Annals of Telecommunications (2023) 78:617–625
https://doi.org/10.1007/s12243-023-00968-5

Telephony speech system performance based on the codec effect

Mohamed Hamidi[1] · Ouissam Zealouk[2] · Hassan Satori[2]

Received: 21 April 2022 / Accepted: 13 May 2023 / Published online: 31 May 2023

Abstract
This paper is a part of our contribution to research on the enhancement of network automatic speech recognition system performance. We built a highly configurable platform by using hidden Markov models, Gaussian mixture models, and Mel frequency spectral coefficients, in addition to VoIP G.711-u and GSM codecs. To determine the optimal values for maximum performance, different acoustic models are prepared by varying the hidden Markov models (from 3 to 5) and Gaussian mixture models (8–16-32) with 13 feature extraction coefficients. Additionally, our generated acoustic models are tested by unencoded and encoded speech data based on G.711 and GSM codecs. The best parameterization performance is obtained for 3 HMM, 8–16 GMMs, and G.711 codecs.

Keywords Interactive system · Hidden Markov model · Speech recognition · Codecs · Feature extraction

1 Introduction

Speech is the main communication style of humans and the most natural way to exchange information. Therefore, several studies have been performed in past decades to design an ideal automatic speech recognition (ASR) system that is capable of understanding speech and sounds in real time under different conditions. However, this capability remains an established requirement for newly developed speech systems. The significant differences in speech cues, such as the absence of distinct boundaries between words or phonemes, and unwanted noise cues caused by the variability of speakers and their surroundings, such as speed of speech, style of speaking, and accents, renders this task more challenging [1, 2]. In addition, the degradation of speech recognition performance over IP networks was one of the main challenges faced by network speech recognition (NSR) researchers. In the NSR case, the client–server architecture was implemented by placing a server-side recognizer using a standard speech encoder. A speech signal is encoded by a conventional speech codec and transmitted to the server for decoding, feature extraction, and recognition phases [3]. The network dependency, coding, and transmission of data degrade the recognition performance due to the impact of data compression, transmission errors, or transcoding [4]. Table 1 presents the automatic speech recognition performance based on VoIP codecs. Table 2 presents automatic speech recognition systems based on audio codec and interactive voice response (IVR) technology.

On the other hand, we present some ASR systems based on hidden Markov model (HMM) and Gaussian mixture model (GMM) approaches.

T. K. DAS et al. [5] designed a speech information system based on HMMs and mel-frequency cepstral coefficients (MFCCs). Their best-achieved result is approximately 90%. H. Satori and F. El Haoussi [6] implemented an Amazigh speech system including digits and letters based on CMU Sphinx tools. Their achieved system performance was 92.8%. The authors in [7] presented an automatic speech recognition system by using the Odia language. The Kaldi toolkit is used to realize the automatic recognition system. Mono-phone and triphone models are investigated for Odia speech recognition, and Odia acoustic modeling is performed using the HMM and GMM.

Voice signal quality plays a major role in augmenting speech recognition system performance. In the case of the network ASR system, the speech signal is encoded by an audio VoIP codec and then transmitted to the server

✉ Mohamed Hamidi
mohamed.hamidi.5@gmail.com

1 Team of Modeling and Scientific Computing, FPN, UMP, Oujda, Morocco

2 Department of Mathematics and Computer Science, LISAC, FSDM, USMBA, Fez, Morocco

Table 1 Automatic speech recognition performance based on VoIP codecs

Ref	Description	Codec	Results
[23]	Using an automated speech synthesis pipeline to encode speech samples instead of regular speech encoders, in situations requiring high data compression with high packet loss scenarios	PCM A-law	TTS (Perfect) 88.84% PCM (0% loss) 91% PCM (5% loss) 89.40% PCM (10% loss) 84.05%
[24]	Analyzing the effect of Opus compression on a combined beamforming and ASR system, gives guidelines about the optimization of compression parameters for far-field ASR. In addition, the authors have proposed a microphone-independent multichannel coding scheme	Opus	Bitrate reduction of 37.5% or a 5.1% relative error rate (WER)
[25]	Evaluation of the perceived quality of commonly used VoIP codecs in the presence of background noise at different loudness levels	PCMUPCMAGSM iLBC G729Speex8K	VoIP speech using the PCMU and Speex8K codecs are the most consistent in terms of perceived quality performance under different loudness conditions
[26]	Proposition of a packet loss concealment method to increase the robustness of ASR for speech encoded with the G729 codec, over Voice over Internet Protocol (VoIP)	G729	G729 (0% loss) 90% G729 (30% loss) 70%
[27]	Evaluation of Amazigh speech recognition system through wireless network based on a combination of both ASR and IVR technologies	G.711, GSM Speex	The best performance is 84.14% achieved by using the GSM audio codec

