ADI17: A FINE-GRAINED ARABIC DIALECT IDENTIFICATION DATASET

Suwon Shon[1]*, *Ahmed Ali*[2], *Younes Samih*[2], *Hamdy Mubarak*[2], *James Glass*[3]

ASAPP Inc., New York, NY, USA[1]
Qatar Computing Research Institute, HBKU, Doha, Qatar[2]
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA[3]

swshon@csail.mit.edu amali@qf.org.qa glass@mit.edu

ABSTRACT

In this paper, we describe a method to collect dialectal speech from YouTube videos to create a large-scale Dialect Identification (DID) dataset. Using this method, we collected dialectal Arabic from known YouTube channels from 17 Arabic speaking countries in the Middle East and Northern Africa. After a refinement process, a total of 3,000 hours of speech was available for training DID systems, with an additional 57 hours of speech for development and testing. For detailed evaluations, the DID data was divided into three sub-categories based on the segment duration: short (less than 5s), medium (5–20s), and long (over 20s). We compare state-of-the-art DID techniques on these data, and also analyze a DID system trained on these data. Since the training and test data share the same channel domain, we also used the Multi-Genre Broadcast 3 (MGB-3) test set to evaluate on domain mismatched condition.

Index Terms— Dialect Identification, Arabic dialect, Language Identification, Dataset, Large-scale

1. INTRODUCTION

Language identification (LID) has become increasingly important for many speech processing tasks such as automatic speech recognition, machine translation, and speech synthesis. LID research has benefited tremendously through the years by regular Language Recognition Evaluation (LRE) challenges that have been organized by the National Institute of Standards and Technology (NIST). The NIST LRE series has established baselines of LID performance and provided datasets of conversational telephone speech.

Dialect identification (DID) can be regarded as a special case of LID. Compared to LID however, DID is arguably more challenging because dialects usually belong the same language family. Thus, subtle differences among dialects and accents are the only cue to identification. Despite these challenges, DID is relatively unexplored compared to LID. One of the main reasons is the lack of a common dataset for research support. For example, the NIST LRE 2015 challenge provided Arabic language sets with 4 regional Arabic dialect labels, which is limited if we consider all the varieties of Arabic.

Arabic is an attractive language to explore DID, due to its uniqueness and widespread use. While 22 countries in the Arab world use Modern Standard Arabic (MSA) as their official language, citizens speak their local dialect in everyday life. Arabic dialects are historically related and share Arabic characters. However, they are not mutually comprehensible. Arabic DID, therefore, poses different challenges compared to other language dialects containing comprehensible vernacular [1]. The previous Multi Genre Broadcast (MGB-3) challenge provided a dataset containing five regional Arabic dialects and resulted in studies covering diverse DID topics such as domain adaptation [2, 3, 4], semi-supervised learning [5, 6, 7, 8], and linguistic feature extraction [1, 9]. Nonetheless, the limitation of prior studies is obvious because there are only five regional dialect classes and each class still needs to cover a large variety of Arabic dialects. Also, the MGB-3 dataset is a very small dataset which has only 53 hours of speech for the training set. NIST LRE series datasets could be used to explore DID, but these datasets are not freely available and also have less than five regional classes. For these reasons, we decided to collect a freely available Arabic Dialect Identification dataset for 17 countries (ADI17) with a large-scale and fine-grained label set.

The previous Arabic dialect datasets were limited under five regional dialect classes. To extend the task to a fine-grained analysis of dialectal Arabic speech, we collected from YouTube about 3 000 hours of Arabic dialect speech data from 17 countries. A further 280 hours of data was collected which was processed using automatic speaker clustering and dialect labeling by human annotators, resulting in 58 hours of speech selected for use as development and test sets. To provide a benchmark performance test, we defined two subtask conditions for a supervised task and semi-supervised task. For two sub-tasks, we evaluated the state-of-the-art systems such as end-to-end DID system and x-vector. Note that this dataset was also used for the 5th edition of MGB challenge [10].

