Building a Speech Corpus based on Arabic Podcasts for Language and Dialect Identification

Khaled Lounnas
USTHB University, Algeria
klounnas@usthb.dz

Mourad Abbas and Mohamed Lichouri
Computational Linguistics Dept., CRSTDLA, Algeria
{m.abbas, m.lichouri}@crstdla.dz

Abstract

In this paper, we present ArPod, a new Arabic speech corpus made of Arabic audio podcasts. We built this dataset, mainly for both speech-based multi-lingual and multi-dialectal identification tasks. It includes two languages: Modern Standard Arabic (MSA) and English, and four Arabic dialects: Saudi, Egyptian, Lebanese and Syrian. A set of supervised classifiers have been used: Support Vector Machines (SVM), Multi Layer Perceptron (MLP), K-Nearest Neighbors (KNN), Extratrees and Convolutional Neural Networks (CNN), using acoustic and spectral features. For both tasks, SVM yielded encouraging results and outperformed the other classifiers. Language Identification, Dialect Identification, CNN, Acoustic features, spectral features, SVM, Arabic Podcast

1 Introduction

The most popular researches on spoken audio language/dialects identification has been conducted based on acoustic information, Phonotactic and prosodic approaches and other techniques. Acoustic information is the lowest and nevertheless simplest level of features that can denote a speech waveform. Indeed, in (Koolagudi et al., 2012), MFCC features have been extracted to study the impact of MFCC's coefficients on Indian language recognition . Phonotactic and prosodic information have been used in(Biadsy et al., 2009) and (Biadsy and Hirschberg, 2009). The authors applied a phonotactic approach to automatically detect Arabic dialects by using phone recognizer followed by dialect modeling using trigram models. They also examined the role of prosodic features (intonation and rhythm) for identification of dialects from four Arabic regions: Gulf, Iraq, Levantine and Egypt. In other researches like in (Alshutayri and Albarhamtoshy, 2011), authors trained HMM to characterize part of speech, to implement a dialect identification system.

In order to establish robust systems for Language/dialect identification, spoken corpora have been developed by research community for several languages, but many other languages still lack such resources such as Arabic. That is why we developed a new speech corpus, Arpod-1.0, which is a Multilingual Arabic spoken dataset extracted from the web podcast. This dataset is composed of more than 8 hours, devoted for Arabic and some of its dialects: Saudi, Lebanese, Egyptian and Syrian, in addition to English. The dataset has been separated to two categories: Languages and dialects without code switching, and dialects with code switching. We trained SVM, Extratrees and kNN using acoustic and spectral features, and CNN using spectorgram. In addition, we conducted experiments to find the impact of duration on speech utterances language identification. Indeed, three duration values have been considered: 6 sec, 30 sec and 1 min.

This paper is organized as follows, we present an overview of the works on speech based language identification in section 2. In section 3 we give a description of the the collected dataset. In section 4 and 5, we present the models used as well the experimental setup and results, respectively and we conclude in section 6.

2 Speech based Language Identification: an Overview

For Spoken Language Identification, we cite the work done in (Ali et al., 2015) where authors investigated different approaches for dialect identification in Arabic broadcast speech, based on phonetic and lexical features obtained from a speech recognition system, and bottleneck features using the i-vector framework. By using a binary classifier to discriminate between MSA and dialectal Arabic, they obtained an accuracy of 100% . While, they obtained an accuracy of 59.2% to discriminate five Arabic dialects, namely: Egyptian, Gulf, Levantine, North African, and MSA. In (Moftah et al., 2018), the authors have introduced a new technique for extracting the characteristics of different Arabic dialects from speech by discovering the repeated sequences (motifs) that characterize each dialect. They adopted an extremely fast parameter-free Self-Join motif discovery algorithm called Scalable Time series Ordered search Matrix Profile (STOMP) and extracted 12 Mel Frequency Cepstral Coefficients (MFCC) from each motif, which were used to train the Gaussian Mixture Model-Universal Background Model (GMM-UBM) classifier. This approach was applied on three different motif lengths 500 ms, 1000 ms, and 1500 ms on a data set that

was downloaded from Qatar Computing Research Institute domain and carried out some experiments on Egyptian (EGY) and Levantine (LEV). Whereas in (Bougrine and Abdelali, 2018), a system based on prosodic speech information, for intra-country dialects has been proposed. DNN and SVM have been used to evaluate KALAM'DZ, a Web-based corpus dedicated to Algerian Arabic Dialectal varieties. The authors have obtained results that show the close-performance between the DNNs and SVM. In (Lounnas et al., 2019), the problem of identifying languages as Persian, German, English, Arabic and Kabyl [1], has been addressed using Voxforge speech corpus [2].