for recognition. This process degrades the quality of the received speech, which affects system performance. In this paper, we have implemented a network ASR system based on IVR and ASR technologies, where a degradation of recognition rates was observed with the integration of the IVR method that is based on the network ASR process. For this reason, we evaluated the performances of the VoIP-ASR system by varying the values of their respective parameters as codecs, HMMs and GMMs, to determine the optimal values for maximum performance.

The remainder of the paper is organized as follows: "Section 2" presents the IVR service and ASR technology. "Section 3" presents the system preparation. Section 4 presents the system architecture. Section 5 presents the conducted experiments. "Section 6" presents the results. "Section 7" presents the comparisons. "Section 8" concludes the paper.

Table 2 IVR-ASR systems performance based on VoIP audio codecs

Ref	Description	Approach	Results
[28]	S. Ayaz et al. have presented a pattern recognition method based on the interactive voice response system and neural networks approach. Their implemented system is aimed at identifying the user's voice by using telephony secure access	IVR MFCC MLP PCM	84%
[29]	Evaluation of the performance of various modern classification methods and adjusting their parameters to aid in the selection of optimal classification methods for gender recognition tasks	IVR SVM KNN NB MLP RF	The SVM is the best classifier among all the five schemes for gender recognition
[30]	Researchers describe the Amazigh speech recognition performance via interactive voice response under noisy conditions. The experiments were conducted for unencoded speech and then repeated for decoded speech in the noisy environment of a train for different signal-to-noise ratios (SNR)	HMMs GMMS G711 GSM	The most affected digits are those containing the consonant "S", which rapidly drops to 0% in 30 dB and 27 dB for unencoded speech and decoded speech, respectively
[31]	The proposed system offers a methodology to remotely extract data from a distance database using the combined interactive voice response (IVR) and automatic speech recognition (ASR) technologies	HMMs GMMS G711	The best-obtained performance is 89.64% by using 3 HMMs and 16 GMMs
[32]	The authors evaluate the influence of G711 and GSM audio codecs on the speech recognition performance based on IVR-ASR vocabulary system that includes the Amazigh letters	HMMs GMMS G711	Unencoded voice 88.99% G711 85.76% GSM 82.19%

2 IVR service and ASR technology

2.1 Audio codecs

Codecs are techniques used for encoding or compressing analog voice signals into digital bitstreams and then back to analog voice signals. There are different codecs, varying in complexity, necessary bandwidth, and voice quality, where better voice quality requires more bandwidth. One problem that emerges in the distribution of high-quality speech is network performance. In this study, our IVR implementation is based on the SIP signaling protocol [8] and RTP protocol [9] with G.711 and GSM codecs, which are employed as VoIP parameters [10].

2.1.1 G711 codec

G.711 [11] is a pulse code modulation (PCM) scheme that generates one 8-bit value every 125.ls, assured in a 64 kb/s bitstream. Speech data are encoded as 8 bits after logarithmic scaling. This audio codec includes two versions, u-Law, which is utilized in North America/Japan, and A-Law, which is exploited in Europe and the rest of the world. The A-Law version permits the conversion of 13-bit linear PCM samples into 8-bit compressed PCM samples, and the decoder performs the conversion, and vice versa, while the u-Law version allows the conversion of 14-bit linear PCM samples into 8-bit compressed PCM samples.

2.1.2 GSM codec

The ETSI GSM 06.10 Full Rate (FR) codec is the first digital speech-coding standard utilized in the Global System for Mobile Communications digital mobile phone systems, operating on an average bitrate of 13 kb/s. This audio codec was introduced in 1987 and exploits the RPE-LTP (regular pulse excitation–long term prediction) linear predictive coding principle [12].

2.2 Automatic speech recognition

Automatic speech recognition (ASR) is defined as the independent computer-driven transcription of spoken language into readable text [6]. Figure 1 shows a typical ASR architecture. Recently, our lab researchers targeted the applications of automatic speech recognition for the Moroccan Amazigh language [13–19].