2. ADI17 DATASET AND COLLECTION PIPELINE

The Arabic Dialect Identification 17 country (ADI17) dataset consists of videos from 17 Arabic countries. The dataset provides 11-character YouTube video IDs, timestamps (i.e., start times and end times) and dialect labels. Since the original videos are subject to copyright, we do not make them available directly. We instead provide the YouTube IDs, timestamps, and annotations[1]. YouTube is freely available, anyone can download and segment the videos using the time stamp information. For this paper, we only used the audio segment and discarded the video. A total of 3,033 hours of speech are provided for the training set, but the proportions of each language are severely unbalanced. For example, Iraq has 815 hours while Jordan has only 25 hours. This unbalance could be problematic for training, but we release the dataset as is without balancing the proportions to provide as much data as possible. The labels of the training set are noisy, so they could contain other dialects or languages while the development and test set were labeled by a human annotator. The test set has three sub-categories with segments of different durations, i.e., short (under 5sec), medium (between 5 and 20 sec) and long (over

*Work done at MIT CSAIL

[1] Available: http://groups.csail.mit.edu/sls/downloads/adi17/

978-1-5090-6631-5/20/$31.00 ©2020 IEEE

Name	Free	Channel	Dialect labels	Duration
MGB-3 [11]	✓	Broadcast News	5 (Regional)	74h
VarDial2018 [12] (only test set is available)	✓	Multimedia (YouTube)	5 (Regional)	26h
GALE Phase 2 Arabic Broadcast Conversation Speech		Broadcast News	2 (MSA or dialect)	251h
Multi-Language Conversational Telephone Speech 2011		Telephone	4 (Regional)	117h
NIST LRE 2017 (most recent from the series)		Telephone	4 (Regional)	-
MADAR [13, 14] (25 Arabic city dialects in the travel domain)		Only text	15 (Arabic countries)	-
ADI17	✓	Multimedia (YouTube)	17 (Arabic countries)	3,091h

Table 1: Comparison of existing Multi-Arabic dialect identification speech data

Country (ISO 3166-1 format)		Training		Dev		Test	
alpha-3 code	English short name	Dur	Utt.	Dur	Utt.	Dur	Utt.
DZA	Algeria	115.7h	32,262	0.6h	246	1.9h	745
EGY	Egypt	451.1h	151,052	1.9h	680	2.1h	760
IRQ	Iraq	815.8h	291,123	1.5h	646	1.9h	760
JOR	Jordan	25.9h	5,514	1.7h	422	2.0h	721
SAU	Saudi Arabia	186.1h	69,350	1.2h	393	2.1h	760
KWT	Kuwait	108.2h	32,654	1.2h	450	2.0h	760
LBN	Lebanon	116.8h	38,305	1.3h	409	1.9h	760
LBY	Libya	127.4h	35,692	2.3h	683	2.0h	760
MRT	Mauritania	456.4h	138,706	0.5h	219	1.3h	509
MAR	Morocco	57.8h	18,530	1.1h	397	1.9h	760
OMN	Oman	58.5h	27,188	1.7h	655	1.8h	760
PSE	Palestine, State of	121.4h	39,129	1.4h	456	2.1h	760
QAT	Qatar	62.3h	26,650	2.0h	929	1.7h	760
SDN	Sudan	47.7h	18,883	0.7h	216	2.0h	760
SYR	Syrian Arab Republic	119.5h	47,606	1.3h	470	2.0h	760
ARE	United Arab Emirates	108.4h	49,486	2.2h	1,144	1.8h	760
YEM	Yemen	53.4h	21,139	1.3h	540	1.8h	760
Total		3033.4h	1,043,269	24.9h	8,955	33.1h	12,615

Table 2: ADI17 dataset statistics

20 sec) duration. Also, the test set duration per dialect was balanced, so each dialect has an average of 2 hours of speech.

2.1. Related datasets

We summarize Arabic speech datasets that have dialect labels in Table 1. For the Arabic dialect datasets, most of the large-scale datasets are not publicly available and their channel condition is relatively clean such as telephone calls [15] and broadcast news [16]. Publicly available datasets usually contain less that 15 hours of speech per dialect which is relatively small. The ADI17 dataset not only provides a large-scale dataset but also has 17 dialect labels, an unprecedented number of dialect classes.

