The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech

Citation for published version:
Ali, A, Shon, S, Samih, Y, Mubarak, H, Abdelali, A, Glass, J, Renals, S & Choukri, K 2020, The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech. in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).* Institute of Electrical and Electronics Engineers (IEEE), pp. 1026-1033, IEEE Automatic Speech Recognition and Understanding Workshop 2019, Sentosa, Singapore, 14/12/19. https://doi.org/10.1109/ASRU46091.2019.9003960

Digital Object Identifier (DOI):
10.1109/ASRU46091.2019.9003960

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

THE MGB-5 CHALLENGE: RECOGNITION AND DIALECT IDENTIFICATION OF DIALECTAL ARABIC SPEECH

Ahmed Ali[1], *Suwon Shon*[2], *Younes Samih*[1], *Hamdy Mubarak*[2], *Ahmed Abdelali*[1]
James Glass[2], *Steve Renals*[3], *Khalid Choukri*[4]

[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
[2]Computer Science & Artificial Intelligence Laboratory, Cambridge, MA, USA
[3]Centre for Speech Technology Research, University of Edinburgh, UK
[4]European Language Resources Association, Paris, France

ABSTRACT

This paper describes the fifth edition of the Multi-Genre Broadcast Challenge (MGB-5), an evaluation focused on Arabic speech recognition and dialect identification. MGB-5 extends the previous MGB-3 challenge in two ways: first it focuses on Moroccan Arabic speech recognition; second the granularity of the Arabic dialect identification task is increased from 5 dialect classes to 17, by collecting data from 17 Arabic speaking countries. Both tasks use YouTube recordings to provide a multi-genre multi-dialectal challenge in the wild. Moroccan speech transcription used about 13 hours of transcribed speech data, split across training, development, and test sets, covering 7-genres: comedy, cooking, family/kids, fashion, drama, sports, and science (TEDx). The fine-grained Arabic dialect identification data was collected from known YouTube channels from 17 Arabic countries. 3,000 hours of this data was released for training, and 57 hours for development and testing. The dialect identification data was divided into three sub-categories based on the segment duration: short (under 5 s), medium (5–20 s), and long (>20 s). Overall, 25 teams registered for the challenge, and 9 teams submitted systems for the two tasks. We outline the approaches adopted in each system and summarize the evaluation results.

***Index Terms*—** Speech recognition, broadcast speech, multigenre, under-resource, dialect identification, multi-reference WER

1. INTRODUCTION

The MGB challenge is a series of evaluations of speech recognition, speaker diarization, lightly supervised alignment, and dialect identification using TV recordings from the BBC and Al Jazeera, as well as YouTube videos. The first edition of the MGB challenge (MGB-1) [1] focused recognition, diarization, and alignment of BBC English TV output across four channels. A total of 1,600 hours of broadcast audio and several hundred million words of BBC subtitle text were provided to train speech recognition systems. The second edition of the MGB challenge (MGB-2) [2] emphasised handling the diversity in the Arabic broadcast news domain, using audio data obtained from 19 distinct programmes broadcast on the Al Jazeera Arabic TV channel. A total of 1,200 hours of acoustic training data was released (with lightly supervised transcriptions) along with over 130M words crawled from the Al Jazeera Arabic website `aljazeera.net`. Finally, the third edition of the MGB challenge (MGB-3) [3] focused on dialectal Arabic (DA) using a multi-genre collection of Egyptian YouTube videos. Seven genres were used for the data collection. A total of 16 hours of videos, split evenly across the different genres, were divided into adaptation, development and evaluation data sets. The MGB-3 challenge had three targets: a) dealing with languages which do not have well-defined orthographic systems, Egyptian Arabic in particular; b) Multi-genre scenarios – seven different genres are included in the challenge; and c) low-resource scenarios – only 16 hours of in-domain data was provided.

The MGB-5 challenge is an evaluation of speech recognition and dialect identification techniques using YouTube recordings. The data is highly diverse, spanning the whole range of YouTube genres. Our aim is to encourage researchers to evaluate the latest research techniques using large quantities of realistic data with immediate real-world applications, as well as encouraging the investigation of novel approaches to lightly-supervised, semi-supervised, and unsupervised learning.