2.3 MFCC feature extraction technique

The extraction of mel-frequency cepstral coefficients (MFCC) [20] includes an analysis based on the frames of an input speech, where the speech signal is segmented into a sequence of frames. Each frame offers a sinusoidal transformation (fast Fourier transform) to generate certain parameters, which then undergo a scale of perception on the mel-scale and decorrelation. The obtained output was a sequence of feature vectors that describe a logarithmically useful compressed amplitude and simplified frequency information. Figure 2 presents the detailed technique on the principle of cepstral analysis.

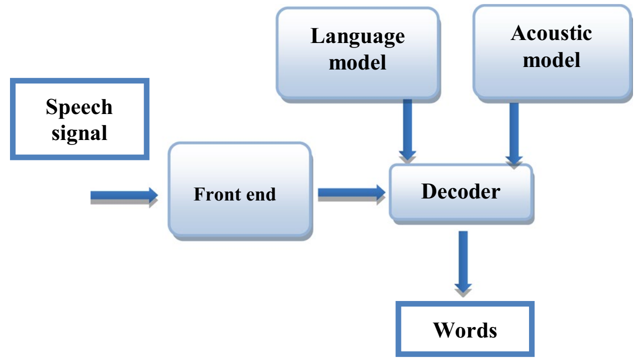

Fig. 1 ASR system architecture

3 System preparation

3.1 Database preparation

The utterances used to evaluate and compare the system performance are collected from 24 Amazigh native speakers aged between 14 and 40 years old. The speech data were recorded in wave format. The applied sampling rates are 8 and 16 kHz. The corpus consists of 10 Amazigh spoken digits (0–9). Each digit is pronounced 10 times in detached data files, and each file includes one pronounced word. The selected digits and their transcription are shown in Table 3. More technical details about our system are shown in Table 4.

3.2 Files preparation

To prepare our acoustic model, we classed a set of input data and processes by exploiting the SphinxTrain tool. The following list presents the input files and data.

- Audio wave dataset
- List of fillers
- List of files for training and testing
- Transcription for training and testing
- Dictionary that determines the pronunciation of selected digits (Table 5)

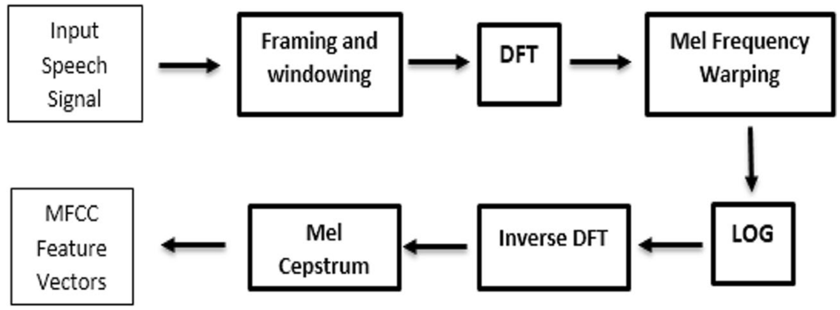

Fig. 2 The MFCC process [12]

Table 3 Ten Amazigh digits with their English transcription

English transcript-ion	Transcrip-tion	Arabic transcr-iption	Number correspond-ence	*N* of syll-abes
AMYA	A M Y A	اميا	0	2
YEN	Y E N	يان	1	1
SIN	S I N	سين	2	1
KRAD	K R A D	كراض	3	2
KUZ	K O Z	كوز	4	1
SMMUS	S M U S	سموس	5	2
SDES	S D E SS	سضيس	6	1
SA	S A	سا	7	1
TAM	T A M	تام	8	1
TZA	T Z A	تزا	9	2

Table 4 System parameters

Parameters	Values
Sampling rate	8/16 kHz
Number of bits	16 bits/8 bits
Audio format	WAV
Number of speakers—training	17
Number of speakers–test	7
Number of repetitions	10
MFCC	13
Recording conditions	Normal environment

