Global Journal of Computer Science and Technology
Volume 11 Issue 15 Version 1.0 September 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Speech Recognition System Based On Hidden Markov Model Concerning the Moroccan Dialect DARIJA

By A. El Ghazi, C. Daoui, N. Idrissi, M. Fakir , B. Bouikhalene

Abstract - In this work, we present a system for automatic speech recognition on the Moroccan dialect. We used the hidden Markov model to model the phonetic units corresponding to words taken from the training base. The results obtained are very encouraging given the size of the training set and the number of people taken to the registration. To demonstrate the flexibility of the hidden Markov model we conducted a comparison of results obtained by the latter and dynamic programming.

Keywords : *Hidden Markov Model (HMM), MFCC, DTW, Acoustic vectors .*

GJCST Classification : *I.2.7, G.3*

Strictly as per the compliance and regulations of:

Speech Recognition System Based On Hidden Markov Model Concerning the Moroccan Dialect DARIJA

A. El Ghazi[α], C. Daoui[Ω], N. Idrissi[β], M. Fakir [Ψ], B. Bouikhalene[¥]

Abstract - In this work, we present a system for automatic speech recognition on the Moroccan dialect. We used the hidden Markov model to model the phonetic units corresponding to words taken from the training base. The results obtained are very encouraging given the size of the training set and the number of people taken to the registration. To demonstrate the flexibility of the hidden Markov model we conducted a comparison of results obtained by the latter and dynamic programming.

Keywords : Hidden Markov Model (HMM), MFCC, DTW, Acoustic vectors.

I. Introduction

The system of automatic speech recognition (ASR) can transcribe a voice message, extracting linguistic information from an audio signal. The system uses hidden Markov model [13] (Hidden Markov Model: HMM) to model words units and sentence of a language. In this work, the interest is to model the Moroccan dialect and implement a recognition system that converts a signal into a meaningful message that can be used thereafter. There is a several applications for a speech recognition system of Moroccan dialect. Most interesting are the man-machine dialogue, ie the passage of oral telephone calls; learning Moroccan dialect and systems helping people with disabilities [1]. The Moroccan dialect is a very important part of popular culture and covers almost the different regions.

The significance of ASR, several free systems have been developed, among the best known: HTK [11] and CMU Sphinx [2-3]. We used the last, it is based on Hidden Markov Model [3] and widely used in the field of speech recognition. In this context, our work focuses on the establishment of foundations for building a system of automatic speech recognition concerning the Moroccan dialect based on Sphinx4 [1].

In the following, we will outline the work done by starting with a theoretical approach to the hidden Markov model and dynamic programming (Section 2). Then, we present in brief a description of Moroccan dialect (section 5). The comparison results obtained from the hidden Markov model and dynamic programming are given in Section 6. And it ends with a conclusion and outlook in Section 8.

Author α Ω β Ψ ¥: Traitement de l'information, Faculté des Sciences et Techniques PB 523, Béni Mellal, Maroc.
E-mails : hmadgm@yahoo.fr, daouic@yahoo.com, najlae_idrissi@yahoo.fr ,fakfad@yahoo.fr, bbouikhlene@yahoo.fr.

II. Theoretical Bases

a) Hidden Markov Model

The hidden Markov model is a stochastic system capable [19], after a learning phase, to estimate the likelihood of observation sequence was generated by this model. The case represents a set of acoustic vectors of a speech signal. The hidden Markov model can be seen as a set of discrete states and transition between these states, it can be defined by all of the following parameters:

N : the number of model states

$A = \{a_{ij}\} = P(q_j/q_i)$: is a matrix of size N * N. It characterizes the transition matrix between states of the model. The transition probability to state j depends only on the state i :

$$P(q_t = j/q_{t-1} = i, q_{t-2} = k, \ldots) = P(q_t = j/q_{t-1} = i) \quad (1)$$

