International Journal of Speech Technology (2020) 23:885–892
https://doi.org/10.1007/s10772-020-09764-1

Noise effect on Amazigh digits in speech recognition system

Ouissam Zealouk[1] · Hassan Satori[1] · Naouar Laaidi[1] · Mohamed Hamidi[1] · Khalid Satori[1]

Received: 28 January 2020 / Accepted: 21 October 2020 / Published online: 5 November 2020

Abstract
Automatic Speech Recognition (ASR) for Amazigh speech, particularly Moroccan Tarifit accented speech, is a less researched area. This paper focuses on the analysis and evaluation of the first ten Amazigh digits in the noisy conditions from an ASR perspective based on Signal to Noise Ratio (SNR). Our testing experiments were performed under two types of noise and repeated with added environmental noise with various SNR ratios for each kind ranging from 5 to 45 dB. Different formalisms are used to develop a speaker independent Amazigh speech recognition, like Hidden Markov Model (HMMs), Gaussian Mixture Models (GMMs). The experimental results under noisy conditions show that degradation of performance was observed for all digits with different degrees and the rates under car noisy environment are decreased less than grinder conditions with the difference of 2.84% and 8.42% at SNR 5 dB and 25 dB, respectively. Also, we observed that the most affected digits are those which contain the "S" alphabet.

Keywords Automatic speech recognition system · Amazigh language · Hidden markov model · Sphinx4 · Noise

1 Introduction

Speech Recognition is the process of converting a speech signal to a sequence of words based on algorithms. Recently, it has become more popular as an input mechanism in several computer applications. The Automatic Speech Recognition systems performance degrades considerably when speech is corrupted by background noise not seen during training where the reason is the observed speech signal does no longer match the distributions derived from the training material. There have been many approaches that aim at solving this mismatch, such as speech features normalization or improvement to remove the corrupting noise from the observations prior to recognition (Yu et al. (2008)), acoustic models compensation (Moreno et al. 1996; Gales and Young 1996) and using the recognizer architectures that use only the least noisy observations (Raj and Stern 2005). Lee et al. (2009) have combined enhancement of speech with endpoint detection and discrimination of the speech/non-speech in a commercial application. Authors in (Kim and Stern (2009)) have presented a new noise robust frontend method and compared to different noise conditions. Model adaptation methods staying the observations unaltered and make updating the recognizer model parameters for giving a more observed speech representative, e.g. (Li et al. 2007; Hu et al. 2006; Seltzer et al. 2010). These approaches can be further enhanced by using different conditions training data and adaptive training techniques. The researchers in (Kalinli et al. (2010)) have developed an algorithm which permits adapting the training noise that can be used to all training data which contains environmental noise. The noise adaptive training estimates the implicit "pseudo-clean" model parameters without based on the intermediate step as the clean speech features. In another study (Janicki and Wawer 2013) a computer game was created for the Polish language based on CMU Sphinx4, The authors tested the performance of driven voice continuous automatic speech recognition under clean and different noise conditions. Their obtained results show that the achieved accuracy with clean speech is about 97.6% and a minor degradation of performance was observed in-car environment, however, accuracy decreased severely for babble and factory noises for SNR below 20 dB. Alotaibi et al. (Alotaibi et al. 2009) aim to recognize the alphabets and digits of the Arabic language in noisy condition. Their work focused on analysis and investigation of the Arabic alphabets

✉ Hassan Satori
Hassan.satori@usmba.ac.ma

[1] Laboratory Computer Science, Image Processing and Numerical Analysis, Faculty of Sciences Dhar Mahraz, Sidi Mohammed Ben Abbdallah University, B.P. 1796, Fez, Morocco

and digits in the noisy environment from an ASR perspective. As a noisy speech simulation, they added white Gaussian noise to the clean speech at different signal to noise ratio (SNR) levels. In (Addarrazi et al. (2017)) the authors have implemented an Amazigh audio-visual speech recognition system that integrates both visual information and speech recognition in order to improve the performance of the system in the noisy environment. The authors (Hamidi et al. 2020) describe the Amazigh speech recognition performance via an IVR system in noisy conditions. Their experiments were conducted for the uncoded speech and then repeated for decoded speech in a noisy environment for different signal noise ratios (SNR). In (Feng et al. 2014) researchers have presented a deep denoising auto encoder (DDA) framework that is able to produce robust speech features for noisy reverberant ASR. Their system is estimated on the CHiME-WSJ0 database and presents a 16–25% absolute improvement on the recognition rate under various SNRs.