Table 5 Dictionary file

AMYA	A M Y A
YEN	Y E N
SIN	S I N
KRAD	K R A D
KUZ	K O Z
SEMUS	S M U S
SEDISS	S D E SS
SA	S A
TAM	T A M
TZA	T Z A

Table 6 Prepared acoustic systems

HMM states	GMMs	Systems	Acoustic model
3	8	Amsystem 3–8	Amacous 3–8
	16	Amsystem 3–16	Amacous 3–16
	32	Amsystem 3–32	Amacous 3–32
5	8	Amsystem 5–8	Amacous5–8
	16	Amsystem 5–16	Amacous 5–16
	32	Amsystem 5–32	Amacous 5–32

- Language model that gives a representation of the occurrence probability for each digit

The phonetic dictionary was prepared in such a way that it consists of all expected digits with possible variants of their pronunciation. The careful and serious preparation of the input data and files plays a crucial role in designing a speech recognition system.

3.3 ASR parametrization

To evaluate the ASR system performances, several ASR configurations were prepared using HMM-states and GMMs. We prepared 6 acoustic models by varying the HMM states from 3 to 5 and the Gaussian distributions from 8 to 32. Table 6 presents different acoustic models utilized in our work.

4 Telephony spoken system architecture

The telephony-spoken system is an interactive pattern system in which a dialog between the user and the system is realized. As shown in Fig. 3, the main modules of our telephony spoken system architecture are IVR and ASR.

In the IVR part, the system receives voice input when the user interacts with the server by voice commands, the codec converts the analog waveforms into digital signals for the transmission as IP packets over the network, and

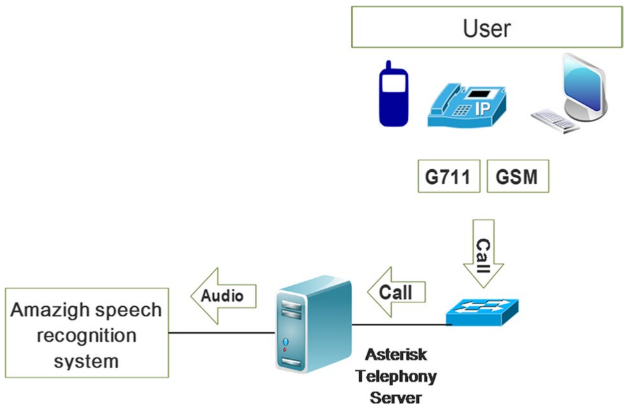

Fig. 3 Model for establishing speech recognition via the Asterisk server

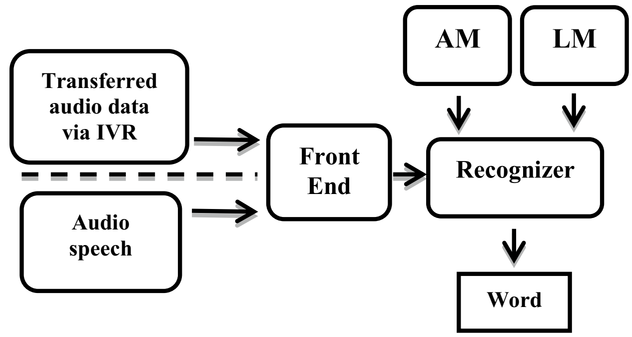

Fig. 4 Speech recognition process

then it converts the digital signals back to analog waveforms. In our study, we focused on voice traffic coding by using G.711u and GSM speech codecs.

In the ASR part, the Amazigh speech recognition system receives the transferred voice data from the Asterisk server. The received data were modeled as a sequence of phonemes, while each phoneme was modeled as a sequence of HMM states. We have used 3 and 5 HMM architectures for each phoneme, one emitting state (or three emitting states) and two non-emitting states as the entry and exit, which join the models of HMM units in the ASR engine. Each emitting state includes Gaussian mixtures trained on 13-dimensional MFCC coefficients, their delta and delta-delta vectors, which are extracted from the signal. The distribution of features for a phone was modeled with 8, 16, and 32 GMMs. Table 7 presents the feature extraction parameterization.

Our principal aim is to find a balance between an acceptable recognition rate and the choice of optimal parameters (HMMs, GMMs, and codecs). Figure 4 shows our process of speech recognition.

Table 7 Feature extraction parameterization

Parameter	Value
Hamming	25.6 ms
Filter:	Mel-frequency filter bank
Frame rate	100 frames per second
Cepstra number	13
Mel filters number	25
DFT size	256
MFCC feature vector	13
Overall feature vector dimension	39

5 Experiments

In this section, all phases of the system (training and recognition) were based on the CMU Sphinx system, which is based on the HMM-GMM combination.

