Annals of the University of Craiova, Mathematics and Computer Science Series
Volume 46(1), 2019, Pages 56–65
ISSN: 1223-6934

Amazigh speech recognition using triphone modeling and clustering tree decision

Safâa El Ouahabi, Mohamed Atounti, and Mohamed Bellouki

Abstract. The main objective of this paper is to develop an Amazigh automatic speech recognition system using a speech corpus composed on 187 distinct Amazigh words. The speech corpus was recorded by 50 (25 male and 25 female) Amazigh-Tarifyt native speakers. The system was evaluated on a speaker-independent approach using Hidden Markov Models (HMMs). The tests were carried out basing essentially on the Gaussian mixture distributions (GMMs), tied states (senons), triphone modeling and clustering tree decision. The recognition rate increases significantly and reached 92, 2% which is a high and satisfactory recognition rate comparing to the systems developed for this language especially in relative to the size of the corpus used on our system.

2010 Mathematics Subject Classification. 68T10; 68T50.
Key words and phrases. Amazigh automatic speech recognition system, Amazigh speech corpus, HMMs, GMMs, tied states (senons), triphone modeling, clustering tree decision.

1. Introduction

Speech is certainly the most natural way of communication that humans use to interact with each other. This can be justified by the fact that the speech signal allows the intelligible transmission of a large amount of information. Speech processing as scientific discipline has known since the 1960s a rapid expansion, linked to the development of the means and techniques of telecommunications. Several problems make the automatic processing of speech a difficult field, because there is a very large amount of variability presented in the speech: intra-speaker variability, due to the mode of speech; interlocutor variability (different timbres, masculine, feminine voices, children's voices ...); variability due to the acquisition of the signal (microphone type), or to the environment interference(noise, crosstalk ...). It is necessary to study, or to process, a large amount of data if we want to discover, or obtain, what makes a sound elementary, in spite of the different contexts, the different modes of speech, the different speakers, and different environments. Fortunately, this is means that the information in the signal will be redundant, and that the different information contained in the same signal will cooperate to allow the understanding of the signal, despite the ambiguities and the "noise" that can be found at each level. The biggest change made in the field of speech recognition is defined by the passage from rules-based systems to systems based on statistical models [1]. The speech signal begins to be represented in terms of probability with HMM (Hidden Markov Models) [2, 3], which makes it possible to combine linguistic information with temporal acoustic achievements of speech

This paper has been presented at the Conference MOCASIM, Marrakesh, 26-27 November 2018.

sounds [20, 21]. This also motivates the appearance of statistical models of language, called n-grams [17, 16]. Innovation in acoustic signal analysis consists of the combination of cepstral coefficients and their temporal derivatives of first and of second order [13]. For acoustic modeling based on hidden Markov models, the sequences of words are divided into basic units, frequently phonemes. For the use of monophones, we speak of acoustic modeling independent of context. For the use of triphone, we speak of acoustic modeling dependent on context, in the sense that the pronunciation of a phoneme is characterized by its previous and following phonemes. Of course, this generates a very large amount of units. In actual conditions where the number of triphone representatives in the corpus learning is insufficient, we cannot model all the possible contexts of phonemes. As a result, we need to group together similar polyphones in a set of polyphones using a procedure named clustering tree decision [15]. The continuous evolution of information and communication technologies has been marked by major advances in the deployment of human language processing, including automatic speech recognition, for the promotion and development of under resourced languages. Nowadays, automatic speech recognition is introduced in many applications, like language learning systems to improve learners pronunciation, Automatic transcription applications for radio and television documents and voice server-type, telephone applications for access to services or access to information by searching in a voice databases. However, speech technologies are not sufficiently exploited for the Amazigh language. In order to reap the benefits of these technologies, we have devoted this study to the development of Amazigh speech recognition system for Amazigh language. In Morocco, however; there is a lot of effort to develop an ASR for the Amazigh language, It is essentially a language with an oral tradition used to communicate between family and in form of poetry, songs, riddles... This language has begun to become a written language with the adoption of the Tifinagh alphabet and its codification by IRCAM in 2003 [6]-[5]. There is a lot of effort to develop an ASR system for the Amazigh language in Morocco. Some work is done in this direction for Amazigh digits and alphabets [9]-[11]. The main objective of this paper is to develop an Amazigh automatic speech recognition system using a speech corpus composed on 187 distinct Amazigh words. The speech corpus was recorded by 50 (25 male and 25 female) Amazigh-Tarifyt native speakers. The system was evaluated on a speaker-independent approach using Hidden Markov Models. The tests were carried out basing essentially on the Gaussian mixture distributions, tied states (senons), triphone modeling and clustering tree decision.

