A Transfer Learning Approach For Identifying Spoken Maghrebi Dialects

Khaled Lounnas[1,2], Mourad Abbas[1], Mohamed Lichouri[1], Hocine Teffahi[2]
Mohamed Hamidi[3], and Hassan Satori[3]

[1] Computational Linguistics Dept, CRSTDLA, Algiers, Algeria,
[2] LCPTS-USTHB University, Algiers, Algeria,
[3] LISAC-FSDM-USMBA University, Fes, Morocco
klounnas@usthb.dz, m.abbas@crstdla.dz, m.lichouri@crstdla.dz,
hteffahi@gmail.com, mohamed.hamidi.5@gmail.com, hsatori@yahoo.com

Abstract. This paper investigates a transfer learning approach to solve the spoken dialects identification problem for some under-resourced dialects of the Maghrebi region, including Algerian Arabic Dialect (AAD), Algerian Berber Dialect (ABD), Moroccan Arabic Dialect (MAD), and Moroccan Berber Dialect (MBD). In our experiments, we used different Transfer learning models, namely: Residual Neural Network (Resnet50, Resnet101), and Visual Geometric Group (VGG16, VGG19) using an in-house corpus that we built for each dialect. The corpus is composed of ten digits recorded for each of the aforementioned dialects, repeated ten times by six native speakers.
The results vary according to different reasons: the number of epochs, neurons, batch size, and also the datasets combinations used in training and test phases. The best score found is 90.4% by the VGG19 model. Overall, the results show the robustness of our system based on the VGG16 model with an average identification rate of 62.7%.

Keywords: Dialect Identification, Resnet50, Resnet101, VGG16, VGG19

1 Introduction

Spoken Dialect Identification (SDI) is a machine learning topic that refers to the assignment of a spoken utterance to its proper dialect among a set of dialects. On one hand, dealing with SDI problems becomes more challenging, especially when dialects are very close to each other and similar in different contexts, including phonological, morphological, lexical, and syntactic levels. On the other hand, building resources for under-resourced dialects are getting more attention in this field and more challenging too. For this purpose, we focus, first, on building a resource for the Maghrebi dialects, then using this resource to carry out a set of experiments based on transfer learning approaches in order to present the best performing configuration of a neural network. The remaining of this paper is organized as follows; we present an overview of the works on speech-based dialect identification in section 2. In section 3, we describe the corpus that we prepared. In section 4, we present the proposed system as well as the experimental setup and results, and we conclude our paper in section 5.

2 Related Work

Spoken dialects identification has attracted many types of research, some of them used classical approaches based on statistical classification, however many others tempted to benefit from deep learning techniques. Following the first paradigm and intending to find the most appropriate characteristic vector and hierarchy of classification, it has been shown in [1–3] that the combination of both acoustic and prosodic information can be used to better distinguish between Arabic and Kabyl. In the same context, in [4], authors exploited the prosodic cues and observed its effectiveness across four major Arabic dialects, namely: Gulf, Iraqi, Levantine, and Egyptian, in which they prove that using such kind of descriptors to train Gaussian Mixture Model (GMM) combined with the Universal Background Model (UBM) can significantly improve the identification of these dialects of 2 minutes utterances. Following the same research focus, authors addressed in [5] Arabic accent and dialect identification; They used phonetic segmentation supra-vector, which consists in building a kernel function that computes phonetic similarities to train the Support Vector Machine classifier. They achieved an Equal Error Rate (EER) of 12.9%. In [6], authors have been interested in identifying spoken Arabic dialects that belong to five regions, namely: Egyptian, Gulf, Levantine, North-African (Maghrebi), and Modern Standard Arabic (MSA). The authors reported that despite the small size of the used data, the Linear Support Vector Machine (LSVM) classifier trained with a feature vector containing textual details, outperformed the other systems getting an accuracy equal to 51.36%.

Deep learning systems have also been implemented to identify spoken Arabic dialects/sub-dialects. Indeed, in [7] authors proposed prosodic parameters to model some under-resourced Algerian dialects using deep learning. The authors carried out a comparative study between statistical and neural approaches. Lounnas et al.[8], collected more than 8 hours of speech data to prepare ArPod corpus which contains Modern Standard Arabic and some of its dialects: Syrian, Lebanese, Egyptian, and Saudi. Using Arpod for spoken language and dialect identification, the authors showed that using Convolutional Neural Network (CNN) with specific features outperforms the classical approach.

