Evolutionist approach and MFCC modeling for Arabic automatic recognition

Maouche Fadila [1], Benmohamed Mohamed [2]
[1] Oum El Bouaghi University, Algeria.
[2] Constantine University, Algeria.
Mifad_5@yahoo.fr
ibnm@yahoo.fr

Abstract: In this article, we suggest a system for automatic recognition of isolated Arabic words, it is a multi-speaker system, even independent of speaker and robust in a noisy environment, it uses a genetic algorithm for recognition, and the Mel frequency cepstral coefficients (MFCC) to modelise the speech signal, it was implemented with the matlab7 platform language. A new mutation method (injection mutation) is proposed and used in the genetic algorithm. To evaluate the performance of the system, we have made an oral corpus that represents the most Arabic language characteristics, it could be used by other researchers to test and validate their systems working on Arabic language.

Keywords: Automatic speech recognition, genetic algorithm, Arabic language, Mel frequency cepstral coefficients (MFCC), language corpora.

Introduction

Nowadays, automatic speech recognition's system are increasingly widespread and used in very different acoustic conditions, and by very different speakers, but despite the spectacular progress, the ideal system does not exist yet, and to overcome the current performance of ASR systems, many works are carried out in various laboratories in the world. The most studied methods in recent years are those inspired by nature as genetic algorithms.

The Arab States have at least 300 million people, in addition to Arabic minorities scattered throughout the world, despite this large number, Arabic's speech processing is still in the beginning, that's why we suggest a system for automatic recognition of isolated Arabic words. This system uses a genetic algorithm for recognition, and MFCC coefficients to modal the speech signal.

1. Architecture of the suggested system

Our system is a multi-speakers ASR system, and even independent of the speaker, it uses the global approach. When a word is pronounced, the acoustic image of this word is facing all references in the dictionary and the most resemble word, by calculating a distance is then chosen [1]. Our system consists of several processes; its scheme is represented in the Figure below:

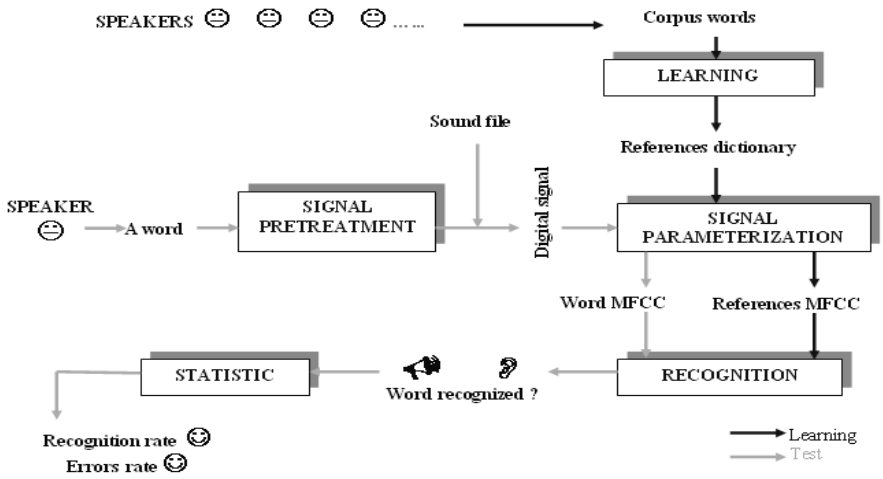

Figure 1 : System architecture.

2. The learning process

The realization of ASR system requires the provision of oral corpora. Several corpora were built for different languages as TIMIT for English language and BDSONS for French language. Because of the absence of Arabic oral corpora, we were obliged to make our own corpus.
The creation of corpora is a difficult task, it requires:

1. Defining the content of the corpus: (choose a few words pronounced by a large number of speakers or a large number of words pronounced by a small number of speakers and what corpus's size can be considered sufficiently? Must corpora depend on their application or not? Shall we select speakers? If yes, what are the criteria?).
2. Making records.