Our approach for modeling the encoded Amazigh sounds consisted of generated and trained acoustic models by using the unencoded voice and testing the system by an encoded voice by varying the audio codecs, HMMs and GMMs.

Seventy percent of the database (collected audio) was utilized for training to ensure speaker independence and the reliability and validity of our system. In the recognition phase, we test the system by 30% of the database (coding data with G711 and GSM codecs). The experimental setups are.

- Software: our setup is based on the open source software Asterisk 1.6, Ekiga is used in the IVR part, CMU Sphinx Tools are used in the ASR part, and the operating system is the Ubuntu 14.04 LTS.
- Hardware: The hardware consists of a laptop with an Intel Core i3 CPU with a speed of 2.4 GHz speed and 4G of RAM.

6 Experimental results

This section presents the results of proposed systems.

6.1 Case 1: Testing the unencoded data with unencoded trained models

Table 8 shows our achieved accuracies of the system, which is trained and tested by using the unencoded voice with three and five HMM states related to 8, 16, and 32 Gaussian mixture distributions. The best result of 91.57% was obtained with 3–16 HMMs–GMMs, where the lowest result of 85.86% was achieved by 5–32 HMMs–GMMs.

By considering the individual word performance of the IVR-ASR system, the highest recognition rate is 92.86% for the words "krad," "smmus," "sdes," and "tza" for Amsystem3-16.

Table 8 System recognition rates based on unencoded data

	Unencoded data					
	3 HMM			5 HMM		
Amazigh digit	8 GMM	16 GMM	32 GMM	8 GMM	16 GMM	32 GMM
AMYA	90.00	90.00	88.57	88.57	90.00	87.14
YEN	88.57	90.00	85.71	87.14	88.57	82.86
SIN	90.00	91.43	88.57	87.14	90.00	87.14
KRAD	91.43	92.86	88.57	90.00	88.57	85.71
KOZ	91.43	91.43	88.57	87.14	91.43	88.57
SMMUS	91.43	92.86	90.00	90.00	91.43	85.71
SDES	91.43	92.86	87.14	88.57	88.57	85.71
SA	88.57	90.00	85.71	88.57	85.71	85.71
TAM	90.00	91.43	87.14	87.14	90.00	82.86
TZA	91.43	92.86	90.00	90.00	90.00	87.14
Total	90.43	91.57	88.00	88.43	89.43	85.86

Based on this finding, we suggest that the number of syllables probably plays a positive role in the accuracy rate increment. Therefore, the lower performance word achieved by the Amsystem5-32 model is "yen." A comparison of the results indicates that our work is in accordance with the results of [6].

6.2 Case 2: Testing the coded-decoded data (G.711 codec) with trained models

In this case, we keep the same trained acoustic models but change the test corpus by an encoded audio test database. In the case of 3 HMM states, the obtained results are 89.71, 88.71, and 87.86% by adopting 8, 16, and 32 Gaussian mixtures, respectively. In 5 HMMs, the system correct rates were 88.28, 87.86, and 85.86%, corresponding to 8, 16, and 32 GMMs, respectively. A higher recognition rate of 89.71% was achieved by the combination of 3–8 HMMs–GMMs (Table 9). The results that we obtained through experiments show that there is a difference in speech recognition for the two categories (unencoded and G 711). The lower recorded recognition rate is 85.86%, which was obtained by testing the system via Amsystem 5–32.

The analysis of the individual word performance showed that the best performance for "Amya" and "tza" words is achieved by the 3HMMs-8GMMs, 3HMMs-16GMMs, and 5HMMs-16GMMs combinations.

For the "krad" and "smmus" digits, the best accuracy is obtained by 3HMMs-8GMMs, 3HMMs-16GMMs, and 5HMMs-8GMMs.

The ASR parameter comparison between the first case and the second case shows that for the unencoded voice, the best results are obtained by testing data with the Amacous 3–16 trained model, and for the G.711-coded data, the higher results are obtained by testing data with the Amacous 3–8 trained model.

