AUTOMATIC SPEECH RECOGNITION SYSTEM CONCERNING THE MOROCCAN DIALECTE (Darija and Tamazight)

A. EL GHazi, C. DAOUI, N. IDRISSI.
Equipe: 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.

ABSTRACT. In this work we present an automatic speech recognition system for Moroccan dialect mainly: Darija (Arab dialect) and Tamazight. Many approaches have been used to model the Arabic and Tamazight phonetic units. In this paper, we propose to use the hidden Markov model (HMM) for modeling these phonetic units. Experimental results show that the proposed approach further improves the recognition.

Keywords: Hidden Markov Model (HMM), MFCC, Tifinaghe

Introduction

The development of human-machine systems based on speech recognition in recent years has growing rapidly and acknowledged a large development. On the one hand, this is due to technological advances and research in the field, the diversity of applications and the fields of interest and on the other hand, the worthwhile offered by the different studied languages. In this work, we are interested in Moroccan dialect composed especially by Darija (Arabic dialect)[25] and Tamazight (Berber dialect)[9].

Morocco is the main Berber-speaking state. Around 40% of the population is Berber grouped in three main variants of the Berber dialect [Tamazight][9]:
• The tachelhit (or chleuh) is spoken in the High Atlas, the Anti-Atlas to the south and the plain of Souss, the Sahara and elsewhere around the kingdom. This is the Berber dialect as spoken. The Central Morocco Tamazight, spoken in the Upper and Middle Atlas, Azilal in the countryside south of Taza, and the center of the kingdom. The Tarifit (or Rif) is spoken by the Rif, Rif people in eastern, north-eastern Morocco: Nador, Al Hoceima, Kebdana.

Darija is the Moroccan Arabic dialect spoken by the Moroccan Arabic (it can be understood by the Arabic-speaking Berber people). It belongs to the North African dialects like Algerian, Tunisian and Maltese. It covers most areas of the country and it is almost standard in all with a few differences in intonations. These Moroccan dialects contain different components according to the history and traditions of each region.

Given the difficulty of collecting a consistent basis of these three Berber dialects. In this work, we focused on the recognition of the Moroccan dialect of the central Tamazight.

The automatic speech recognition of the Moroccan dialect presents a great importance because it allows expanding the use of systems man-machine dialogue, helping the disabled, the illiterate and the uneducated. Given the great interest the development of human-machine systems based on speech recognition as the case of our proposed oral composition of a phone number from a terminal, our work focuses on the implementation of an automatic speech recognition system of the Moroccan dialect namely based on Darija and central Tamazight.

Thus, we will outline the work done by starting with a theoretical approach to the hidden Markov model (section 2). Then, we show the steps of extracting parameters of speech signals (Section 3). Next, we present a brief description of the Moroccan dialect (section 6). And it ends with a conclusion and outlook in section 7.

1. THEORETICAL BASES

2.1 Hidden Markov Model (MMC):

The hidden Markov model is a stochastic automaton [19] can after a learning phase, estimate the probability of observation sequence was generated by this model. The observation is represented by acoustic
vectors of the speech signal. The hidden Markov model can be seen as a set of discrete states and transitions between these states, it can be defined by all of the following:

N: the number of states of the model;

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

\[ P(q_t = j | q_{t-1} = i, q_{t-2} = k, ...) = P(q_t = j | q_{t-1} = i) \]

\[ B = \{b_j(o_t)\} = P(o_t|q_j), j \in \{1, N\} : \text{is the set of emission probabilities of the observation } o_t \text{ knowing that the system is in state } q_j. \text{ The shape of this probability determines the type of HMM used. In this work, we use a continuous probability density [19] defined by the normal distribution:} \]

\[ b(o, m, v) = \frac{1}{\sqrt{(2\pi)^n|C|}} e^{-\frac{1}{2}(o_m - m)^T C (o_m - m)} \]

With:

O: the frame of observation

C: covariance matrix (diagonal)

\[ C = \frac{1}{n-1} \sum_{k=0}^{n} (o_k - m_k)' (o_k - m_k) \]

m: the mean of each coefficient.

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

The inclusion of several pronunciations of a word requires the use of a multi-Gaussian probability density [21] that the resulting probability is given as:

\[ B(o_t) = \sum_{i=1}^{k} C_{ij} * b_i(o_t) \quad (A.1) \]

k: number of Gaussian

I_{ij}: weight of Gaussian i in state j

B_{ij}(o_t): probability of the observation at time t on the state j.