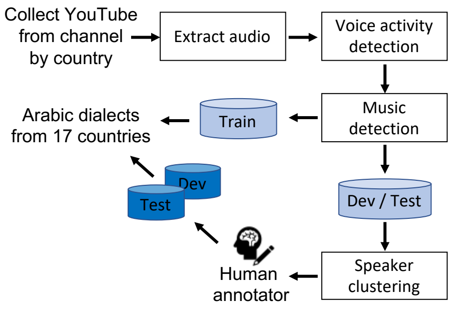


Fig. 1: Arabic dialect speech collection pipeline

2.2. Collection Pipeline

Step 1. Collect Arabic video channels: We compiled an average of 30 YouTube channels per country. The list of YouTube channels was compiled and reviewed by a native speaker from each country. Initially, we asked native speakers from each country to list channels that could be of interest. We tried to diversify the channels across multiple genres per country.

Step 2. Download audio: All the videos related to each channels were downloaded from YouTube. For each channel, we crawled up to 100 hours, if available, to avoid the data being biased to a specific channel or genre. We checked for duplicate videos for each dialect class to prevent the same video from appearing in multiple dialect classes. Initially, we downloaded a total of 7,554 hours, which reduced to 4,248 hours after removing duplicates.

Step 3. VAD and Music detection: To segment out speech from the audio channel of each video, we used WebRTC's Voice Activity Detection (VAD)[2]. We found that there was still much music and other background sound after VAD. To filter out segments that contained music or background sounds, we used Ina's music and speech detection system [17].

Step 4. Divide into Train and Dev/Test sets: We divided speech segments into two partitions, i.e. Train and Dev/Test set. For the Dev/Test set, we randomly picked YouTube IDs to have an average 15 hours for each dialect (Totaling 280 hours and 99,967 utterances). This set was annotated by human dialect experts. The rest of the speech segments are used for the Train set. We did not have any further processing on the Train set, as shown in Table 2.

Step 5. Speaker clustering: Before annotation, we decided to reduce the human annotation effort by using speaker clustering. We assumed that the same person in a video would speak a single dialect. If we can cluster the speakers in each video under this assumption, fewer annotations will be needed for each video because the other segments in same speaker cluster can be regarded as the same dialect. For clustering, we followed a similar approach as speaker diarization. First, we trained a speaker verification system to extract speaker embeddings. The system was trained using Voxceleb 1 and 2 which has a total of 7,205 speakers in the training set. We used the same speaker verification as described in [18] and the system showed 4% EER on the Voxceleb 1 test set using Cosine distance between speaker embeddings. We used this system to extract speaker embeddings from all utterances and used Agglomerate Hierarchical Clustering (AHC) to cluster speakers for each video.

For reliability, we used a conservative number of clusters because over-clustering is not an issue on this task. Over-clustering means that the number of clusters is greater than the expected number of speakers. We used 10 clusters for each video. Then, we gave the first and last segment for each cluster to a human annotator. If the first and last segment in a cluster were spoken in the same dialect from same person, we used all the segments in the cluster. If the dialect was not the same, we discarded all the segments in the cluster. Through this conservative approach, we could increase the reliability against erroneous speaker embeddings.

Step 6. Annotation: Using speaker clustering, we reduced the human annotation effort from 99,967 segments to 11,254 segments, so we saved 90% of the cost for human annotation. Since annotating 17 dialects is a significantly hard task even for native Arabic people, we gave 2 binary tasks such as "EGY dialect or not" and "speech or not" to each annotator. These binary tasks were possible because we already collected the video using channels per country as described in step 1.

[2] https://webrtc.org/

Step 7. Finalize Dev and Test sets For the final step, we divided the Dev/Test set into Dev and Test set for validation and evaluation of the identification system. We divided the set into three subcategories, under 5sec, between 5sec and 20sec and over 20sec to represent short, medium, and long utterances. The number of utterances in all three sub-categories in the test set are balanced across the dialects except Mauritania.

Condition	Training	Validation	Evaluation
Supervised	Train set(labeled)	Dev set(labeled)	Test set (labeled)
Semi-supervised	99% Train set (unlabeled) 1% Train set (labeled)	Dev set (labeled)	Test set (labeled)

Table 3: Evaluation conditions for ADI17 dataset.