3 Corpus Preparation

As the purpose of this paper is to implement a spoken Maghrebi dialects identification system, we built our own dataset from scratch. This resource contains four dialects from the Maghreb region namely: Algerian Arabic Dialect (AAD), Algerian Berber Dialect (ABD), Moroccan Arabic Dialect (MAD) and Moroccan Berber Dialect (MBD). The Moroccan dialects are taken from the corpus used in [10]. The first step for collecting the corpus is to ask the native Maghrebi speakers (Algerian and Moroccan) to utter and record ten times the numerical digits from zero to nine according to the real-life conditions. The next step consists

Table 1: The corpus' characteristics.

Sampling rate	16 Khz
Number of bits	16 bits
Number of Channels	1, Mono
Audio data file format	.wav
# speakers	24
# speakers per dialect	6
# dialect	4
# tokens per speaker	100
# speakers according to gender	12 males and 12 females
Total number of tokens	2400
Number of digits	10 digits (ABD) 10 digits (MBD) 10 digits (AAD) 10 digits (MAD)
Number of repetitions per word	10
Condition of noise	normal life
Preemphased	$1 - 0.97z^{-}1$
Window type Hamming	25.6 ms
Frames overlap	10 ms

of re-sampling the output spoken digits to get a standardized frequency using Praat [4] software, given that the recording conditions vary from one speaker to another. The last step is to segment the original files into tiny fragments (with ten repetitions for every digit). This task is conducted using Audacity[5]. Table 1 outlines the corpus's characteristics.

4 Experiments and Results

4.1 System Description

Our system is based on transfer learning models, including Residual Neural Network (ResNet) and Visual Geometric Group (VGG) as illustrated in Figure 1. We used the Keras's applications included in the tensorflow library[6] to conduct our experiments using different parameters and different combinations of the dataset. First, we extracted spectrogram images to be used as an input of our system.

We run experiments using multiple combinations of speakers to form ten different learning and test sets. Each combination presents four speakers representing 70% for training and two speakers representing 30% for test, as shown in table 2, where S_i denotes speaker number i.

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

[5] https://www.audacityteam.org

[6] https://www.tensorflow.org/api_docs/python/tf/keras/applications/

Table 2: Multiple combinations of speakers datasets

set	Training Speakers' set	Test Speakers' set
01	S1,S2,S3,S4	S5,S6
02	S6,S5,S1,S2	S3,S4
03	S4,S3,S6,S5	S1,S2
04	S4,S3,S6,S2	S1,S5
05	S4,S2,S1,S5	S3,S6
06	S1,S2,S3,S5	S4,S6
07	S1,S3,S4,S6	S2,S5
08	S2,S4,S5,S6	S1,S3
09	S2,S3,S5,S6	S1,S4
10	S1,S3,S5,S6	S2,S4

Table 3: Configurations parameters (Number of epochs, neurons, Batch Size, in the first layer of the NN2 network)

Configuration	# epochs	Batch size	# Neurons
Cf1	2	20	128
Cf2	2	20	256
Cf3	10	10	128
Cf4	10	10	256

4.2 Dialect Identification using Deep Residual Network

Deep Residual Network is almost identical to the networks that are stacked over each other with convolution, pooling, activation and fully-connected layers. The identity connection between the layers is the only building of the simple network to make it a residual network[12].

In this section, we conducted a series of experiments to test different hyper-parameters of the last two layers network (NN2). This network is used to retrain both pre-trained ResNet50 and ResNet101 models. In the framework of these experiments, We varied the number of neurons, epochs, and batch size, as presented in table 3. The results are presented in terms of F1 score in tables **??** and 5. For the Resnet50+NN2 model, the best results obtained with the first and second configurations (cf1 and cf2) are 60.3% and 39% using the eighth and third sets, respectively. Compared to cf1 and cf2 configurations, cf3 and cf4 yielded an improvement by 16% and 25%, respectively. This improvement is achieved using the second set.

Overall the dialect identification system performs better using the cf3 and cf4 configurations than the cf1 and cf2 with an average rate of 53.97% and 61.05%, respectively. There is no clear interpretation on why the system's performance is high when using some sets and low for some others. The reason is probably the different environments in which the data was recorded.