Recently, Amazigh speech recognition has become an important focus area of speech research. However, ASR and speech analysis for Moroccan Amazigh is a less researched area. Some efforts to develop ASR for Moroccan dialect have been reported in the literature as Interactive and diagnostic systems (Satori and Elhaoussi 2014; Hamidi et al. 2019).

This paper focuses on analysis and investigation of Amazigh digits in the noisy environment from an ASR perspective. As a simulation, the car and grinder noisy speech was added to the clean speech at different signal-to-noise ratio (SNR) level. In the best of our knowledge, it is the first attempt towards developing a noise-robust Amazigh speech recognition system.

The rest of this paper is organized as follows: Sect. 2 presents the speech recognition in a noisy environment. The speech recognition system is introduced in Sect. 3. Section 4 gives an overview of the Amazigh language. Section 5 shows the technology and method used in this work. Finally, Sect. 6 investigates the experimental results. We finished by a conclusion.

2 Speech recognition in noisy environments

Performance of speech recognition systems used in noisy environments is usually decreasing, this phenomenon is observed in many studies (Benesty et al. 2007). Different techniques were studied to develop robustness against noise. Among them is the use of the speech ameliorate algorithms—in this process, before the submission of the speech signal to the ASR system, it undergoes to a denoising method, e.g. by Wiener filtering or spectral subtraction, or using a different method as developing new auditory models which are lower sensitive to noise.

Other researchers suggest advanced feature processing, like cepstral normalization techniques (e.g., cepstral mean normalization—CMN, variable cepstral mean normalization—VCMN), or other techniques which try to assessment cepstral parameters of undeformed speech, given cepstral parameters of the noisy speech, this is integrated occasionally with multi-condition training, i.e., training acoustic models with speech distorted with several noisy kinds and signal-to-noise (SNR) ratios (Hansen et al. 2001; Deng et al. 2001). Using sparse representation based classification permits for improving robustness, though it requires a lot of processing power. For some kinds of noise using perceptual properties proved to enhance the accuracy of the ASR system (Haque et al. 2009). In traditional methods for noise robust automatic speech recognition, the acoustic models are typically trained using clean speech or using multi-condition data that is processed by the same feature enhancement algorithm expected to be used in decoding.

3 Speech recognition system

An Automatic Speech Recognition (Gaikwad et al. 2010) is a process of decoding speech signals captured by the microphone and converts it into words in real-time. The recognized words can be used such as commands, data entry, or application control. Recently, this technology has reached a higher level of performance. The applications of recognition speech are often found in several domains like Healthcare, Military, commercial/industrial applications, Telephony, Personal Computers and many other devices. Figure 1 shows the speech recognition system structure.

The acoustic and language models are used by the speech recognition system to determine the word recognition rate of input speech. The acoustic model plays a principal role in improving the performance of ASR systems. For the given acoustic observation A, the speech recognition aims to find the most probable word sequence, M, that has the maximum posterior probability P (M/A), that is

$$\hat{M} = \underset{M}{argmax}\, P(A/M)P(M)$$

where A is the acoustic feature of the word sequence M, P(M) represents the language model. The language model contains a set of rules for a language that is used as the primary context for the recognized text. It plays an important role where it allows reducing the search space and resolving acoustic ambiguity (Huang et al. 2001).

3.1 Acoustic analysis

The speech signal is featured by many parameters that present more difficulties for an ASR system as the intra-speaker