3. EVALUATION CONDITIONS

To evaluate the effectiveness of the dialect labels and to promote unsupervised representation learning, we defined two sub-tasks, a *supervised task* and a *semi-supervised task*. For the supervised task, all the labels in the training set are used to train a system. For the semi-supervised task, only 1% of the labels in the training data are available and the rest are regarded as unlabeled data. For the 1% labeled data, we selected 700 utterances per dialect randomly, a total of 35 hours for 17 dialects. This semi-supervised task mimics the real-world challenge of dialect and language recognition in that there are not enough data to build systems, as we mentioned in section 2. The validation and evaluation set is the same for the two sub-tasks, so we can directly compare the performance between the two tasks.

4. EXPERIMENTS

4.1. Baseline systems

Several DID approaches were examined, as described below.

i-vector: We followed the i-vector training approach using Kaldi's recipe (sre08/v1). We used 20 MFCCs and delta and delta-delta as a feature for the Gaussian Mixture Model-Universal Background Model (GMM-UBM) with 2048 mixture components. A Total Variability (TV) matrix was trained to extract 600-dimensional i-vectors. Logistic Regression (LR) was used to calculate a posterior probability for each dialect.

x-vector: We followed the x-vector training approach using Kaldi's recipe (sre16/v2). We used 23 MFCCs as input to a time-delayed Deep Neural Network (DNN). Details of the DNN structure were described in [19]. An x-vector was extracted from the first fully connected layer in the DNN structure. LR was used to calculate a posterior probability for each dialect.

E2E(x-vector): We also have a variant of the x-vector system which operated in an end-to-end manner. We used the same system as the x-vector system and the only difference is how to get the posterior probability. Rather than extract embedding from one of the fully connected layers, we used the softmax layer output as the posterior probability for each dialect.

E2E(Softmax): This system was trained using CNNs with a softmax output. We used 40-dimensional MFCCs as input and used four 1-dimensional Convolution Neural Network (CNN) layers. The filter sizes are 40×5 - 1000×7 - 1000×1 - 1000×1 with 1-2-1-1 strides and the number of filters is 1000-1000-1000-1500. A global statistic pooling layer, that calculates mean and standard deviation of the last CNN layer outputs to produce a fixed output size of 3,000, is used to connect the CNN and Fully-Connected (FC) layers with 1500-600 nodes. Then the FC layer output is fed into a Softmax output layer. Other details are described in [1]. We use the Softmax output as a posterior probability for each dialect. The baseline system code and pre-trained model is publicly available.[3]

E2E(Tuplemax): This system substituted the E2E softmax output layer with Tuplemax [20]. Other settings are the same, except that the learning rate is slightly increased.

E2E(AM-Softmax): This system substituted the E2E softmax layer with Additive Margin Softmax [21]. Other settings are all same except that the learning rate is slightly decreased. The optimal hyperparameters (margin, scale) we found were (0.02, 5) for the supervised task and (0.05, 10) for the semi-supervised task.

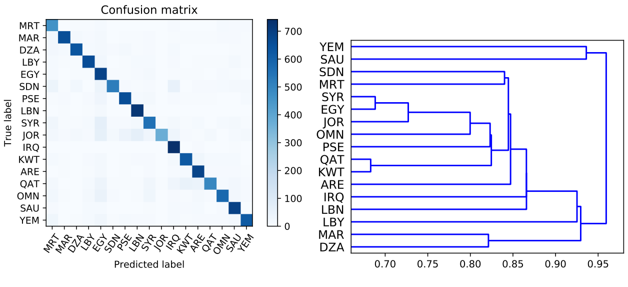
4.2. Evaluation result

To evaluate fine-grained DID, we used overall accuracy and cost. We regard the ADI17 task as a closed-set identification task, so we pick the maximum score among 17 dialects scores for each test utterance to calculate the accuracy. We also used average cost performance C_{avg} for each target/non-target pair defined in NIST LRE 2017 [22] with P_{target} as 0.5.

