Investigation of Amazigh Speech Recognition Performance Based on G711 and GSM Codecs

Mohamed Hamidi, Hassan Satori, Ouissam Zealouk, and Khalid Satori

Abstract In this chapter, we evaluate the Amazigh alphabets speech performance through an interactive voice response (IVR) system based on the losses in the G711 and GSM coding schemes. To investigate the effect of voice coding on the speech recognition rate, our designed system was trained to recognize the 33 Amazigh alphabets by using 3 and 5 HMM states with different Gaussian mixture models (GMMs) and the mel-frequency spectral coefficients (MFCCs) used to extract the feature. The speech corpus contains a total of 15,840 alphabets recorded by 30 (15 male and 15 female) Amazigh native speakers. The proposed Amazigh speech recognition system is based on the Carnegie Mellon University (CMU) Sphinx tools. Our results indicate that the best system performance is found for the G711 codec, 3 HMM, and 16 GMMs.

Keywords Speech recognition · IVR · Telephony server · Asterisk server · Amazigh language

1 Introduction

In the last decade, modern human life was totally dependent on technologies such as interactive voice response (IVR) systems which become a central data source for natural language processing applications. Also, with the evolution of automatic speech recognition (ASR) and natural language processing, several spoken dialogue systems (SDS) appear in our lives as information assistants [1].

IVR is now a commonplace in different fields like medicine where Lieberman et al. [2] investigate the ability to track the illness progression by using interactive voice response technology. They developed a system which allows symptom monitoring and as an adjunct to the treatment of chronic pain. On the other hand, the authors investigate the possibility of conducting a high-quality survey about

M. Hamidi (✉) · H. Satori · O. Zealouk · K. Satori
LISAC, Department of Mathematics and Computer Science, FSDM, USMBA, Fez, Morocco

M. Elhoseny et al. (eds.), *Distributed Sensing and Intelligent Systems*, Studies in Distributed Intelligence, https://doi.org/10.1007/978-3-030-64258-7_59

problematic alcohol and drug use in the general population based on Internet and IVR technologies [3]. In addition, the IVR solution is used as a proactive outreach for engaging the low-income relapsed smokers in a new treatment cycle [4].

Rose et al. [5] present the integration of the ASR system with VoIP-based Asterisk PBX server. In their work, HTK is used with Asterisk server. IAX and SIP protocols are utilized to make calls and communicate naturally. Aust et al. [6] have created an automatic system that permits users to ask for train traffic information using the telephone. This system connects 1200 German cities. The caller can retrieve information talking fluently with the system which behaves like a human operator. The important components of their system are speech recognition, speech understanding, dialogue control, and speech output which is executed sequentially. Bhat et al. [1] created the Speech Enabled Railway Enquiry System (SERES) which is a system that permits users to get the railway information considering the Indian scenario, as a case study to define issues that need to be fixed in order to enable a usable speech-based IVR solution.

In another ASR study [7], authors have studied the classification of speech communication channels using MFCC features and GMM-based ML classifier. They utilize various databases from several sources to build and test models. Their obtained accuracy is about 95%. The researchers have examined the robust speech recognition in the GSM mobile environment. They are focused on voice degradation due to the losses in the GSM coding platform [8, 9]. Basu et al. [10] have described the real-time challenges of designing the telephonic automatic speech recognition system. In their study, authors have used the asterisk server to design a system that asks some queries, and the spoken responses of users are stored and transcribed manually for ASR system training. In this work, the speech data are collected from West Bengal.

Satori et al. [11] have created a system based on HMM (Hidden Markov Models) using the CMU Sphinx tools. The aim of this work is the creation of automatic Amazigh speech recognition system that includes digits and alphabets of Amazigh language. The system performance achieved was 92.89%. In [12, 13], we present our first experiment to integrate the ten first digits of Amazigh language in an interactive voice response (IVR) server where the users use speech (ten first Amazigh digits) to interact with the system. In general, the Amazigh speech recognition for different fields was targeted by researchers [14–17].

In this work, we compare the VoIP Amazigh ASR system performance by varying the values of their respective parameters as codecs, HMMs, and GMMs in order to determine the influence of codecs on the system recognition rates. Also, we aim to increase the recognition accuracy by varying the different automatic speech recognition parameters for both speaker-independent and speaker-dependent.

