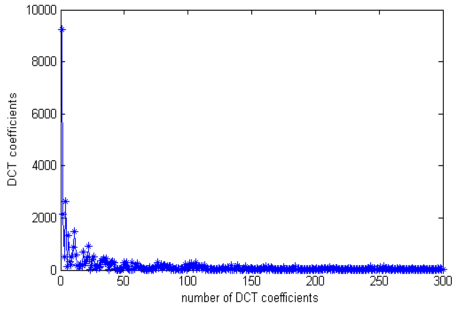


Figure 8. DCT coefficients variation.

For the 100 training images we'll get a matrix of (100*100) containing the 100 relevant DCT coefficients of the 100 images.

After calculating the DCT transform, these feature vectors are calculated for the training set and then used to train the neural network, this architecture is called DCT-MLP.

The dimension of DCT feature vectors fed into the MLP neural networks is essential for recognition. In [16], experimental results showed that the best results are achieved when the feature dimension is 25-30. If the feature dimension is too small, the feature vectors do not contain sufficient information for recognition. However, it does not mean that more information will result in higher RR. It has been indicated that if the dimension of the network input is comparable to the size of the training set, the system is liable to overfitting and result in poor generalization [16].

Moreover, the addition of some unimportant information may become noise and degrade the performance. Their experiments also showed that the best RRs are achieved when the feature dimension is about 30 [16].

From our experiments, the best performance is obtained when 60-100 DCT coefficients are used in our recognition system. These vectors are fed to the MLP and are trained. The MLP with single hidden layer gives better performance. The number of processing elements in the hidden layer is varied.

Various transfer functions were tested for training the network and average minimum MSE on training (MSEA) is measured; log sigmoid is the most suitable transfer function. The MLP neural network is trained using learning rules namely CG. TRAINSCG is a network training function that updates weight and bias values according to the scaled CG method. Finally, network is tested on training and testing dataset.

The RR is 96% obtained by retaining 100 DCT components and using an MLP with 30 hidden neurons. A complete training run typically takes about 2~3 minutes on a Pentium4 PC with a 775MHz CPU.

The performance of DCT-MLP approach is evaluated (and compared) by using the RR standard defined by: $R(i)=\frac{N_i}{N}$

Where N is the total number of test sample test, N_i is the number of test images recognized correctly.

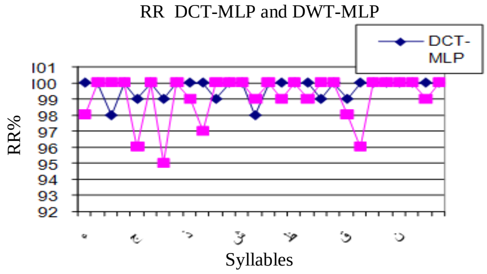


Figure 9. RR for DCT-MLP and DWT-MLP.

From our experiments, the DCT-MLP classifier achieves high RR as well as high training and recognition speed. It can be seen that the DCT-MLP algorithm exhibits better performance than DWT-MLP, results obtained are also better than the DCT-RBF classifier.

We conclude that the MLP neural network performs well than the RBF classifier for the facial database.

We have also demonstrate that the visual modality ameliorate the RR of the audiovisual speaker recognition compared to the audio modality where the acoustic recognition varied from 75% to 100%, the results obtained depend on a lot of parameters like:

Noise, recording conditions, speed of elocution and emotional state, etc.

The speaker characterization is a very important step in the development of audiovisual speaker identification. It permits to construct a set of pattern vectors that are uncorrelated and discriminants, in addition it ameliorate the performance and the robustness of the identification system.

7. Bimodal Speaker Recognition using Lip and Acoustical Information

In the speaker/speech recognition literature, audio is generally modelled by MFCC. However, for lip information, there are several approaches reported in the literature such as texture-based, motion-based, geometry based and model-based. In texture-based approaches, DCT-domain lip image intensity is used as features. Motion-based approaches compute motion vectors to represent the lip movement during speaking [4].

Cetingul *et al.* [4] present in their work a multimodal speaker/speech recognition system that integrates audio, lip texture and lip motion modalities. The aim of their work is to investigate the benefits of inclusion of lip motion modality for two distinct cases: Speaker and speech recognition.

Geometry-based and model-based approaches, in fact, utilize similar processing methods such as active shape models, active contours or parametric models to segment the lip region.