The Moroccan Arabic automatic speech recognition (ASR) task in MGB-5 used a data set comprising 13 hours of speech extracted from 93 YouTube videos distributed across seven genres: comedy, cooking, family/children, fashion, drama, sports, and science clips. This amount of data is not enough by itself to build robust speech recognition systems, but could be useful for adaptation, and for hyper-parameter tuning of models built using the MGB-2 data. Therefore, we suggested that the MGB-2 training data was reused in this challenge, with the provided in-domain data considered to be (supervised) adaptation data. In addition to the transcribed 13 hours, the complete videos were also provided, amounting to a total 48 hours data across the 93 programs. This additional untranscribed data can be used for in-domain speech or genre adaptation.

The fine-grained Arabic Dialect Identification (ADI) task involved dialect identification of speech from YouTube across 17 dialects. The previous MGB-3 challenge resulted in studies covering diverse dialect identification topics such as domain adaptation [4, 5], semi-supervised learning [6, 7, 8, 9], and linguistic feature extraction [10, 11]. However MGB-3 was limited to 5 dialects. To extend the task to a finer-grained analysis of dialectal Arabic speech, for MGB-5 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 linking and dialect labeling by human annotators, resulting in 58 hours of speech selected for use as development and test sets.

2. MGB-5 DATA

2.1. Data for Speech Recognition

As discussed above, the 13 h of multi-genre Moroccan Arabic speech data provided for MGB-5 is used for supervised adaptation, development, and testing. Since dialectal Arabic does not have a clearly defined orthography, different people write the same word in slightly different forms. Therefore, instead of developing strict guidelines to ensure a standardized orthography, variations in spelling are allowed. Thus multiple transcriptions were produced, allowing transcribers to write the transcripts as they deemed correct. Each file was segmented and transcribed by four different Moroccan annotators – inter-annotation agreement and transcription differences are discussed in section 3.1, and summarised in 3. The 93 YouTube clips were manually segmented and labelled as speech or non-speech. About 12 minutes from each program was selected for transcription, and the resulting 13 h of speech segments were divided into training, development and test sets (table1).

In addition to the transcribed 13 hours, the complete recordings are also provided, amounting to a total of 48 h across the 93 programs. This data can be used for in-domain speech or genre adaptation. Transcription of the data was shared in both Arabic as well as Buckwalter[1] format.

2.2. Data for Dialect Identification

The MGB-3 data previously used for Arabic dialect identification has been successfully investigated by many researchers. The MGB-3 dataset has several challenges because the training set is comparatively small (53 h) and test set domain is mismatched with the training set. Most significantly, the MGB-3 dataset has only five regional dialect classes which only partially covers the variety of dialectal Arabic. For this reason, for MGB-5 we collected an Arabic dialect identification dataset comprising about 3,000 hours of speech from 17 Arabic countries (ADI17), obtained from YouTube. Since we collected the speech by considering the YouTube channels in a specific country, the dataset will include some labeling errors. The presence of this noisy labeling potentially would benefit from unsupervised learning. When constructing the ADI17 development and test sets, about 280 h speech data was collected from YouTube. After automatic speaker linking and dialect labeling by human annotators, we selected 58 h of this data to use as development and test sets for performance evaluation. The test dataset was divided into three sub-categories based on the segment duration corresponding to short (<5 s), medium (5–20 s), and long (> 20 s). Detailed statistics of the ADI17 dataset are presented in table 2. Since the original videos are subject to copyright, we do not make them available directly. We instead provide the YouTube URLs, timestamps, and annotations.[2]

Genre	Adapt/train	Dev	Test
Comedy	1.4/10	0.2/1	0.4/2
Cooking	1.5/13	0.3/2	0.2/3
Family/Kids	1.7/10	0.3/2	0.1/1
Fashion	1.5/11	0.4/2	0.2/2
Drama	1.4/8	0.2/1	0.3/2
Science	1.4/8	0.3/1	.1/2
Sports	1.3/9	0.2/1	0.6/2
Total transcribed speech segments	10.2/69	1.3/10	1.4/14
*Overall speech segments	32.5/69	8.2/10	7.5/14

Table 1: MGB-5 data distribution across the three classes, duration in hours/number of programs (12 minutes each roughly). * is the duration for the complete recordings including speech and non-speech segments

3. BASELINE SYSTEMS

3.1. Performance measurements

Similar to previous MGB challenges, we provided an open source baseline system for the challenge for both the speech transcription and dialect identification tasks. Word Error Rate (WER) continues to be the most commonly used metric for evaluating ASR. For English broadcast news there is about 3% inter-annotator disagreement [12], hence a single gold reference transcript is adequate for WER estimation. However, for the MGB-5 ASR task there is about 45% inter-annotator disagreement across the four annotators in dev and test (as

[1] Buckwalter is a one-to-one mapping allowing non Arabic speakers to understand Arabic scripts, and it is also left-to-right, making it easy to render on most devices.