The paper is organized as follows: Section 2 presents some theoretical basis of the Hidden Markov Models. Section 3 describes the Amazigh speech recognition system and all implementation requirements and components required to adapt the system to Amazigh language as well as the details of experiments and results. We finally present our conclusions and future directions in section 4.

2. Theoretical basis

The Hidden Markov Model is a powerful statistical method for characterizing the observed samples data of a discrete time process. It provides not only an efficient way of constructing parametric models, but it also incorporates the dynamic programming principle to unify time-varying segmentation and sequence classification.

In modeling a process with an HMM, the samples can be characterized by a random parametric process whose parameters can be estimated in a well-defined framework. The basic theory of HMM has been published in a series of papers by Leonard E. Baum. HMMs have become the most widely used method for speech signal modeling in several applications like automatic speech recognition, fundamental frequency and formant tracking, speech synthesis, machine translation, syntax tagging, oral language understanding, automatic translation... In a Markov chain, each state corresponds to a deterministic observation event. A natural extension to the Markov chain introduces a non-deterministic process that generates output symbols for each state. Observation is therefore a probabilistic function of the state. The new model is called HMM, which can be seen as double stochastic processes, which is unobservable directly. This underlying process is therefore probabilistically associated with another process producing the sequence of frames, which is observable. There are three basic problems to solve for the application of this method:

- Evaluation problem: Determine the probability of a model generating an observation sequence, this problem is solved by applying the FORWARD algorithm.
- The decoding problem: Determine the most probable sequence of states for a given model and observation sequence, the VITERBI algorithm is used to perform this task.
- The learning problem: Adjust the parameters of the model to maximize the likelihood (joint probability) of generation of an observation sequence; the algorithms of BAUM-WELCH and VITERBI make it possible to carry out the learning. In speech applications, continuous HMMs are frequently used, where the observation does not belong to a discrete set but to a distribution. Thus, a left-right topology for a continuous HMM makes it possible to model the successive states of a phoneme for a speech signal. More generally, the objective to be achieved is the determination from acoustic vectors of the pronounced phonetic sequence.

The topology we have chosen for our Amazigh speech recognition system is the Bakis topology, left-right with the authorization of the looping transitions and to the next state (see figure 1). It is suitable for signal modeling and also for the learning phase. There are fewer parameters to estimate. This topology is widely used in the automatic speech recognition systems.

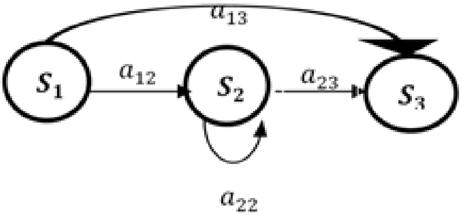


FIGURE 1. Bakis topology.

3. Amazigh Speech Recognition System using triphone Modeling and clustering tree decision

In this section, we present our Amazigh automatic speech recognition system. We propose an approach based on Hidden Markov Models evaluated on a speech corpus developed for Amazigh language [19]. CMU Sphinx [7] is the toolkit that we have used to develop our system. In the context of continuous speech recognition independently of the speaker, our goal is to develop a reference system based on hidden Markov models; it should be enable to evaluate the new techniques aimed to improving the ASR systems developed for Amazigh language. The main components and processes of the Amazigh speech recognition system using CMU Sphinx, as shown in Figure 2, are described in more detail in the following sections.

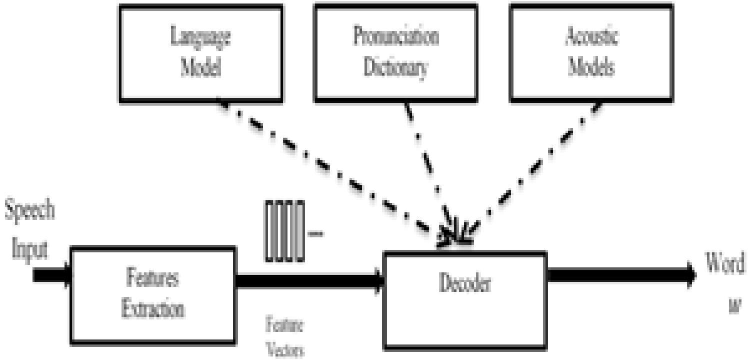


FIGURE 2. Architecture of the Amazigh speech recognition system using CMU Sphinx.