2.1. The vocabulary

The 30 words of our corpus were selected by linguists of the Arabic language institution at Constantine University. All phonetics' characteristics of Arabic language are taken into account, without binding to a specific field. Our corpus is made up of 6 sub corpora according to the Arabic language characteristics, each sub corpus contains 5 words. The words of our corpus are grouped in the following table:

Corpus type	**Corpus words**
Simple corpus	وزن – كتب – عرف – دحرج – شحذ
Stress corpus	قيد – الشمس – فسر – كرس - رد
Emphatic corpus	نظر – طبع – صرف – ضرب - قرب
Duration corpus	هتاف – غروب – يؤول – مماليك - نادى
Tanwine corpus	عزف – فرش – منزل – لون - مكتب
Mixture corpus	مواد – خزاعة – اضطر – قض - مشتط

Table 1 : Corpus words.

2.2. The registration condition

The recording environment is very suitable for reaching good quality corpus. The speech is recorded at a sampling rate of 44100 Hertz coded in 16-bit. The digitalization of the signal is made by the professional software "Sound Forge version 8", it is the famous well-known in the digital audio editing [2].

The vocabulary is pronounced by 4 women and 4 men; their age is between 22 years and 55 years. Each speaker repeats 3 times each word, so every word has 24 different occurrences in the corpus, the entire corpus contains 720 sound files. This corpus is used to define:

1. The learning corpus (540 sound files), is pronounced by 6 speakers,
2. Test1 corpus (180 sound files), is made by two speakers who participate in learning process,
3. Test2 corpus (180 sound files), made by two speakers who have not participated in learning process,
4. Test3 corpus which contains the same files as test1 corpus but adding a few sound effects (echo, noise).

3. The signal pre-processing

To be used by a computer, signal must first be digitalized. The transition process from analogue sound to the digital sound is called "sampling". We measure the voltage of analogue signal at regular intervals. The value obtained is finally encoded in binary. Another very important parameter of sampling is the precision with which the voltage of analogue signal is read and coded (2nbits) [3].

4. The signal parameterization process

The speech is a signal consisting of infinite information, we must extract the most important ones. A direct comparison treatment on this kind of signal is impossible, because there is too much information. Several techniques were used to represent the speech signal [4]:

- RCC : Real Cepstral Coefficients,
- LPC : Linear Prediction Coefficients,
- LPCC : Linear Predictive Cepstral Coefficients,
- MFCC : Mel Frequency Kestrel Coefficients,
- PLP : Perceptual Linear Prediction,

All these methods are reasonable solutions for speech signal parameterization, but the famous MFCC are better than the other candidates. The MFCC paradigm, introduced by Davis and Mermelstein has maintained its dominance since its introduction in 1980, because of its effectiveness, and even in noisy conditions it retains its strength. The purpose of MFCC is to reduce the number of data characterizing the signal and shows a limited number of parameters or coefficients, discriminating and robust [3,4,5,6].
To transform an audio file in MFCC cepster several steps are necessary [4,5,7]:

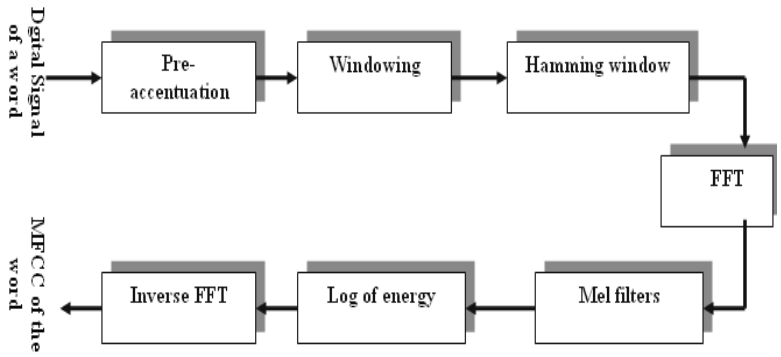

Figure 2 : From digital signal to MFCC.

4.1. The signal pre-emphasis

In speech recognition by MFCC, the signal must first undergo pre-emphasis to remedy the fact that high frequencies are less powerful than the low frequencies.
The signal pre-emphasis formula is [3.4]:

$$h_n = 1 - \alpha . Z_n^{-1} \qquad (5.1)$$

α is the pre-emphasis factor, commonly taken to 0,970. h is the pre-emphasis of the signal Z.