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

Country (ISO 3166-1 format)		Training		Dev					Test				
alpha-3 code	English short name	Dur	Utterances	Dur	Utterances				Dur	Utterances			
					Total	<5sec	5sec∼20sec	>20sec		Total	<5sec	5sec∼20sec	>20sec
DZA	Algeria	115.7h	32,262	0.6h	246	86	139	21	1.9h	745	285	400	60
EGY	Egypt	451.1h	151,052	1.9h	680	223	395	62	2.1h	760	300	400	60
IRQ	Iraq	815.8h	291,123	1.5h	646	254	350	42	1.9h	760	300	400	60
JOR	Jordan	25.9h	5,514	1.7h	422	101	230	91	2.0h	721	261	400	60
SAU	Saudi Arabia	186.1h	69,350	1.2h	393	115	235	43	2.1h	760	300	400	60
KWT	Kuwait	108.2h	32,654	1.2h	450	161	247	42	2.0h	760	300	400	60
LBN	Lebanon	116.8h	38,305	1.3h	409	127	220	62	1.9h	760	300	400	60
LBY	Libya	127.4h	35,692	2.3h	683	181	393	109	2.0h	760	300	400	60
MRT	Mauritania	456.4h	138,706	0.5h	219	78	125	16	1.3h	509	194	267	48
MAR	Morocco	57.8h	18,530	1.1h	397	121	235	41	1.9h	760	300	400	60
OMN	Oman	58.5h	27,188	1.7h	655	265	347	43	1.8h	760	300	400	60
PSE	Palestine, State of	121.4h	39,129	1.4h	456	148	244	64	2.1h	760	300	400	60
QAT	Qatar	62.3h	26,650	2.0h	929	398	479	52	1.7h	760	300	400	60
SDN	Sudan	47.7h	18,883	0.7h	216	64	108	44	2.0h	760	300	400	60
SYR	Syrian Arab Republic	119.5h	47,606	1.3h	470	165	264	41	2.0h	760	300	400	60
ARE	United Arab Emirates	108.4h	49,486	2.2h	1,144	536	567	41	1.8h	760	300	400	60
YEM	Yemen	53.4h	21,139	1.3h	540	219	279	42	1.8h	760	300	400	60
Total		3033.4h	1,043,269	24.9h	8,955	3,242	4,857	856	33.1h	12,615	4,940	6,667	1,008

Table 2: ADI17 dataset statistics

shown in table 3), which is much higher than the one observed for the MGB-3 data (table 3 in [3]). When we apply surface normalization[3], the inter-annotator disagreement goes down by 1–2%. We also measured the character error rate (CER) in the inter-annotator disagreement, which is about 17% across the four annotators.

For the MGB-5 Challenge, we continue to consider the multi-reference WER – MR-WER [13]. This metric is based on comparing the recognized text against multiple manual transcriptions of the speech signal, which are all considered valid references. This approach thus accepts a recognized word if any of the references include it in the same form. The code for computing the MR-WER is available on GitHub.[4]

To evaluate fine-grained dialect identification, we used overall accuracy and cost average . We regard the 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 Langauge Recognition Evaluation (LRE) 2017 [14] with P_{target} as 0.5.

