# Towards Phone Number Recognition For Code Switched Algerian Dialect

**Khaled Lounnas**
LCPTS-FEI, USTHB
Algiers, Algeria
klounnas@usthb.dz

**Mourad Abbas**
High Council of Arabic Language
Algiers, Algeria
abb.mourad@gmail.com

**Mohamed Lichouri**
LCPTS-FEI, USTHB
Algiers, Algeria
mlichouri@usthb.dz

## Abstract

This paper addresses the problem of phone number recognition taking into account some of the peculiarities of dialectal Arabic used in the daily life of Algerian people as code-switching and accent variety. Accordingly, we have set up an ASR system aiming to have the capacity to cope with these peculiarities, i.e. to recognize sequences of digits that may have been spoken in Algerian dialect, in French, or in both. For this purpose, we built an in-house corpus composed of 100 couples of digits (from '00' to '99') spoken in French and dialectal Arabic by two persons. The findings show that, our ASR system behaves more or less effectively when dealing with one language. In fact, it yielded a WER of 1.4 % for French, and 7.1 % for both Blida and Algiers dialect, using 13 MFCC's Coefficients and 32 GMMs. Unfortunately, due to the code-switching phenomenon between these dialects and French, and the limited data size, the performance degrades drastically and reaches a WER of 17.3% with 13 MFCCs and 64 GMMs.

## 1 Introduction

Automatic Speech recognition (ASR) system provides users with the ability to use their voices rather than the keyboard to search for information. Nowadays, mobile phones have emerged as a trendy area offering attractive platforms for speech recognition-based functions that can help in solving various issues in mobile telephony (Varga et al., 2002), automated contact centres and consumer electronics items. Many applications have been implemented as intelligent telephone answering systems (Lobanov et al., 1997) that use standard speech modems to respond to incoming calls and recognize the call recipient and caller's name. Interactive voice response (IVR) systems can be used for mobile purchases, banking payments, services, retail orders, travel information.

A number of manufacturers currently offer mobile phones with built-in voice interfaces (Wu et al., 1998; Tabani et al., 2017). Most of these interfaces are developed to support particular languages (Salimbajevs, 2018), for example English,French and Hindi (Deka et al., 2018), whereas there are many other languages and dialects, especially those with low resources such as Arabic dialects.

The purpose of this work is to implement an application to recognize digits spoken in Algerian dialects, taking into account the code switching phenomenon that characterizes these dialects. In this direction, we first developed a new in-house speech corpus composed of one hundred spoken digits (from "00"-"99"). This corpus was recorded by two native speakers from two different cities in Algeria: Algiers, Blida. Furthermore, the same speakers recorded the digits (from "00"-"99") in French . This corpus is used for training the acoustic and language models in different configurations in different situations provided that users may pronounce the sequence of digits in both dialectal Arabic and French spoken with different accents. We will give more details about this in the following sections.

This paper is organized as follows: literature review is presented in section 2. In section 3, a brief description of the linguistic material has been given. Section 4 is devoted to experiments and results. Finally, the conclusion is presented in section 5.

## 2 State of the art

### 2.1 Spoken Digits Corpora

Unlike their previous work on designing ASR based on romanized characters (Satori et al., 2007), Satori et al in (Satori et al., 2009) built an in house speech based digits corpora to train the CMU Sphinx tool in an entirely Arabic environment. Other researches focused on applying Deep Neural Network (DNN) technology to solve Spo-

