Towards an Automatic Dialect Identification System for Algerian Dialects Using YouTube Videos

Khaled Lounnas
USTHB, Algeria
klounnas@usthb.dz

Mohamed Lichouri
CRSTDLA, Algeria
m.lichouri@crstdla.dz

Mourad Abbas
HCLA, Algeria
m_abbas04@yahoo.fr

Thissas Chahboub, Samir Salmi
USDB, Algeria
chahboubthissas@gmail.com
samirsalmi.eg@gmail.com

Abstract

Many academics are becoming more interested in Spoken Arabic Dialect Identification. Nonetheless, most under-resourced languages suffer from a lack of data, such as the common Algerian dialect, which provides an intriguing case study. As a result, the purpose of this research is to compare the performance of two techniques for the automated identification of Algerian dialects. The first is based on acoustic features whereas the second is based on spectral components extracted from audio sequences collected from YouTube. The experiments were carried out in two setups: raw data and noiseless data (applied noise filter) on 23 Algerian dialects using machine, deep, and transfer learning models while selecting three duration: 5s, 10s, and 20s. The CNN classifier performed the best, enabling us to generate an average F1score of 97.09% with raw data and 96.5% with noiseless data, independent of duration. However, the 20s duration result, which had an F1score of 98.09%, was the best duration that produced the best results for us.

1 Introduction

In all communication technologies, speech is the most natural mode for individuals to make direct contact. With the progress of technology, the scientific community has grown increasingly interested in the field of speech processing, seeking to explore and examine language and the process of voice generation. These tasks are very interesting, especially for low-resourced languages like Arabic and its dialects.

Algeria's dialect, with its richness and diversity, does not correspond to linguistic criteria since it differs from standard Arabic and is composed of a vocabulary with several sources. As a result, we have chosen a phonetic idea of the language rather than a linguistic one. Thus, automatic dialect recognition is the initial stage in performing numerous tasks in NLP (speech translation, opinion mining, etc.), and this study will be the first step in breaking down communication barriers between several Algerian areas.

The contribution of this paper is the development of an automatic spoken language identification system, employing a variety of machine and deep learning approaches to cover 23 classes of Algerian dialects acquired from YouTube videos.

This paper is organized as follows: we present an overview of both speech-based dialect identification and recognition of dialectal speech, and the related work in sections 2. In section 4, we present the system architecture. In section 3, we describe the corpus used to run different experiments. Sections 5 and 6 is devoted to experiments and results regarding dialect identification. The conclusion is presented in section 7.

2 Related Work

Many different sorts of studies have been conducted on spoken dialect identification; some have employed traditional approaches based on statistical classification (Korkmaz and Boyacı, 2022), while others have been enticed to apply deep learning techniques (Garain et al., 2021). However, relatively little study has been conducted on Arabic dialects (Biadsy et al., 2011; Bougrine and Abdelali, 2018; Lounnas et al., 2022).

In the case of Arabic spoken dialect identification, we refer to the research published in (Biadsy and Hirschberg, 2009), the authors used prosodic cues and demonstrated their efficacy across four main Arabic dialects, including Gulf, Iraqi, Lev-

