Building HMM Independent Isolated Speech Recognizer System for Amazigh Language

Safâa El Ouahabi, Mohamed Atounti and Mohamed Bellouki

Abstract This paper describes the implementation of Hidden Markov Model based speaker independent spoken digits and letters speech recognition system for Amazigh language which is an official language in Morocco. The system is developed using HTK. The system is trained on 33 Amazigh alphabets and 10 first digits by collecting data from 60 speakers and is tested using data collected from another 20 speakers. This document details the experiment by discussing the implementation using the HTK Toolkit. Performance was measured using combinations of HMM 8-states and various number of Gaussian mixture distribution. The experimental results show that the system have given better recognition rate 85.95 % with 4 Gaussian Mixture. The results obtained are improved in comparison with our previous work.

Keywords Speech recognition system · Hidden Markov model · Gaussian mixture distribution · Hidden Markov model toolkit (HTK) · Amazigh letters and digits

1 Introduction

Speech Recognition is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone. It is the process of converting an acoustic signal into a sequence of words. To recognize the underlying symbol sequence given a spoken utterance, the continuous speech waveform is first converted into a sequence of equally spaced feature vectors. The task of the recognizer is to map between sequences of feature vectors and the wanted underlying symbol sequences. Automatic speech recognition (ASR) is one of the fastest growing areas

S. El Ouahabi (✉) · M. Atounti · M. Bellouki
Faculty Polydisciplinary of Nador, Laboratory of Applied Mathematics and Information System, Mohammed First University, Oujda, Morocco
e-mail: safaa.elouahabi@gmail.com

Á. Rocha et al. (eds.), *Europe and MENA Cooperation Advances in Information and Communication Technologies*, Advances in Intelligent Systems and Computing 520, DOI 10.1007/978-3-319-46568-5_31

of engineering and technology [1] and there is lot of systems which are developed for English and other major languages spoken in developed countries [2–7], in the other hand, spoken alphabets and digits for different languages were targeted by ASR researchers [8, 9], as for the Amazigh language, many ASR were developed for digits and letters [10, 11]. Automatic speech recognition systems have been implemented using various toolkits and software, the most commonly used amongst them are the Hidden Markov Model ToolKit, Sphinx Toolkit, ISIP Production System, Julius Open-Source Large Vocabulary CSR Engine, HMM Toolbox for Matlab etc, among all these tools the HTK toolkit is the most popularly used tool to design ASR systems, since it is used in building and manipulating hidden Markov Models, it has applications in other research areas as well. HTK is well documented and provides guided tutorials for its use. The work presented in this paper aims to build a HMM independent isolated speech recognizer system for Amazigh language based on HMM toolkit (HTK). The system is trained on 33 Amazigh alphabets and 10 first digits by collecting data from 60 speakers and it is tested using data collected from another 20 speakers including both males and females. This paper details the experiment by discussing the implementation using the HTK Toolkit, performance was measured using combinations of HMM 8-states and various number of Gaussian mixture distribution. The article is organized as follows: Sect. 2 presents a brief description of the Amazigh language. In Sect. 3 we describe the hidden markov models. While Sect. 4 emphasizes on the description of HTK toolkit. In Sect. 5, we describe the Amazigh ASR developed, experiments and results. Finally, Sect. 6 concludes this paper.