ken Arabic digits recognition issue, in (Mahfoudh Ba Wazir and Huang Chuah, 2019) authors used a corpus made of 1040 spoken digits. The aim is to design Long Short-Term Memory (LSTM) model which has the capability to treat problems associated with temporal dependencies requiring long-term learning and to solve the vanishing gradient problems associated with RNN. their reported findings show that the LSTM model can achieve 69% in accuracy when recognizing spoken Arabic digits. In (Sharmin et al., 2020), authors carried out experiments on the first ten digits spoken in Bengali. They achieved classification by the way of Convolutional Neural Network (CNN), where they obtained an accuracy of 98.37%. In the same context, it has been shown in (Sharan, 2020) that using Wavelet Scalogram and CNN performed on a dataset including 56,290 segments belonging to ten spoken digits, their proposed approach surpass other methods and achieve a test error of 2.84% . In (Djellab et al., 2017), the authors conducted a study on the classification of regional accents in complex linguistic environments in the case of Algerian dialects, tested on their prepared corpus entitled Algerian Modern Colloquial Arabic Speech Corpus.

### 2.2 Speech recognition

(Alotaibi et al., 2008) addressed the process of automatic digits recognition of Saudi dialect by implementing a Hidden Markov Model using from SAAVB corpora (Alghamdi et al., 2008). The findings show that the proposed system reach 93.67% overall correct rate of digit recognition. In (Lounnas et al., 2020), authors investigated at what extent language identification can improve the performance of the Moroccan Automatic Speech Recognition (ASR) system, they found that their proposal greatly improved the overall accuracy, and outperform the baseline system by 33%. Likewise, (Ghangam et al., 2021) propose an approach which consists in designing a compact multilingual speech recognition system based on language identification, the results show that their approach is low memory consumption with an improvement of WER by 30%, compared to the baseline approach.

## 3 Linguistic Material

Building a typical corpus is a fundamental step for any engineering system. For this reason, we designed our own corpus by soliciting two speakers from two different Algerian cities: Algiers and Blida, to record one hundred digits from "00" to "99", each digit being repeated 15 times. The recordings have been divided into short segments using Praat tool [1]. In Table 1, we show some statistics related to the developed corpus .

| Features | Value |
|---|---|
| Sampling rate | 16 KHz |
| Number of bits | 16 bits |
| Number of Channels | 1, Mono |
| Audio data file format | .wav |
| # Speakers | 2 |
| #Speakers per dialect | 2 |
| # Dialect | 2 |
| # Language | 2 |
| # Tokens per speaker | 1500 |
| # speaker's gender | Male |
| # Total number of tokens | 6000 |
| #Number of digits | 100 digits (AAD) |
| | 100 digits (BAD) |
| | 100 digits ($FR_{AAD}$) |
| | 100 digits ($FR_{BAD}$) |
| # Repetitions per word | 15 |
| Condition of noise | normal life |
| Preemphased | $1 - 0.97z^{-}1$ |
| Window type Hamming | 25.6 ms |
| Frames overlap | 10 ms |

Table 1: Details on the corpus. AAD stands for Algerian Arabic Dialect, BAD: Blida Arabic Dialect, $FR_{AAD}$ and $FR_{BAD}$: French spoken in AAD and BAD accents, respectively.

## 4 Experiments and Results

We achieved a number of experiments to measure the performance of the spoken digit recognition system according to the following three situations:
- Building four (acoustic and language) models by training individually each of the following corpus categories: (BAD, AAD, $FR_{AAD}$ and $FR_{BAD}$).
- Building two (acoustic and language) models based upon (BAD, AAD) and ($FR_{AAD}$, $FR_{BAD}$) respectively.
- Building one global (acoustic and language) model based on BAD, AAD, $FR_{AAD}$ and $FR_{BAD}$ in front of speech comprising BAD, AAD and the code switched sequences.

Note that we used CMU-sphinx in the system de-

[1] http://www.fon.hum.uva.nl/praat/

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 9.5 | 23.2 | 10.2 | 21.4 | 23.4 | 46.8 |
| 4 | 5.5 | 13.6 | 4.1 | 10.2 | 14.6 | 30.4 |
| 8 | **2.4** | 6.2 | 2.6 | 6.6 | 11 | 23.2 |
| 16 | 3.2 | 7.2 | 4.7 | 11.6 | 8.6 | 19.2 |
| 32 | 4.9 | 11.6 | 4.5 | 10.6 | 8.8 | 19 |
| 64 | 10.6 | 23 | 8.8 | 19.6 | 10.6 | 21.8 |

Table 2: Performance of the ASR system for AAD dialect.