Table 4 shows the performances of the baseline systems on the ADI17 development and test sets. For supervised conditions, the entire train set is used to estimate the parameters in the systems. For semi-supervised conditions, the i-vector system used 99% of the unlabeled part of the train set to train a GMM-UBM and the TV matrix, then the 1% labeled part was used to estimate the parameters for logistic regression. For the other systems such as x-vector and E2E systems only the 1% labeled part of the train set was used to estimate NN parameters. The other 99% unlabeled part was not used at all. For test set evaluation, we used the parameters that showed the best EER on the dev set. We did not apply any dataset augmentation.

On both conditions, the E2E approaches that use the output layer as a posterior probability for each dialect are better than using latent representation such as i-vector or x-vector systems. For the supervised condition, the E2E(Softmax) system showed the best performance on both Accuracy and C_{avg}. When the labeled data are limited such as for the semi-supervised condition, the i-vector and E2E(Tuplemax) systems showed high efficiency. In particular, the E2E(Tuplemax) system showed significantly low C_{avg} compared to other E2E systems. The i-vector system is still efficient when the utterance is long, which is a common observation for generative models such as i-vectors.

We also applied AM-Softmax [21] which has shown competitive performance on speaker verification tasks. However, we observed that the margin in the angular softmax could not introduce



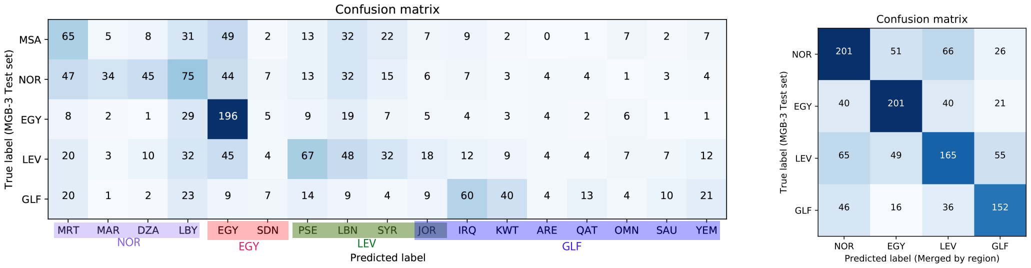
(a) (b)

Fig. 2: (a) Confusion matrix of DID result on E2E(Softmax) system (b) AHC dendrogram using mean of each dialect embeddings.

[3] https://github.com/swshon/arabic-dialect-identification

Conditions	System	Test set										Dev set	
		Overall				<5sec		5sec $\sim$ 20sec		>20sec		Overall	
		Accuracy	Precision	Recall	Cost	Accuracy	Cost	Accuracy	Cost	Accuracy	Cost	Accuracy	Cost
Supervised task	i-vector	60.3	60.7	60.5	29.1	51.7	36.5	64.5	25.8	75.3	15.0	59.7	28.7
	x-vector	72.1	72.1	72.7	20.1	65.7	24.0	75.4	18.3	81.9	13.9	71.0	20.2
	E2E(x-vector)	77.8	77.8	78.7	16.4	72.7	19.8	80.0	14.9	88.6	9.0	76.6	16.0
	E2E(Softmax)	**82.0**	**82.1**	**83.3**	**13.7**	**76.2**	**18.8**	**85.1**	**10.9**	**90.4**	**6.7**	**83.0**	**11.7**
	E2E(Tuplemax)	78.6	78.7	80.9	14.2	71.9	18.8	82.1	11.9	88.7	8.2	78.6	13.9
	E2E(AM-Softmax)	63.7	63.8	62.9	36.1	57.5	40.1	66.5	34.0	75.0	30.5	62.5	36.5
Semi-supervised task	i-vector	47.4	47.4	47.3	40.7	39.3	49.2	50.4	37.0	**67.2**	23.9	46.8	39.4
	x-vector	39.3	39.2	38.7	49.3	32.3	56.4	42.5	45.9	52.4	36.8	41.2	48.0
	E2E(x-vector)	40.5	40.3	40.0	49.7	33.1	58.3	43.6	45.8	56.2	33.5	42.1	48.0
	E2E(Softmax)	48.8	48.6	48.8	48.2	40.5	57.1	52.7	44.3	63.6	30.7	47.5	46.7
	E2E(Tuplemax)	**50.4**	**50.2**	**49.9**	**38.6**	**42.3**	**46.2**	**54.2**	**35.2**	64.7	**23.8**	**48.7**	**37.3**
	E2E(AM-Softmax)	49.8	49.6	48.7	51.0	41.3	55.8	53.5	49.0	66.2	41.1	48.1	50.0

Table 4: Performance evaluation using ADI17 test set. Note that Cost is equal to C$_{avg} * 100$.