2. EXTRACTION OF ACOUSTIC PARAMETERS

2.1 Pretreatment

The speech signals used were acquired using a microphone. Noise intra sentence was manually removed using the tool wavsurfer[23]. The digitized signals are represented by a family \((xn) n \in [1, k]\) where k is the total number of samples in the signal. The latter is sampled using the sound card of the computer with a frequency \(Fs = 16kHz\) here we take a values follows a period \(1/Fs\) seconds.

2.2 Mel coefficients

The parameterization of speech signals consists by extracting the Mel coefficients. This approach is based on the Mel scale to model the speech perception in a way similar to the human ear linear in bottom of 1000 Hz [22] and logarithmic above. The importance of the logarithmic scale appears when using a bench of values broad as it helps to space the small value and approach the large ones.
The digitized signals must be further treated for use in the recognition phase. To do a pre-emphasis is performed to meet the high frequencies:

\[ h_n = 1 - 0.97 \cdot z_n^{-1} \]

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

\[ w(n) = 0.54 + 0.46 \cdot \cos\left(2\pi \cdot \frac{n}{N-1}\right) \]

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

\[ X(n) = \frac{1}{N} \sum_{k=0}^{N-1} x(n)e^{\frac{2\pi i}{N} kn} \]

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

\[ \text{Mel}(f) = 2595 \cdot \log\left(1 + \frac{f}{700}\right) \]

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

\[ x(n) = g(n) \cdot h(n) \]

The application of the logarithm of the module gives:

\[ \log|X(k)| = \log|G(k)| + \log|H(k)| \]

Finally, to obtain the Mel coefficients we use the inverse Fourier transform defined by:

\[ \text{FFT}^{-1}\{X(i,n)\} = x(n) = \frac{1}{N} \sum_{k=-N/2}^{N/2} X(i,n)e^{\frac{2\pi i n}{N}} \]

We obtain a vector of coefficients for each Hamming window. The number of filters in this work is taken 13 were added first and second derivatives of these coefficients, resulting in total 39 coefficients. Figure 4 gives a summary on the extraction of Mel coefficients (MFCC).
3. TRAINING

After the extraction of Mel parameters phase, the speech signal is represented by a matrix whose size is 39*N such that N is the number of windows in the signal. Audio files used in the training phase should 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 of means for a state i, it is given by:

$$m_i = \frac{1}{n} \sum_{k=1}^{n} O_k$$,

n: number of vectors for each state.

$$O_k$$: Vector observation number k.

Covariance matrix for state i:

$$C_{oi} = \frac{1}{n-1} \sum_{k=1}^{n} (O_k - m_i)'(O_k - m_i)$$

The calculation of mean vector and the covariance matrix is done for each Gaussian. In this paper we use five gaussian so there will be five mean vectors and five covariance matrices for each state. The calculation of result probability of observation for each state is done by the following relations:

The training model consists 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- Initialization of the model
2- Creation of the HMM for each state
   - Initialize the vector of initial probabilities \( \pi \) with a higher probability for the first state and non-zero for the other states.
   - Initialize the matrix of transition probabilities with any respecting the sum of transitions is equal to 1 and the model type is left-right (diagonal above)
3- Maximizing: In this step, each iteration updates the model parameters and the likelihood is recalculated and again. The updating of the model parameters is done via the following relations:

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

$$m_j = \frac{\sum_{t=1}^{T} \alpha_t(j,k)}{\sum_{j=1}^{N} \gamma_t(j)}$$, \( 1 \leq j \leq N, 1 \leq k \leq M \)

$$C_{o_j} = \frac{\sum_{t=1}^{T} (o_t - m_j)(o_t - m_j) \gamma_t(j,k)}{\sum_{t=1}^{T} \gamma_t(j)}$$, \( 1 \leq j \leq N, 1 \leq k \leq M \)

with:

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

with:

$$\gamma_t = \frac{a_t(i)\beta_t(j)}{\sum_j a_t(i)\beta_t(j)}$$

$$\gamma_t(j,k) = \frac{c_{jk}(o_t,m_k,C_{o_j})}{\sum_{k=1}^{M} c_{jk}(o_t,m_k,C_{o_j})}$$
Cjk is the weight of Gaussian k relative to the state j and the coefficient α and β are calculated by the Forward-Backward algorithm [15].