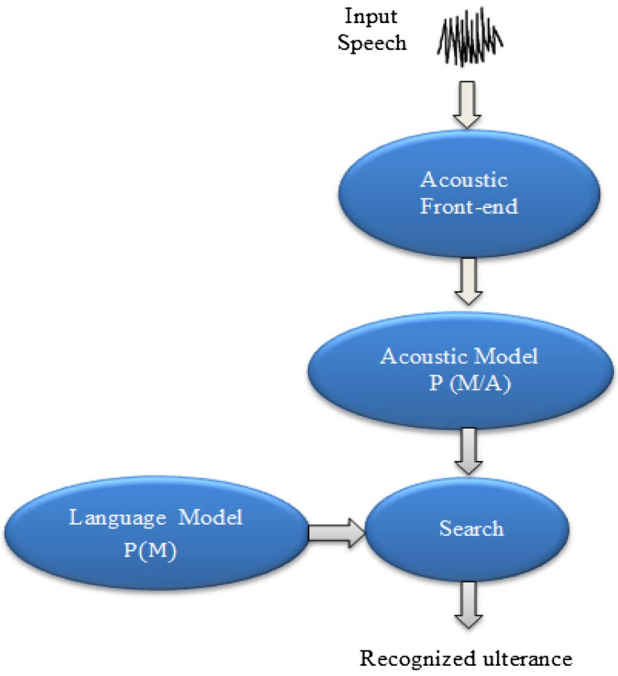

Fig. 1 Architecture of an automatic speech recognition system

variability and the redundancy of the acoustic signal. The acoustic analysis phase includes three main stages, which are analog filtering, an analog / digital conversion and a calculation of coefficients.

• Analog filtering: It is observed that the pertinent acoustic data of the speech signal is in the frequency band [50 Hz–8 kHz], the analog filtering allows to eliminate any data out of the used band.

• Analog/digital conversion: This process successively needs: guard filtering, sampling and quantization. To satisfy Shannon's theorem, the voice signal must be sampled at a sampling frequency better than or equal to twice the most frequency component.

• MFCC Coefficients: MFCC Coefficients have been vastly utilized in automatic speech recognition. These coefficients are calculated by using the Mel scale. In the form of a set of triangular bandpass filters.

3.2 Pre-processing

The speech signals analogic captured by a transducer such as a microphone or a telephone must be digitized according to the Nyquist theorem. Form this last, the signal must be sampled more than twice the rate of the highest frequency intended in the analysis. Mostly, a sampling rate between 8 and 20 kHz is utilized for speech recognition application. For normal microphones, 16 kHz sampling rate is used while 8 kHz is recommended for the telephonic channel (Kumar et al. 2012).

3.3 Feature extraction

Feature extraction is a method that allows finding a set of utterance properties which have acoustic connexion with the speech signal. In this process, the feature extractor keeps useful information and it discards the irrelevant one. To do this, successively some speech signal portion is used for processing, called window size. Data acquired in a window is called a frame. Generally, frame size ranges between 10 and 25 ms with an overlap of about 50%–70% among consecutive frames. The data in this analysis interval is multiplied with a windowing function. Several windows kinds like Rectangular, Hamming, Bartlett, Blackman or Gaussian can be utilized. Then, the characteristics are extracted on the frame by frame basis. There are various techniques to extract features like LPCC, MFCC and PLP (Kumar et al. 2012).

3.4 Hidden Markov Model

The Hidden Markov Model (HMM) is a popular statistical tool for modeling a wide range of time series data. It provides functional algorithms for state and parameter estimation, and it automatically precedes dynamic signals time warping that is locally stretched. HMMs are based on the well-known chains from probability theory that can be utilized to model a sequence of events in time. The Markov chain is deterministically an observable event. The probable word with the largest probability is generated as the result of the given speech waveform. A natural extension of the Markov chain is the Hidden Markov Model, where the internal states are hidden and any state produces observable symbols or observable evidence (Zealouk et al. 2019). 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.

Here all three symbols represent 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 $\lambda = (A, B, \pi)$ (Zealouk et al. 2018).

The basic HMM model used in this work is 5-states HMMs architecture for each Amazigh phoneme, three emitting sates and two non-emitting ones as entry and exit which join models of HMM units together in the ASR engine, as shown in Fig. 2, each emitting state consists of Gaussian mixtures trained on 13 dimensional Coefficients MFCC, their delta and delta vectors, which are extracted from the signal.

4 Overview of Amazigh language

The Amazigh language or Tamazight which is a branch of the Afro-Asiatic (Hamito-Semitic) languages nowadays, it covers the Northern part of Africa which is extended from the Red Sea to the Canary Isles and from the Niger in the Sahara to the Mediterranean Sea. In Morocco, the Amazigh language is spoken by some 28% of the population, gathered in three principal regions which are: north of Morocco where we found Tarifit, Central Morocco and South-East which speak Tamazight and southern Morocco which speaks Tachelhit (Gales 1998; Rabiner 1989).