antine, and Egyptian, to demonstrate how employing these descriptors to train the Gaussian Mixture Model (GMM) in conjunction with the Universal Background Model (UBM) may greatly enhance the identification of these dialects of 2-minute utterances. In keeping with their same area of study, the authors tackled the identification of the Arabic accent and dialect in (Biadsy et al., 2011). To do this, they employed phonetic segmentation supravector, which entails creating a kernel function that computes phonetic similarities in order to train the Support Vector Machine classifier. Their Equal Error Rate (EER) was 12.9 %.In (Ali et al., 2015), where researchers looked at several methods for dialect identification in Arabic broadcast speech based on phonetic and lexical characteristics received from a voice recognition system and bottleneck features created using the i-vector framework. They achieved 100% accuracy by employing a binary classifier to distinguish between Modern Standard Arabic (MSA) and dialectal Arabic. While they were able to distinguish between five Arabic dialects—Egyptian, Gulf, Levantine, North African, and MSA—with an accuracy of 59.2%. Authors in (Eldesouki et al., 2016) were concerned with recognizing spoken Arabic dialects from five regions, namely Egyptian, Gulf, Levantine, North-African (Maghrebi), and MSA. Despite the modest quantity of data employed, the researchers claimed that the Linear Support Vector Machine (LSVM) classifier trained with a feature vector incorporating textual features beat the other systems, achieving an accuracy of 51.36%. (Shon et al., 2020) supplied vast dialectal Arabic corpora encompassing 17 dialects to provide more resources for Arabic and its dialects. A total of 3000 hours of speech were provided for training a fine-grained Arabic dialects recognition system, which was divided into three groups based on time ($<$ 5 sec, 5 sec $\sim$ 20 sec, and $>$20 sec). Furthermore, several cutting-edge approaches were developed utilizing the aforementioned dataset, the results reveal that the longer the duration of the speech (in this case more than 20 seconds), the better its identification. Concerning the same issue, and to emphasize the use of the X-Vector approach in the identification of Arabic-spoken dialects, Hanani et al. (Hanani and Naser, 2020) created an X-Vector model utilizing a collection of relevant characteristics (acoustic, lexical, and phonetic) derived from VarDial 2018 and VarDial 2017 and shown that it outperforms existing state-of-the-art models, such as those based on i-vectors, Bottleneck features, and GMM-tokens.

However, for a vernacular Arabic dialect like the Algerian dialect, there are not many works have been done on it. We can cite the contribution of Bougrine et al. (Bougrine et al., 2016) in which she introduced the first Algerian spoken dialect corpora where six Algerian dialects have been modeled in (Bougrine et al., 2018) utilizing prosodic information, which is comprised of rhythm and intonation, and SVM based on the Universal Pearson VII Kernel function (PUK). The authors discovered that prosodic cue was appropriate even for brief utterances with a precision of over 69%. In (Terbeh et al., 2018), the authors suggested a statistical method based on phonetic modelling to determine the relevant Arabic dialect for each input acoustic signal by computing the necessary phonetic model, which was then compared to all referred Arabic dialect models using cosine similarity. (Lounnas et al., 2018) conducted a series of tests using various feature configurations to distinguish between Standard Arabic and one of the Berber dialects known as Kabyl [1]. They demonstrated that a combination of acoustic (Mel Frequency Cepstral Coefficients) and prosodic (melody and stress) characteristics are the best way to distinguish these dialects. A further extension of this work is the one developed in (Lounnas et al., 2019b), in which The difficulty of recognizing languages such as Persian, German, English, Arabic, and Kabyl has been handled by using the Voxforge voice corpus where different systems have been built to identify Persian, German, English, Arabic, and Kabyl dialects. Despite the small size of the data, the system produced an encouraging accuracy of 84.6%.

3 Comparison methodology

As described before, in this paper we focus on identifying Algerian dialects(Bougrine et al., 2016) using our own private dataset. To begin, we converted the videos into audio files. Following the conversion, the audios are pre-processed to reduce noise; this step is only performed for the acoustic approach. The audio files will be divided into 5s, 10s, and 20s segments. We use the segmented data to execute two distinct methods: the first entails extracting the dialects' acoustic parameters and assigning them to a single numerical vector (Lounnas et al., 2020) in order to classify them