3 Dataset

We downloaded more than 8 hours of speech data from "Arab podcast" website[3]. This dataset covers MSA and some of its dialects from the following regions: Saudi Arabia (KSA), Syria (SYR), Egypt (EGY), Lebanon (LEB) in addition to English (ENG). The language/dialects are of duration ranging from 50 min to 1 h 30 min. Note that LEB, EGY and KSA-E dialectal corpora include some English expressions along with the conversations. Accordingly this may cause performance degradation compared to the remaining corpora. For training requirements and system design it was necessary to split the downloaded speech files into a smaller segments of around five minutes each, using MKVToolNix GUI $v31.0.0$ [4]. The whole corpus is sampled at 44.1 khz and encoded on 16 bits. Each language/dialect involves conversations spoken by two speakers or more (male and female). Table 1 summarizes the overall statistics of Arpod-1.0 corpus, describing the duration per language/dialect.

Language/Dialect	**Duration** (hours)
KSA	00:50:05
MSA	00:50:05
SYR	00:50:05
ENG	00:50:05
EGY	01:30:00
KSA-E	01:30:00
LEB	01:30:00
Total	**08:10:00**

Table 1: ArPod dataset used for language/dialect identification

The targeted applications that will be trained using Arpod-1.0 are several and not only for the two aforementioned tasks. Since it might be of great help for researchers, we will make it available next[5].

[1] Kabyl is an Algerian Berber dialect.

[2] https://github.com/computational-linguistics-department/Spoken-Language-and-Topic-Identification-Datasets

[3] https://ar-podcast.com/

[4] https://mkvtoolnix.download

[5] https://www.kaggle.com/corpora4research/arpod-corpus-based-on-arabic-podcasts

4 General System

The system includes two types of data representation: acoustic and spectral ones. We used many acoustic features as MFCC, Entropy of Energy, Zero Crossing Rate, Spectral centroid and many others. We used two schemes according to the work mentioned in (Giannakopoulos, 2015). The second type of speech data representation is by using spectogram. We give more details in the following subsections. In our experiments, We used a set of classifiers, namely: kNN, SVM, MLP and Extratrees.

4.1 Acoustic Features based Classification

Scheme 1

In this scheme, 34 features are selected.

1. MFCC coefficients (13)
2. energy(1) & energy of entropy(1)
3. Zero Crossing Rate(1) & Spectral Centroid(1)
4. Spectral Spread (1) & Spectral Entropy(1)
5. Spectral Rolloff(1) & Chroma Vector(12)
6. Spectral Flux(1) & Chroma Deviation(1)

Scheme 2

We have used a framework[6] on the basis of Librosa (McFee et al., 2015), which includes spectral features and rhythm characteristics. We present in the following the features used in this framework, with a total of 193 components:

1. MFCC coefficients (40)
2. Mel spectrogram (128) & Chroma Vector (12)
3. Spectral contrast (7) & Tonnetz(6)

4.2 Spectogram based Classification

In this approach, We opted for an image recognition process to solve the problem of spoken language identification. The idea is to extract the spectogram for our speech dataset which is under .wav format. Then, we applied a CNN classifier to identify languages and dialects based on their respective spectograms.

5 Experiments and Results

In this study, we divided Arpod-1.0 dataset into two parts according to their content: the first one includes 3 hours and 40 minutes of speech, covering two languages: MSA and English (ENG) and two dialects: Saudi (KSA) and Syrian (SYR). The second part -4 hours 30 minutes- is composed of three dialects characterized by language alternation or code switching: Egyptian (EGY), Lebanese (LEB) and Saudi (KSA-E). Note that, in this second part of dataset, speakers alternate between their dialects and English. Experiments have been achieved on speech segments with different durations: 6, 30 and 60 sec.