Since 2003, Tifinagh-IRCAM has become the official graphic system for writing Amazigh in Morocco since 2003 (Zealouk et al. 2018).

- 27 consonants including: the labials ((ⴱ , ⴼ , ⵎ)), the dentals (ⵜ , ⴷ , ⵟ , ⴹ , ⵏ , ⵔ , ⵕ , ⵍ), the alveolars (ⵙ , ⵣ , ⵚ , ⵥ), the palatals (ⵛ , ⵊ), the velar (ⴽ, ⴳ) the labiovelars (ⴽⵯ , ⴳⵯ), the uvulars (ⵇ , ⵅ , ⵖ), the pharyngeal (ⵃ , ⵄ) and the laryngeal (ⵀ)
- 2 semi-consonants: (ⵢ , ⵡ)
- 4 vowels: three full vowels: (ⴰ , ⵉ, ⵓ) and ⴻ neutral vowel (or schwa) which has a rather special status in Amazigh phonology.

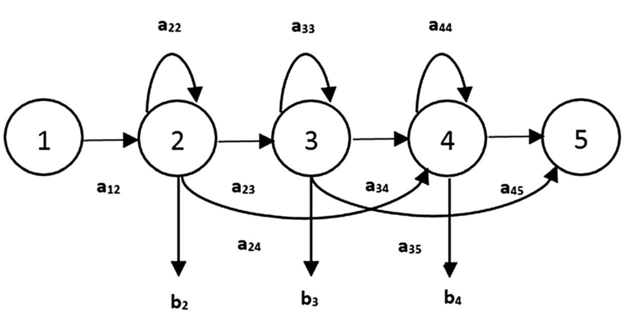

Fig. 2 Hidden Markov Model (HMM)—5-states

The allowed syllables in Amazigh language are: V, CV, VC, CVC, C, CC and CCC where V indicates a vowel while C indicates a consonant.

The first ten Amazigh digits used in our study was pronounced in Tarifit accent according to IRCAM approved language. As following, we present the used digits and their English equivalent.

Amya	0	Zero	Yen	1	One	Sin	2	Two
Krad	3	Three	Kuz	4	Four	Smmus	5	Five
Sdes	6	Six	Sa	7	Seven	Tam	8	eight
Tza	9	nine						

5 Technology and method

5.1 CMU sphinx framework

CMU Sphinx framework was designed and has been continuously developed at Carnegie Mellon University (CMU). In our work, we used Sphinx4 recognizer and SphinxTrain toolset for preparing training data and constructing acoustic and language models. Sphinx4 is a multi-threaded CPU-based recognizer written entirely in the java programming language, which utilizes HMMs for speech recognition. The N-gram language model used by the ASR system guide the search for correct word sequence by predicting the likelihood of the nth word, using the $n-1$ preceding words. The Sphinx was originally designed for English, but nowadays it supports several languages, e.g. French and Arabic. Recently, the Amazigh language was integrated successfully in this framework by the researchers (Hoffman 2006; Fadoua and Siham 2012).

5.2 Amazigh corpus preparation

The database Amazigh digits was created in the framework of this work and it contains a corpus of speech and their transcription of 40 Berber Moroccan speakers. The corpus consists of spoken 10 Amazigh firsts' digits (0–9). Thus, the task of labelling speech signals after segmentation is easy. The sampling rate of the recording is 16 kHz, with 16 bits resolution, 25.6 ms Hamming Window with consecutive frames overlap by 10 ms and Mel-Frequency Cepstral Coefficients (MFCC). Table 1 shows more speech corpus technical details. During the recording sessions, speakers were asked to utter the 10 times each digit respecting the numerical order. Audio recordings for a single speaker were

Table 1 System parameters

Parameter	Value
Sampling rate	16 kHz
Number of bits	16 bits
Number of Channels	1,Mono
Channel	Wav
Corpus	Amazigh_10 digits
Speakers' age	24–44 years-old
Hamming	25.6 ms
Frames overlap	10 ms
Number of MFCC coefficients	39
Types of noise	Car-grinder

saved into one ".wav" file and sometimes up to four ".wav" files depending on number of sessions the speaker spent to finish recording. It is time consuming to save every single recording once uttered. Hence, the corpus consists of 10 repetitions of every digit produced by each speaker. Depending on this, the corpus consists of 4000 tokens. During the recording session, the waveform for each utterance was visualized back to ensure that the entire word was included in the recorded signal. Therefore, there was a need to segment manually these bigger ".wav" files into smaller ones each having a single recording of a single word and manual classification of those ".wav" files into the corresponding directories was done. Wrongly pronounced utterances were ignored and only correct utterances are kept in the database. The software used to the voice with speakers in wavesurfer (Hamidi et al. 2018).