The speaker recognition schemes proposed in [2, 11, 15, 23, 28] are basically opinion fusion techniques that combine multiple expert decisions through adaptive or non-adaptive weighted summation of scores, whereas in [5, 6], fusion is carried out at feature-level by concatenating individual feature vectors so as to exploit the temporal correlations that may exist between audio and video signals [4].

In audio-visual speech recognition [3] authors concatenates audio and lip data, while in [29] unimodal decisions are combined to obtain the fused result. Furthermore, recent works show the success of multistream HMM structures in speech recognition [4].

After the unimodal analysis, we will investigate the audiovisual analysis using the acoustic and the visual feature vectors with serial concatenation. The objective of our work is to design audiovisual speaker recognition depending on Arabic phonemes in which the visual modality has a great importance by analyzing the score obtained. The following paragraph describes the bimodal audiovisual speaker recognition system that integrates the cepstral acoustic parameters MFCC and the visual parameters DCT-DWT.

The design of a multimodal recognition system requires addressing three basic issues: Which modalities to fuse; how to represent each modality with a discriminative and low-dimensional set of features; and how to fuse existing modalities.

Speech content and voice can be interpreted as two different, though correlated, information existing in audio signals. Likewise, video signal can be split into different modalities, such as face/lip texture and lip motion [4].

The second issue, representative feature selection, also includes modelling of classifiers through which each class is represented with a statistical model or a representative feature set. Curse of dimensionality, computational efficiency, robustness, invariance and discrimination capability are the most important criteria in selection of the feature set and the recognition methodology for each modality [4].

As for the final issue, that is, the fusion problem, different strategies are possible: in the so-called “early integration”, modalities are fused at data or feature level, whereas in “late integration” decisions or scores resulting from each unimodal recognition are combined to give the final conclusion. Multimodal decision fusion can also be viewed from a broader perspective as a way of combining classifiers, which is a well-studied problem in pattern recognition [4, 5].

The main motivation for multimodal fusion is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision. Misclassification errors are in general inevitable due to numerous factors such as environmental noise, measurement and modelling errors or time-varying characteristics of signals. A comprehensive survey and discussion on classifier combination techniques can be found in [4].

In our work, we studied the fusion of bimodal information at feature level using the cepstral analysis MFCC and PLP combined with the frequency analysis DCT and DWT can improve the RR of the speaker identification system depending on the Arabic syllables.

The audio and visual streams are processed in serial by its corresponding module for feature extraction, the cepstral analysis MFCC and PLP for the acoustic modality and the DCT and DWT for the visual modality. The fusion module combines the two feature vectors extracted from the two modalities to construct the novel vector AV=[AD; VD] to make the final recognition decision. In addition, a recognition decision is made separately for each modality.

We note that AV defines the audiovisual vector where the audio descriptors AD and the visual descriptors VD are concatenated in a single vector. Figure 10 shows the fusion at feature level, recognition is done by applying a MLP and RBF classifier.

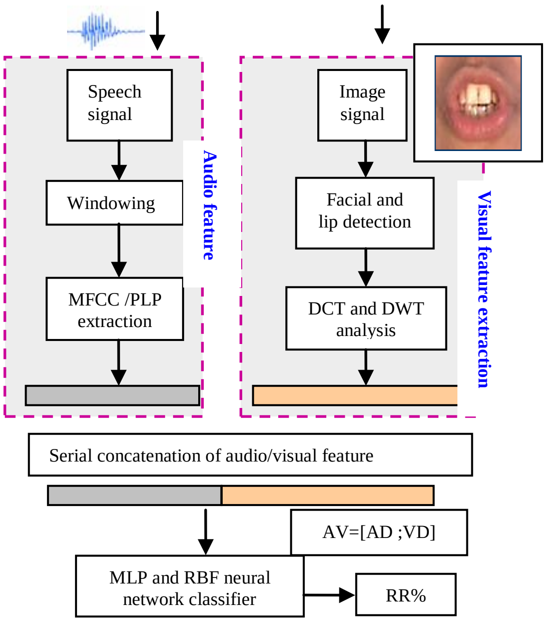


Figure 10. Overview of the bimodal speaker/speech recognition system.

From our experiments, we conclude that the visual modality ameliorate significantly the performance of the audiovisual speaker identification depending of the Arabic syllables compared to the acoustic modality. The visual modality demonstrates the role of vision in speaker identification, particularly in acoustical noise. The combination of the audio and visual streams shows that the resulting system performance is better than the acoustic modality recognizer.