3.1. Features Extraction. The goal of speech parameterization is to reduce the negative environmental influences on speech recognition. Speech varies according to a certain number of aspects:

- Differences in the pronunciation of speakers vary by sex, dialect, voice etc.
- Environmental noise.
- The recording channel.

Different parameterization methods exist to improve the robustness of the speech recognition system for different recording conditions. The two most effective methods used for parameterization of speech are MFCCs [8] and perceptual linear prediction (PLP) [14]. We chose in our work to use the MFCC coefficients because they are efficiently calculated by CMUSPhinx that we chose for our Amazigh speech recognition system. The statistical MFCC coefficients are calculated for each window of the sampled signal. So, for a $25ms$ window shifted by $10ms$ with a frequency sample of $16Khz$. Static MFCC coefficients are generally extended by their temporal derivatives $\Delta + \Delta\Delta$. As a result, MFCC $\Delta + \Delta\Delta$ extract $13 + 13 + 13 = 39$ acoustic characteristics for a frame. The original vector for the 46750 audio samples in a frame is reduced to the 39 characteristics vector Acoustics MFCC $\Delta + \Delta\Delta$.

3.2. Dictionary. Dictionary provides pronunciation information (sequences of phonetic units) and the classification of words. It provides the sequences of units in words and makes the optional classification of it. The meaning of a unit varies according to the task of recognition. For the recognition of isolated words, the unit could be an entire word. For the recognition of the big vocabulary, the unit could be a phoneme or a triphone.

3.3. Amazigh acoustic model. The acoustic model estimates the probability $P(a|w;\theta)$ of generation of acoustics characteristics for a given word w. The most successful acoustic modeling methods do not directly estimate the probability $P(a|w)$, but estimate the probability $P(a|f_1f_2f_3f_4)$ of the generation of characteristics a for the phones $w = f_1f_2f_3f_4$ which form the pronunciation of the word w. The phone is the smallest unit of speech. Acoustic characteristics of a phone depend on its context. The previous phone and the next one influence strongly on the sound of the phone in the middle. The triphone is a sequence of three phones that capture the context of a single phone. Therefore, the acoustic properties of triphones vary much less depending on the context of the phone. triphones will be put together to reduce the number of triphones for the acoustic modeling. In speech recognition, a multivariate Gaussian distribution is usually used to model the probabilities of HMM states. The parameters of the distribution Gaussian are individually estimated for each state. However, states are typically grouped together during acoustic model learning and states within of the same group share the same parameters for the Gaussian distribution. The procedure to create the acoustic model is done using SPHINXTRAIN tool, the different inputs files needed to generate it are described in the figure 3.

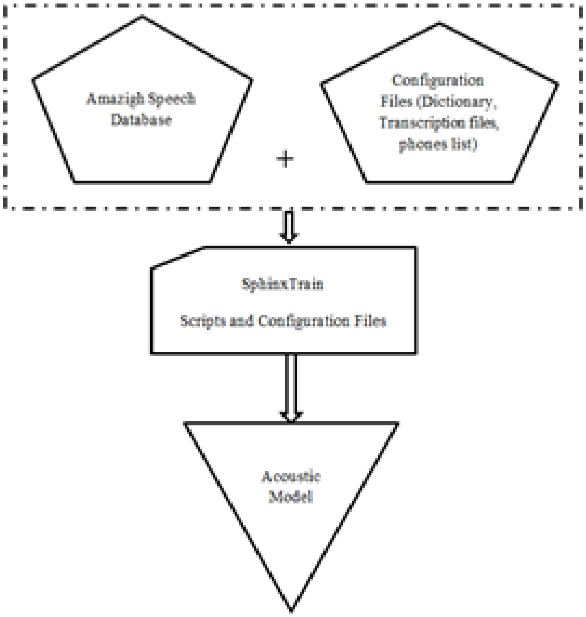

Amazigh Speech Database
+
Configuration Files (Dictionary, Transcription files, phones list)
SphinxTrain
Scripts and Configuration Files
Acoustic Model

FIGURE 3. procedure to create the acoustic model is done using SPHINXTRAIN tool.