Our noises databases were recorded from the low to the height-noisy environment. Several sets of recordings (car and grinder kinds) were obtained, with the SNR ranging from + 5 dB to + 45 dB, depending on the record distance among noisy source and recording device. The original noise data were sampled at 16 kHz and stored as 16-bit in order to fit with the already clean speech fixed parameter. The finale noisy speech samples were obtained by mixing the clean and noisy speech using the audio processing Sox tool (Satori et al. 2017).

6 Amazigh Acoustic Model Preparations

The acoustic model is a statistical representation of an acoustic image for the speech signal. In this work, our Statistic model was performed using a speech signal from the first ten Amazigh digits training database. Every recording in the training corpus is transformed into a series of feature vectors with MFCC coefficients. For each obtained wave data, a set of features files are computed using the front-end. The designed model is trained by using continuous state probability density is 16 Gaussian mixture distributions. The procedure to create the acoustic model is grouping a set of input data and treats them with Sphinxtrain tool. The training input data are:

- Audio wave data (sound file).
- Amdigits_test.fileids
- Amdigits _test.transcription
- Amdigits _train.fileids
- Amdigits _train.transcription
- Amdigits.phone list
- Amdigits.filler
- Amdigits.dic
- Amdigits.lm.DMP

6.1 Amazigh pronunciation dictionary

The phonetic dictionary is considered as an intermediary between the acoustic and the language models where it allows giving a symbolic representation for each utilized word. Also, this file named lexicon it includes all Amazigh words we want to recognize followed by their pronunciation. Substitutional transcripts marked with parentheses as (1) means for the second pronunciation.

6.2 Language model

The language model includes the grammar used in the speech systems. Based on probabilistic information, this model was designed to detect the connections among the words in a sentence with the help of the phonetic dictionary (Satori et al. 2017). It is the single largest set trained with several words, including a big number of parameters. The word sequence occurrence probability W is calculated as:

$$P(W) = P(w_1, w_2, .., w_{n-1}, w_n) = P(w_1).P(w_2|w_1).P(w_3|w_1w_2) \ldots P(w_n|w_1w_2 \ldots w_{n-1})$$

In our work, the language model is constructed statistically by using CMU-Cambridge Statistical Language Modelling toolkit that is founded on the modelization of unigrams, bi-grams, and tri-grams of the language to recognize the subject text.

7 Experimental results

7.1 Amazigh noisy recognition testing

Our ASR system was first tested with the car noisy environment and then it was tested with grinder noisy environment. On the other hand, the training database for the speaker independent experiments includes isolated words (digits) pronounced by 30 speakers (70% of the database) in a clean environment. The rest of the speakers (30% of the database)