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 16.4 | 32.2 | 16.1 | 30.4 | 19.8 | 36.2 |
| 4 | 8.8 | 17.8 | 11.1 | 24.8 | 14.4 | 27.6 |
| 8 | 5.8 | 12.2 | 5.9 | 13.8 | 9.5 | 19 |
| 16 | 4.8 | 11 | 5 | 11 | 9.5 | 18.4 |
| 32 | **3.8** | 8.4 | 6.4 | 14.8 | 9.8 | 20 |
| 64 | 5 | 11.2 | 8.4 | 19.8 | 11.7 | 24.8 |

Table 3: Performance of the ASR system for BAD dialect.

sign [2]. We defined the size of the acoustic features (MFCC's) to 13, 26 and 39 coefficients, in addition to different values of GMM 2, 4, 8, 16, 32, 64.

### 4.1 Monolingual Spoken Digits Recognition

Table 2 and 3 present the WER and SER obtained through different setup for dialectal Arabic (dialects spoken in Algiers and Blida). The best WER for AAD was 2.4% using a default configuration (8 GMM and 13 MFCC) followed by a WER of 3.2% with 16 GMM's and 13 MFCC's . The best score achieved for BAD was with (32 GMM and 13 MFCC). It should be noted that for both Dialects 13 MFCC's coefficients is usually enough to get the highest performance.

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 6 | 12.6 | 6.4 | 13.8 | 9.3 | 19.8 |
| 4 | 4.3 | 9.4 | 4.8 | 10.6 | 5.9 | 12.4 |
| 8 | **2.6** | 5.4 | 4 | 8.2 | 5.2 | 10.4 |
| 16 | 4 | 6.8 | **1.8** | 3.6 | 4.8 | 9.8 |
| 32 | 5.3 | 10 | 6.6 | 11.4 | 8 | 14.2 |
| 64 | 9.6 | 14 | 11.8 | 17.4 | 15.4 | 23.6 |

Table 4: Performance of the ASR system for $FR_{AAD}$.

In a similar way, we present in tables 4 and 5, performance obtained through multiple configurations for French digits spoken by both Algiers and Blida's people ($FR_{AAD}$, $FR_{BAD}$), respectively. It can be noticed for $FR_{AAD}$, in table 4, that the best WER (1.8%) is obtained using the configuration (16 GMM and 26 MFCC). The second best result is achieved using the configuration (8 GMM and 13 MFCC) with a WER of 2.6%. In the case of $FR_{BAD}$, WER of 1.9% is achieved with 16 GMM and 13 MFCC followed by and WER of 2.1% achieved by the default setup (8 GMM and 13 MFCC). The performance obtained for French is slightly higher than that obtained for Blida and Algiers dialects.

[2]https://cmusphinx.github.io/

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 13.7 | 28.8 | 12.4 | 26.4 | 15.2 | 31 |
| 4 | 4.6 | 10 | 7.3 | 15.2 | 8.4 | 17.6 |
| 8 | 2.1 | 4.4 | 5.5 | 11.6 | 7.1 | 15 |
| 16 | **1.9** | 3.8 | 3.7 | 7.4 | 6.1 | 12.6 |
| 32 | 2.5 | 5.4 | 4.1 | 8.6 | 5.9 | 12 |
| 64 | 7.1 | 13 | 9.1 | 17.4 | 11.2 | 21.4 |

Table 5: Performance of the ASR system for $FR_{BAD}$.

