The Fifth Edition of the Multi-Genre Broadcast Challenge: MGB-5

(www.mgb-challenge.org)

Suwon Shon, Ahmed Ali, Younes Samih, Ahmed Abdel Ali,
Hamdy Mubarak, James Glass, Steve Renals, Peter Bell, Khalid Choukri

MIT CSAIL, Cambridge, MA, USA
Qatar Computing Research Institute, Doha, Qatar
Centre for Speech Technology Research, University of Edinburgh, UK
European Language Resources Association, Paris, France

Organizers

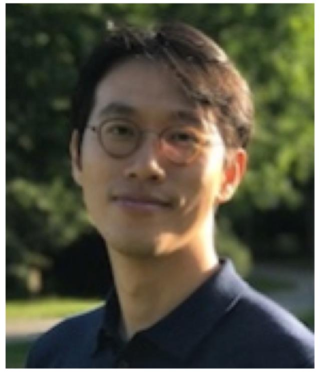
Ahmed Ali

Younes Samih

Ahmed Abdel Ali

Hamdy Mubarak

QCRI



Suwon Shon

James Glass

MIT CSAIL

Steve Renals

Peter Bell

Univ. of Edinburgh

Khalid Choukly

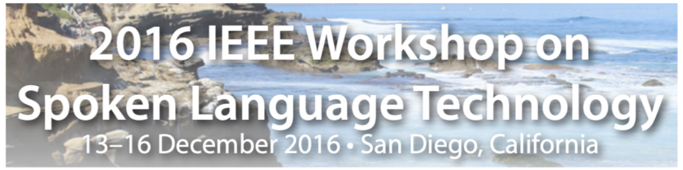
ELRA

History of the MGB challenges

MGB-1

- English Speech
- Recorded from BBC
- 1,600h, 8 Genre
- Subtasks
  - ASR
  - Alignment(word level)
  - Speaker diarization

ASRU 2015



SLT 2016

MGB-2

- Arabic Speech
- Recorded from Al Jazeera TV
- 3,000h, News Genre
- Subtasks
  - ASR
  - Alignment

MGB-3

- Dialectal Arabic Speech
- YouTube, Jazeera TV
- 1,200 h for ASR, 7 Genre
- 70h for dialect ID
- Subtasks
  - ASR (Egyptian dialect)
  - Dialect ID (5 classes)

ASRU 2017

ASRU 2019

ASRU 2019

MGB-5

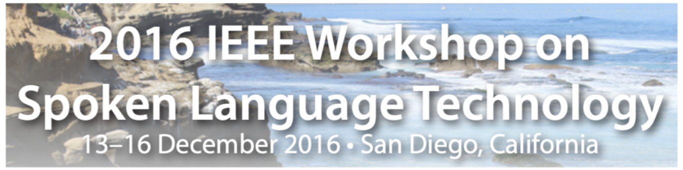
- Moroccan ASR
- Arabic Dialect ID

History of the MGB challenges

MGB-1 **(Transcribed)**
- English Speech
- Recorded from BBC
- 1,600h, 8 Genre
- Subtasks
  - ASR
  - Alignment(word level)
  - Speaker diarization

ASRU 2015

SLT 2016



MGB-2 **(Light Alignment)**
- Arabic Speech
- Recorded from Al Jazeera TV
- 3,000h, News Genre
- Subtasks
  - ASR (1,200h)
  - Alignment

MGB-3 **(Supervised Youtube)**
- Dialectal Arabic Speech
- YouTube, Jazeera TV
- 1,200 h for ASR, 7 Genre
- 70h for dialect ID
- Subtasks
  - ASR (Egyptian dialect)
  - Dialect ID (5 classes)