[6] https://github.com/mtobeiyf/audio-classification

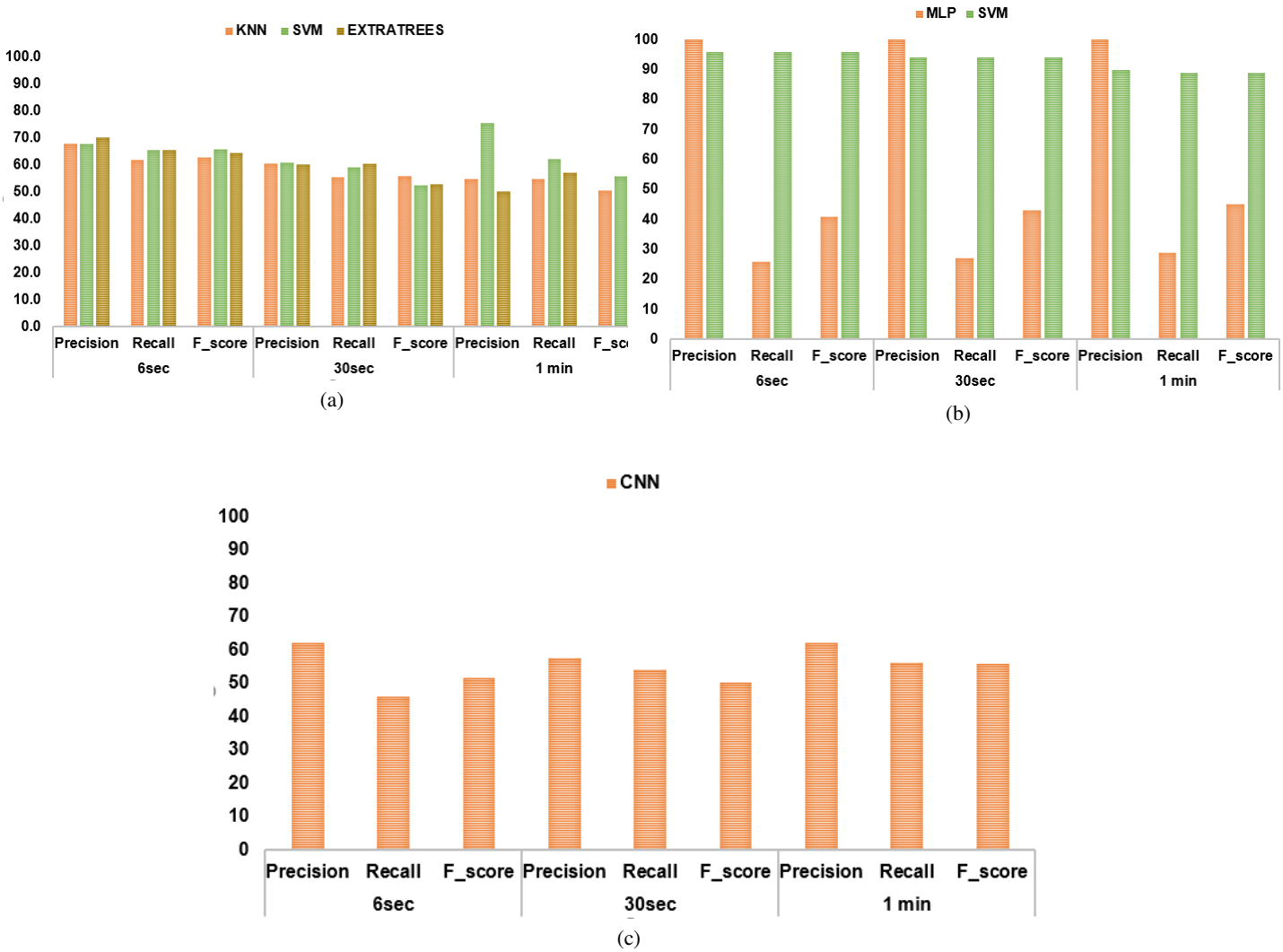

Figure 1: Languages and dialects without code switching, (a) System's performance with scheme 1, (b) System's performance with scheme 2, (c) System's performance using spectrogram based approach.