### 4.2 Bilingual / Multilingual Spoken Digits Recognition

Unlike the Monolingual ASR system where we trained four models using the four corpora ($FR_{AAD}$, $FR_{BAD}$, BAD, AAD) separately, we built two models for bilingual ASR based on training the merged corpora of the couples ($FR_{AAD}$, $FR_{BAD}$) and (BAD, AAD), in addition, we trained one single model for Multilingual ASR based upon the four corpora.

For bilingual ASR, as can be noticed in table 6, results show that merging ($FR_{AAD}$, $FR_{BAD}$) improved the performance (reduction of WER by 0.4%). On the contrary, recognition of the digits spoken in the two Arabic dialects has been degraded. In fact, the best obtained WER is 7.1% (table 7), which is less than WER recorded for monolingual ASR: (AAD, 2.4%) and (BAD, 3.8%). The reason is that French digits are spoken in standard way by the two speakers which makes the related corpus bigger unlike the two Arabic dialects. However, the difference in pronunciation of the two speakers makes the AAD and BAD corpora different than the French one is. Note that we found more than 50 couple of digits spoken differently in the two Arabic dialects.

For multi-lingual ASR, a global model has been trained based on the whole corpus comprising

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 12.8 | 26.9 | 13.4 | 27.8 | 14.1 | 28.7 |
| 4 | 6.5 | 13.7 | 6.5 | 13.9 | 9.3 | 19.7 |
| 8 | 3.5 | 7.4 | 4.8 | 10.2 | 6.9 | 14.9 |
| 16 | 1.9 | 4 | 4.4 | 9.3 | 6.3 | 13.5 |
| 32 | **1.4** | 2.8 | 3.5 | 7.3 | 6.4 | 13.4 |
| 64 | 2.3 | 4.5 | 4.3 | 8.9 | 6 | 12.4 |

Table 6: Performance of the ASR system for French spoken in both $FR_{AAD}$ and $FR_{BAD}$.

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 26.3 | 47.9 | 21.6 | 42.6 | 27.7 | 49.5 |
| 4 | 17.4 | 34.3 | 15.1 | 32 | 19.7 | 37.6 |
| 8 | 12.6 | 26.2 | 10.7 | 21.6 | 15.1 | 27.9 |
| 16 | 8.8 | 18 | 8.5 | 18.1 | 11.5 | 22.4 |
| 32 | **7.1** | 15 | 8.8 | 17.3 | 10.9 | 20.9 |
| 64 | 7.4 | 15.9 | 8.2 | 16.8 | 11 | 21.2 |

Table 7: Performance of the ASR system for both AAD and BAD dialects.

BAD, AAD, $FR_{AAD}$ and $FR_{BAD}$ corpora. This is to deal with code switching phenomenon that characterizes the sequences to be recognized. The best WER obtained is about 17.3% using 13 MFCC and 64 GMM. These results show the necessity to use Monolingual ASR approach which is best performing on condition to integrate a language identification component.

## 5 Conclusion

In this paper, we tackled the digits recognition issue for code switched Algerian dialect. The main challenge is that Algerian people use introduce French sequences in their conversations. The results of our experiments show that it is possible to deal with this task using a global model, on condition that larger corpora are used.

| MFCC | 13 | | 26 | | 39 | |
|---|---|---|---|---|---|---|
| GMM | WER | SER | WER | SER | WER | SER |
| 2 | 44.6 | 69.3 | 44.4 | 70 | 45.8 | 71.8 |
| 4 | 36.1 | 57.9 | 35.7 | 58.7 | 38.6 | 62.4 |
| 8 | 30 | 51.9 | 30.1 | 49 | 31.7 | 52.5 |
| 16 | 24.1 | 43 | 25.9 | 44.4 | 27.4 | 46.2 |
| 32 | 19.9 | 36.4 | 22.1 | 38.9 | 24.8 | 42.8 |
| 64 | 17.3 | 32.2 | 19.2 | 34.8 | 22.9 | 40.6 |

Table 8: Performance of the ASR system for code switched sequences (AAD, BAD, $FR_{AAD}$, $FR_{BAD}$.