Table 9 System recognition rates based on the G711 codec

	G.711					
	3 HMM			5 HMM		
Amazigh digit	8 GMM	16 GMM	32 GMM	8 GMM	16 GMM	32 GMM
AMYA	87.14	87.14	85.71	85.71	87.14	84.29
YEN	90.00	88.57	88.57	87.14	88.57	87.14
SIN	88.57	87.14	87.14	85.71	87.14	84.29
KRAD	91.43	91.43	90.00	91.43	87.14	85.71
KOZ	92.86	90.00	91.43	92.86	90.00	88.57
SMMUS	90.00	90.00	88.57	90.00	88.57	87.14
SDES	87.14	87.14	85.71	85.71	85.71	84.29
SA	88.57	85.71	85.71	87.14	87.14	85.71
TAM	90	88.57	87.14	88.57	85.71	84.29
TZA	91.43	91.43	88.57	88.57	91.43	87.14
Total	89.71	88.71	87.86	88.28	87.86	85.86

Table 10 System recognition rates based on the GSM codec

	GSM					
	3 HMM			5 HMM		
Amazigh digit	8 GMM	16 GMM	32 GMM	8 GMM	16 GMM	32 GMM
AMYA	88.57	88.57	87.14	87.14	85.71	84.29
YEN	84.29	85.71	84.29	85.71	84.29	82.86
SIN	87.14	88.57	85.71	87.14	85.71	85.71
KRAD	91.43	91.43	90.00	87.14	90.00	87.14
KOZ	88.57	90.00	87.14	85.71	82.86	82.86
SMMUS	90.00	90.00	90.00	88.57	90.00	88.57
SDES	88.57	87.14	87.14	85.71	84.29	82.86
SA	84.29	85.71	82.85	82.85	82.86	82.86
TAM	85.71	88.57	84.29	87.14	85.71	84.29
TZA	88.57	88.57	87.14	85.71	84.29	84.29
Total	87.71	88.43	86.57	86.28	85.57	84.57

6.3 Case-3: Testing the coded-decoded data (GSM Codec) with trained models

For the GSM case (Table 10), the obtained accuracy is lower than that of G711. When the models were trained by the unencoded speech and tested by GSM decoded speech, the best and lowest recognition rates were 88.43% for Amsystem 3–8 and 84.57% for Amsystem 5–32, respectively. By considering the words' individual performance, our finding shows that the higher recognition rate is 91.40% for "krad" obtained with the Amsystem 3–16 model. Table 9 shows the measured recognition rate for the GSM codec. The GSM-decoded signal causes degradation in speech recognition rates due to the distortions introduced to cepstral representations.

7 Best-case comparison

In this section, we present the confusion matrices of our best-obtained accuracies that are achieved with unencoded and G711-decoded speech. The testing set includes 700 utterances from seven speakers. Table 11 presents the performance comparison of our proposed method with some of the existing works in the same field.

Table 12 shows the confusion matrix of the system based on the unencoded speech. The global accuracy from this experience is 91.57%. Table 13 presents the system confusion matrix for the encoded speech using the G 711 audio codec. The overall performance of the G 711 codec was 89.71%, which is similar to the overall performance

Table 11 Comparison of our proposed method with some existing works

Ref	Year	Method	Results
[23]	2019	PCM A-law	PCM (0% loss) 91% PCM (5% loss) 89.40% PCM (10% loss) 84.05%
[26]	2018	G729	G729 (0% loss) 90% G729 (30% loss) 70%
[27]	2023	G.711 GSM Speex	84.14%
[28]	2009	IVR MFCC MLP PCM	84%
[31]	2020	Varied Amazigh speech system based on HMMs, GMMS and G711 codec	89.64%
[32]	2019	Interactive system based on Amazigh alphabets HMMs	Unencoded voice 88.99% G711 85.76% GSM 82.19%
Our work	-	IVR-ASR based on digits	Unencoded 91.57 G711 89.71 GSM 88.43

Table 12 Confusion matrix for the best recognition rates (unencoded voice)

	AMYA	YEN	SIN	KRAD	KOZ	SMMUS	SDES	SA	TAM	TZA	Omitted	Substitutions
AMYA	63	-	-	-	-	-	-	-	-	-	7	0
YEN	-	63	-	-	-	-	-	-	-	-	7	0
SIN	-	-	64	-	-	-	-	-	-	-	6	0
KRAD	-	-	-	65	-	-	-	-	-	-	5	0
KOZ	-	-	-	-	64	-	-	-	-	-	6	0
SMMUS	-	-	-	-	-	65	-	-	-	-	5	0
SDES	-	-	-	-	-	-	65	-	-	-	5	0
SA	-	-	-	-	-	-		63	-	4	3	4
TAM	-	-	-	-	-	-	-	-	64	-	6	0
TZA	-	-	-	-	-	-	-	-	-	65	5	0