5.1 Languages and Dialects without Code Switching

As aforementioned, the first experiment has been devoted to identifying languages and dialects that do not contain any kind of code switching. It is about MSA, English, Syrian and Saudi dialects. We should note that all the experiments have been conducted by taking into account the different durations of utterances which are: 6, 30 and 60 seconds.

Based on the results reported in figures 1, we conclude that SVM based on scheme 2 outperforms scheme 1 and spectrogram based approaches, with F1 measure equal to 96%, through short utterances (6 sec). The spectogram based approach yielded an F1 score of 56 % for utterances with 1 min of duration. We should emphasize that performance based on schemes 1 and 2 is inversely proportional to duration, and it is better when dealing with shorter utterances. This is true for kNN, SVM and Extratrees classifiers, except for MLP performance which increases slightly whith duration.

5.2 Dialects with Code Switching

In this experiment, we study whether the system is robust to the code switching phenomenon or not. The speech corpora selected to be used are in Egyptian, Saudi and Lebanese dialects where speakers alternate between English and these dialects. Figure2, shows that the best result was achieved by SVM using the second scheme with an F1 of 98%, for the shortest utterances (6 seconds).

However, unlike experiments dealing with languages and dialects without code switching, performance obtained using the two schemes and the spectrogram based approach is not influenced by the duration of the test utterances.

6 Conclusion

In this paper, we presented the dataset Arpod-1.0 that we collected from Arabic podcasts and prepared to be used for Arabic dialect identification. we conducted a set of experiments to find the model giving the best performance for our language identification system. We have taken into consideration different circumstances like duration of speech utterances and the presence of code switching phenomenon. The findings showed, in the absence of code switching, that shorter utterances are well identified and performance decrease when utterances are longer. Surprisingly, utterances taken from datasets including code