The rest of this chapter is organized as follows: Sect. 2 presents an overview of the VoIP system and protocols. Section 3 gives an overview of automatic speech recognition system. In Sect. 4, Amazigh language is explained. In Sect. 5, Telephony Amazigh speech recognition will be discussed. Finally, Sect. 6 is experimental results. We finish with some conclusion.

2 VoIP System and Protocols

VoIP (voice over Internet protocol) is a technology during the last decade. It provides audio and video streaming facility on successful implementation in the network.

2.1 Asterisk Server

Telephony Server Asterisk is an open source and a development environment for various telecommunication applications programmed in C language. It provides establishment procedures enabling to manipulate communication sessions in progress. Asterisk supports the standard protocols: SIP, H.323, and MGCP and transformations between these protocols. It can use the IAX2 protocol to communicate with other Asterisk servers [18, 19].

2.2 Session Initiation Protocol

The session initiation protocol (SIP) is a signaling protocol which is responsible for creating media sessions between two or more participants. SIP was defined by Internet Engineering Task Force (IETF) and is simpler than H.323 and adapted more specifically for session establishment and termination in VOIP [20]. In our word, SIP was used to create the user account and to assure internetwork communication.

2.3 Real-Time Transport Protocol

The real-time transport protocol (RTP) is an Internet protocol that allows transmitting real-time data such as audio and video. RTP is a protocol which facilitates the transport of data over a network in real-time applications. It is intended to be used for applications such as audio and video conferencing, real-time systems control, and unicast or multicast services [21].

2.4 Codec

The codecs are basically different tools of mathematical used for encoding or compressing the analogue voice signal into digital bit streams and back. The various

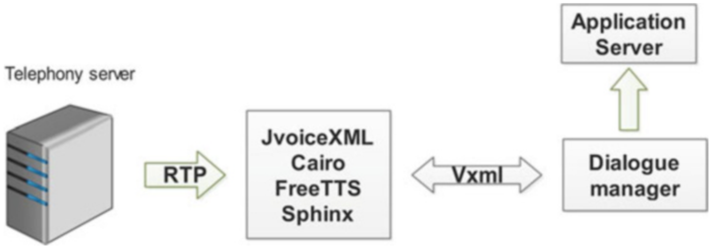

Fig. 1 Spoken dialogue system architecture

codecs are based on an algorithm of compression, data rate, and the sampling rate [22]. The VoIP codecs used in this work are G.711-u and GSM.

2.5 Spoken Dialogue System

A spoken dialogue system (SDS) is an interactive computer system, where a dialogue between the user and the computer is achieved [23]. The speech is received by the system from the user, and the response is given as an action or information. The spoken dialogue system architecture includes several components integrated together such as telephony, speech technologies, and web technologies. The scheme of spoken dialogue system is shown in Fig. 1.

3 Automatic Speech Recognition System

3.1 Speech Recognition

Speech recognition is the process of decoding the speech signal captured by the microphone and converting it into words [24]. The recognized words can be used as commands, data entry, or application control. Recently, this technology has reached a higher level of performance. The applications of speech recognition are found in several domains like healthcare, military, commercial/industrial applications, telephony, personal computers, and many other devices. Figure 2 shows the speech recognition system structure.

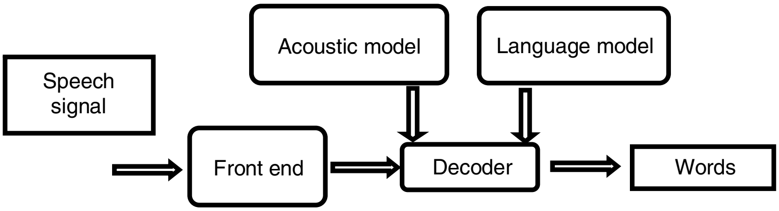

Fig. 2 Speech recognition system

3.2 MFCC Features

The mel-frequency cepstral coefficients are used in stressful speech and speaker recognition fields. These coefficients provide a high-level approximation of a human auditory perception, and they play a mainly role in the voice identification. The MFCC process details are explained in [25].

3.3 Hidden Markov Model

The Hidden Markov Model (HMM) [26] is a popular method in machine learning and statistics for modeling sequences like speech. This model is a finite ensemble of states, where each set is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. Markov models are excellent ways of abstracting simple concepts into a relatively easily computable form. It is used in data compression to sound recognition. The Markov Model makes the speech recognition systems more intelligent.