2 Amazigh Language

The Amazigh language, known as Berber or Tamazight, stretches over an area that is so vast in Africa, from the Canary Islands to the Siwa Oasis in the North, and from the Mediterranean coast to Niger, Mali and Burkina-Fasso in the South. Historically, the Amazigh language has been autochthonous and was exclusively reserved for familial and informal domains [12]. In Morocco, we may distinguish three big regional varieties, depending on the area and the communities: Tarifiyt in the Northern, Tamazight in the Center and Tashlhiyt in the South-west and the High Atlas of the country. As regards the Amazigh varieties, there are 2.5 million Tachelhit-speakers in the south of Morocco known as Sus, 3 million Tamazight-speakers in the Atlas Mountains, and about 1.7 million Tarifiyt-speakers in the Riff [13]. In Morocco, the status of Amazigh has achieved the most advanced level, especially by the foundation of the Royal Institute of Amazigh Culture (IRCAM), in the Dahir on October 17th 2001, which stipulates that the Amazigh culture is a "national matter" and, this being the case, is a concern of all the citizens. Like any language passes throw oral to a written mode, the Amazigh language has been in need of a graphic system. In Morocco, the choice ultimately fell on Tifinaghe for

Arabic correspondance	Tifinaghe Transcription [IRCAM]	Arabic correspondance	Tifinaghe Transcription [IRCAM]	Arabic correspondance	Tifinaghe Transcription [IRCAM]	Arabic correspondance	Tifinaghe Transcription [IRCAM]	Arabic correspondance	Tifinaghe Transcription [IRCAM]
صفر	ⴰ	ا	ⴰ	ه	ⵀ	-	ⵉ	ى	ⵢ
واحد	ⵉ	ب	ⴱ	ح	ⵃ	ر	ⵔ	ز	ⵣ
اثنان	ⵊ	-	ⵜ	ع	ⵄ	-	ⵍ	-	ⵎ
ثلاثة	ⵏ	-	ⵅ	ج	ⵊ	غ	ⵖ	ف	ⵏ
أربعة	ⵓ	د	ⴷ	ق	ⵇ	س	ⵙ	ص	ⵚ
خمسة	ⵔ	ض	ⴹ	ي	ⵉ	ش	ⵛ	ك	ⴽ
ستة	ⵕ	ي	ⵉ	ج	ⵊ	ش	ⵛ	ك	ⴽ
سبعة	ⵖ	ف	ⴼ	ل	ⵍ	ت	ⵜ	ة	ⵜ
ثمانية	ⵎ	ك	ⴽ	م	ⵎ	ظ	ⵚ	ع	ⵄ
تسعة	ⵏ	-	ⵅ	ن	ⵏ	ا	ⴰ	و	ⵓ

Fig. 1 Table of the 10 Amazigh digits and 33 alphabets

technical, historical and symbolic reasons. Since the Royal declaration on February 11th 2003, Tifinaghe has become an official graphic system for writing Amazigh, particularly in schools. Thus, the IRCAM has developed an alphabet system called Tifinaghe-IRCAM. This alphabet is based on a graphic system towards phonological tendency. This system does not retain all the phonetic realizations produced, but only those that are functional [14–16].

Figure 1 shows the ten Amazigh digits and the 33 Amazigh letters, along with their transcription in Tifinagh and Arabic correspondence.

3 Hidden Markov Model

Hidden Markov Model (HMM) is very powerful mathematical tool for modeling time series. It provides efficient algorithms for state and parameter estimation, and it automatically performs dynamic time warping of signals that are locally stretched. Hidden Markov models are based on the well-known chains from probability theory that can be used to model a sequence of events in time. Markov chain is deterministically an observable event. The most likely word with the largest probability is produced as the result of the given speech waveform. A natural extension of Markov chain is Hidden Markov Model (HMM), the extension where the internal states are hidden and any state produces observable symbols or observable evidences [17].

Mathematically Hidden Markov Model contains five elements.

1. Internal States: These states are hidden and give the flexibility to model different applications. Although they are hidden, usually there is some kind of relation between the physical significance to hidden states.
2. Output: $O = \{O_1, O_2, O_3, \ldots, O_n\}$ an output observation alphabet.
3. Transition Probability Distribution: $A = a_{ij}$ is a matrix. The matrix defines what the probability to transition from one state to another is.
4. Output Observation: Probability Distribution $B = b_i(k)$ is probability of generating observation symbol $o(k)$ while entering to state i is entered.
5. The initial state distribution ($\pi = \{ \pi_i \}$) is the distribution of states before jumping into any state.