$B = \{b_j(o_t)\} = P(o_t/q_j)$, where $j \in [1, N]$ is the set of emission probabilities of the observation o_t when the system is in the state q_j. The shape of this probability determines the type of HMM used. In this work, we use a continuous probability density [19] defined by the bellow relation:

$$b(o, m, v) = N(o, m, v) = \frac{1}{\sqrt{(2\pi)^n |c|}} e^{-\frac{1}{2}(o_n - m_i)c^{-1}(o_n - m_i)'} \quad (2)$$

Where:

O: Observation trame

C: covariance matrix (diagonal)

$$C = \frac{1}{n-1} \sum_{k=0}^{n} (o_k - m_k)' * (o_k - m_k)c$$

m : the mean of each coefficient

$$m = \frac{1}{n} \sum_{k=1}^{n} o_k$$

Taking into account several pronunciations of a word requires the use of a multi-Gaussian probability density [21] that the resulting probability is given by:

$$B_j(o_t) = \sum_{i=1}^{k} C_{ij} * b_j(o_t) \quad (3)$$

k : number of Gaussian
C_{ij} : Gaussian weight of i in j
$B_j(o_t)$:observation probability at time t for state j

b) Dynamic programming

Dynamic programming [18] is one of the algorithms used in speech recognition domain, the principle is to compare two speech signals based on the distance between two matrices corresponding to the coefficients of Mel [18] of the two signals. Calculates the Euclidean distance between two vectors is sound by using the relationship:

$$d(i,j) = \sqrt{\sum_{k=1}^{K}(x_k - y_k)^2}\ ,\ i = \begin{pmatrix} X_1 \\ \vdots \\ X_N \end{pmatrix}, j = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$$

Then, calculate the minimum distance by traversing the element of the matrix obtained using the relation:

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2.d(i,j) \\ g(i,j-1) + d(i,j) \end{cases}$$

The final distance is:

$$G = \frac{g(I,J)}{I+J}$$

Where:

I, J: Length acoustic arrays corresponding two signals.

III. Extraction of Acoustic Parameters

a) Pretreatment

The speech signals used were acquired using a microphone. The noise intra sentence was deleted manually using the tool wavsurfer. The digitized signals will be represented by a family $(xn)\ n \in [1, k]$ where k is the total number of samples. After, the signal is sampled using the computer's sound card with a frequency Fs = 16kHz ie taking values follows a period 1/FS seconds.

i. *Mel coefficients*

Parameterization of speech signals is to extract the coefficients of Mel. This stage is based on the Mel scale to model the perception of speech in a manner similar to the human ear, linear up to 1000 Hz and logarithmically above [22]. The importance of the logarithmic scale appears when using a broad bench of values as it helps to space the small value and approach large values. The digitized signals must be further processed for use in the recognition phase. To do this the pre-emphasis is performed to meet the high frequencies:

$$h_n = 1 - 0.97 * z_n{}^{-1} \quad (5)$$

Then the signal is segmented into frames each frame contains N sample of speech and includes almost 30ms of speech, to do this we use a sliding time window of size 256. The successive windows overlap by half of their size ie 128 points in common between two successive windows. In this work we used the Hamming window [23]:

$$w(n) = 0.54 + 0.46 * \cos(2\pi * \frac{n}{N-1}) \quad (6)$$

In the next step the signal spectrum is calculated, it can introduce the signal (time domain) in frequency domain using the fast Fourier transform FFT:

$$X(n) = \frac{1}{N}\sum_{k=0}^{N-1} x(n) e^{jk\,2\pi(\frac{n}{N})} \quad (7)$$

To simulate the functioning of the human ear, we filter the signal through a bank of filters that each have a triangular response bandwidth. The filters are spaced so that their evolution is the Mel scale [22]. The approximate formula of the scale of Mel:

$$Mel(f) = 2595 * \log(1 + \frac{f}{700}) \quad (8\text{-}1)$$

$$X(i,k) = \sum_{n=0}^{N/2} X(n,k) * Mel(n,k) \quad (8\text{-}2)$$

The speech signal can be seen as the convolution in the time domain excitation signal g (n) and the vocal tract impulse response h (n):

$$x(n) = g(n) * h(n) \quad (9)$$

The application of the logarithm of the model on this equation gives:

$$Log|X(k)| = Log|G(k)| + Log|H(k)| \quad (10)$$

Finally, to obtain the coefficients of Mel applying the inverse Fourier transform defined by:

$$FFT^{-1}\{X(i,n)\} = x(n) = \frac{1}{N}\sum_{k=\frac{N}{2}}^{\frac{N}{2}-1} X(i,n) e^{jk2\pi(\frac{n}{N})} \quad (11)$$