4.2. The windowing

The audio signal can not be treated as a whole because this would require a lot of calculations for the machine, so we cut the signal into slices called windows that have the particularity of overlap in half with the aim of have a better treatment for FFT (Fast Fourier Transform). It typically uses a window of N samples, N is a number that is a power of 2, it is because the FFT algorithm is much faster for these numbers [3].

4.3. Applying Hamming window

A Hamming window is applied to each window in order to decrease the spectral distortion created by the overlap, and minimize errors produced by FFT. The Hamming window improves the sharpness of harmonics and removes discontinuities on the edges. To create the Hamming signal, we use the following formula:

$$0{,}54 - 0{,}46 . cos\left(\frac{2\pi . n}{N-1}\right) . n \in [0, N-1] \qquad (5.2)$$

N is the size of the signal (the number of samples) [8.3].

4.4. Application of the FFT (Fast Fourier Transform)

To transform the signal from time domain to frequency domain, we must calculate the discrete Fourier transform (DFT). The Fast Fourier Transform (FFT) is a very powerful algorithm for calculating the discrete Fourier transform. The FFT calculation time is about 10 times lower than a classic DFT [3]. The result of this step is the signal specter, horizontal axis represent time, vertical axis represent frequency and intensity is represented by the color [9].

4.5. Mel filters bank

The frequencies range in the FFT spectrum is very wide, so much data to process, we must use a filter bank in the Mel scale. We passes the speech signal through a filter bank, the Mel filter bank is built from triangular filters, each filter will give a cepstral coefficient, commonly we use 12 factors, but we have to use 13 filters because the 0ème coefficient is not needed for speech recognition[10, 3].

4.6. The cepstrals coefficients

This is the final step, we transform data from Mel scale to time scale. We make the inverse of the Fourier transform. The result of this step will be the MFCC itself [3].

The most popular MFCC implementation is written by Malcolm Slaney in the "Auditory toolbox" of MATLAB toolbox.

5. The recognition process

The recognition is made by a genetic algorithm that compares MFCCs references and MFCC of the word to recognize, the outcome of this process is the sound recognized. In the test phase, this recognition must be confirmed by the user to avoid false recognitions.

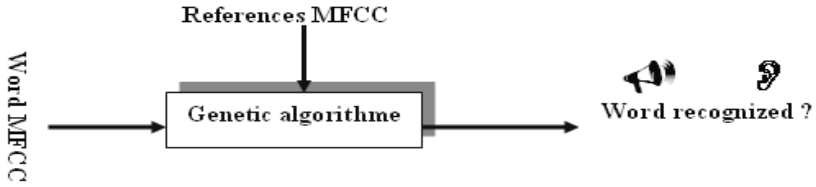

Figure 3 : Recognition process

The main procedures of our genetic algorithm are mentioned in what follows:

5.1. Initialization of the population

The reference dictionary (learning corpus) is the population managed by our genetic algorithm. This dictionary is made up of 540 entries, so 540 individuals. This population is divided into 30 sub-populations (the number of words), the choice of the initial population is random for each word to recognize. An initial population is made up of all occurrences of a word, then 18 individuals. If we do not reach the

recognition threshold in this sub population after 5 generations, we change initial population, and so on, until we reach the recognition threshold of the word to recognize. Each individual of this population contains the word it represents, its MFCC cepster, his fitness and his recognition threshold.

A MFCC cepster is a matrix, the number of lines is the number of MFCC coefficients, the number of columns is the same of the signal windows. The following Figureure represents the MFCC cepster of the word 'وزن' :

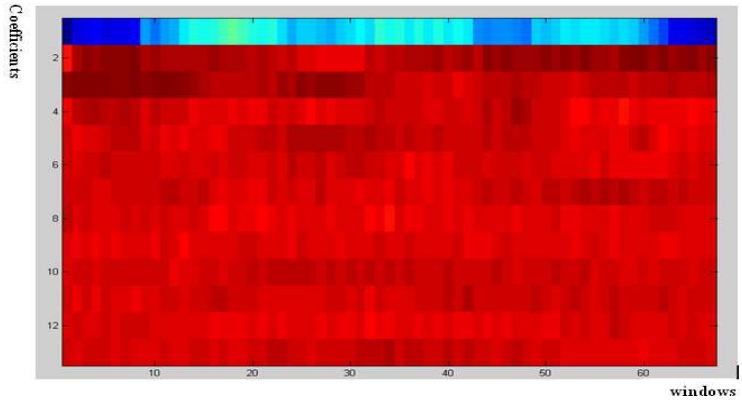

Figure 4. MFCC cepster of the word 'وزن'

The number of columns in this matrix differs from one word to another. The cepsters used in our system are matrices of (12 * 160), we choose a fixed size for all words depending on the size of the longest word in our vocabulary, the first line is unnecessary, it is ignored in our treatment.