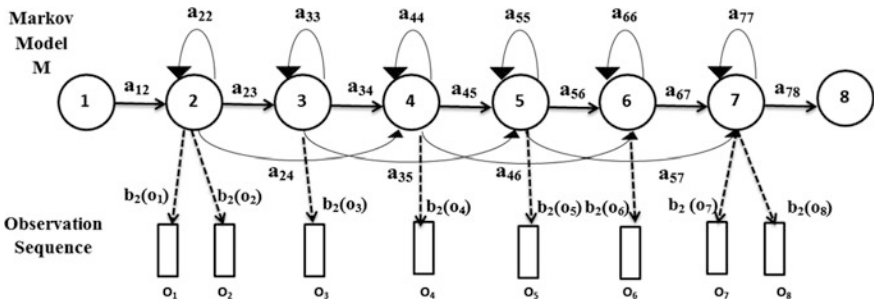

Fig. 2 Hidden markov model (HMM)—8-states

Here all three symbols represents probability distributions i.e. A, B and π. The probability distributions A, B and π are usually written in HMM as a compact form denoted by lambda as $\lambda = (A, B, \pi)$.

In our work, we are using the same kind of HMM. We are using a file called 'proto' which contains all necessary information and specifications. This file has been taken as it is from the HTK book [17]. This prototype 'proto' has to be generated for each digit and letter in the dictionary. Same topology is used for all the HMMs and the defined topology consists of 6 active states (observation functions) and two non-emitting states (initial and the last state with no observation function), See Fig. 2.

4 HTK Toolkit

Hidden Markov Model Toolkit i.e., HTK is a portable toolkit developed by the Cambridge University Engineering Department (CUED) freely accessible for download after registration at the URL http://htk.eng.cam.ac.uk. It consists of several library module and C program code and with good documentation (HTK Book [17]). Precompiled binary versions are also available for download (for Unix/Linux and Windows operating systems). Apart from speech recognition it has been applied to character recognition, speech synthesis, DNA sequencing etc. The toolkit provides tools for data preparation, training, testing and analysis, see Fig. 3.

5 Amazigh Speech Recognition System Using HTK

This section presents the different phases of the development of our system which are: data preparation, training, test and the analysis of results. It highlights on the system performance based on HTK.

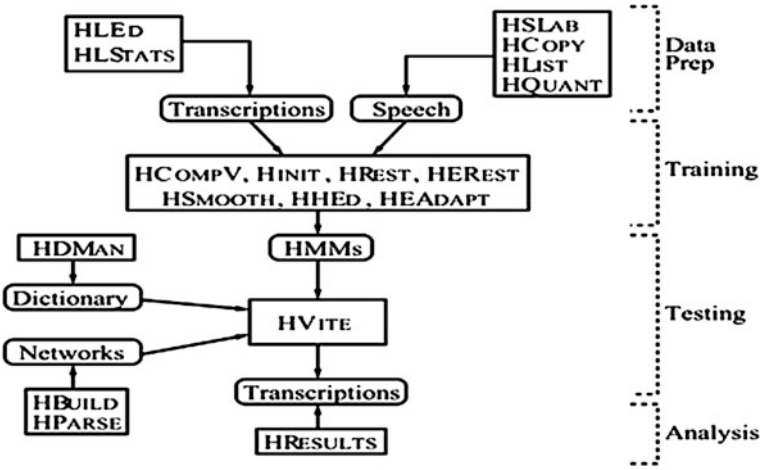

Fig. 3 HTK tools

5.1 Data Preparation

This phase consists of recording the speech signal. Firstly the data is recorded with the help of a microphone using a recording tool wavesurfer [18] in .wav format. The sampling rate used for recording is 16 kHz. The corpus consists of 10 Amazigh spoken digits (0–9) and 33 Amazigh alphabets each digits and alphabet is uttered 10 times in separated data files, each file containing 1 utterance and so the 10 distinct digits and 33 alphabets spoken by 80 person results in 19,500 file. Each utterance was visualized back to ensure that the entire word was included in the recorded signal, wrongly pronounced utterances were ignored and only correct utterances are kept in the database.