4 Amazigh Language

Before the implementation of a speech recognition voice response server system for any language, it is necessary to have a preliminary study of this language. In our case, we choose Amazigh which is a less-resourced Moroccan language. In the best of our knowledge, this is the first IVR using this language. The Amazigh language is widely spoken in a vast geographical area of North Africa. It is spoken by 28% of the Moroccan population. Our work is based on the 33 alphabets of Amazigh language which are consecrated by the foundation of the Royal Institute of Amazigh Culture (IRCAM).

5 Alphabets Speech Recognition

In this section, we describe our experience to create and develop a telephony Amazigh voice recognition system-based alphabets using Asterisk server and CMU Sphinx tools. Our experiments, both training and recognizing, were based on CMU Sphinx system, which is HMM-based, speaker-independent and speaker-independent, and continuous recognition system.

The system is created using Oracle virtual box tool on the host machine with 2 GB of RAM and an Intel Core i3 CPU of 1.2 GHz speed. The operating system used in our experiment was Ubuntu 14.04 LTS.

5.1 Speech Preparation

In this section, we describe our speech recognition database. The corpus consists of 33 spoken Amazigh letters collected from 30 Amazigh Moroccan native speakers aged between 16 and 50 years old. The audio data are recorded in wave format by using the recording tool WaveSurfer [27]. Each alphabet is pronounced 10 times. In our work, the database is partitioned to training 70% and testing 30% in order to ensure the speaker-independent aspect. For the speaker-dependent aspect, we use the same speakers in both phases training and testing.

5.2 Training Phase

In order to determine their optimal values for maximum performance, different acoustic models are prepared by varying HMMs (3–5) and GMMs (8–16–32–64). The wave recorded audio data is used in the training phase where the database is partitioned to 70% training and 30% testing in order to ensure the speaker independent aspect. In our work, we used a grammar file that includes the 33 isolated Amazigh alphabets.

5.3 Telephony Recognizing Phase

Our idea is to acquire and process audio signals from the transferred audio where the system was tested by coding audio data-based G711 VoIP audio codec. The prepared system includes two major threads: a coding–decoding audio stream and the alphabets speech recognition process.

In the first step, the audio is transferred via IVR service. In the next step, the audio signal is split into frames, and the MFCCs are calculated for each of them.

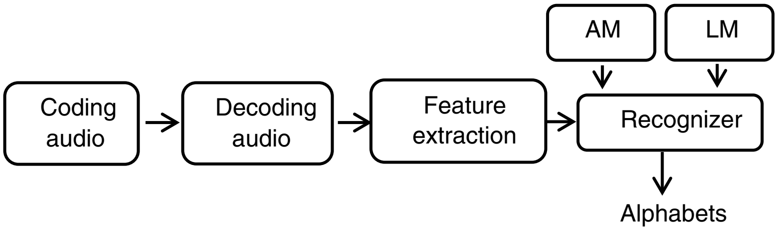

Fig. 3 The scheme of the telephony alphabets Amazigh speech recognition

Table 1 The training and testing data

Parameters	Values	Parameters	Values
Sampling rate	8–16 kHz	Token number	15,840
Number of bits	16 bits	HMMs	3–5
Audio format	WAV	GMMs	8–16–32–64
Number of speakers—Training	21	Number of g711 coded samples	2970
Number of speakers—Test	9	Number of GSM coded samples	2970

These MFCCs are expressed with GMMs parameters and compared with the stored database. The recognition rate for each alphabet was observed and recorded for each experiment. Figure 3 presents the main telephony ASR system components.

In this work, The CMU-Cambridge Statistical Language Modeling toolkit is used to generate language model of our system. Sphinx was utilized for the recognizer implementation because it is free and has been applied by many researchers all over the world.

6 Experimental Results

In order to evaluate the performances of the telephone ASR system, we train our system by the uncoded voice and we perform three recognition experiments (Exper1, Exper2, and Exper3). These experiments focused on testing the system using different subsets of the Amazigh alphabets corpus. In the first experiment, the system was tested by using the uncoded voice; the second experiment was working on the Amazigh alphabets voice decoded by the G711 audio codec; and in the third, the speech was decoded by using the GSM codec. More technical details about our system are shown in Table 1. The recognition rate for each alphabet was observed and recorded for each experiment. Our systems were trained using different HMMs and GMMs. The numbers of HMMs were 3 and 5, and Gaussian mixtures per model were 8, 16, 32, and 64. Table 2 presents the overall system recognition rate for different parameters for speaker-independent and speaker-dependent systems.