3.2. ASR Baseline

The ASR baseline system was trained using the MGB-5 training data, 10.2 hours transcribed by four different annotators, this gives us more than 40 hours in total. This data was augmented by applying speed and volume perturbation [15], increasing the number of training frames by a factor of three to about 120 hours. The code recipe is available on the Kaldi repository[5]. The acoustic modeling is similar to the QCRI submission to the MGB-2 Challenge [16]. The lexicon was grapheme-based, covering 950,000 words[17] collected from a set of shared lexicons, as well as the training data text. The systems used a single-pass decoding with a trigram Language Model (LM), along with a purely sequence trained Time Delay Neural Network (TDNN) acoustic model [18]; i-vector were used for speaker adaptation. We report results for the MGB-5 development set on which we achieve an average WER of 75.1% and MR-WER 57.0%. Results are detailed in table 4; this is a weak baseline compared to the MGB-3 results, owing to the limited training data.

	ref2	ref3	ref4
ref1	44/43/15	49/48/17	48/47/17
ref2	–	47/46/17	47/46/17
ref3	–	–	47/45/17

Table 3: The inter annotator disagreement on the development and test data across the four different human references before and after normalization (in %). Note that the three numbers (in order) are: word-level word error rate / normalized text word-level error rate / character-level error rate.

3.3. ADI Baseline

The baseline system for the ADI task was trained using the ADI17 training set. We used an end-to-end dialect identification system based on a deep neural network Mel-frequency cepstral coefficents (MFCC) input features. This system is based on the system in [10]. We used four 1-dimensional

[3]Surface orthographic normalization for three characters; alef, yah and hah, which are often mistakenly written in dialectal text. This normalization is standard for dialectal Arabic pre-processing and reduces the sparseness in the text.

[4]`https://github.com/qcri/multiRefWER`.

[5]`https://github.com/kaldi-asr/kaldi/tree/master/egs/mgb5`

	WER1	WER2	WER3	WER4	AV-WER	MR-WER
Comedy	72.9	72.0	72.0	73.5	72.6	56.6
Cooking	70.8	69.2	70.2	70.1	70.1	49.3
FamilyKids	73.5	70.4	73.2	71.4	72.1	51.4
Fashion	74.9	73.9	74.8	74.4	74.5	54.4
Drama	66.3	66.9	68.3	67.5	67.3	48.4
Science	74.0	73.7	75.2	76.2	74.8	55.6
Sports	97.1	97.2	97.6	97.0	97.2	95.4
Overall WER	**75.5**	**74.2**	**75.6**	**75.0**	**75.1**	**57.0**

Table 4: Baseline results in % for the development data after applying surface text normalization

Evaluation set	Overall	<5sec	5sec∼20sec	>20sec
Dev	83.0	76.5	85.5	93.7
Test	82.0	76.2	85.1	90.4

(a) Accuracy

Evaluation set	Overall	<5sec	5sec∼20sec	>20sec
Dev	11.7	17.2	9.8	4.6
Test	13.7	18.8	10.9	6.7

(b) Cost (C_avg*100)

Table 5: Baseline performance evaluation for ADI task

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 average pooling layer which averages the last CNN layer outputs to produce a fixed output size of 1,500 is used to connect to CNN and Fully-Connected (FC) layers with 1500-600 nodes. Then the FC layer output is fed into a Softmax output layer and the cross entropy loss is calculated against the ground truth dialect label. The MFCC frames were extracted every 10 ms using a 25 ms window; the CNN layer spans a total 11 frames, a width of 110 ms. For the baseline we did not apply any dataset augmentation method and using the unaugmented original training set.

Table 5 shows the performance of the baseline system on development and test sets. The overall accuracy is around 83% for both sets and it showed much better accuracy compared to the previous MGB-3 task (57.2%) which has only five dialect class. The baseline system code and pre-trained model is publicly available.[6]

4. SUBMISSION RESULTS

4.1. ASR

In the ASR task, participants submitted one primary submission and as many contrast submissions as they wished. We scored and ranked results based on the primary submissions. The test set was manually segmented, and only non-overlapping speech was used for scoring. Over 25 teams registered for this task, and three systems were submitted. Table 6 and 7 summarize the results for the ASR track. In addition the standard WER, we report the multi-reference WER and avergae WER across the multiple manual transcriptions of the speech signal.