Before the corpus can be used to train HMMs it must be converted to an appropriate parametric form and the transcriptions must be converted to the correct format and use the required labels. Each wav file in corpus was labeled with the Wavesurfer tools, which results in to 19600 file (.lab). Due to the significant amount of files, a long time was required to validate this step of labeling signals.

To parameterize audio HCopy is used. In this work, we have used the acoustic parameters type MFCC. The parameters necessary for acoustic analysis such as format of input speech files, features to be extracted, window size, window function, number of cepstral coefficients, pre-emphasis coefficients, number of filter bank channels and length of cepstral filtering is provided to the HCopy in a configuration parametrisation.conf. The values to these parameters used for our experiment are given in Table 1.

Table 1 HCopy configuration parameters

Parameter	Value
SOURCEFORMAT	WAV
TARGETKIND	MFCC_0_D_A
WINDOWSIZE	250000.0
TARGETRATE	100000.0
NUMCEPS	12
USEHAMMING	T
PREEMCOEF	0.97
NUMCHANS	26
CEPLIFTER	22

5.2 Training

Before starting the learning phase, the HMM parameters must be properly initialized. The Hinit command in HTK tool is used for initializing the HMM by using the time alignment algorithm Viterbi from prototypes, and the training data in their MFCC form and their associated labeled file. This prototype has been generated for each word in the dictionary. Same topology is used for all the HMMs and the defined topology consists of 6 active states (observation functions) and two non-emitting states (initial and the last state with no observation function), see Fig. 2. This topology is used for all the HMMs. various numbers of Gaussian distributions with diagonal matrices are used as observation functions and these are described by a mean vector and variance vector in a text description file known as prototype. After initialization of models, Hcompv and HRest tool are applied in several iterations to estimate simultaneously all models on all sequences of acoustic vectors not labeled. The resulting models are improved by increasing the number of Gaussians for estimating the observation emission probability in a state. The models have been re-estimated by Hrest and Herest.

5.3 Testing

Once the models are trained, they are tested using the HTK command called HVite. We have created a dictionary and the word network before executing HVite. HVite uses the Virterbi algorithm to test the models. As input, HVite requires a network describing the allowable word sequences, a dictionary defining how each word is pronounced and a set of HMMs in addition to the result of decoding phase (.mfcc). HVite will convert the word network to a phone network and attach the appropriate HMM definition to each phone instance, after which recognition can be performed on direct audio input or on a list of stored speech files.

5.4 Analysis

Analysing an HMM-based recognizer's performance is done by the tool HResults. It uses dynamic programming to align the transcriptions output and correct reference transcriptions. The recognition results of the test signals are compared with the reference labels by a dynamic tracking performed by HResult.

5.5 Results

The evaluation of the performance of the speech recognition system was done by using HTK tool HResults. In HMM training, The system was tested on Baum Welch algorithm and Viterbi algorithm. Performance was measured using combinations of HMM 8-states and various number of Gaussian mixture distribution (1, 2, 4, 8, 16, 32), the state was modeled using Left Right (LR) HMM topologies. After initialization of models (Single Gaussian Mixture), Hcompv and HRest tool are applied in several iterations to estimate simultaneously all models on all sequences of acoustic vectors not labeled. The resulting models are improved by increasing the number of Gaussians for estimating the observation emission probability in a state. The models have been re-estimated by Hrest and Herest. In each mixture number, evaluation of the performance of the speech recognition system was measured by using HResults. The system is relatively successful, as it can identify the spoken digit and alphabets at an accuracy of 85.95% for speaker independent approach when the system trained using 8 states-HMM and 4 Gaussian Mixture, see Table 2. There is possibility that by using more robust model, we can increase the accuracy further.