To enhance the performance of our system, we retrained the ResNet with higher number of layers i.e.101 layers. The best results achieved with cf1 and cf2

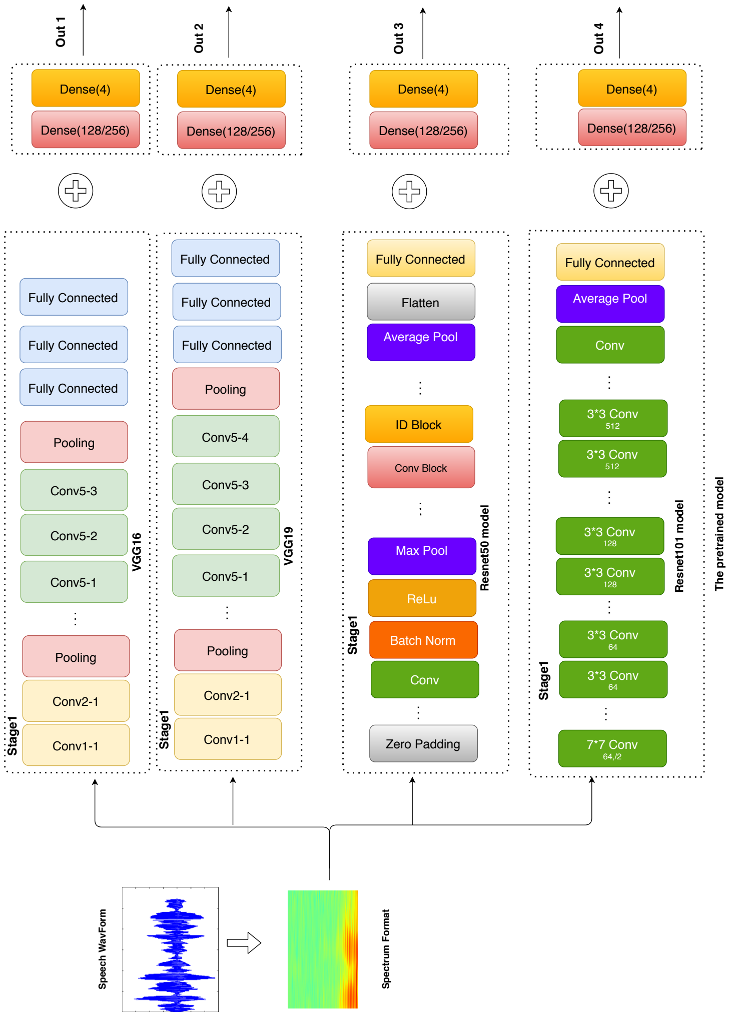

Fig. 1: Transfer Learning based Approach for Spoken Digits Dialects Identification.

configurations are 49.56% and 63.59% using the third and last set, respectively. For cf3 and cf4 configurations, we noted a significant improvement in both cases at the second set compared with the cf2 configuration, with an enhancement of 15% and 20%, respectively. The results are reported in table 5. Overall, while observing the results obtained in the four configurations (cf1,cf2, cf3 and cf4), we noted that the performance of the ResNet based Dialect identification is related to the network hyper-parameters mentioned above in addition to the number of neurons in the first layer of NN2 network. In fact, increasing the number of neurons (from 128 to 256) has a positive impact on the system's

Table 4: Performance of ResNet50 based system

	Configurations			
Set	Cf1	Cf2	Cf3	Cf4
01	40,52	35,22	60,06	50,77
02	41,00	34,00	**75,78**	**85,22**
03	38,00	**39,00**	50,29	61,97
04	48,00	28,00	50,29	63,13
05	26,36	33,61	54,18	64,24
06	43,74	27,71	60,46	60,20
07	38,82	31,64	36,86	49,01
08	**60,30**	34,33	33,09	61,75
09	40,93	34,15	58,17	52,60
10	27,45	30,38	60,50	61,56
Average	40,51	32,80	53,97	**61,05**

Table 5: Performance of ResNet101 based system

	Configurations			
Set	Cf1	Cf2	Cf3	Cf4
01	33.59	35.32	54.99	56.58
02	33.95	51.40	**77.54**	**83.83**
03	**49.56**	48.05	62.42	52.37
04	33.30	50.94	61.04	60.15
05	33.66	33.99	61.73	58.45
06	23.90	33.27	56.27	59.38
07	41.11	32.97	60.84	34.35
08	39.76	37.50	39.59	62.03
09	25.05	58.26	60.87	56.98
10	39.73	**63.59**	38.50	58.52
Average	35,36	44,53	57,38	**58,26**

performance. Expanding the number of layers has the same effect when using the cf2 configuration. The score jumped from 39%(ResNet50+NN2) for the third set (ResNet101+NN2) to 63.59% for the tenth set. In general, the cf2 and cf3 configurations seem to perform better than cf1 and cf4.