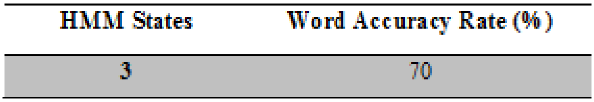
3.4. Language model. The language model effectively reduces the hypotheses of the acoustic model. A probability of acoustic characteristics of a given word transcription $P(a|w)$ estimated by an acoustic model is combined with the probability of a word transcription P(w) estimated by a language model in order to calculate the posterior probability of transcription $P(a|w) = \frac{(P(a|w) \times P(w))}{(P(a))}$. The probabilities for each word are estimated using frequencies relative to n-grams, $(n-1)$-grams, $(n-2)$-grams, on data learning. The language model estimates probability by counting relative frequencies on the corpus of texts that is usually chosen.

3.5. Decoder. Recognition is done by constructing word patterns from rules, the phrases are subsequently formed by concatenating the word patterns. Word patterns are connected via an HMM model. For isolated word recognition, the HMM model is evaluated for every possibility of words. The HMM model integrates the decomposed each units in 3 states. In general, a phoneme model is composed of 3 transmitter states serving to describe the shape (pattern) of the phoneme to be recognized. For the recognition of speech, we generally consider that phonemes (contextual or non-contextual) have three parts, each corresponding to a state:

- a left context corresponding to the transition between the previous phoneme and the current phoneme,
- a central part considered as stationary,
- a right context corresponding to the transition between the current and the next phoneme.

From the probabilities generated by the acoustic model and probabilities sequences of words obtained with the language model and the lexicon, the algorithm of decoding helps to produce the best sentence hypothesis. The purpose of the algorithm is to find the optimal path to maximize the probability of the following observations given the model used without generating the entire lattice.

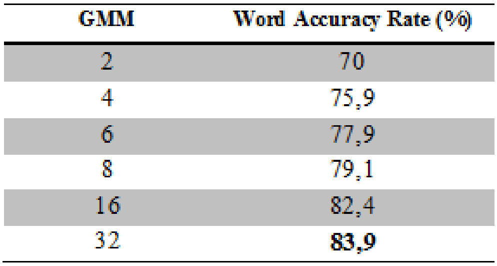
3.6. Experience and results. In Our experiments, we have used two sets of data: one for training the system and the other for testing. Database is composed on 187 spoken Amazigh words. A total of 50 speakers uttered the 187 Amazigh words with five repetitions. The first 40 speakers were used in training, while the rest 10 speakers were reserved for testing. 46750 is the total of files used in our experiments. The first step to build HMM triphone models is to use a simple cloning of independent context models (monophones) already learned. The vectors means and covariance matrices, as well as the transition probabilities will be identical for all triphones associated with the appropriate monophone. The 31 phones representing the phonetic vocabulary of the Amazigh database must first be initialized. The word sequences are modeled by a set of acoustic units, frequently the phonemes. For the development of a monophonic recognition system (independent of the context), each phoneme must be modeled by a single HMM Left-to-right with three-state (see Figure 1). The initial state and the final state have for objective to be used only for connecting models in continuous speech without transmitting observation. The probabilities of emission are calculated by a weighted sum of multivariate Gaussian (GMM), characterized by their mean vector and their covariance matrix. The recognition rates achieved for the database using HMM Left-to right with three-state for monophone recognition system (independent-context) is shown in the table below (Figure 4).



HMM States	Word Accuracy Rate (%)
3	70

Figure 4. Recognition Accuracy Rate for monophone recognition system (independent-context).

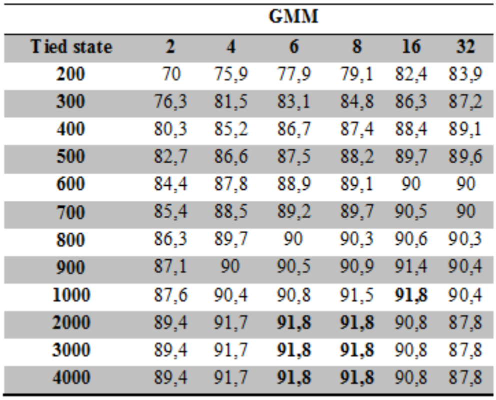
From the results obtained in Figure 4, it is clearly seen that the results obtained are not very satisfactory. We can improve monophonic models by increasing the number of Gaussian to estimate the probability of emission of a vector in a state. However it is essential to choose the necessary number of Gaussian awarded to each state, making a better adaptation between an adequate modeling of HMM monophones and the number of learning data. The problem that arises then is to find the number of components that is best suited to the data available. From the results obtained in Figure 5, it is clearly seen that the results are improved and the best and high number of Gaussians leads to improve the accuracy rate, because the training data has a sufficient number of samples for each phoneme.