Figure 11 shows that the audiovisual speaker recognition is improved by adding the visual vector descriptor to the acoustic one. The visual modality has a great importance in the design of an audiovisual speaker identification, especially in noisy environments.

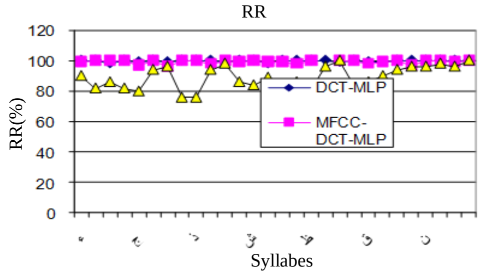


Figure 11. RR for the audio modality (MFCC-PLP), visual modality DCT and the audiovisual modality (MFCC-DCT).

8. Conclusions

In this investigation, a bimodal speaker/speech recognition system has been presented. The proposed approach integrates both acoustic and video features, such as: MFCC, PLP, DCT, and DWT where lip characteristics are directly extracted from the facial modality. Furthermore, a fusion technique combining audio and lip information has been proposed. As described previously, for the visual modality, the DCT and DWT characterization have been used, whereas for the speech modality, the MFCC and PLP characterization were employed to build our mono-modal speaker identification system with Arabic syllables. For the bimodal speaker identification system, a combination of the two audio and video features using artificial neural networks was investigated.

As noticed in the experimental results, good recognition performances were obtained by using the fusion and multimodality techniques. That is, once again we show that the fusion is an interesting approach in pattern recognition, in general, and in speaker/speech recognition more particularly. We demonstrate the benefit of the visual information to the acoustic information for our task.

For future work, another classifier like SVM and HMM could be investigated to ameliorate speech or speaker RR.

References

[1] Abushariah M., Ainon R., Zainuddine R., Elshafei M., and Khalifa O., "Arabic Speaker Independent-Continuous Automatic Speech Recognition Based on Phonetically Rich and Balanced Speech Corpus," *The International, Arab Journal of Information Technology*, vol. 9, no. 1, pp. 84-93, 2012.

[2] Brunelli R. and Falavigna D., "Person Identification Using Multiple Clues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955-966, 1995.

[3] Bregler C. and Konig Y., "Eigenlips for Robust Speech Recognition," *in Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, pp. 669-672, 1994.

[4] Cetingul H., Erzin E., Yemez Y., and Tekalp A., "Multimodal Speaker/Speech Recognition Using Lip Motion, Lip Texture and Audio," *Signal Processing*, vol. 86, no. 12, pp.3549-3558 2006.

[5] Civanlar M. and Chen T., "Password-Free Network Security through Joint Use of Audio and Video," *in Proceedings of SPIE Photonic*, Boston, pp. 120-125, 1996.

[6] Chaudhari U., Ramaswamy G., Potamianos G., and Neti C., "Information Fusion and Decision Cascading for Audio-Visual Speaker Recognition Based on Time-Varying Stream Reliability Prediction," *in Proceedings of the International Conference on Multimedia and Expo ICME*, pp. 9-12, 2003.

[7] Chakraborty P., Ahmed F., Monirul M., Shahjahan M., and Murase K., "An Automatic Speaker Recognition System," *in Proceedings of International Conference on Neural Information*

Processing, Berlin Heidelberg, pp. 517-526, 2007.

[8] Chelali F. and Djeradi A., "Face recognition System based on DCT and Neural Network," *in Proceedings of Artificial Intelligence and Pattern Recognition (AIPR-10)*, pp.13-18, Florida, 2010.

[9] Chelali F. and Djeradi A., "Face Recognition System using Discrete Cosine Transform Combined with MLP and RBF Neural Networks," *International Journal of Mobile Computing and Multimedia Communication (IJMCMC)*, vol. 4, no. 4, pp. 1-35, 2012.

[10] Eleyan A. and Demirel H., PCA and LDA based Neural Networks for Human Face Recognition, INTECH Open Access Publisher, 2007.

[11] Frischholz R. and Dieckmann U., "BioID: a Multimodal Biometric Identification System," *Computer*, vol. 33, no. 2, pp. 64-68, 2000.

[12] Gerasimos P., *Audio-Visual Automatic Speech Recognition: An Overview*, MIT Press, 2004.