4.3 Dialect Identification using Visual Geometric Group (VGG)

The Visual Geometric Group is a convolutional neural network with a specific architecture that was proposed in [13]. As shown in Figure 2, its structure consists of blocks, where each block is made of 2D Convolution and Max Pooling layers. The VGGNet comes in two versions, VGG16 and VGG19, where numbers 16 and 19 reflect the number of layers. In this part, we present the experiments that we conducted using the VGG16 and VGG19 pre-trained models, for which

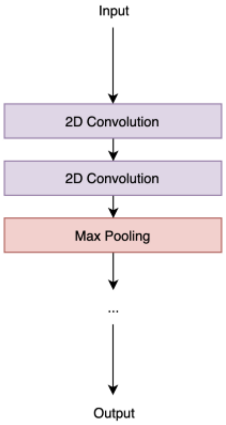

Fig. 2: Visual Geometric Group Network Structure

we adopted the same configurations reported in table 3. The obtained results are reported in tables 6 and 7.

Table 6: Performance of VGG 16 based system

	Configurations			
Set	Cf1	Cf2	Cf3	Cf4
01	48,41	44,97	57,45	49,39
02	45,07	27,52	**79,43**	**84,95**
03	20,50	44,42	60,64	37,61
04	49,35	**61,48**	53,31	49,93
05	18,43	34,08	58.06	55.40
06	37,65	26,02	59.53	51.10
07	18,83	35,93	57.49	47.27
08	25,39	23,39	49.71	62.97
09	**54,37**	48,30	57.47	59.60
10	34,50	35,61	40.79	33.08
Average	35,25	38,17	**62,70**	55,47

For the VGG16+NN2 model, the two best results are obtained with the fourth and third configuration using the second set, with a score of 84.95% and 79.43%, respectively. Whereas, using the 9^{th} set yielded to a score of 54.37% by adopting the first configuration. Furthermore, the second configuration allowed an improvement of around 7% using a different set, the fourth one. Having the performance achieved by the four aforementioned configurations, we note clearly that the lower the number of neurons, the worse the performance of the system. Note that the best result is obtained with the fourth configuration (Max: 84.95%, Avg: 55.47%) whereas the best average performance was obtained with the third configuration (Max: 79.43%, Avg: 62.70%). However, the VGG16+NN2 with the third configuration seems to be more appropriate and more stable in comparison to the three other configurations.

Table 7: Performance of VGG 19 based system

	Configurations			
Set	Cf1	Cf2	Cf3	Cf4
01	27.15	10.30	50.90	38.54
02	**60.64**	**55.81**	**72.15**	**90.40**
03	54,20	25,25	55.42	60.66
04	19,54	09,84	59.54	44.69
05	39,66	28,6	45.03	53.55
06	26,69	34,12	54,24	54,43
07	31,70	12,81	26,37	47,26
08	15,16	23,80	59,15	35,42
09	44,55	32,55	59,82	53,46
10	31,11	18,45	56,99	31,78
Average	32,82	23,17	**51,31**	44,47

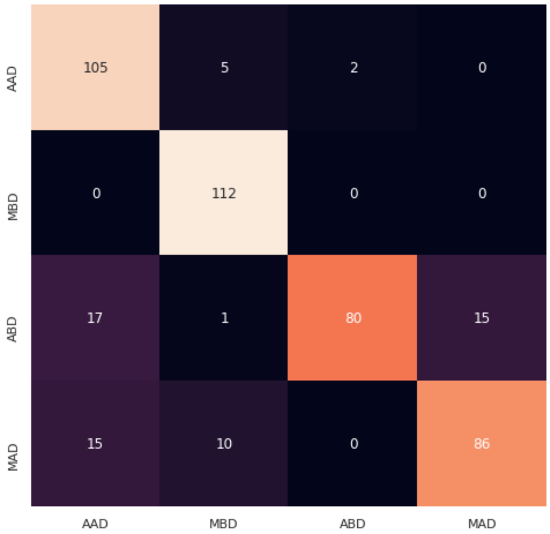

Fig. 3: Confusion matrix: ResNet50 (Cf4, 2^{nd} set)