GMM	Word Accuracy Rate (%)
2	70
4	75,9
6	77,9
8	79,1
16	82,4
32	**83,9**

Figure 5. Recognition Accuracy rate for different GMMs.

The same phoneme is pronounced differently depending on its context. The variability of speech signal is not perfectly represented by context independent HMM models (monophones). In order to take into account the effects related to the phenomena of coarticulation, several contextual models have been proposed. The recognition rates of speech can be significantly improved through these models. It is better to work with triphones models taking phonetic contexts into account left and right. Figures 6 and 7 shows Word recognition correction rate (%) in reference to tied triphone and GMM.



	GMM					
Tied state	**2**	**4**	**6**	**8**	**16**	**32**
200	70	75,9	77,9	79,1	82,4	83,9
300	76,3	81,5	83,1	84,8	86,3	87,2
400	80,3	85,2	86,7	87,4	88,4	89,1
500	82,7	86,6	87,5	88,2	89,7	89,6
600	84,4	87,8	88,9	89,1	90	90
700	85,4	88,5	89,2	89,7	90,5	90
800	86,3	89,7	90	90,3	90,6	90,3
900	87,1	90	90,5	90,9	91,4	90,4
1000	87,6	90,4	90,8	91,5	**91,8**	90,4
2000	89,4	91,7	**91,8**	**91,8**	90,8	87,8
3000	89,4	91,7	**91,8**	**91,8**	90,8	87,8
4000	89,4	91,7	**91,8**	**91,8**	90,8	87,8

Figure 6. Table of word recognition rate (%) in reference to tied triphone and GMM.

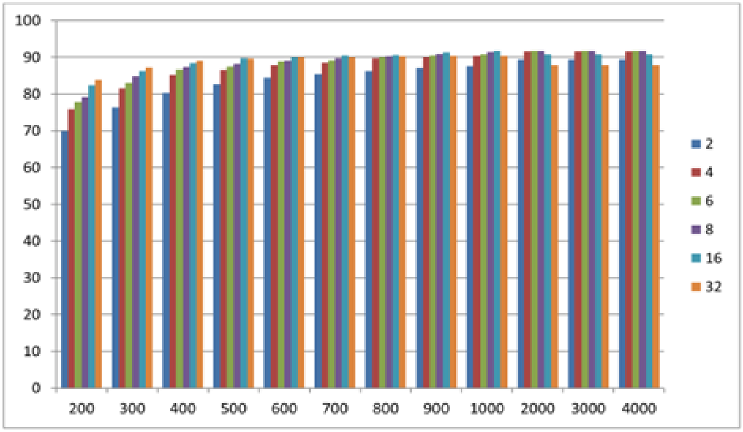
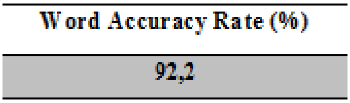


Figure 7. Graph of word recognition rate (%) in reference to tied triphone and GMM.

From the Figures 6 and 7, it is clearly seen that the results are improved significantly and the best combination for high accuracy rate is 6 and 8 Gaussian mixture distributions with 2000, 3000 and 4000 senons obtaining $91,80\%$ word recognition accuracy. The number of HMM models has increased from 31 monophones to several thousand triphone (5529 triphones). It is unthinkable to have sufficient data to make a correct apprenticeship of all of these triphones models. Indeed, some appear only sometimes in the learning database. To get around this difficulty, we have chosen to use the state-sharing method by decision tree. Decision trees are built for each phonetic class using a sequential optimization procedure from top to bottom. Initially all triphone models belonging to the same phonetic class are placed in a single group at the root of the tree. A series of binary language questions (QS) is executed to partition the states that maximize the likelihood. This approach relies on linguistic knowledge by exploiting a specific decision tree at each state. A binary language question is asked at each node of the tree which deals with the phonetic context left or right of the phoneme taken into account.



Word Accuracy Rate (%)
92,2

FIGURE 8. Word recognition correction rate (%) using clustering tree decision.

The recognition rate increases and reached 92, 2% as seen in Figure 8 using tree decision which is a high and satisfactory recognition.

