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

Citation for published version:

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)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
THE MGB-5 CHALLENGE:
RECOGNITION AND DIALECT IDENTIFICATION OF DIALECTAL ARABIC SPEECH

Ahmed Ali\textsuperscript{1}, Suwon Shon\textsuperscript{2}, Younes Samih\textsuperscript{1}, Hamdy Mubarak\textsuperscript{2}, Ahmed Abdelali\textsuperscript{1}
James Glass\textsuperscript{2}, Steve Renals\textsuperscript{3}, Khalid Choukri\textsuperscript{4}

\textsuperscript{1}Qatar Computing Research Institute, HBKU, Doha, Qatar
\textsuperscript{2}Computer Science & Artificial Intelligence Laboratory, Cambridge, MA, USA
\textsuperscript{3}Centre for Speech Technology Research, University of Edinburgh, UK
\textsuperscript{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) 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) 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) 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 stud-
ies covering diverse dialect identification topics such as do-
main 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 Ara-
bic 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 standardzed
orthography, variations in spelling are allowed. Thus mul-
tiple transcriptions were produced, allowing transcribers to
write the transcripts as they deemed correct. Each file was
segmented and transcribed by four different Moroccan an-
otators – inter-annotation agreement and transcription dif-
ferences 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 pro-
gram was selected for transcription, and the resulting 13 h of
speech segments were divided into training, development and
test sets (table 1).

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 iden-
tification has been successfully investigated by many re-
searchers. The MGB-3 dataset has several challenges because
the training set is comparatively small (53 h) and test set do-
main 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 construct-
ing 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 per-
formance 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].

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 in-
cluding speech and non-speech segments.

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

3. BASELINES 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 ref-
ence 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

1Buckwalter 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.

2http://groups.csail.mit.edu/sls/downloads/adi17/
shown in Table 3, which is much higher than the one observed for the MGB-3 data (Table 3 in [13]). When we apply surface normalization 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.

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 Language 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 QCRF 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.

### 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 coefficients (MFCC) input features. This system is based on the system in [10]. We used four 1-dimensional

#### Table 2: ADI17 dataset statistics

<table>
<thead>
<tr>
<th>Country (ISO 3166-1 format)</th>
<th>Training</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>alpha-3 code</strong></td>
<td>Dur Utterances</td>
<td>Dur Utterances</td>
<td>Dur Utterances</td>
</tr>
<tr>
<td>DZA</td>
<td>Algeria</td>
<td>115.7h</td>
<td>32,262</td>
</tr>
<tr>
<td>EGY</td>
<td>Egypt</td>
<td>451.1h</td>
<td>151,052</td>
</tr>
<tr>
<td>IRQ</td>
<td>Iraq</td>
<td>815.8h</td>
<td>291,123</td>
</tr>
<tr>
<td>JOR</td>
<td>Jordan</td>
<td>25.9h</td>
<td>5,514</td>
</tr>
<tr>
<td>SAU</td>
<td>Saudi Arabia</td>
<td>186.1h</td>
<td>69,350</td>
</tr>
<tr>
<td>KWT</td>
<td>Kuwait</td>
<td>108.2h</td>
<td>32,654</td>
</tr>
<tr>
<td>LBN</td>
<td>Lebanon</td>
<td>116.8h</td>
<td>38,305</td>
</tr>
<tr>
<td>LBY</td>
<td>Libya</td>
<td>127.4h</td>
<td>35,692</td>
</tr>
<tr>
<td>MRT</td>
<td>Mauritania</td>
<td>456.4h</td>
<td>138,706</td>
</tr>
<tr>
<td>MAR</td>
<td>Morocco</td>
<td>57.8h</td>
<td>18,530</td>
</tr>
<tr>
<td>OMN</td>
<td>Oman</td>
<td>58.5h</td>
<td>27,188</td>
</tr>
<tr>
<td>PSE</td>
<td>Palestine, State of</td>
<td>121.4h</td>
<td>39,129</td>
</tr>
<tr>
<td>QAT</td>
<td>Qatar</td>
<td>62.3h</td>
<td>26,650</td>
</tr>
<tr>
<td>SDN</td>
<td>Sudan</td>
<td>47.7h</td>
<td>18,883</td>
</tr>
<tr>
<td>SYR</td>
<td>Syrian Arab Republic</td>
<td>119.5h</td>
<td>47,606</td>
</tr>
<tr>
<td>ARE</td>
<td>United Arab Emirates</td>
<td>108.4h</td>
<td>49,486</td>
</tr>
<tr>
<td>YEM</td>
<td>Yemen</td>
<td>53.4h</td>
<td>21,139</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3,033.4h</strong></td>
<td><strong>1,043,269</strong></td>
<td><strong>24.9h</strong></td>
</tr>
</tbody>
</table>