RDI & Cairo University : The RDI-CU submission [19] achieved the lowest error rates in the speech-to-text task. Their submission is based on a combination of two acoustic models. They trained a CNN with a factorized TDNN (CNN-TDNN-f), they also trained a TDNN-f acoustic model. The acoustic model was trained using the MGB-2 data to train background models and the MGB-5 training and development data to transfer to the MGB-5 task. They applied data augmentation in three steps: speed and volume perturbation, data reverberation and music-noise-speech injection. This increased the amount of training data by a factor of nine. They combined 100 dimensional i-vector with the 512 dimensional x-vector per frame. They applied dimensionality reduction on the combined vector to a 200-dimensional vector per frame for speaker adaptation. Finally, they used the MGB-5 untranscribed data and applied semi-supervised learning for genre adaption which gave them more than 2% an absolute gain. Their final system benefited from language model interpolation and system combination.

Dialectal Arabic Transcription System (DARTS): The DARTS system [20] was mainly developed to study the MGB-3 task. The authors analyzed the following: transfer learning from high resource broadcast domain to low-resource dialectal domain, and semi-supervised learning where they used in-domain unlabeled audio data collected from YouTube. Key features of their system are: A deep neural network acoustic model that consists of a front end CNN followed by several layers of TDNN Network and LSTM; sequence discriminative training of the acoustic model; n-gram and recurrent neural network language model for decoding and N-best list rescoring. The system was trained on the combined MGB-2 and MGB-5 datasets. They achieved significantly better results on the dev set with respect to the baseline that was trained only on the MGB-5 training data (65% versus 75%).

Zhengzhou Xinda Institute of Advanced Technology (ZXIAT): ZXIAT submitted two systems, one system using a hybrid HMM/DNN, using a TDNN with trigram LM the other using an end-to-end ASR system based on transformer models [21, 22, 23]. They used subword output symbols [24], rather than graphemes or words. The model was first trained with the MGB-2 dataset, obtaining about 22.7% on the MGB-2 development set. Then the MGB-5 dataset is used for fine-tuning, with the encoder fixed or the decoder fixed. In their experiments, they found the system obtains the best performance when the decoder is fixed, however the performance is still worse than the baseline.

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

	Baseline	RDI-CU	DARTS	ZXIAT
Comedy AV-WER	78.0%	74.2%	79.6%	80.3%
Comedy MR-WER	60.0%	59.8%	61.5%	62.9%
Cooking AV-WER	66.8%	52.6%	63.6%	66.9%
Cooking MR-WER	49.4%	32.4%	44.9%	50.0%
FamilyKids AV-WER	68.7%	59.6%	63.2 %	67.7%
FamilyKids MR-WER	48.8%	37.1%	39.8%	48.0%
Fashion AV-WER	60.6%	49.89%	56.6%	60.2%
Fashion MR-WER	42.2%	26.7%	35.9%	42.3%
Drama AV-WER	64.5%	58.3%	64.5%	65.2%
Drama MR-WER	46.1%	37.2%	44.7%	47.1%
Science AV-WER	71.1%	58.5%	62.5%	70.7%
Science MR-WER	55.2%	38.3%	43.4%	54.0%
Sports AV-WER	65.5%	60.0%	56.4%	65.6%
Sports MR-WER	45.6%	38.5%	32.7%	46.4%
MGB5 AV-WER	67.1%	59.4%	62.7%	67.5%
MGB5 MR-WER	48.4%	37.6%	41.8%	49.3%

Table 6: Error rates (AV-WER and MR-WER over four reference transcriptions) per genre for Arabic speech-to-text transcription for the MGB-5 Moroccan Arabic test set.

	MGB5 WER per transcriber				**MGB5**	
	WER1	**WER2**	**WER3**	**WER4**	**AV-WER**	**MR-WER**
RDI-CU	59.1	58.0	60.1	60.1	**59.4**	**37.6**
DARTS	62.3	62.2	62.9	63.6	62.7	41.8
Baseline	66.8	66.9	67.2	67.6	67.1	48.4
ZXIAT	67.3	67.2	67.7	67.8	67.5	49.25

Table 7: Summary of speech-to-text transcription results for the MGB-5 data. WERs are given for each of the four references (produced by different transcribers), as well as AV-WER and MR-WER across the four references.

4.2. ADI

For the ADI task, 15 teams registered and we received a total of 15 submissions from 6 teams. All participants could submit a maximum of 3 systems and only the primary submission was used for the evaluation. Only two teams showed better results than the baseline. Participants used various approaches and we summarized the main features of the submitted systems by their system description in table 9. Most of submission used end-to-end approach to identity the dialect by using the last softmax layer output. Since the task is closed set identification with 17 classes, it seems using softmax output directly has more benefit than extracting an embedding from hidden layer for a scoring module. Below, we briefly summarized the top two teams' approaches.