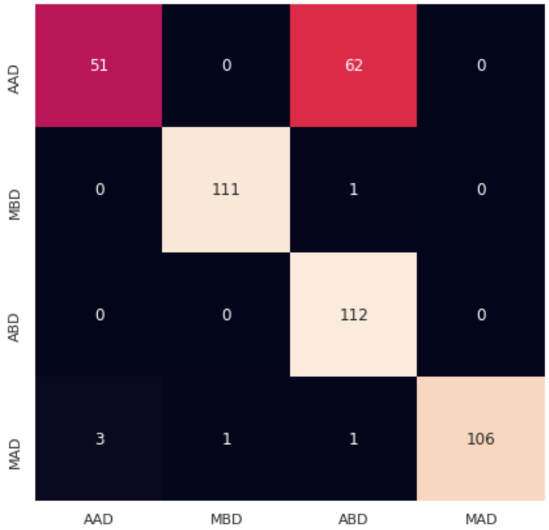

Fig. 4: Confusion matrix: ResNet101 (Cf4, 2^{nd} set)

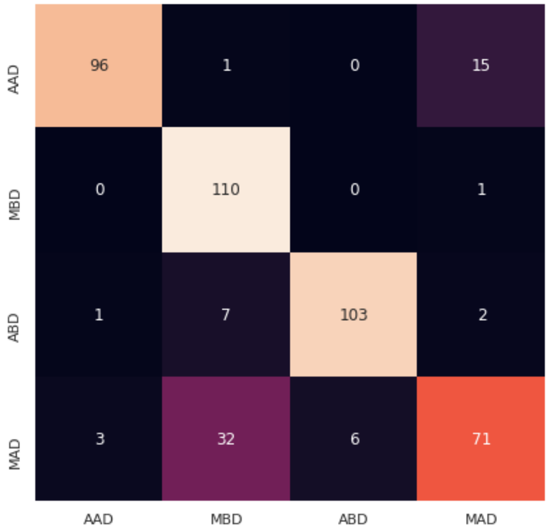

Fig. 5: Confusion matrix: VGG16 (Cf4, 2^{nd} set)

The VGG19+NN2 model is better than VGG16+NN2 in some cases: (4^{th} configuration, 2^{nd} set) where we achieved the best performance (90.40% versus 84.95%). This is due to the learning parameters {144M (VGG19+NN2) Versus 138M (VGG16+NN2)}. Despite the best performance achieved in such indi-

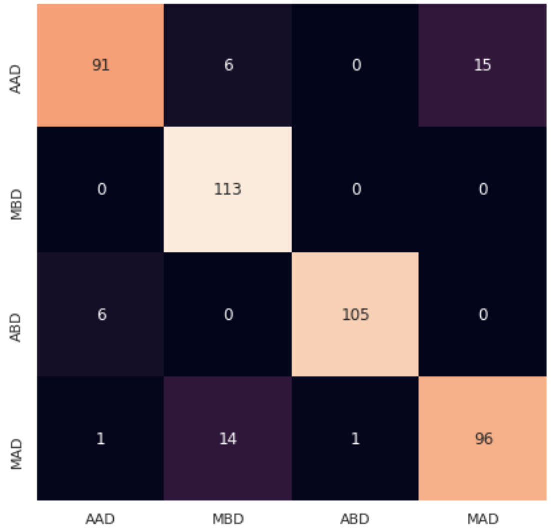

Fig. 6: Confusion matrix: VGG19 (Cf4, 2^{nd} set)

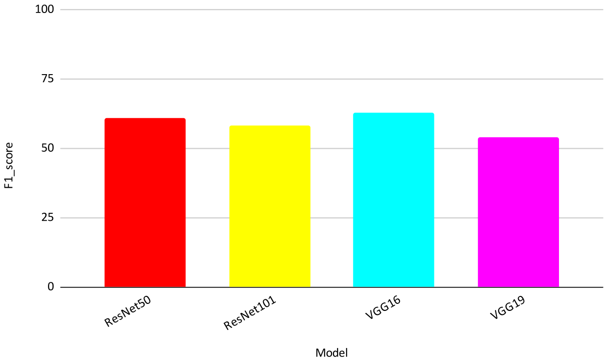

Fig. 7: The best scores achieved with VGG16, VGG19, ResNet50, and ResNet101 models

vidual (more than 90%), the overall performance is around 45%. Results are summarized in table 7.

4.4 Summary and Discussion

In this section, we present the best results achieved by the appropriate configurations. The system works better using the values epoch=10 and batch size=10 (configurations three and four) than using epoch=2 and batch size=20 (configurations first and second). Confusion matrices are plotted for the pre-trained

Table 8: Spoken digits differences in AAD, ABD, MAD, and MBD.