Table 13 Confusion matrix for the best audio codec performance (G711)

	AMYA	YEN	SIN	KRAD	KOZ	SMMUS	SDES	SA	TAM	TZA	Omitted	Substitutions
AMYA	61	-	-	-	-	-	-	-	-	-	9	0
YEN	1	63	3	-	-	-	-	-	-	-	3	4
SIN	-	-	62	-	-	-	-	-	-	-	8	0
KRAD	-	-	-	64	-	-	-	-	-	-	6	0
KOZ	-	-	-	1	65	-	-	-	1	-	3	2
SMMUS	-	-	-	-	-	63	-	-	-	-	7	0
SDES	1			1	2	-	61	-	-	-	5	4
SA	-	-	-	2	-	-	-	62	-	4	2	6
TAM	-	-	-	-	-	-	-	-	63	-	7	0
TZA	-	-	-	-	-	-	-	-	-	64	6	0

achieved by the noncoded speech. However, the confusion matrices of both experiences show important differences.

The analysis of the substituted words showed the following findings:

- For unencoded voice, the exchange error involves two symmetrical substitutions that can be schematically represented [21] as SA ~ TZA, where inclusion of SA would bias the matrix toward symmetry.
- For decoded voice, the substitution words increase, especially for the digits YEN, KOZ, SDES, and SA, and all these words are monosyllabic.

Generally, the omitted and substitution words increase in encoded voice. This behavior may be attributed to the effect on the ASR system when the actual pronunciation is different from what the recognizer expects or the deviation of the pronounced consonants via the telephonic channel that is in accordance with those of [22].

8 Conclusions

In this paper, we have evaluated the performances of an interactive speech recognition system based on the ASR and IVR technologies. We have searched for a balance between an acceptable recognition rate and the choice of optimal parameters (HMMs, GMMs, and codecs). The best system performance by considering the IVR parameterization is observed for the G711 codec. On the other hand, the best ASR parameterization for the combined system is three HMMs and 8–16 GMMs. Moreover, we have observed that the substitution words increase for the monosyllabic words in the case of encoded speech. Despite these results, certain limitations, such as background noise or speaking accent, influence the recognition rate of our proposed system.

In our future work, we will exploit the deep learning approach to improve the performance of the IVR-ASR system with a large voice database.