[13] Hermansky H., "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[14] Husmeier D., *Neural Networks for Conditional Probability Estimation, Perspectives in Neural Computation*, Springer-Verlag, 1999.

[15] Jourlin P., Luettin J., Genoud D., and Wassner H., "Acousticlabial Speaker Verification," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 853-858, 1997.

[16] Joo M., Chen W., and Wu S., "High-Speed Face Recognition Based on Discrete Cosine Transform and RBF Neural Network," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 679-691, 2005.

[17] Minh D., "An Automatic Speaker Recognition System," *Digital Signal Processing Mini-Project, Audio Visual Communications Laboratory, Swiss Federal Institute of Technology*, Switzerland, pp.1-14, White paper, 1996.

[18] Nocedal J., "Theory of Algorithms for Unconstrained Optimization," *Acta Numerica1*, pp. 199-242, 1992.

[19] Parizeau M., "Le Perceptron Multicouche et Son Algorithme de Retropropagation des Erreurs," Technical Report, 2004.

[20] Sumby W. and Pollak I., "Visual Contribution to Speech Intelligibility in Noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212-215, 1954.

[21] Shivappa S., Trivedi M., and Rao B., "Audiovisual Information Fusion in Human-Computer Interfaces and Intelligent Environments: A Survey," *the IEEE*, vol. 98, no. 10, pp. 1692-1715, 2010.

[22] Sheela K. and Prasad K., "Linear Discriminant Analysis F-Ratio for Optimization of TESPAR and MFCC Features for Speaker Recognition," *Journal of Multimedia*, vol. 2, no. 6, pp. 34-43, 2007.

[23] Sanderson C. and Paliwal K., "Noise Compensation in a Person Verification System Using Face and Multiple Speech Features," *Pattern Recognition*, vol. 36, no. 2, pp. 293-302, 2003.

[24] Satori H., Hiyassat H., Harti M., and Chenfour N., "Investigation Arabic Speech Recognition Using CMU Sphinx System," *The International, Arab Journal of Information Technology*, vol. 6, no. 2, pp. 186-190, 2009.

[25] Senthil G. and Dandapat S., "Speaker Recognition under Stressed Condition," *International Journal of Speech Technology*, vol. 13, no. 3, pp. 141-161, 2010.

[26] Shih F., Chuang C., and Wang P., "Performance Comparisons of Facial Expression Recognition in JAFFE Database," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 3, pp. 445-459, 2008.

[27] Tsuhan C., "Audiovisual Speech Processing, Lip Reading and Lip Synchronization," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9-21, 2001.

[28] Wang Y., Guan L., and Venetsanopoulos A., "Kernel Cross-Modal Factor Analysis for Multimodal Information Fusion," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, pp. 2384-2387, 2011.

[29] Wark T. and Sridharan S., "Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification," *Digital Signal Processing*, vol. 11, no. 3, pp. 169-186, 2001.

[30] Zhang D., Automated Biometrics, Springer US, 2000.

Fatma Zohra Chelali received her engineering degree in Electronic engineering from University of science and technology Houari Boumedienne of Algiers; ALGERIA (USTHB) in 1994. She works as Assistant teacher in the high school of Aeronautical Technicians (Ecole supérieure des techniciens de l'Aéronautique ESTA) from 1997 to 2008, she received an academic certificate for teaching from the Algerian institute of management (Institut international de management d'Alger) in 1999. She spent a year of post graduation from 2002 to 2003. Then, she received a "magister" degree in speech communication in 2006 and Doctorate degree in speech communication and signal processing laboratory (LCPTS, USTHB,Algiers) in 2012, the subject of her thesis treats audiovisual speaker recognition applied to Arabic phonemes. She teaches courses with telecommunications department on Electromagnetic waves, transmission lines and digital electronics since 2007 in Electronic engineering and computer science Faculty, university of science and technology (USTHB). Her interests include audiovisual analysis and recognition, pattern recognition and classification, speech and image processing.

Amar Djeradi received his engineering degree in Electronics in 1984, his magister degree in applied electronics, and Doctorate degree in 1992.He teaches since 1985 in different modules for graduation and post graduation such as Electronics, Television, digital electronics, and principal functions of electronics, pattern recognition, and human-machine communication. His current research interests are in the area of speech communication, human-machine Communication, multimodal interfaces and signal analysis.