## References


Mansour Alghamdi, Fayez Alhargan, Mohammed Alkanhal, Ashraf Alkhairy, Munir Eldesouki, and Ammar Alenazi. 2008. Saudi accented arabic voice bank. *Journal of King Saud University - Computer and Information Sciences*, 20:45–64.

Yousef Ajami Alotaibi, Mansour Alghamdi, and Fahad Alotaiby. 2008. Using a telephony saudi accented arabic corpus in automatic recognition of spoken arabic digits. In *Proceedings of 4th International Symposium on Image/Video Communications over Fixed and Mobile Networks*, pages 43–60. Citeseer.

Barsha Deka, Joyshree Chakraborty, Abhishek Dey, Shikhamoni Nath, Priyankoo Sarmah, SR Nirmala, and Samudra Vijaya. 2018. Speech corpora of under resourced languages of north-east india. In *2018 Oriental COCOSDA-International Conference on Speech Database and Assessments*, pages 72–77. IEEE.

Mourad Djellab, Abderrahmane Amrouche, Ahmed Bouridane, and Noureddine Mehallegue. 2017. Algerian modern colloquial arabic speech corpus (amcasc): regional accents recognition within complex socio-linguistic environments. *Language Resources and Evaluation*, 51(3):613–641.

Sangeeta Ghangam, Daniel Whitenack, and Joshua Nemecek. 2021. Dyn-asr: Compact, multilingual speech recognition via spoken language and accent identification. *arXiv preprint arXiv:2108.02034*.

Boris M Lobanov, Simon V Brickle, Andrey V Kubashin, and Tatiana V Levkovskaja. 1997. An intelligent telephone answering system using speech recognition. In *Fifth European Conference on Speech Communication and Technology*.

Khaled Lounnas, Hassan Satori, Mohamed Hamidi, Hocine Teffahi, Mourad Abbas, and Mohamed Lichouri. 2020. Cliasr: a combined automatic speech recognition and language identification system. In *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–5. IEEE.

Abdulaziz Saleh Mahfoudh Ba Wazir and Joon Huang Chuah. 2019. Spoken arabic digits recognition using deep learning. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 339–344.

Askars Salimbajevs. 2018. Creating lithuanian and latvian speech corpora from inaccurately annotated web data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

H Satori, M Harti, and N Chenfour. 2007. Arabic speech recognition system based on cmusphinx. In *2007 International Symposium on Computational Intelligence and Intelligent Informatics*, pages 31–35. IEEE.

Hassan Satori, Hussein Hiyassat, Mostafa Haiti, and Noureddine Chenfour. 2009. Investigation arabic speech recognition using cmu sphinx system. *International Arab Journal of Information Technology (IAJIT)*, 6(2).

Roneel V Sharan. 2020. Spoken digit recognition using wavelet scalogram and convolutional neural networks. In *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 101–105. IEEE.

Riffat Sharmin, Shantanu Kumar Rahut, and Mohammad Rezwanul Huq. 2020. Bengali spoken digit classification: a deep learning approach using convolutional neural network. *Procedia Computer Science*, 171:1381–1388.

Hamid Tabani, Jose-Maria Arnau, Jordi Tubella, and Antonio González. 2017. Performance analysis and optimization of automatic speech recognition. *IEEE Transactions on Multi-Scale Computing Systems*, 4(4):847–860.

Imre Varga, Stefanie Aalburg, Bernt Andrassy, Sergey Astrov, Josef G Bauer, Christophe Beaugeant, Christian Geißler, and H Hoge. 2002. Asr in mobile phones-an industrial approach. *IEEE transactions on speech and audio processing*, 10(8):562–569.

Su-Lin Wu, Brian Kingsbury, Nelson Morgan, and Steven Greenberg. 1998. Performance improvements through combining phone-and syllable-scale information in automatic speech recognition. In *ICSLP*, volume 1, pages 160–163.