4. Conclusion

The goal of our research is to build a speech recognition system independent of the speaker implemented from HMMs models. The system is firstly constructed using context-independent phonetic models (monophones). Several experiments have been carried out with this system. We also examined the evolution of decoding rates after using tied state language model and context-dependent phonetics appraoch (triphones) and clustering tree decision. Experimental results show that our system provide significant improvements of the phonetic recognition rate (Accuracy) using 187 words of the Amazigh speech corpus AMZWRD [19]. The recognition rate increases significantly and reached 92, 2% which is a high and satisfactory recognition rate comparing to the systems developed for this language especially in relative to the size of the corpus used on our system.

References

[1] S.J. Aroba, R.P. Singh, Automatic Speech Recognition: A Review, *International Journal of Computer Applications* **60** (2012), no. 9, 34–44.

[2] L.E. Baum, J.A. Eaghon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, *Bulletin of the American Mathematical Society* **73** (1967), 360–363.

[3] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, *Annals of Mathematical Statistics* **37** (1966), 1554–1563.

[4] A. Boukous, *Phonologie de l'Amazigh*, Institut royal de la culture Amazigh, Rabat, 2009.

[5] A. Boukous, The planning of Standardizing Amazigh language The Moroccan Experience, *IRCAM* **6** (2014), no. 1, 7–23.

[6] A. Boumalk, K. Nait-Zerrad, *La Vocabulaire grammatical amazighe*, IRCAM, CAL, Publications de l'IRCAM, Rabat, 2009.

[7] CMU Sphinx Open Source Speech Recognition Engines, *http://www.cmusphinx. sourceforge.net/html/cmusphinx.php*, Retrieved February 10, 2013.

[8] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28** (1980), no. 4, 357–366.

[9] A. El Ghazi, C. Daoui, N. Idrissi, Automatic Speech Recognition for tamazight enchained digits, *World Journal Control Science and Engineering* **2** (2014), no. 1, 1–5.

[10] S. El Ouahabi, M. Atounti, M. Bellouki, Amazigh Isolated-Word speech recognition system using Hidden Markov Model toolkit (HTK), *International Conference on Information Technology for Organizations Development (IT4OD)* (2016), DOI: 10.1109/IT4OD.2016.7479305, 1–7.

[11] S. El Ouahabi, M. Atounti, M. Bellouki, Building HMM Independent Isolated Speech Recognizer System for Amazigh Language, *Europe and MENA Cooperation Advances in Information and Communication Technologies, of the series Advances in Intelligent Systems and Computing* **520** (2016), 299–307.

[12] A. Fadoua, B. Siham, *Natural language processing for Amazigh language: Challenges and future directions*, Language Technology for Normalisation of Less-Resourced Languages, 2012.

[13] S. Furui, Speaker-independent isolated word recognition using dynamic features of speech spectrum, *IEEE Transactions on Acoustics, Speech and Signal Processing* **34** (1986), no. 1, 52–59.

[14] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America* **87** (1990), 1738–1752.

[15] X. Huang, A. Acero, H-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[16] F. Jelinek, B. Merialdo, S. Roukos, M. Straussi, Self-organized language modeling for speech recognition,*Readings in Speech Recognition. Morgan Kaufmann* (1990), 450–506.

[17] F. Jelinek, The Development of an Experimental Discrete Dictation Recognizer, *Proceedings of the IEEE* **73** (1985), no. 11, 1616–1624.

[18] M. Outahajala, L. Zenkouar, P. Rosso, *Building an annotated corpus for Amazigh*, Proceedings of the 4th International Conference on Amazigh and ICT, Rabat, Morocco, 2011.

[19] S. El Ouahabi, M. Atounti, M. Bellouki, A database for Amazigh speech recognition research: AMZSRD, *3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat* (2017), DOI: 10.1109/CloudTech.2017.8284715, 1–5.

[20] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE. T.* **77** (1989), 257–286.

[21] L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[22] H. Satori, F. El Haoussi, Investigation amazigh speech recognition using CMU tools, *Int. J. Speech Technol.* **17** (2014), no. 3, 235–243.

(Safâa El Ouahabi, Mohamed Atounti, Mohamed Bellouki) Laboratory of Applied Mathematics and Information Systems, Multidisciplinary Faculty of Nador, University Mohammed First, 62702 Selouane - Nador, Morocco.
E-mail address: `safaa.elouahabi@gmail.com`