4. RECOGNITION

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

According to Bayes’s formula the probability P (X / S) can be written:

\[
P(W|S) = \frac{P(W)P(S|W)}{P(S)}
\]

With:
P (W): a priori probability of sequence of word W:
(Language model).
P (S / W): The probability of the signal S, given the sequence of words W (acoustic model).
P (S): probability of the acoustic signal S (independent of W).

Figure 1 shows the various stages of recognition, at first the signal is treated to extract acoustic vectors, based on these vectors the acoustic model is loaded from the phoneme HMM learned on the training corpus. The succession of HMM of phonemes forms a model of words.

5. PRESENTATION OF MOROCCAN DIALECT

The Moroccan dialect called Darija is the most popular spoken dialect broadcasted in all regions of the country. This dialect is the mean of communication most used and standardized everywhere. Darija is the dialect that contains the most Arabic word added to the regional components. The difference between classical Arabic and Darija [24] is the level of pronunciation and a conversion of some letters as shown in the following figure [tab.1]:

---

**Fig. 1: Recognition steps**

- Extraction of Mel parameters
- Calculation of probability P(O/m)
- Dictionnaire WAHD : W A H D
- Recognition
Tab. 1 Difference between classical Arabic and Darija

<table>
<thead>
<tr>
<th>Rule</th>
<th>Pronunciation</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>The succession of two consonants (sokoun) is not permitted. Two successive consonants come together to provide Chedda.</td>
<td>OKTOB</td>
<td>أكتب</td>
</tr>
<tr>
<td>Most letters are pronounced with 'sokoun'. Most words are pronounced without vowels (SAKINA)</td>
<td>KTB</td>
<td>كتب</td>
</tr>
<tr>
<td>In Darija we pronounce ق (the consonant number 21) gua</td>
<td>Gua</td>
<td>ق</td>
</tr>
</tbody>
</table>

Moroccan Tamazight language prevails in rural areas, it is the oldest language in North Africa. He began to spread in Algeria and Morocco, Tamazight now beginning to take the official way by the creation of the Royal Institute of Amazigh Culture (IRCAM) and integration into primary school books. The writing of the Tamazight language is Tifinagh with the letters presented in Table 2.

Tab.2: phonetic symbols used for recognition of Amazigh figures

<table>
<thead>
<tr>
<th>Symbole</th>
<th>représentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ṡ</td>
<td>C</td>
</tr>
<tr>
<td>ṣ</td>
<td>K</td>
</tr>
<tr>
<td>ṡ</td>
<td>U</td>
</tr>
<tr>
<td>ṣ</td>
<td>Z</td>
</tr>
<tr>
<td>ṡ</td>
<td>SS</td>
</tr>
<tr>
<td>ś</td>
<td>TT</td>
</tr>
<tr>
<td>ṣ</td>
<td>T</td>
</tr>
<tr>
<td>ṡ</td>
<td>D</td>
</tr>
<tr>
<td>ṡ</td>
<td>MM</td>
</tr>
</tbody>
</table>

6. EXPERIMENTAL RESULTS

6.1 Training database

The training database used in our system is described in the table below:
Tab. 3: Features of Tamazight Database

<table>
<thead>
<tr>
<th>Length of Tamazight Database</th>
<th>Number of peoples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1h45min of pronunciation.</td>
<td>- 3 adults</td>
</tr>
<tr>
<td></td>
<td>- 1 doesn’t speak Tamazight.</td>
</tr>
<tr>
<td></td>
<td>- 8 girls</td>
</tr>
<tr>
<td></td>
<td>- 14 boys</td>
</tr>
</tbody>
</table>

Tab. 4: Features of DARIJA Database

<table>
<thead>
<tr>
<th>Length of Tamazight Database</th>
<th>Number of peoples</th>
</tr>
</thead>
<tbody>
<tr>
<td>3h30min of pronunciation.</td>
<td>6 Women</td>
</tr>
<tr>
<td></td>
<td>2 Man</td>
</tr>
</tbody>
</table>

The construction of the training database was made by taking the pronunciation of the Arabic numerals 0 through 9 in Moroccan dialect (Darija and Tamazight), Table 5 and Table 6 show the formation of the training database and the phonetic transcription used.