This gives a vector of coefficients on each Hamming window. The number of filter adopted in this work is 12, it added the first and second derivatives of these coefficients, which gives in total 39 coefficients. Figure 1 gives a summary of the extraction of Mel coefficients (MFCC).

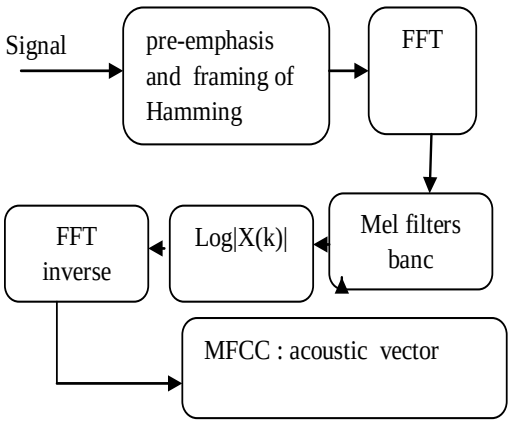

Fig 1 : Stage for acoustic parameters extraction

IV. Learning

After the extraction phase, the speech signal is represented by a matrix N* 39 which N is the number of windows in the signal. The audio files used in the learning phase must be segmented into phonemes; each word corresponds to a sequence of phonemes. Each of these will be represented by a hidden Markov model with three states, each state is characterized by:

- Vector averages for a state i, is given by:

$mi = \frac{1}{n}\sum_{k=1}^{n} O_k$, n: number of vectors for each state
O_k : Observation vector number k.

- Covariance matrix for state i:

$$Coi = \frac{1}{n-1}\sum_{k=1}^{n}(O_k - m_i)' * (O_k - m_i) \qquad (12)$$

The calculation of the mean vector and covariance matrix is performed for each Gaussian. In this paper we use five Gaussian so there will be five vehicles and five averages covariance matrices for each state. The calculation of the probability of resulting observation for each state is realized by the relationship 3.

Learning the model tends to maximize the logarithm of the probability of observation called the likelihood, to do this we use the Baum-Welch algorithm [15], whose steps are:

1- Initializing the model

- Creation of HMM for each state
- Initialization of the initial probability vector π with a higher probability for the first state and non-zero for the other two remaining states.
- Initialization of the transition matrix with probabilities respecting any transitions that the sum is equal to 1 and the model is a left-right (upper diagonal).

2 - Maximization: In this step, each iteration updates the model parameters and calculate again the likelihood. The updating of the model parameters is done via the following relations:

$$C_{ij} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{t=1}^{T}\gamma_{(j)}},\ 1 \le j \le N, 1 \le k \le M \qquad (13)$$

$$mj = \frac{\sum_{t=1}^{T} o_t\gamma_t(j,k)}{\sum_{t=1}^{T}\gamma_t(j)}, 1 \le j \le N, 1 \le k \le M \qquad (14)$$

$$Co_j = \frac{\sum_{t=1}^{T}(o_t - m_j)(o_t - m_j)'\gamma_t(j,k)}{\sum_{t=1}^{T}\gamma_t(j)}, 1 \le j \le N, 1 \le k \le M \qquad (15)$$

With:

M: number of Gaussian.
N: number of acoustic vectors for each state.
With:

$$\gamma_t = \frac{\alpha_t(j)\beta_t(j)}{\sum_i \alpha_t(i)\beta_t(j)}$$

$$\gamma_t(j,k) = \gamma_t\left(\frac{C_{jk}\mathcal{N}(o_t, m_{jk}, Co_{jk})}{\sum_{k=1}^{M} C_{jk}\mathcal{N}(o_t, m_{jk}, Co_{jk})}\right)$$

Cjk is the weight of the Gaussian k relative to the state j and the coefficients α and β are calculated by the Forward-Backward algorithm [15].