Duke Kunshan University (DKU) - The DKU system pipeline consists of three main components: dataset augmentation, frame-level feature extraction, and utterance-level modeling. First they performed speed perturbation to increase the diversity and amount of training data. They applied using factors of 0.9, 1.0 and 1.1 as implemented in the Kaldi toolkit. For frame-level feature extraction, they used 64-dimensional mel-filterbank energy (Fbank) vectors, with a frame length of 25ms. Short-time Cepstral Mean Subtraction (CMS) is applied with 3 s sliding window. For the end-to-end network, they use a residual network (ResNet) system with a global statistics pooling layer and a fully connected layer and each output layer is represented as target dialect class [25]. The model was trained with standard cross-entropy loss with a softmax layer. During training the input utterance length was sub-sampled between 200 to 400 frames. They trained 4 types of system by varying the size of dataset and residual block size. Fusion was done for 4 systems. Their best single system achieves accuracy of 94.7% on the development set, as well as the accuracy of 93.8% on the evaluation set. Finally, with score-level fusion, primary systems achieved accuracy of 97.4% on the development set and 94.9% on the test set.

University of Kent (UKent) - The UKent system combines CNN and Long Short Term Memory (LSTM) layers in an end-to-end neural network architecture. They also investigated Time-Scale Modification (TSM) approach to balance for the low-resource dialect (Jordan) in the training set. The

Affiliation name	Test set										Dev set
	Overall				<5sec		5sec∼20sec		>20sec		Overall
	Accuracy	Precision	Recall	Cost	Accuracy	Cost	Accuracy	Cost	Accuracy	Cost	Accuracy
DKU	**94.9**	**94.9**	**94.9**	**4.3**	**93.3**	**5.5**	**95.6**	**3.7**	**97.7**	**2.0**	97.4
UKent	91.1	91.1	91.1	6.2	88.4	8.3	92.3	5.3	96.1	2.5	92.3
Baseline	82.0	82.1	83.3	13.7	76.2	18.8	85.1	10.9	90.4	6.7	83.0
UWB	81.9	82.0	83.3	34.0	76.1	36.5	85.0	32.7	90.7	29.8	-
NUS	81.5	81.7	82.5	18.5	75.2	22.4	84.8	16.4	90.8	12.7	-
IDIAP	67.3	67.5	67.9	28.3	58.3	35.6	71.9	25.1	80.9	13.9	65.1
UCD	42.5	42.4	45.2	52.0	41.4	53.4	42.9	51.2	44.7	50.5	**100.0**

Table 8: Evaluation of submitted systems for ADI task. Note that Cost is equal to $\mathrm{C}_{avg} * 100$. (DKU: Duke Kunshan University, UKent: University of Kent, UWB: University of Western Bohemia, IDIAP: Idiap Research Institute, UCD: University of Chouaib Doukkali)

	Baseline	DKU	UKent	UWB	IDIAP
Acoustic feature	MFCC	Filterbank	MFCC	-	MFCC
DNN structure	CNN	CNN (ResNet)	CNN+ LSTM	-	TDNN
Scoring	softmax output	softmax output	softmax output	Various	PLDA
Score normalization	-	-	-	-	z-norm
Augmentation	-	perturbing speed	time-scaling	-	-
Fusion	-	score-level	-	score-level	-
Label usage	Train	Train+dev	Train	Train	Train

Table 9: Main Features in the submitted systems for ADI task

TSM generates several version of time-stretched and time-compressed utterance. And they also investigated dataset augmentation using noise and Room Impulse Response (RIR) convolutions. Finally, the primary system archived 93.5% on the development set and 93.1% accuracy on the test set.

5. DISCUSSION

In this section, we discuss some limitation of the MGB-5 task. The main limitation that we had a tight schedule. We released Train/Dev set on April 25, 2019, Test set on June 10, 2019, and end of the evaluation was June 24, 2019. Thus, All the participants have only 2 months for the development period. Considering the tasks cover topics such as low-resource problem and semi-unsupervised learning, participants could not have enough time to explore new approaches on this new dataset. For this reason, many of participants used state-of-the-art approaches which were already verified in the previous studies. After the evaluation was done, the dataset is now publicly available, so we expect active investigation from many researchers.