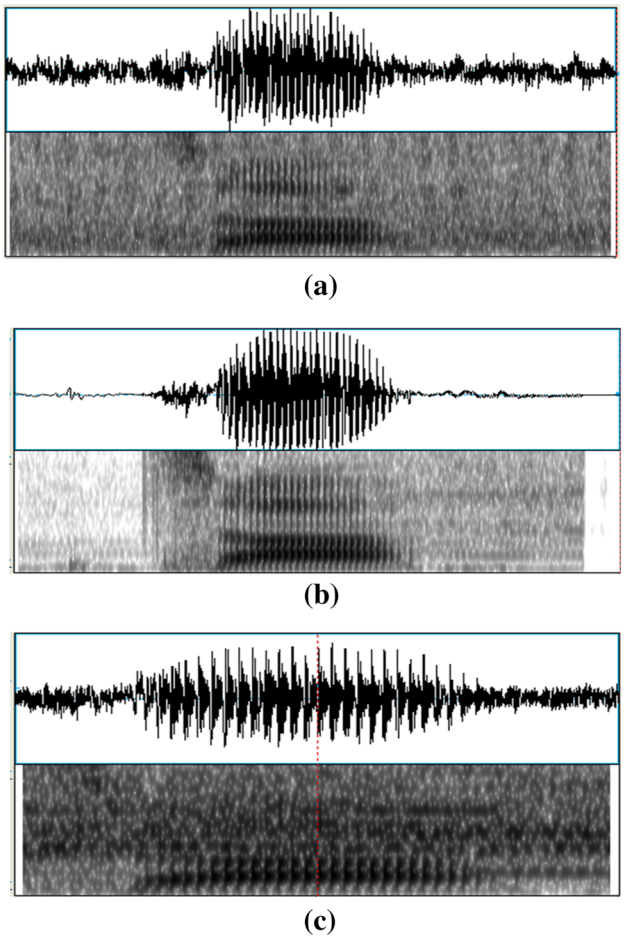

Fig. 3 **a** Spectrogram of the Kuz digit in normal environment. **b** Spectrogram of the kuz digit at 25 SNR under-car noise. **c** Spectrogram of the kuz digit at 25 SNR under grinder noise

were utilized in the test. Two testing databases sets were created with the speech in noise starting from 5 dB and with an increase of 10 dB at each setup until we reached 45 dB for each type of noisy. Figure 3 presents the used Kuz digit spectrograms under noisy. We observe that the car noise is a lower-band stationary noise, whilst the grinder has much wider bandwidth and it contains sudden sharp sounds.

A speech signal that is given in a noisy environment is less intelligible than the same signal given in a clean environment due to the fact that the spectral distance among the speech signal and the noise signal is reduced.

7.2 System evaluation

The evaluation of our systems was made according to the obtained recognition accuracy and computed using WER as follow:

Table 2 Overall recognition rates

Noise	5 dB	15 dB	25 dB	35 dB	45 dB
Car	72,92	56,50	31,75	4,83	0,17
Grinder	70,08	51,42	23,33	3,58	0,00

$$\mathrm{WER} = (\mathrm{S} + \mathrm{D} + \mathrm{I})/\mathrm{N}.$$

where

- Substitutions S
- Deletions D
- Insertions I
- Number of words N

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead:

$$\begin{aligned}\mathrm{WRR} &= (1 - \mathrm{WER}) * 100 = 100 * (\mathrm{N} - \mathrm{S} - \mathrm{D} - \mathrm{I})/\mathrm{N} \\ &= 100 * (\mathrm{H} - \mathrm{I})/\mathrm{N}.\end{aligned}$$

where H is N-(S + D), the number of correctly recognized words.

7.3 Amazigh speech system performances

The results provided in this paper depend mainly on the outcomes of the designed Amazigh digits recognition system in noisy conditions. All noisy environments tests were performed based speaker independent acoustic model. For the car noisy environment the overall performances are 72.92%, 56.50%, 31.75%, 4.83 and 0.17% for SNR = 5 dB, 15 dB, 25 dB, 30 dB and 35 dB, respectively. For the grinder conditions, the overall performances are 70.08%, 51.42%, 23.33%, 3.58 and 0.00% for the same SNR values respectively. Table 2 gives the Amazigh digits speech recognition rates by using the same testing data under car and grinder noise conditions for various SNR levels.

As a results comparison between the two use types of noise, a difference in the recognition rates were observed for the same SNR value. For example, the difference of overall performances are 2,84%, 5,08%, 8,42%, 1,25 and 0,17 for SNR = 5 dB, 15 dB, 25 dB, 35 dB and 45 dB, respectively. The performance levels of most current speech recognizers decrease significantly when environmental noise occurs during use. Such performance degradation is mainly caused by mismatches in training and operating environments.

From Fig. 4, the Krad digit has got 79.17% accuracy at 5 dB and it degraded to 70,00%, 41,67%, 11,67% and