[1] Kabyl is an Algerian Berber dialect

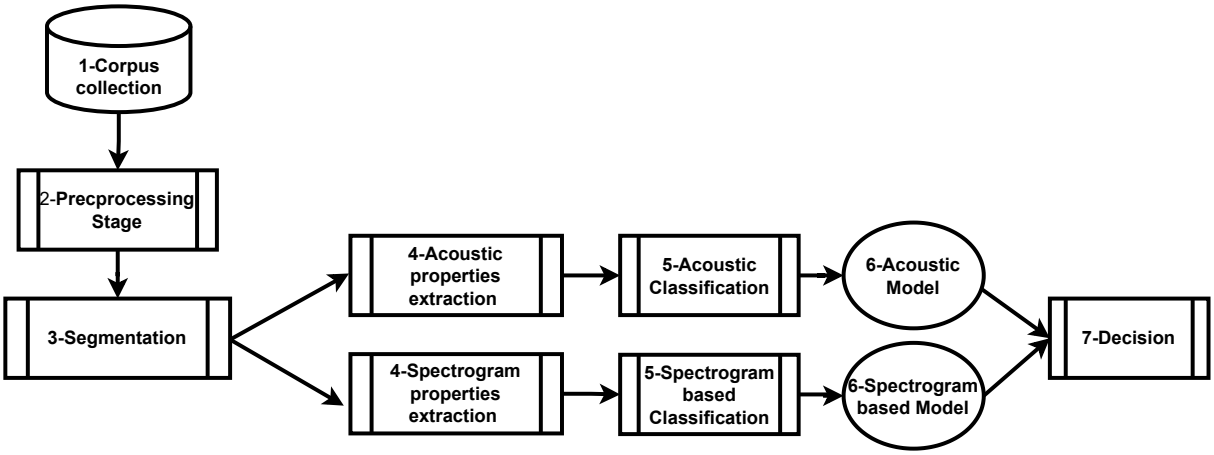

Figure 1: Architecture of our proposed spoken Algerian dialect identification system

using two models, a Support Vector Machine and a Convolutional Neural Network model. The second technique entails transforming the audios of each segment duration into spectrograms (Lounnas et al., 2022), and the resulting image will be pre-processed before being classified using a pre-trained Transfer Learning model VGG16, which has a high accuracy for image classification. Finally, we use assessment metrics to evaluate the trained model.

The schematic presented in Figure 1 depicts the overall design of our approach for the automated identification of Algerian dialects. The architecture is structured as a pipeline of multiple processes which are run sequentially: (1) collection of data; (2) preprocessing; (3) segmentation; (4) features extraction (acoustic vs spectrogram); (5) model training and finally (6) system evaluation. All these processes will be presented in the following sections.

4 Corpus presentation

We selected YouTube as a source for our work since it contains videos on a range of topics created by authors from all across the country. This allowed us to create an Algerian voice datasets with accentual and dialectal variations with around 30 hours of spoken audio. After selecting some YouTube channels that are related to food recipes, daily life, education, and monologue videos, we downloaded all the videos and saved them in MP4 format. The choice is explained by the nature of this video where they don't contain music in the background and with low noise levels. Because our corpus only has 23 cities, there are only 23 sub-dialect. We have various dialects from the center, west, and east that are presented in Table 1 with clips ranging in length from 2h7min to 2h42min for each dialect.

It should be noted that this is one of the fewer works, to our knowledge, that build an Algerian speech corpus for dialect identification, where the first corpus is named KALAM'DZ (Bougrine et al., 2016). In Table 2, which highlights the number of sub-dialects, overall duration, duration per sub-dialect, preprocessing processes, source of data, number of speakers per file, and use cases of each corpus, we compared our corpus to KALAM'DZ. Even though our corpus is smaller than KALAM'DZ's, there is one difference: our emphasis was on YouTube videos with extremely little background noise and no music.

Region	Departments
The North Centre	Algiers, Blida, Tipaza, AinDefla, Tizi Ouzou, Ténès
The North West	Tlemcen, SidiBelabbes, Oran, Maghnia
The North East	Jijel, Annaba, Guelma, Constantine
Central Highlands	M'sila , Laghouat, Bousaada
Eastern Highlands	Batna, Tebessa, Setif, Khenchela
Hoggar-Tassili	Tamanrasset

Table 1: Collected Algerian speech corpora department classification by geographical region (ONS, 2011)

Corpus	KALAM'DZ	Our Corpus
# sub-dialect	43	23
Duration H	104.4	around 47
DpSd H	13.05 (average)	around 2
Preproc	Non-speech segments removal; Speaker Diarization	Noise Reduction
Source	Algerian radio; Algerian TVs; YouTube; Podcast	YouTube
# speaker	Multi speaker	Monologue
Use cases	NLP	Dialect Identification