In general, the best rates are obtained when the system is tested by speaker-dependent speech using uncoded speech.

Table 2 System overall recognition rate

3 HMMs								
	8 GMMs		16 GMMs		32 GMMs		64 GMMs	
Audio codec	Diff speakers	Same speaker	Diff speakers	Same speaker	Diff speakers	Same speaker	Diff speakers	Same speaker
Exper1	88.99	97.00	87.91	96.68	86.84	94.44	86.63	93.88
Exper2	82.96	95.77	85.76	96.55	80.54	93.11	79.76	92.22
Exper3	79.39	94.44	82.19	96.22	76.97	92.22	76.20	90.78
5 HMMs								
	8 GMMs		16 GMMs		32 GMMs		64 GMMs	
Audio codec	Diff speakers	Same speaker	Diff speakers	Same speaker	Diff speakers	Same speaker	Diff speakers	Same speaker
Exper1	87.37	96.66	87.00	94.77	86.80	91.11	86.06	90.78
Exper2	80.67	94.44	83.47	93.11	78.25	90.44	77.47	90.33
Exper3	78.28	92.22	81.08	91.66	75.86	90.00	75.08	90.00

In the case of Exper1, the speaker-dependent system reaches the highest rate of 97.00% with 3 HMMs and 8 GMMs in comparison to the speaker-independent system rate which is lower by 8.01% in the same parametrization.

Concerning the second experience, the recognition rates have witnessed a decrease where the best rate for speaker-dependent amounting is 96.55% while the speaker-independent dropped at 3.23%. The lower recognition rates are obtained in the second experiment compared to the first experiment. That may be due to speech coding which degrades the voice quality.

For the Exper3, the speaker-independent system recognition rates were 79.39, 82.19, 76.97, and 76.20% for 8, 16, 32, and 64 GMMs, respectively. For 5 HMMs, the system performances were 78.28, 81.08, 75.86, and 75.08% for 8, 16, 32, and 64 GMMs, respectively. For speaker-dependent system recognition rates, the best rate is 96.22 with a difference of 0.78% with the first experiment.

Also, it is noted that in the first experiment, the best and higher accuracy was found with 3 HMMs and 8 GMMs. In the two last experiments, the system performance was better for 3 HMMs and 16 GMMs but lower for 5 HMMs and 64 GMMs.

By comparing the obtained results, we found that the voice coding has an effect on the recognition rates where a vast difference is observed between the rates achieved by using uncoded speech and decoded speech. Probably the degradation of recognition performance is due to the impact of data compression, transmission errors, or bandwidth. Based on the voice codec results, we can say that the G.711 audio codec performs better than GSM codec.

7 Conclusions

In this chapter, the Amazigh ASR system via the interactive voice response service was investigated using the sounds database corresponding to the Moroccan Amazigh language. The designed system was implemented based on the IVR system and ASR system. The results in this chapter compare the recognition performance of two audio codecs G711 and GSM based on speaker-independent and speaker-dependent approaches. The main aim of this chapter is studying the impact of data coding on the Amazigh alphabets speech recognition performance. Our findings show that the best performances for both speaker-independent and speaker-dependent are 85.76% and 96.55%, which were found by using the G711 codec with 3 HMMs and 16GMMs.