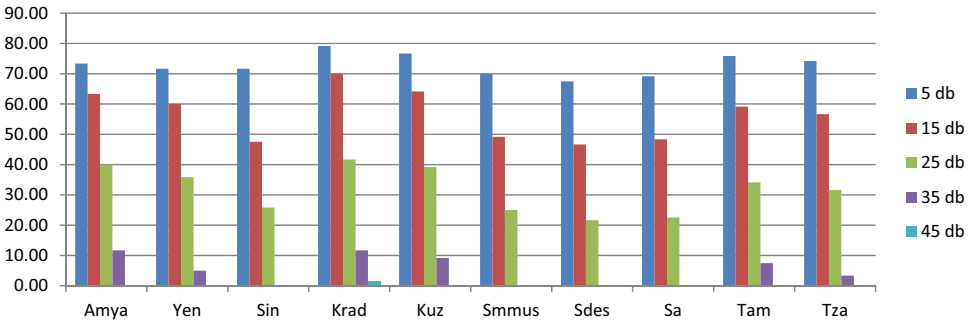

Fig. 4 Digits recognition rates under car noisy conditions

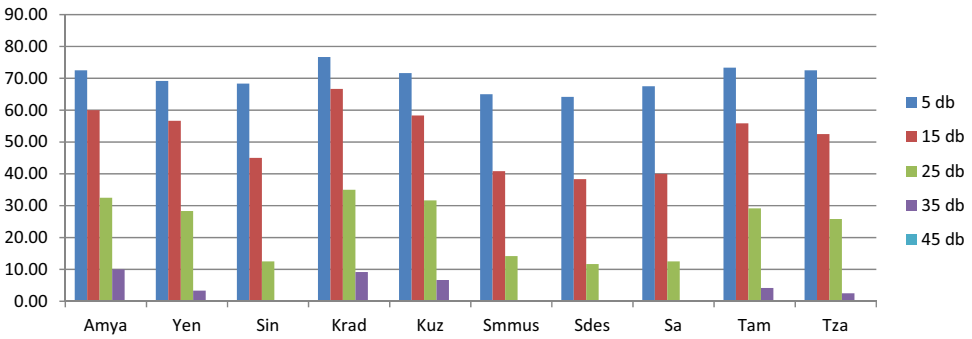

Fig. 5 Digits recognition rates under grinder noisy conditions

1,67% in the noisy environment at SNR 15 dB, 25 dB, 35 dB and 45 dB, respectively. The confusion of Krad with Kuz gradually decreases with the increase of noise level. The similar situation also happened with the all used digits. Further, lower rates were observed for Sin, Smmus, Sdes and Sa where these digits have got the accuracies lower than the other with all used SNR where the noisy influence was clearly observed with 25 dB. Figure 5 shows the system recognition rates for grinder noisy speech with some SNR values used in the first experiment. The high accuracy has got from the Krad digit and Amya, Kuz, Tam and Tza digits maintain the recognition rates more than 70% while the others digits reach accuracy below 70% up to SNR 5 dB. For SNR 15 dB and more the recognition decreases again for all digits. The studied digits have got a lower accuracy from 25 dB and a very low accuracy was achieved at 35 dB. Moreover, we noted that the digits which contain the S alphabets are not recognized at 35 dB and these digits possess a very high dissimilarity compared to all other spoken digits. For the most resisted digit is Krad, due to his included strong consonants and number of syllables.

8 Conclusion

This paper describes our experiments for Amazigh digits speech recognition system under noisy conditions. The designing and implementing of acoustic and language models based on CMU Sphinx tools were described. Experiments with speech recognition in noisy conditions presented that the performance degradation was observed if recognition was tested at 5 dB and the recognition rate was hardly affected if SNR exceeded 25 dB for both noisy kinds. However, major degradation of accuracy was observed if the speech signal was distorted with noise and the SNR exceeded 35 dB. In this investigation, we found that the digits which include the S alphabet are affected more than others digits for different SNR values. In future, we will try to improve the performance of this ASR system based on the combined HMMs and Deep learning techniques.

References

Addarrazi, I., Satori, H., & Satori, K. (2017, April). Amazigh audiovisual speech recognition system design. In: 2017 intelligent systems and computer vision (ISCV). IEEE, 2017 (pp. 1–5).

Alotaibi, Y., Mamun, K., & Ghulam, M. (2009, July). Noise effect on arabic alphadigits in automatic speech recognition. In: IPCV. 2009 (pp. 679–682).

Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.). (2007). *Springer handbook of speech processing*. Berlin: Springer.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56,* 85–100.

Deng, L., Acero, A., Jiang, L., Droppo, J., & Huang, X. (2001). *high-performance robust speech recognition using stereo training data. In: Proceedings of ICASSP*. Salt Lake City, Utah: ICASSP.