Table 2: A comparative study between our proposed corpus and the most useful and existing one KALAM'DZ
DpSd: Duration per Sub dialect; Preproc: Preprocessing

4.1 Pre-processing the corpus

In Machine Learning, data pre-processing is an important step that helps improve data quality and promotes the extraction of relevant information from data. It is the process of preparing (cleaning and organizing) raw data so that it may be used to build and train Machine Learning and Deep Learning models. Simply stated, data preprocessing is a data mining approach that converts raw data into a comprehensible and legible format.

4.1.1 Audio preprocessing

We eliminate noise by utilizing Python's "NoiseReduce" library [2], which lowers noise in temporal data such as voice. It is based on a mechanism known as a "spectral gate," which is a type of Noise Gate. It works by computing a signal's spectrogram and predicting a noise threshold (or gate) for each frequency band of this signal/noise. This threshold is used to create a mask that filters out noise below the frequency variation threshold (Sainburg et al., 2020). As seen in Figure 2, the NoiseReduce library removes a considerable number of contaminants from our signal.

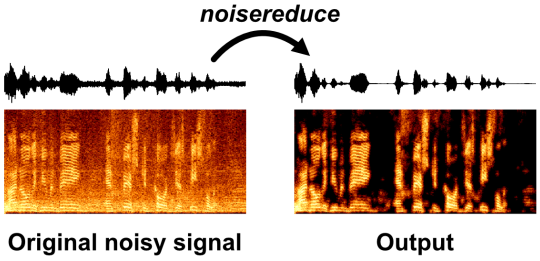

Figure 2: Removing noise from an audio file using the NoiseReduce library (Sainburg, 2019)

5 Experimental Setup

Our classification tasks begin with annotated audio data. There are several forms of audio classifications, but for the sake of our research, we are only interested in two: classification based on acoustic characteristics and classification based on spectrograms.

5.1 Acoustic-based approach

Acoustic information is commonly regarded as the initial level of processing in speech production. It is one of the simplest types of information that may be collected straight from raw speech during the speech parameterization process. Higher level speech information, such as phonotactic and word information, may also be retrieved from acoustic data. Linear Prediction, Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP), and Linear Prediction Cepstral Coefficient (LPCC) are the most often utilized parameterization approaches. We utilized a process based on Librosa (McFee et al., 2015), which incorporates spectral and rhythm properties. As in our previous work in (Lounnas et al., 2019a), we will use the same characteristics, with a total of 193 components:

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

[2] https://pypi.org/project/noisereduce/

These features served as the training data for both an SVM model as well as CNN.

5.2 Spectogram-based approach

We chose to implement the spectrogram method in order to compare the efficiency of the acoustical features extraction approach with that of the spectrogram feature extraction approach for audio clips. The goal is to learn how to classify audio and predict which category they belong to (dialect). We classify the audio from the image by looking at the spectrogram, which relates an intensity or a power to each frequency. The classification task in this situation seems as an image classification. This issue can be applied to a variety of practical applications, such as classifying music videos to determine the genre of music (Nirmal and Mohan, 2020; KM et al., 2021) or classifying short utterances by a group of speakers to identify the speaker based on voice (Liu et al., 2018).

6 Results and discussions

We used the **Sklearn** package with the default parameters to build our first SVM model. Concerning the CNN we will initialize our model as follows (Figure 3):

```
model.add(Conv1D(64, 3, activation='relu', input_shape = 193))
model.add(Conv1D(64, 3, activation='relu'))
model.add(MaxPooling1D(3))
model.add(Conv1D(128, 3, activation='relu'))
model.add(Conv1D(128, 3, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(23, activation='softmax'))
```