References

1. Bhat, C., Mithun, B. S., Saxena, V., Kulkarni, V., & Kopparapu, S. (2013, August). Deploying usable speech enabled ivr systems for mass use. In *Human computer interactions (ICHCI), 2013 international conference* (pp. 1–5). IEEE.
2. Lieberman, G., & Naylor, M. R. (2012). Interactive voice response technology for symptom monitoring and as an adjunct to the treatment of chronic pain. *Translational Behavioral Medicine, 2*(1), 93–101.
3. Sinadinovic, K., Wennberg, P., & Berman, A. H. (2011). Population screening of risky alcohol and drug use via internet and interactive voice response (IVR): A feasibility and psychometric study in a random sample. *Drug and Alcohol Dependence, 114*(1), 55–60.
4. Carlini, B. H., McDaniel, A. M., Weaver, M. T., Kauffman, R. M., Cerutti, B., Stratton, R. M., & Zbikowski, S. M. (2012). Reaching out, inviting back: Using interactive voice response (IVR) technology to recycle relapsed smokers back to Quitline treatment—A randomized controlled trial. *BMC Public Health, 12*(1), 507.
5. Rose, G. L., MacLean, C. D., Skelly, J., Badger, G. J., Ferraro, T. A., & Helzer, J. E. (2010). Interactive voice response technology can deliver alcohol screening and brief intervention in primary care. *Journal of General Internal Medicine, 25*(4), 340–344.
6. Aust, H., Oerder, M., Seide, F., & Steinbiss, V. (1995). The Philips automatic train timetable information system. *Speech Communication, 17*(3), 249–262.
7. Gao, D., Xiao, X., Zhu, G., et al. (2008). Classification of speech transmission channels: Landline, GSM and VoIP networks. In *Signal processing, 2008. ICSP 2008. 9th International conference on* (pp. 671–675). IEEE.
8. Salonidis, T., & Digalakis, V. (1998). Robust speech recognition for multiple topological scenarios of the GSM mobile phone system. In *Acoustics, speech and signal processing, 1998. Proceedings of the 1998 IEEE International conference* (Vol. 1, pp. 101–104). IEEE.
9. Kim, H. K., & Cox, R. V. (2001). A bitstream-based front-end for wireless speech recognition on IS-136 communications system. *IEEE Transactions on Speech and Audio Processing, 9*(5), 558–568.
10. Basu, J., Bepari, M. S., Roy, R., & Khan, S. (2013). Real time challenges to handle the telephonic speech recognition system. In *Proceedings of the fourth international conference on signal and image processing 2012 (ICSIP 2012)* (pp. 395–408). Springer.
11. Satori, H., & ElHaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. *International Journal of Speech Technology, 17*(3), 235–243.
12. Hamidi, M., Satori, H., & Satori, K. (2016). Amazigh digits speech recognition on IVR server. *Advances in Information Technology: Theory and Application, 1*(1).
13. Hamidi, M., Satori, H., & Satori, K. (2016). Implementing a voice interface in VOIP network with IVR server using Amazigh digits. *The International Journal of Multi-Disciplinary Sciences, 2*(2), 38–43.
14. Satori, H., Zealouk, O., Satori, K., & ElHaoussi, F. (2017). Voice comparison between smokers and non-smokers using HMM speech recognition system. *International Journal of Speech Technology, 20*(4), 771–777.
15. Zealouk, O., Satori, H., Hamidi, M., Laaidi, N., & Satori, K. (2018). Vocal parameters analysis of smoker using Amazigh language. *International Journal of Speech Technology, 21*(1), 85–91.
16. Hamidi, M., Satori, H., Zealouk, O., Satori, K., & Laaidi, N. (2018). Interactive voice response server voice network administration using Hidden Markov Model Speech Recognition System. In *2018 Second world conference on smart trends in systems, security and sustainability (WorldS4)* (pp. 16–21). IEEE.
17. Hamidi, M., Satori, H., Zealouk, O., & Satori, K. (2019). Amazigh digits through interactive speech recognition system in noisy environment. *International Journal of Speech Technology, 23*, 101–109.

18. Spencer, M., Allison, M., & Rhodes, C. (2003). *The asterisk handbook*. Asterisk Documentation Team.
19. Madsen, L., Van Meggelen, J., & Bryant, R. (2011). *Asterisk: The definitive guide*. O'Reilly Media.
20. Handley, M., Schulzrinne, H., & Schooler, E., et al. (1999). *SIP: Session initiation protocol*.
21. Schulzrinne, H. (1996). *RTP: A transport protocol for real-time applications*. Internet Engineering Task Force, Audio-Video Transport Working Group. RFC 1889.
22. Karapantazis, S., & Pavlidou, F. N. (2009). VoIP: A comprehensive survey on a promising technology. *Computer Networks, 53*(12), 2050–2090.
23. Sharma, S. R. (2009). U.S. patent no. 7,502,737. U.S. Patent and Trademark Office.
24. Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development* (Vol. 95). Prentice Hall PTR.
25. Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. In *arXiv preprint arXiv:1003.4083*.
26. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.
27. Wavesurfer. (2016). https://sourceforge.net/projects/wavesurfer/