Table 2 Results of test

Number of mixture	Number of state	Word correction rate (%)
1	8	82.66
2	8	84.14
4	8	**85.95**
8	8	85.83
16	8	85.30
32	8	84.25

6 Conclusion and Future Works

The system was tested using testing corpus data and the system scored up to 85.95 % word recognition for speaker independent approach. The work has catered for only an Isolated Digit-Letters Speech data. As much as it has created a basis for research, it can be expanded to cater for more extensive language models and larger vocabularies. The system can be enhanced to a larger vocabulary including commonly used words; it can be made robust by using larger database for training.

References

1. Anusuya, M.A., Katt, S.K.: Speech recognition by machine: a review. Int. J. Comput. Sci. Inf. Secur. **6**(3) (2009)
2. Kimutai, S.K., Milgo, E., Milgo, D.: Isolated Swahili words recognition using Sphinx4. Int. J. Emerg. Sci. Eng. **2**(2) (2013). ISSN:2319–6378
3. Ananthi, S., Dhanalakshmi, P.: Speech recognition system and isolated word recognition based on Hidden markov model (HMM) for Hearing Impaired. Int. J. Comput. Appl. **73**(20), 30–34 (2013)
4. Kumar, K., Jain, A., Aggarwal, R.K.: A Hindi speech recognition system for connected words using HTK. Int. J. Comput. Syst. Eng. **1**(1), 25–32 (2012)
5. Sameti, H., Veisi, H., Bahrani, M., Babaali, B., Hosseinzadeh, K.: A large vocabulary continuous speech recognition system for Persian language. EURASIP J. Audio Speech Music Process. (2011)
6. Abushariah, M.A., Ainon, M.R.N., Elshafei, R.M., Khalifa, O.O.: Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools. In: International Conference Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malayzia, doi:10.1109/ICCCE.2010.5556829 (2010)
7. Gales, M.J.F., Diehl, F., Raut, C.K., Tomalin, M., Woodland, P.C., Yu. K.: Development of a phonetic system for large vocabulary Arabic speech recognition. In: IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), Kyuto, Japan, pp. 24–29, doi:10.1109/ASRU.2007.4430078, 9–13 Dec 2007
8. Alotaibi, Y.A.: Investigating spoken Arabic digitsin speech recognition setting. Inf. Sci. **173**, 115–139 (2005)
9. Chapaneri, S.V.: Spoken digits recognition using weighted MFCC and improved features for dynamic time warping. Int. J. Comput. Appl. (0975–8887) **40**(3) (2012)
10. EL Ghazi, A., Daoui, C., Idrissi, N.: Automatic speech recognition for Tamazight enchained digits. World J. Control Sci. Eng. **2**(1), 1–5 (2014)
11. Satori, H., El Haoussi, F.: Investigation amazigh speech recognition using CMU tools. Int. J. Speech Technol. **17**(3), 235–243 (2014). doi:10.1007/s10772-014-9223-y
12. Boukous, A.: Société, langues et cultures au Maroc: Enjeux symboliques. Najah El Jadida, Casablanca, Maroc (1995)
13. Moustaoui, A.: The Amazigh language within Morocco's language policy, Dossier 14, University of Autònoma de Madrid (2003)
14. Boukous, A.: Phonologie de l'amazighe. Institut Royal de la Culture Amazighe, Rabat (2009)
15. Outahajala, M., Zenkouar, L., Rosso, P.: Building an annotated corpus for Amazighe. In: 4th International Conference on Amazigh and ICT, Rabat, Morocco (2011)

16. Fadoua, A., Siham, B.: Natural language processing forAmazigh language: Challenges and future directions. Language Technology for Normalisation of Less-Resourced Languages, (2012)
17. Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (2002). http://htk.eng.cam.ac.uk
18. http://sourceforge.net/projects/wavesurfer/