V. Recognition

The principle of recognition can be explained as the calculation of the probability P (W / O): the probability that a sequence of words W is the signal S and to determine the word sequence that maximizes this probability.

According to Bayes formula the probability P (W / S) can be written:

$$\mathbf{P(W/S)=P(w).P(S/W)/P(S)} \quad (2)$$

With:

- P (W): Prior probability of word sequence W:(Sample language).
- P (S / W): Probability of signal S, given the sequence of words W (Acoustic Model).
- P (S): probability of the acoustic signal S (independent of W).

The figure 2 shows the various stages of recognition, as a first step the signal is treated to extract acoustic vectors, based on these vectors the acoustic model is built from the HMM of phonemes learned on the training corpus . The succession of phonemes HMMS form the words models.

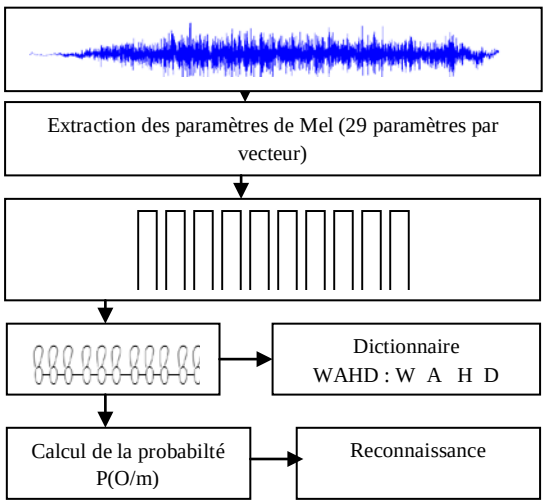

Fig. 2 : Stages of the recognition

VI. Presentation of Moroccan Dialect

The Moroccan dialect called Darija is the popular language broadcast in almost all regions of the country. This dialect is a communication tool widely used and is different from one region to another. The dialect Darija contains almost Arabic words in addition to a regional component, the difference between classical Arabic and dialect Darija is at the pronunciation. The figure below shows an illustration of the difference:

bases	Pronunciation	Scripture	
The succession of two consonants (sokoun) is permitted. Two successive consonants come together.	OKTOB	أكتب	Classical Arabic
Most letters are pronounced with 'sokoun'. Most words are pronounced without vowels (SAKINA)	KTB	كتب	Darija

Fig 1. Difference between classical Arabic and Darija

VII. Experimental Results

a) Learning base

The learning base used in our system contains 2500 pronunciation, the characteristics of the training set are illustrated in the following table:

Duration of the training set	Number of pronunciations
1h40min of pronunciations.	2500 recorded pronunciation independently and in different situations

Table1 : Characteristics of the learning base

The construction of the training set was made by taking the pronunciation of Arabic numerals 0 to 9 in the Moroccan dialect, Table 2 shows the formation of the learning base.

numbers	Phonetic Transcription
0	S I F R
1	W A H D
2	J U J
3	T L A T A
4	R B 3 A
5	X M S A
6	S T T A
7	S B 3 A
8	T M N Y A
9	T S 3 U D

Tab.2 : Phonetic transcription used for the recognition of digits in dialect Darija

b) Results

The test database contains 300 different pronunciations including noisy audio files. The recognition quality is measured by calculating the rate of recognition given by equation (3):

$$t = \frac{\text{number of words recognized}}{\text{size of the test database}}$$

The results obtained are shown in Table 4.

Test database	Results
300 different pronunciations introducing more noisy audio files	T=91%

Tab.4 : Results for the recognition system of the dialect Darija

The comparison of the results was made on noisy audio data. Table 5 illustrates the results obtained.