The other limitation is that the channel and speaker of the training and test set could be matched. Since we collected the data from YouTube, they would share a somewhat similar channel domain which is a comparably easier task than the domain mismatched case. Furthermore, although we divided the training, dev and test set considering that the unique YouTube ids are not to be overlapped across the set, some of the speakers could appear in a different setting such as popular new presenters or actors. We speculate this as a major reason to have very high performance than expected on the ADI task by over-fitted on the training dataset domain. In the future, we will collect the dialect speech from another domain which is not matched with YouTube such as broadcasting system or telephone, so that we expect more objective evaluation of the task.

6. CONCLUSION

The MGB-5 Arabic Challenge continued our efforts to evaluate speech recognition systems for diverse broadcast media, using fixed training sets. This year's challenge is an extension of the previous MGB-3 challenge in two aspects: A) Studying Moroccan Arabic which is very difficult Arabic dialect, which is even challenging in the orthographic rules, where we reported about on average more 45% inter-annotation disagreement, the best system in the speech-to-text track achieved 59% average WER and 38% multi-reference WER; B) Increasing the granularity of Arabic dialect identification from 5 classes to 17 by collecting data from 17 Arabic speaking countries. By using YouTube channels, we could collect more than 3,000 hours for Arabic dialect. Compare to the previous MGB-3 ADI task which has only 5 regional dialect class, the overall accuracy has been greatly improved. The main reason is that the domain is matched between trained and test set. We also speculate the fine-grained label helps to learn dialects although it inherently have a noise in the label on train set. We plan to continue the challenge by adding more dialects and potentially collect more YouTube recording to explore transfer learning using a large pool of in-domain un-transcribed speech data.

7. REFERENCES

[1] Peter Bell, Mark JF Gales, Thomas Hain, Jonathan Kilgour, Pierre Lanchantin, Xunying Liu, Andrew McParland, Steve Renals, Oscar Saz, Mirjam Wester, and Philip Woodland, "The MGB challenge: Evaluating multi-genre broadcast media recognition," in *ASRU*, 2015.

[2] Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang, "The MGB-2 Challenge: Arabic multi-dialect broadcast media recognition," in *SLT*, 2016.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Samuel Thomas, Masayuki Suzuki, Yinghui Huang, Gakuto Kurata, Zoltan Tuske, George Saon, Brian Kingsbury, Michael Picheny, Tom Dibert, Alice Kaiser-Schatzlein, et al., "English broadcast news speech recognition by humans and machines," in *IEEE ICASSP*, 2019, pp. 6455–6459.

[13] Ahmed Ali, Walid Magdy, Peter Bell, and Steve Renals, "Multi-reference WER for evaluating ASR for languages with no orthographic rules," in *ASRU*, 2015.

[14] 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.

[15] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition.," in *Interspeech*, 2015.

[16] Sameer Khurana and Ahmed Ali, "QCRI Advanced Transcription System (QATS) for the Arabic multi-dialect broadcast media recognition: MGB-2 Challenge," in *SLT*, 2016.

[17] Ahmed Ali, *Multi-dialect Arabic broadcast speech recognition*, Ph.D. thesis, The University of Edinburgh, 2018.

[18] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," *Interspeech*, 2016.

[19] Hany Ahmed, Hazem Mamdouh, Salah Ashraf, Ali Ramadan, and Mohsen Rashwan, "RDI system for the 2019 Arabic multi-genre broadcast challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.

[20] Sameer Khurana, Ahmed Ali, and James Glass, "DARTS: Dialectal Arabic transcription system," in *ARXIV*, 2019.

[21] Linhao Dong, Shuang Xu, and Bo Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[22] Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," *arXiv preprint arXiv:1804.10752*, 2018.

[23] Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu, "A comparison of modeling units in sequence-to-sequence speech recognition with the transformer on mandarin chinese," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 210–220.

[24] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.

[25] Weicheng Cai, Zexin Cai, Wenbo Liu, Xiaoqi Wang, and Ming Li, "Insights in-to-end learning scheme for language identification," in *IEEE ICASSP*. IEEE, 2018, pp. 5209–5213.