Figure 3: CNN model architecture

6.1 Acoustic features-based classification results

After collecting the video samples, we converted them to ".wav" format using the **fffmpeg** function with a sampling rate of 16k. For noise suppression, we utilized the **NoiseReduce** tools(Sainburg, 2019) at the preprocessing stage. In order To compare three different case studies, we divided our speech data into segments of three different duration (5s, 10s, and 20s), which would serve as the input data for three different models. We utilize the functions presented in **Librosa** library to extract the aforementioned characteristics. Each audio file of our corpus will go through this procedure, with the features being stacked vertically in one array vector and the labels in another. Both vectors will be stored in two files, "Features.npy" for the features and "Labels.npy" for the labels, at the end of the feature extraction procedure for later use. After the feature extraction stage, we divide our data into training data and test data, each comprising 80% and 20% of the corpus respectively. Finally, we were able to compare our 12 categorization models, and the following table presents the obtained accuracy and F1score (see Table 3).

For 5s segments, we see that both algorithms' (SVM and CNN) raw data results (95.55% and 96.89%) are actually superior to those obtained from preprocessed data (90.37% and 95.33%). For the 10-second segments, we see that the CNN model's predictions for preprocessed data (96.49%) are marginally better as compared to the raw data (96.31%). However, the SVM model with raw data surpasses the preprocessed data results by 5% (96.65% vs 91.71%). The raw data results outperform the preprocessed data for the 20s segments for both algorithms. We find a minor decline in performance for the preprocessed data, which is related to the fact that it is not as good as the raw data, which is due to noise removal loss.

6.2 Spectogram features-based classification results

The corpus utilized is identical to the one mentioned in the preceding section. We next turn the segmented audio into their spectrograms template. Given that we were working with spectrograms representation, we couldn't utilize SVMs as it is; otherwise, we'd have to add a feature vector extraction phase, so we opted to work with a pre-trained model called Transfer Learning (TL). As there are many TL models, we selected one of the smaller models in term of parameters numbers which is the visual geometric group (VGG16).

According to the results described in Table 4, identifying dialects using spectrograms achieved its best performance with 20s duration data with a macro F1score of 88%, while the scores for 10s and 5s audios are 84% and 57%, respectively. These results are appropriate since the 20s audios include more information than the 5s and 10s. We noted that the duration of the audio file may affect the number of epochs needed for the CNN, where the 5s, 10s, and 20s have required epochs 3, 5, and 20, respectively. To summarize, the pre-trained VGG-16 model didn't achieve the results that were attended, if we take into account that it is well-known for its great accuracy in image classification. We

	Model\Data	*Raw data*			*Preprocessed data*		
		5s	**10s**	**20s**	**5s**	**10s**	**20s**
Accuracy	**SVM**	95.52	96.63	95.77	90.36	91.7	91.94
	CNN	96.84	96.27	98.08	95.29	96.48	97.67
F1score	**SVM**	**95.55**	**96.65**	**95.79**	90.37	91.71	91.92
	CNN	**96.89**	96.31	**98.09**	95.33	**96.49**	97.69

Table 3: Algerian Dialect Identification Based Acoustic features: Reported Result Before and After Processing (noise reduction). The best F1score obtained are in bold.

	5s	**10s**	**20s**
F1score	54	84	88
Epoch	3	5	20

Table 4: Results obtained by the classification of spectogram images using VGG16 pretrained model. The best epoch for each duration is reported.

think that the problem resides in the architecture of the network that we used to retrain the VGG-16 (2-layer network).

By comparing the performance of both the acoustic and spectrogram approach based on the obtained findings, we noted that the acoustic-based approach performs way better than the spectrogram-based approach (with an increase of about 10%). This suggests that auditory features are more trustworthy descriptors for distinguishing various dialects, reducing confusion, and producing better outcomes. The acquired findings are quite good when compared to what is available in the literature; our addition to this work is that we worked with 23 Algerian dialects.

7 Conclusion

This research focuses on the creation of an automatic system for recognizing Algerian dialects. To attain our aim, we employed two approaches. The first is based on acoustic features derived from audio, while the second is based on spectrogram features. The two approaches have been evaluated using SVM and CNN for the first one whereas a transfer learning technics (VGG-16) was applied for the second approach, respectively.