5.2. Evaluation of the population

To evaluate the population, we calculate the fitness of each individual, it is the distance between the cepster of the word to recognize and the cepster of each individual, if this distance is less than the recognition threshold then the word is recognized, its formula is derived from the Euclidean distance [11]:

$$\text{Distance (A, B)} = \frac{1}{N * M} \sum \sum (A - B)^2 \qquad (6.1)$$

A and B are two **matrices** of size (N * M). In our case, the best fitness is the smallest distance.

The recognition threshold of a word is the maximum distance between the 18 occurrences of this word in the learning corpus.

5.3. The stop criterion

The stop criterion of our genetic algorithm is either the word is recognized or all sub populations have been covered.

5.4. The selection

After evaluating all individuals of the population, we apply the elitist selection method (Many researchers have found that elitism improves the performance of the AG). This method allows the genetic algorithm to retain a number of best individuals for the next generation. These individuals may be lost if they are not selected to reproduce [13.14].

5.5. The crossover

Its fundamental role is to enable the recombination of information contained in the genetic heritage of the population. We applied the one point cross with the probability 0.80 [12.13].

5.6. The mutation

A mutation is simply a change of a gene found in a locus randomly determined. The altered gene may cause an increase or a weakening of the solution value that represents the individual [16].

The principle of the mutation used in our system is new, the injection method, we inject a gene of the word to recognize directly in the reference dictionary, instead of using a random gene (a gene is a column of MFCC cepster). This principle will accelerate the convergence of the algorithm without influence the result [17]. The probability of mutation is 0.01 [12.13].

5.7. The replacement

The elitist replacement is the most suitable in our case, it keeps individuals with the best performance from one generation to the next. In general, a new child takes place within the population if it is more efficient than the less powerful individuals of the previous population, so we replace the worst parents [16].

6. The statistical process

To evaluate our system, we carried out three types of tests:

- Test1: The test's words are pronounced by speakers who participated to learning. This test has proved that our system is a multi speakers system.

- Test2: The words of the test are pronounced by speakers who have not participated to learning. This test has proved that our system is independent of speakers.
- Test3: The words of the test are pronounced by speakers who participated to learning but the environment is noisy. This test has proved the robustness of our system in a noisy environment.

The recognition system errors can be classified into 3 basic types who have not the same weight [18]:

- Substitution: a word is confused with another word (false recognition),
- Elision: a word is not recognized (non-recognition),
- Insertion: a word that does not belong to the vocabulary has been recognized.

We considered two types of errors, substitution and elision. The statistical process gives several results: the recognition rate, the substitutions rate, the elision rate, the number of tested words...

7. The experiment's results

The recognition rate of test1 is 100%, regardless of the corpora type. The recognition rate of the test2 differs from one corpus to another, for simple corpus and stress corpus, it reaches 80%, for tanwine corpus, it exceeds 68%, for the mixture corpus and duration corpus, it exceeds 50%. It is noticeable that the type of words affects the recognition rate, so a good choice of vocabulary, can yield better results.

In the test3, our system has acquired an acceptable recognition level in a pseudo real environment (sound+echo, sound+noise). Test3 has proved the robustness of our system in a noisy environment. The recognition rate of simple corpus is 80%, for emphatic corpus, it reaches 60%, for stress corpus is 55%. So our system has acquired an acceptable recognition rate because there is no system adapted to all real situations.

During the tests, we have noted down that our system does not make the distinction between men and women voice (if a woman pronounces a word of vocabulary, the recognized sound may be a man voice, because our system works on the word signal itself without taking any consideration of speaker's personality).

Conclusion and perspective

In this article, we suggest a system for automatic recognition of isolated Arabic words, with a vocabulary of 30 words that represent the different Arabic language

characteristics. Our system is a multi-speaker system or even independent of speaker, robust in a noisy environment.