Tab. 5: Phonetic symbols used for the recognition of digits in Darija

<table>
<thead>
<tr>
<th>Numbers</th>
<th>Phonetic Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>S I F R</td>
</tr>
<tr>
<td>1</td>
<td>W A H D</td>
</tr>
<tr>
<td>2</td>
<td>J U J</td>
</tr>
<tr>
<td>3</td>
<td>T L A T A</td>
</tr>
<tr>
<td>4</td>
<td>R B 3 A</td>
</tr>
<tr>
<td>5</td>
<td>X M S A</td>
</tr>
<tr>
<td>6</td>
<td>S T T A</td>
</tr>
<tr>
<td>7</td>
<td>S B 3 A</td>
</tr>
<tr>
<td>8</td>
<td>T M N Y A</td>
</tr>
<tr>
<td>9</td>
<td>T S 3 U D</td>
</tr>
</tbody>
</table>

Tab. 6: Phonetic symbols used for the recognition of digits in Tamazight

<table>
<thead>
<tr>
<th>Numbers</th>
<th>Phonetic Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>I L E M</td>
</tr>
<tr>
<td>1</td>
<td>Y E N ou Y A N</td>
</tr>
<tr>
<td>2</td>
<td>S I N</td>
</tr>
<tr>
<td>3</td>
<td>C R A DD</td>
</tr>
<tr>
<td>4</td>
<td>K O Z</td>
</tr>
<tr>
<td>5</td>
<td>SS E M (S MM U S pour quelques régions)</td>
</tr>
<tr>
<td>6</td>
<td>SS E DD (SS DD E SS pour quelques régions)</td>
</tr>
<tr>
<td>7</td>
<td>SS A</td>
</tr>
<tr>
<td>8</td>
<td>TT A M</td>
</tr>
<tr>
<td>9</td>
<td>T Z A</td>
</tr>
</tbody>
</table>
7. Results

The test database contains 300 different pronunciations including noisy audio. The quality of recognition is measured by calculating the rate of recognition given by the relation (eq.1):

\[
t = \frac{\text{number of word recognized}}{\text{Length of database}} \quad \text{(eq.1)}
\]

The results obtained are shown in Table 4.

<table>
<thead>
<tr>
<th>Test Database</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 different pronunciations introducing more noisy audio</td>
<td>T=92.33%</td>
</tr>
</tbody>
</table>

Table 7: Results obtained for the Amazigh speech recognition system

<table>
<thead>
<tr>
<th>Test Database</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 different pronunciations introducing more noisy audio</td>
<td>T=90%</td>
</tr>
</tbody>
</table>

Table 8: Results obtained for the Amazigh speech recognition system

Figure 2 shows our Java interface used for automatic recognition system of the Moroccan dialect:

![Java interface for Moroccan dialect automatic recognition system](image)

We based on this system of recognition of Moroccan dialect to achieve a system of recognition of national telephone numbers, this system is based on recording of discrete numbers one by one, Figure 3 shows the Java interface used.
8. CONCLUSION

This work enables the establishment of a voice recognition system of the Moroccan dialect. Tamazight is one of the most complex languages in phonetics and at the regional differentiation, and Darija Arabic dialect is different in pronunciation from region to region. This article can give an idea about phonetics used for the recognition of these dialects. We have developed a system that is a Java application designed to oral combination of a phone call on the Moroccan dialect, used by persons with disabilities and those not in schools. Relying on this system, we will expand our work with other applications in robotics and human-machine dialogue.

This work reform a basic idea about a stochastic modeling of Hidden Markov Model, it cans decompose the linguistics units (word) into elementary components (phone or syllable). view this strong modeling, the experiments results are satisfies despite the length of the training database and the quality of waves files used.

Based on this work, our speech recognition system can be exploited on the banc systems, robotic simulation and oral messaging. In addition to this applications the speech recognition is one of the researches domains that can help for progress of a regional dialects and there implementation in lot of systems human-machines.

REFERENCES

[17] T. AL ANI "Modèles de Markov Cachés (Hidden Markov Models (HMMs))", Laboratoire
A. El Ghazi et al. / International Journal of Engineering Science and Technology (IJEST)

A2SI-ESIEE-Paris / LIRIS.

[22] SEMET Gaetan & TREFFO 2002, 'Grégory,'Reconnaissance de la parole avec les coefficients MFCC' TIPE.