Fadoua, A. A., & Siham, B. (2012). Natural language processing for Amazigh language. Challenges and future directions. Language technology for normalisation of less-resourced languages, 19.

Feng, A., Zhang, Y., & Glass, J. (2014, May). Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1759–1763). IEEE.

Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications, 10*(3), 16–24.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language, 12*(2), 75–98.

Gales, M. J. F., & Young, S. J. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech Audio Process., 4*(5), 352–359.

Hamidi, M., Satori, H., Zealouk, O., Satori, K., & Laaidi, N. (2018, October). Interactive voice response server voice network administration using hidden markov model speech recognition system. In: Second 8 World conference on smart trends in systems, security and sustainability (WorldS4). IEEE (pp. 16–21).

Hamidi, M., Satori, H., Zealouk, O., & Satori, K. (2019). Speech coding effect on Amazigh alphabet speech recognition performance. *Journal of Advanced Research in Dynamical and Control Systems, 11*(2), 1392–1400.

Hamidi, M., Satori, H., Zealouk, O., & Satori, K. (2020). Amazigh digits through interactive speech recognition system in noisy environment. *International Journal of Speech Technology, 23*(1), 101–109.

Hansen, J. H., Sarikaya, R., Yapanel, U., & Pellom, B. (2001). *Robust speech recognition in noise: An evaluation using the SPINE corpus. In: Proceedings of eurospeech*. Aalborg: Eurospeech.

Haque, S., Togneri, R., & Zaknich, A. (2009). Perceptual features for automatic speech recognition in noisy environments. *Speech Communication, 51*(1), 58–75.

Hoffman, K. E. (2006). Berber language ideologies, maintenance, and contraction: Gendered variation in the indigenous margins of Morocco. *Language & Communication, 26*(2), 144–167.

Huang, X., Acero, A., & Hon, H. (2001). *Spoken language processing: A guide to theory, system and algorithm development*. New Jersey: Prentice Hall.

Hu, Y., & Huo, Q. (2006, December). An HMM compensation approach using unscented transformation for noisy speech recognition. In: ISCSLP (pp. 346–357).

Janicki, A., & Wawer, D. (2013). Voice-driven computer game in noisy environments. *IJCSA, 10*(1), 31–45.

Kalinli, O., Seltzer, M. L., Droppo, J., & Acero, A. (2010). Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(8), 1889–1901.

Kim, C., & Stern, R. M. (2009). Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. *Interspeech, 2009,* 28–31.

Kumar, K., Aggarwal, R. K., & Jain, A. (2012). A Hindi speech recognition system for connected words using HTK. *International Journal of Computational Systems Engineering, 1*(1), 25–32.

Lee, S. H., Chung, H., Park, J. G., Young, H.-J., Lee, Y. (2009). A commercial car navigation system using korean large vocabulary automatic speech recognizer. In: APSIPA 2009 annual summit and conference (pp. 286–289).

Li, J., Deng, L., Yu, D., et al. (2007). High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. In: Automatic speech recognition & understanding, ASRU. IEEE workshop (pp. 65–70).

Moreno, P., Raj, B., & Stern, R. (1996). A vector Taylor series approach for environment-independent speech recognition. In: Proceedings of international conference on audio, speech, signal processing, Atlanta, GA (pp. 733–736).

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Raj, B., & Stern, R. M. (2005). Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine, 22*(5), 101–116.

Satori, H., & Elhaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. *International Journal of Speech Technology, 17*(3), 235–243.

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.

Seltzer, M. L., Acero, A., & Kalgaonkar, K. (2010, March). Acoustic model adaptation via linear spline interpolation for robust speech recognition. In: IEEE International Conference on Acoustics speech and signal processing (ICASSP), 2010 (pp. 4550–4553).

SoX - Sound eXchange. (2019). Retrieved March 2019 from https://sox.sourceforge.net/.

"Wavesurfer". (2018). Version1.8.8p4. Retrieved January 2018 from https://sourceforge.net/projects/wavesurfer.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., & Acero, A. (2008, March). A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition. In: International conference on acoustics, speech and signal processing. IEEE (pp. 4041–4044).

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.

Zealouk, O., Satori, H., Hamidi, M., & Satori, K. (2019). Speech recognition for Moroccan dialects: Feature extraction and classification methods. *Journal of Advanced Research in Dynamical and Control Systems, 11*(2), 1401–1408.