#### 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.

<table>
<thead>
<tr>
<th>ref</th>
<th>ref2</th>
<th>ref3</th>
<th>ref4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ref1</td>
<td>44/43/15</td>
<td>49/48/17</td>
<td>48/47/17</td>
</tr>
<tr>
<td>ref2</td>
<td>–</td>
<td>47/46/17</td>
<td>47/46/17</td>
</tr>
<tr>
<td>ref3</td>
<td>–</td>
<td>–</td>
<td>47/45/17</td>
</tr>
</tbody>
</table>

---

3Surface 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.

https://github.com/qcri/multiRefWER

5https://github.com/kaldi-asr/kaldi/tree/master/egs/mgb5
Table 4: Baseline results in % for the development data after applying surface text normalization

<table>
<thead>
<tr>
<th></th>
<th>WER1</th>
<th>WER2</th>
<th>WER3</th>
<th>WER4</th>
<th>AV-WER</th>
<th>MR-WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comedy</td>
<td>72.9</td>
<td>72.0</td>
<td>72.0</td>
<td>71.5</td>
<td>72.6</td>
<td>56.6</td>
</tr>
<tr>
<td>Cooking</td>
<td>70.8</td>
<td>69.2</td>
<td>70.2</td>
<td>70.1</td>
<td>70.1</td>
<td>49.3</td>
</tr>
<tr>
<td>FamilyKids</td>
<td>73.5</td>
<td>70.4</td>
<td>73.2</td>
<td>71.4</td>
<td>72.1</td>
<td>51.4</td>
</tr>
<tr>
<td>Fashion</td>
<td>74.9</td>
<td>73.9</td>
<td>74.8</td>
<td>74.4</td>
<td>74.5</td>
<td>54.4</td>
</tr>
<tr>
<td>Drama</td>
<td>66.3</td>
<td>66.9</td>
<td>68.3</td>
<td>67.5</td>
<td>67.3</td>
<td>48.4</td>
</tr>
<tr>
<td>Science</td>
<td>74.0</td>
<td>73.7</td>
<td>75.2</td>
<td>76.2</td>
<td>74.8</td>
<td>55.6</td>
</tr>
<tr>
<td>Sports</td>
<td>97.1</td>
<td>97.2</td>
<td>97.6</td>
<td>97.0</td>
<td>97.2</td>
<td>95.4</td>
</tr>
<tr>
<td>Overall</td>
<td>75.5</td>
<td>74.2</td>
<td>75.6</td>
<td>75.0</td>
<td>75.1</td>
<td>57.0</td>
</tr>
</tbody>
</table>

Table 5: Baseline performance evaluation for ADI task

<table>
<thead>
<tr>
<th>Evaluation set</th>
<th>Overall</th>
<th>&lt;5sec</th>
<th>5sec~20sec</th>
<th>&gt;20sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>83.0</td>
<td>76.5</td>
<td>85.5</td>
<td>93.7</td>
</tr>
<tr>
<td>Test</td>
<td>82.0</td>
<td>76.2</td>
<td>85.1</td>
<td>90.4</td>
</tr>
</tbody>
</table>

(a) Accuracy

<table>
<thead>
<tr>
<th>Evaluation set</th>
<th>Overall</th>
<th>&lt;5sec</th>
<th>5sec~20sec</th>
<th>&gt;20sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>11.7</td>
<td>17.2</td>
<td>9.8</td>
<td>4.6</td>
</tr>
<tr>
<td>Test</td>
<td>13.7</td>
<td>18.8</td>
<td>10.9</td>
<td>6.7</td>
</tr>
</tbody>
</table>

(b) Cost (C_avg*100)

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 (ZX-IAT): 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 [23], 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.

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 average 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.

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

---

Table 6 and 7 summarize the results for the ASR track. In addition the standard WER, we report the multi-reference WER and average 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 (ZX-IAT): 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 [23], 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.

---

https://github.com/swshon/arabic-dialect-identification
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.

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

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.

<table>
<thead>
<tr>
<th>MGB5 WER per transcriber</th>
<th>MGB5</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER1</td>
<td>59.1</td>
</tr>
<tr>
<td>WER2</td>
<td>58.0</td>
</tr>
<tr>
<td>WER3</td>
<td>60.1</td>
</tr>
<tr>
<td>WER4</td>
<td>60.1</td>
</tr>
</tbody>
</table>

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

Table 9: Main Features in the submitted systems for ADI 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