	HMM	DTW
Execution Time	Very fast	Plus then 10s for a big wav files
Recognition rate	91%	60%

Tab.5 : Results of comparison between the HMM and DTW

The efficiency of dynamic programming appears on the audio files not noisy. The disadvantage is that the execution time increases proportionally with the length of the file, which influence the time of recognition. In comparison with dynamic programming, hidden Markov model can model a word by a sequence of phonemes and sentence by a sequence of word models, which makes this process more effective and more appropriate to be implemented in systems Recognition advanced.

VIII. Conclusion

This work enables the establishment of a voice recognition system of the Moroccan dialect. This article can give an idea about the phonetics used for the recognition of the language. In comparison with dynamic programming, the results obtained by the

hidden Markov model are very satisfactory despite the limited number of speakers and size of the database. This shows the importance of stochastic and probabilistic modeling in the field of recognition.

Based on what has been achieved in this work, we'll build a system of passing oral phone call on the Moroccan dialect integrated into mobile phones, helping people with disabilities and people who do not dial telephone numbers.

References Références Referencias

1. Ali sadiqui & Noureddine chenfour " Reconnaissance de la parole arabe basé sur CMU Sphinx" , Séria Informatica. Vol VIII fasc. 1 2010.
2. H. Satori & M. Harti " Système de la reconnaissance de la reconnaissance automatique de la parole", Faculté des Sciences, B.P. 1796, Dhar Mehraz Fès, Maroc.
3. Cornijeol and L. Miclet, "Apprentissage artificielle-méthode et concept" 1988.
4. T. Pellegrini et Raphael, "Durée suivi de la voix parlée garce au modèle caché" 1989.
5. R. Gonzales and M. Thomson, "Syntactic pattern recognition" 1986.
6. Divejver and J. Killer, "Pattern recognition" in Pattern Recognition: a statistical approach"; Prentice Hall 1982.
7. Reweis, "Hidden Markov-Modele-Sam" 1980.
8. Robiner and Juang. "Fundamentales of speech recognition" 1993.
9. M. Amour ,A. Bouhjar & F. Boukhris IRCAM: publication : "initiation à la langue Amazigh" 2004.
10. RAP: Thèse 2008-[Benjamin LECOUTEUX].
11. B. Resch "Automatic Speech Recognition with HTK" 2003.
12. Chunsheng Fang (2009) "From Dynamic Time Warping (DTW) to Hidden Markov Model" (HMM) University of Cincinnati
13. A. Cornijeol and L. Miclet , "Apprentissage
14. Artificielle-méthode et concept" 1988.
15. P. Galley, B. Grand & S. Rossier , "reconnaissance vocale Sphinx-4" EIA de Fribourg mai 2006.
16. T. Pellegrini et R. Duée "Suivi de la voix parlée grâce aux modèles de Markov Caché", lieu : IRCAM 1 place Igor Stravinsky 75004 PARIS jiun 2003.
17. S. Sigurdsson, Kaare Brandt Petersen and Tue Lehn-Schiøler "Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3Encoded Music", Informatics and Mathematical Modelling Technical University of Denmark Richard Petersens Plads - Building 321 DK-2800 Kgs. Lyngby – Denmark.
18. T. AL ANI "Modèles de Markov Cachés (Hidden Markov Models (HMMs))", Laboratoire
19. A2SI-ESIEE-Paris / LIRIS.
20. G. SEMET & G. TREFFOT "La reconnaissance de la parole avec les MFCC" TIPE juin 2002.
21. A. Chan, Evandro Gouvêa & Rita Singh "Building Speech Applications Using Sphinx and Related Resources": http://docpp.sourceforge.net , August 2005.
22. Dr. A. Drygajlo "Introduction aux statistiques gaussiennes et à la reconnaissance statistique de formes", Ecole Polytechnique Fédérale de Lausanne.
23. S. Jamoussi , "Méthodes statistiques pour la compréhension automatique de la parole", Ecole doctorale IAEM Lorraine, 2004.
24. SEMET Gaetan & TREFFO , 'Grégory,'Reconnaissance de la parole avec les coefficients MFCC' TIPE jiun 2002.

This page is intentionally left blank