These models were evaluated using audio durations of 5s, 10s, and 20s. The results for the 20s duration data using CNN were extremely good, with an accuracy of 98.09% for the raw data. However, it should be highlighted that while employing a spectrogram to train a VGG deep learning model, our proposal performed best when the size of the audio voice was significant (the 20s). The previous experience with the 23 Algerian dialects stresses the relevance of acoustic parameters and the use of spectrograms in differentiating Algerian dialects. Future research might try to add new dialects and cities to our current corpus in order to eventually encompass all Algerian dialects. In addition, we will investigate deep learning methods to improve the modeling of Algerian dialects. Finally, we will look for the optimal duration that will allow our system to generalize more effectively.

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*.

Fadi Biadsy and Julia Bell Hirschberg. 2009. Using prosody and phonotactics in arabic dialect identification. In *Academic Commons, https://doi.org/10.7916/D8HM5HRV*.

Fadi Biadsy, Julia Bell Hirschberg, and Daniel PW Ellis. 2011. Dialect and accent recognition using phonetic-segmentation supervectors.

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.

Soumia Bougrine, Hadda Cherroun, and Djelloul Ziadi. 2018. Prosody-based spoken algerian arabic dialect identification. *Procedia Computer Science*, 128:9–17.

Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, Abdallah Lakhdari, and Aicha Chorana. 2016. Toward a rich arabic speech parallel corpus for algerian subdialects. In *The 2nd Workshop on Arabic Corpora and Processing Tools*, pages 2–10.

Mohamed Eldesouki, Fahim Dalvi, Hassan Sajjad, and Kareem Darwish. 2016. Qcri@ dsl 2016: Spoken arabic dialect identification using textual features. In

Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), pages 221–226.

Avishek Garain, Pawan Kumar Singh, and Ram Sarkar. 2021. Fuzzygcp: A deep learning architecture for automatic spoken language identification from speech signals. *Expert Systems with Applications*, 168:114416.

Abualsoud Hanani and Rabee Naser. 2020. Spoken arabic dialect recognition using x-vectors. *Natural Language Engineering*, 26:691 – 700.

Athulya KM et al. 2021. Deep learning based music genre classification using spectrogram.

Yunus Korkmaz and Aytuğ Boyacı. 2022. A comprehensive turkish accent/dialect recognition system using acoustic perceptual formants. *Applied Acoustics*, 193:108761.

Zheli Liu, Zhendong Wu, Tong Li, Jin Li, and Chao Shen. 2018. Gmm and cnn hybrid method for short utterance speaker recognition. *IEEE Transactions on Industrial informatics*, 14(7):3244–3252.

Khaled Lounnas, Mourad Abbas, and Mohamed Lichouri. 2019a. Building a speech corpus based on arabic podcasts for language and dialect identification. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing*, pages 54–58.

Khaled Lounnas, Mourad Abbas, Mohamed Lichouri, Mohamed Hamidi, Hassan Satori, and Hocine Teffahi. 2022. Enhancement of spoken digits recognition for under-resourced languages: case of algerian and moroccan dialects. *International Journal of Speech Technology*, pages 1–13.

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

Khaled Lounnas, Lyes Demri, Leila Falek, and Hocine Teffahi. 2018. Automatic language identification for berber and arabic languages using prosodic features. In *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, pages 1–4. IEEE.

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.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P 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, pages 18–25. Citeseer.

MR Nirmal and Shajee Mohan. 2020. Music genre classification using spectrograms. In *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–5. IEEE.

ONS. 2011. Recensement général de la population et de l'habitat – 2008 – (résultats issus de l'exploitation exhaustive).

Tim Sainburg. 2019. timsainb/noisereduce: v1.0.

Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. Adi17: A fine-grained arabic dialect identification dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8244–8248 . IEEE.

Naim Terbeh, Mohsen Maraoui, and Mounir Zrigui. 2018. Arabic dialect identification based on probabilistic-phonetic modeling. *Computación y Sistemas*, 22(3):863–870 .