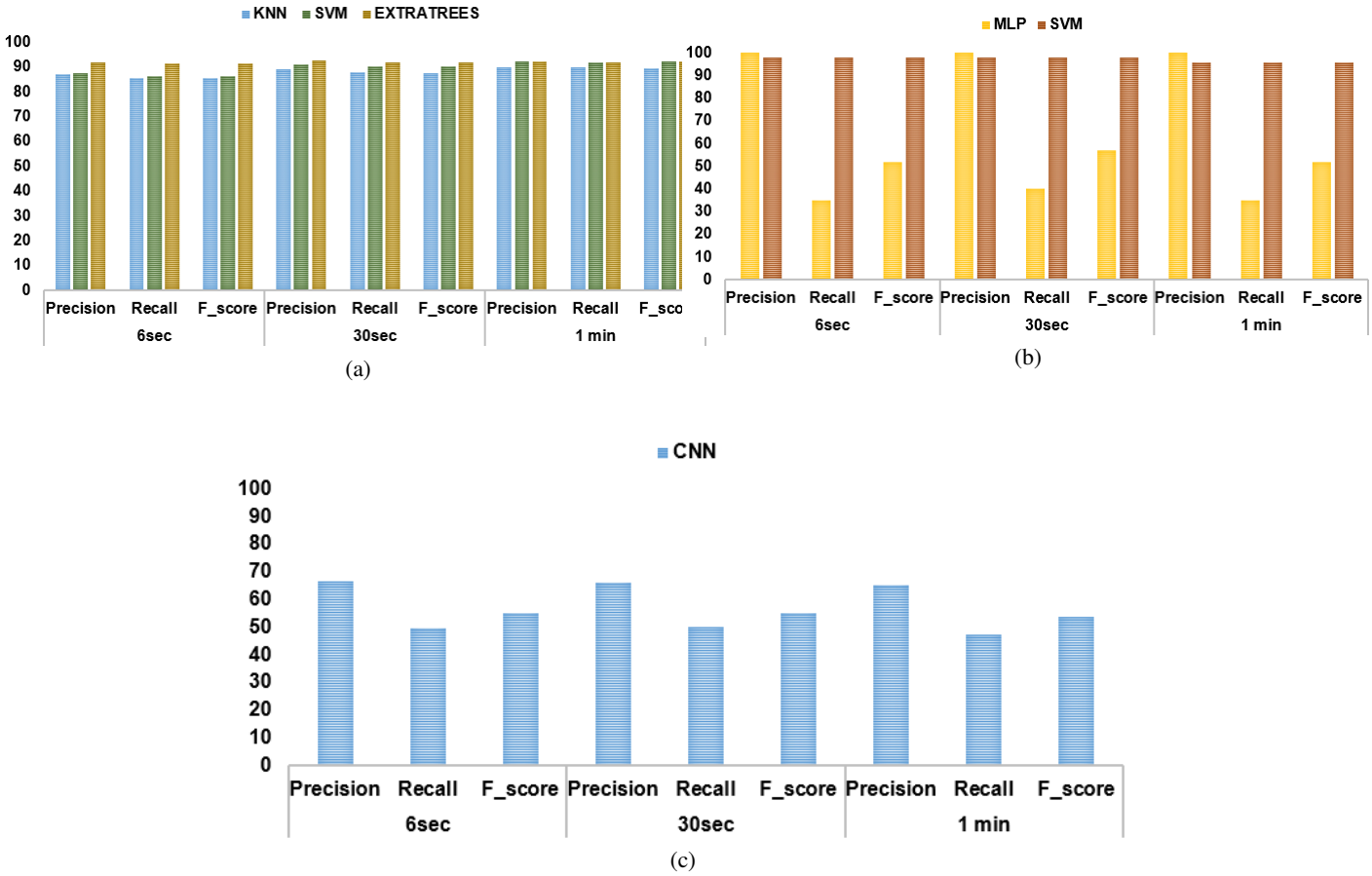

Figure 2: Dialects with code switching, (a) System's performance with scheme 1, (b) System's performance with scheme 2, (c) System's performance using spectrogram based approach.

switched dialects, are well identified using SVM and Extratrees -schemes 1 and 2 - and seem that these models are robust to code switching and duration variation.
In future work, we aim to build a robust model based on other features, like the Shifted delta coefficients (SDCs) which have proven to be efficient in language identification (Lee et al., 2016; Jiang et al., 2014; Ferrer et al., 2015).

Acknowledgment

We thank the reviewers for their valuable comments.

References

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2015. Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

Areej Alshutayri and Hassanin Albarhamtoshy. 2011. Arabic spoken language identification system (aslis): A proposed system to identifying modern standard arabic (msa) and egyptian dialect. In *International Conference on Informatics Engineering and Information Science*, pages 375–385. Springer.

Fadi Biadsy and Julia Hirschberg. 2009. Using prosody and phonotactics in arabic dialect identification. In *Tenth Annual Conference of the International Speech Communication Association*.

Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pages 53–61. Association for Computational Linguistics.

Hadda Cherroun Soumia Bougrine and Ahmed Abdelali. 2018. Spoken arabic algerian dialect identification. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Luciana Ferrer, Yun Lei, Mitchell McLaren, and Nicolas Scheffer. 2015. Study of senone-based deep neural network approaches for spoken language recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1):105–116.

Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610.

Bing Jiang, Yan Song, Si Wei, Jun-Hua Liu, Ian Vince McLoughlin, and Li-Rong Dai. 2014. Deep bottleneck features for spoken language identification. *PloS one*, 9(7):e100795.

Shashidhar G Koolagudi, Deepika Rastogi, and K Sreenivasa Rao. 2012. Identification of language using mel-frequency cepstral coefficients (mfcc). *Procedia Engineering*, 38:3391–3398.

Kong Aik Lee, Haizhou Li, Li Deng, Ville Hautamäki, Wei Rao, Xiong Xiao, Anthony Larcher, Hanwu Sun, Trung Hieu Nguyen, Guangsen Wang, et al. 2016. The 2015 nist language recognition evaluation: the shared view of i2r, fantastic4 and singams.

Khaled Lounnas, Mourad Abbas, Hocine Teffahi, and Mohamed Lichouri. 2019. A language identification system based on voxforge speech corpus. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 529–534. Springer.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

Mohsen Moftah, Mohammed Waleed Fakhr, and Salwa El Ramly. 2018. Arabic dialect identification based on motif discovery using gmm-ubm with different motif lengths. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.