Data Availability The speech database utilized in this study belongs to the laboratory and is its property.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Walid M, Bousselmi S, Dabbabi K, Cherif A (2019) Real-time implementation of isolated-word speech recognition system on raspberry Pi 3 using WAT-MFCC. IJCSNS 19(3):42
2. Hamidi M, Zealouk O, Satori H, Laaidi N, Salek A (2022) COVID-19 assessment using HMM cough recognition system. Int J Inf Technol 1–9
3. Kim HK, Cox RV (2001) A bitstream-based front-end for wireless speech recognition on IS-136 communications system. IEEE Trans Speech Audio Process 9(5):558–568
4. Lilly BT, Paliwal KK (1996) Effect of speech coders on speech recognition performance. In Proceedings of ICSLP, 2344–2347
5. Das TK, Nahar KM (2016) A voice identification system using hidden Markov model. Indian J Sci Technol 9(4)
6. Satori H, Elhaoussi F (2014) Investigation Amazigh speech recognition using CMU tools. Int J Speech Technol 17(3):235–243
7. Karan B, Sahoo J, Sahu PK (2015) Automatic speech recognition based Odia system. In Microwave, Optical and Communication Engineering (ICMOCE), International Conference on (pp. 353–356). IEEE
8. Micolini O, Herrera A, Erlang AM (2013) Traffic analysis over a VoIP server. 11(1):370–375
9. Handley M, Schulzrinne H, Schooler H et al (1999) RFC 2543. Session Initiation Protocol, SIP
10. RFC3550-IETF, R. T. P. (2003) A transport protocol for real-time applications internet engineering Task Force
11. Kumar A, Thorenoor SG (2011) Analysis of IP Network for different quality of service. In International Symposium on Computing, Communication, and Control (ISCCC), Proc. of CSIT Vol. 1
12. Karapantazis S, Pavlidou FN (2009) VoIP: a comprehensive survey on a promising technology. Comput Netw 53(12):2050–2090
13. Zealouk O, Satori H, Hamidi M, Laaidi N, Satori K (2018) Vocal parameters analysis of smoker using Amazigh language. Int J Speech Technol 21(1):85–91
14. Zealouk O, Satori H, Hamidi M, Satori K (2019) Speech recognition for moroccan dialects: feature extraction and classification methods. J Adv Res Dyn Control Syst 11(2):1401–1408
15. Lounnas K, Abbas M, Lichouri M, Hamidi M, Satori H, Teffahi H (2022) Enhancement of spoken digits recognition for under-resourced languages: case of Algerian and Moroccan dialects. Int J Speech Technol 25(2):443–455
16. Zealouk O, Satori H, Hamidi M, Satori K (2018. Voice pathology assessment based on automatic speech recognition using Amazigh digits. In Proceedings of the 2nd International Conference on Smart Digital Environment. ACM, pp. 100–105
17. 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, Secur Sustain (WorldS4) (pp. 16–21). IEEE
18. Zealouk O, Hamidi M, Satori H, Satori K (2020) Amazigh digits speech recognition system under noise car environment. In Embedded systems and artificial intelligence: Proceedings of ESAI 2019, Fez, Morocco (pp. 421–428). Springer Singapore
19. Boutazart Y, Satori H, Anselme RAM, Hamidi M, Satori K (2023) COVID-19 dataset clustering based on K-means and EM algorithms. Int J Adv Comput Sci Appl 14(3):924–934
20. Zheng F, Zhang G, Song Z (2001) Comparison of different implementations of MFCC. J Comput Sci Technol 16(6):582–589
21. Shattuck-Hufnagel S, Klatt DH (1979) The limited use of distinctive features and markedness in speech production: evidence from speech error data. J Verbal Learn Verbal Behav 18(1):41–55
22. Fosler-Lussier E, Morgan N (1999) Effects of speaking rate and word frequency on pronunciations in convertional speech. Speech Commun 29(2–4):137–158
23. Lero RD, Exton C, Le Gear A (2019) Communications using a speech-to-text-to-speech pipeline. In 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (pp. 1–6). IEEE
24. Drude L, Heymann J, Schwarz A, Valin JM (2021) Multi-channel Opus compression for far-field automatic speech recognition with a fixed bitrate budget. Preprint arXiv:2106.07994
25. Das S, Choudhury P (2020) Evaluation of perceived speech quality for VoIP codecs under different loudness and background noise condition. In Proceedings of the 21st International Conference on Distributed Computing and Networking (pp. 1–5)
26. Bakri A, Amrouche A, Abbas M, Bouchakour L (2018) Automatic speech recognition for VoIP with packet loss concealment. Procedia Comput Sci 128:72–78
27. Hamidi M, Zealouk O, Satori H (2023) Automatic speech recognition analysis over wireless networks. In: Bhateja, V., Yang, XS., Chun-Wei Lin, J., Das, R. (eds) Intelligent data engineering and analytics. FICTA 2022. Smart Innovation, Systems and Technologies, vol 327. Springer, Singapore
28. Shah SAA, ul Asar A, Shaukat SF (2009) Neural network solution for secure interactive voice response. World Appl Sci J 6(9):1264–1269, ISSN 1818- 4952
29. Ahmad J, Fiaz M, Kwon SI, Sodanil M, Vo B, Baik SW (2016) Gender identification using MFCC for telephone applications-a comparative study, arXiv preprint arXiv: 1601.01577
30. Hamidi M, Satori H, Zealouk O, Satori K (2020) Amazigh digits through interactive speech recognition system in noisy environment. Int J Speech Technol 23(1):101–109
31. Hamidi M, Satori H, Zealouk O, Satori K (2020) Interactive voice application-based amazigh speech recognition. In Embedded Systems and Artificial Intelligence (pp. 271–279). Springer, Singapore
32. Hamidi M, Satori H, Zealouk O, Satori K (2019) Speech coding effect on amazigh alphabet speech recognition performance. J Adv Res Dyn Control Syst 11(2):1392–1400