	MRT	MAR	DZA	LBY	EGY	SDN	PSE	LBN	SYR	JOR	IRQ	KWT	ARE	QAT	OMN	SAU	YEM
MSA	65	5	8	31	49	2	13	32	22	7	9	2	0	1	7	2	7
NOR	47	34	45	75	44	7	13	32	15	6	7	3	4	4	1	3	4
EGY	8	2	1	29	196	5	9	19	7	5	4	3	4	2	6	1	1
LEV	20	3	10	32	45	4	67	48	32	18	12	9	4	4	7	7	12
GLF	20	1	2	23	9	7	14	9	4	9	60	40	4	13	4	10	21

(a)

	NOR	EGY	LEV	GLF
NOR	201	51	66	26
EGY	40	201	40	21
LEV	65	49	165	55
GLF	46	16	36	152

(b)

Fig. 3: (a) Confusion matrix of MGB-3 test set on adi17 system. (x-axis reflects the maps of the Arabic region from west-to-east to group similar countries together) (b) Confusion matrix of MGB-3 test set on regional class.

an improvement on the DID system. The margin between the classes seems not to be an important factor since the task is closed-set identification and there would be no new class as input.

For the semi-supervised condition, we applied Factorized Hierarchical Variational Autoencoder (FHVAE) for representation learning from the unlabeled part (99%) of the train set. This approach was based on previous work [5]. However, we did not observe that the unsupervised representation learning showed a benefit on the DID task. We suspect that the FHVAE model is too small and only local information is being encoded.

4.3. Discussion

The confusion matrix and the AHC dendrogram as shown in Figure 2. Jordanian (JOR) shows poor performance compared to other dialects, and this can be due to limited training data. The dendrogram shows the closeness of dialects of countries close to each other and with the same geographic region. These indicate that the data collection for the ADI17 is reasonable.

However, the ADI17 dataset has a strong limitation. All speech were collected from the same video sharing platform, so that the channel domain of the training and evaluation set is matched. Moreover, we only considered YouTube ID to partition the train, dev and test sets and eliminate overlap across the sets. For this reason, there might be the same speakers (such as popular actors or broadcasters) appearing in different sets. These issues hinder objective system evaluation and we speculate this is the reason that the 17 DID accuracy is higher than the previous 5 regional DID result [1, 4].

Consequently, we need a more objective dataset to evaluate the performance of a system trained using the ADI17 dataset. Since the country-labeled dialect speech collection from another domain is very difficult, one of the remedies is to use available datasets. For example, the MGB-3 [11] test set[4] was collected from high-quality broadcasting system servers by down-sampling to 16kHz, and can be leveraged for evaluation. As shown in Figure 3(a), we fed the MGB-3 test set in E2E(Softmax) system and generated the confusion matrix. Since the MGB-3 data has only 5 regional classes, we cannot calculate the performance directly. Thus, we merged the 17 Arabic dialect classes into 4 regional classes (excluding Modern Standard Arabic (MSA), which we do not have in the ADI17 system). The result is shown in Figure 3(b). They show 58% accuracy on the MGB-3 test set. When compared to other domain mismatched DID systems which showed 48.79% and 51.27% as reported in [4], the system trained using ADI17 is significantly better even if they do not have same label. In this way, we could have a more objective benchmark to prevent over-fitted performance on the same channel domain.