Our system uses a genetic algorithm for recognition, and MFCC coefficients to modelize the speech signal. It is implemented with the matlab7 platform language. To evaluate the performance of our system, we have made an oral corpus, which could be used by other researchers to test and validate their systems working on Arabic language. The corpus words have been carefully selected and recorded in good conditions in order to create diversity in words type and in phonetics context.

The results acquired by our system confirm that the use of genetic algorithms joined with MFCC parameterization is a very promising method in the ASR field. In addition, we have suggested a new mutation method that accelerates the convergence of genetic algorithm.

There are many prospects of this study; it would be interesting :

- To extend our corpus size and vary the registration conditions.
- To test our system with standards databases of English language and French language.
- To use phonemes as decision unit in the recognition process instead of words, etc.

References

[1] Jean-Paul Haton, Jean Marie Pierrel, Guy perennou, Jean callean, Jean-luc gauvain. Automatic speech recognition, Dunod edition 1999.

[2] Sound Forge 8, http://www.wsystem.com/html/sound_forge.html.

[3] Robot guide par reconnaissance vocale, PPE 2004, http://www.tigen.org/perso/geogeo/TPE_Vocale/Dossier%20(reconnaissance%20vocale).PDF.

[4] Tudor Dimitrov Ganchev. “Speaker recognition”. Dissertation subject at the Patras University for the Doctor degree of philosophy, Greece-2005.

[5] Bruno Jacob. “Un outil informatique de gestion de modèles de markov caches : expérimentations en reconnaissance de la parole », Phd thesis at the Paul Sabatier University in Toulouse III- 1995.

[6] Dan Ellis , “ PLP, RASTA, MFCC et inversion dans Matlab”. Colombia University, identification and organization laboratory of speech and acoustic,2006.
http://72.30.186.56/language/translatesPage?Ip=en_fr&tt=url&text=http%3a%2

f%2fwww.ee.colombia.edu%2f%7edpwe%2fresources%2fmatlab%2frastamat%2f

[7] Yacine Mami, " reconnaissance de locuteurs par localisation dans un espace de locuteurs de références", doctoral thesis, signal and image speciality, telecommunication national school, Paris – 2003.

[8] Anicet Foukou, Sébastien Henaff, Matthieu Lagacherie, Paul Rouget, final presentation of Marvin project (Modest-encoding AlgoRIthm with Vocal IdentificatioN), EPITA – 2002.

[9] José Hernandez. "Algorithmes d'acquisition, compression et restitution de la parole à vitesse variable, étude et mise en place", graduation project,1995.

[10] Daniel Moraru, " Segmentation en locuteurs de documents audio et audiovisuels : application à la recherche d'information multimédia ", thesis for doctor degree of inpg, speciality : signal image parole télécoms, national polytechnic institute Grenoble – 2004.

[11] Mathematic distance, an article from Wikipédia, the free encyclopedia. http://fr.wikipedia.org/wiki/Distance_%28math%C3%A9matiques%29

[12] Renaud Dumeur. "Synthèse de comportements animaux individuals et collectives par algorithms génétiques", science departement, artificial intelligence institute, Paris_8 university – 1995.

[13] Marek Obitko, student technical university. "Genetic algorithms", 1998 at Hochschule fur technih und wirtschaft Dresden (FH) University of applied sciences.

[14] Mitchell Melanie, An introduction to genetic algorithm, a Bradford book the MIT press, Cambridge, Massachusetts, London, England, fifth printing 1999.

[15] Noelle Carbonell, Jean_Paul Haton, Gilles Simon, Henry Poincaré university, publisher Maximilian Buder, Nancy – 2004.

[16] Souquet Amédée, Radet Francois-Gérard. "Genetic algorithms", Thesis of the end of the year, supported on 21/06/2004.

[17] Fadila Maouche, "Automatic speech recognition, evolutionist approach, case of Arabic", Masters Thesis, Oum El Bouaghi university, Algeria – 2008.

[18] F Néel, G Chollet, L Lamel, W Minker. " reconnaissance et compréhension de la parole : evaluation et applications". The LIMSI and CNRS laboratories in France