Dialect	AAD	ABD	MAD	MBD
’0’	SIFER	OULECH	SIFER	ILEM
’1’	WAHED	YEWAN	WAHED	YEN
’2’	ZOUJ	SIN	JOJ	SIN
’3’	TLATHA	THYATHA	THLALATA	KRAD
’4’	REBAA	REBAA	RABAA	KUZ
’5’	KHEMSA	KHEMSA	KHAMSA	SMUS
’6’	SETTA	SETSA	STTA	SEDISS
’7’	SEBAA	SEBAA	SBAA	SA
’8’	THEMANYA	THMANIA	THMANYA	TAM
’9’	TESAA	TESSAA	TSAAOD	TZA

models in Figures 3, 4, 5, 6. Most of the four Maghrebi dialects are well recognized, though the closeness of lexicon between these dialects where they differ in one or two digits only, as indicated in table 8.

Table 9: Number of parameters for each model

Model	# parameters
ResNet50+NN2	26M
ResNet101+NN2	44.5M
VGG16+NN2	138M
VGG19+NN2	144M

We summarize in Figure 7 the best performance achieved with each of the four models (average F1 score). Note that the VGG16+NN2 model outperforms ResNet50+NN2, ResNet101+NN2 and VGG19+NN2. Nevertheless, ResNet models are advantageous in the sense they use a smaller number of ”learning parameters” as shown in table 9.

5 Conclusion

In this work, we applied the ”Transfer Learning” approach on the Automatic Identification of Spoken Dialects in the Maghreb region, namely Algerian and Moroccan dialects. using spectrogram-based features, we retrained multiple pre-trained models. A series of experiments were carried out by varying a model-specific parameter referring to the number of layers (50 and 101 for ResNet and 16 and 19 for VGG, respectively) in addition to the number of iterations (epoch) and batch size. We also varied the number of neurons of a two layers network (NN2) which was used to retrain the four pre-trained models.

The findings show that despite the small amount of the used dataset, VGG

outperforms the other models. Nevertheless, the performance achieved with VGG is in cost of computation time and memory space. As a result, the ResNet pre-trained models seem to be a reasonable alternative to the VGG pre-trained models to overcome this problem.
In perspective, one of the next steps to be done, is to enrich our corpus by adding other dialects, increasing the corpus size and using different models such as AlexNet[14], GoogleNet[15] and DenseNet[16].

References

1. Lounnas, K., Demri, L., Falek, L., & Teffahi, H. (2018, October). automatic language identification for berber and arabic languages using prosodic features. In 2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM) (pp. 1-4). IEEE.
2. Lounnas, K., Abbas, M., Teffahi, H., & Lichouri, M. (2019, March). A Language Identification System Based on Voxforge Speech Corpus. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 529-534). Springer, Cham.
3. Lounnas, K., Satori, H., Teffahi, H., Abbas, M., & Lichouri, M. (2020, April). 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) (pp. 1-5). IEEE.
4. Biadsy, F., & Hirschberg, J. (2009). Using prosody and phonotactics in arabic dialect identification. In Tenth Annual Conference of the International Speech Communication Association.
5. Biadsy, F., Hirschberg, J., & Ellis, D. P. (2011). Dialect and accent recognition using phonetic-segmentation supervectors. In Twelfth Annual Conference of the International Speech Communication Association.
6. Eldesouki, M., Dalvi, F., Sajjad, H., & Darwish, K. (2016, December). 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) (pp. 221-226).
7. Bougrine, H. C. S., & Abdelali, A. (2018, April). Spoken arabic algerian dialect identification. In 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP) (pp. 1-6). IEEE.
8. Lounnas, K., Abbas, M., & Lichouri, M. (2019). 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 (pp. 54-58).
9. Satori, H., Harti, M., & Chenfour, N. (2007). Système de Reconnaissance Automatique de l'arabe basé sur CMUSphinx. Corpus, 22, 25.
10. Satori, H., & Elhaoussi, F. (2014). Investigation Amazigh speech recognition using CMU tools. International Journal of Speech Technology, 17(3), 235-243.
11. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (Vol. 8).
12. ANKIT SACHAN. Detailed Guide to Understand and Implement ResNets. Published in 2019-09-17, Accessed online in 2020-12-09. https://cv-tricks.com/keras/understand-implement-resnets/

13. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
14. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90.
15. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
16. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).