5. CONCLUSION

LID and DID typically do not receive much attention because they are regarded as a variant of speaker recognition, and subsequently lack publicly available datasets. However, the state-of-the-art approach in speaker recognition such as AM-Softmax turns out not to be suited for LID and DID as we showed in our experiments, and we need another approach for closed-set identification. To promote DID research, we presented a large-scale fine-grained labeled dataset collection pipeline. The metadata of the dataset is publicly available, so we expect follow-up studies on this dataset for both supervised and semi-supervised tasks.

In the future, we plan to provide another benchmark protocol to evaluate the system using the speech data which was collected through a different channel domain e.g. MGB-3 test set. We will also consider additional annotation works for the MGB-3 dataset to obtain country-level dialect labels. Additionally, we plan to add other dialect classes to cover all the Arabic speaking countries in the world.

[4] Available: https://github.com/qcri/dialectID/blob/master/data/test.MGB3/wav.lst

6. REFERENCES

[1] Suwon Shon, Ahmed Ali, and James Glass, "Convolutional neural network and language embeddings for end-to-end dialect recognition," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 98–104.

[2] Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals, "Automatic dialect detection in Arabic broadcast speech," in *Interspeech*, 2016, vol. 08-12-Sept, pp. 2934–2938.

[3] Suwon Shon, Ahmed Ali, and James Glass, "Mit-qcri arabic dialect identification system for the 2017 multi-genre broadcast challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017, pp. 374–380.

[4] Suwon Shon, Ahmed Ali, and James Glass, "Domain attentive fusion for end-to-end dialect identification with unknown target domain," in *IEEE ICASSP*, 2019, pp. 5951–5955.

[5] Suwon Shon, Wei-Ning Hsu, and James Glass, "Unsupervised Representation Learning of Speech for Dialect Identification," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 105–111.

[6] Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass, "A factorial deep markov model for unsupervised disentangled representation learning from speech," in *IEEE ICASSP*, 2019, pp. 6540–6544.

[7] Qian Zhang and John HL Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 873–882, 2018.

[8] Chunlei Zhang, Qian Zhang, and John HL Hansen, "Semi-supervised learning with generative adversarial networks for arabic dialect identification," in *IEEE ICASSP*, 2019, pp. 5986–5990.

[9] Maryam Najafian, Sameer Khurana, Suwon Shon, Ahmed Ali, and James Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *IEEE ICASSP*, Calgary, 2018, pp. 5174–5178.

[10] Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, pp. 1026–1033.

[11] Ahmed Ali, Stephan Vogel, and Steve Renals, "Speech Recognition Challenge in the Wild: ARABIC MGB-3," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017, pp. 316–322.

[12] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, and Jörg Tiedemann, "Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign," *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pp. 1–17, 2018.

[13] Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al., "The madar arabic dialect corpus and lexicon," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[14] Mohammad Salameh, Houda Bouamor, and Nizar Habash, "Fine-grained arabic dialect identification," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1332–1344.

[15] Karen Jones, Stephanie Strassel, Kevin Walker, and David Graff, *Multi-Language Conversational Telephone Speech 2011*, Linguistic Data Consortium, University of Pennsylvania, LDC2019S02.

[16] Kevin Walker, Christopher Caruso, Kazuaki Maeda, Denise DiPersio, and Stephanie Strassel, *GALE Phase 2 Arabic Broadcast Conversation Speech Part 1*, Linguistic Data Consortium, University of Pennsylvania, LDC2013S02.

[17] David Doukhan, Eliott Lechapt, Marc Evrard, and Jean Carrive, "Ina's mirex 2018 music and speech detection system," in *Music Information Retrieval Evaluation eXchange (MIREX 2018)*, 2018.

[18] Suwon Shon, Hao Tang, and James Glass, "Frame-level Speaker Embeddings for Text-independent Speaker Recognition and Analysis of End-to-end Model," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1007–1013.

[19] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, and Yishay Carmiel, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Interspeech*, 2017, pp. 999–1003.

[20] Li Wan, Prashant Sridhar, Yang Yu, Quan Wang, and Ignacio Lopez Moreno, "Tuplemax loss for language identification," in *IEEE ICASSP*, 2019, pp. 5976–5980.

[21] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[22] Seyed Omid Sadjadi, Timothee Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero, "The 2017 nist language recognition evaluation.," in *Odyssey*, 2018, pp. 82–89.