ASRU 2017

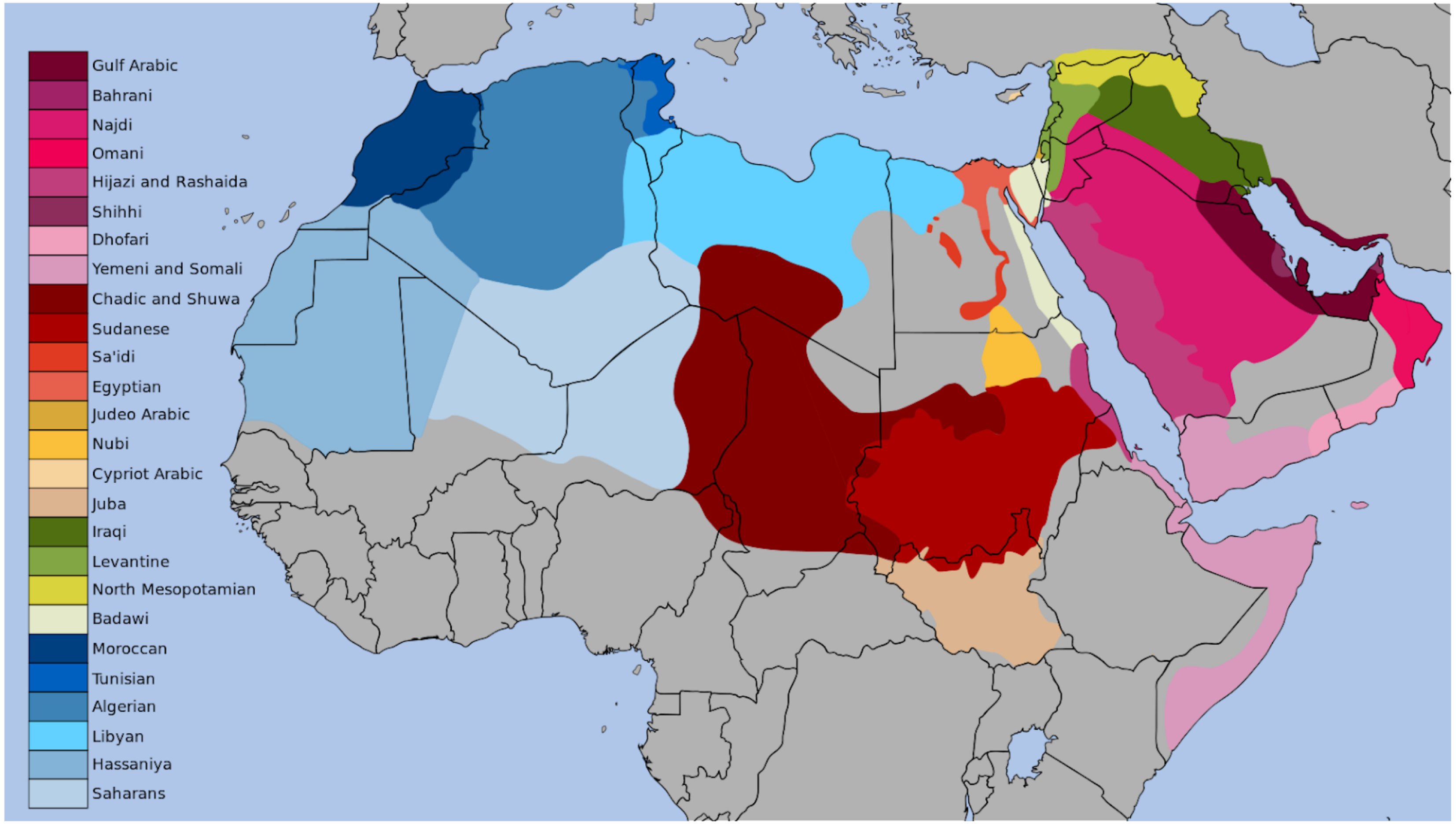
ASRU 2019

ASRU 2019

MGB-5 **(Weakly supervised YouTube)**
- Moroccan ASR
- Arabic Dialect ID

Motivation

- **Variety of Arabic Languages**



26 Dialects from 22 Arabic-speaking countries

Motivation

- **Available Arabic dialect speech corpus**

Name	Free	Channel	Dialect labels	Duration
MGB-3	✓	Broadcast News	5 (Regional)	74h
VarDial2018 (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 (25 Arabic city dialects in the travel domain)		Only text	15 (Arabic countries)	-

Lack of fine-grained labeled data

Motivation

- Previous datasets has 5 regional dialect class

MSA	EGY	LAV	GLF	NOR
Modern Standard Arabic	Egyptian dialect	Levantine dialect	Gulf dialect	North African dialect
	• Egyptian • Sudan	• Lebanon • Syria • Palestine • Jordan	• Iraq • Kuwait • UAE • Qatar • Oman • Saudi • Yemen	• Morocco • Algeria • Libya • Mauritania

MGB-3

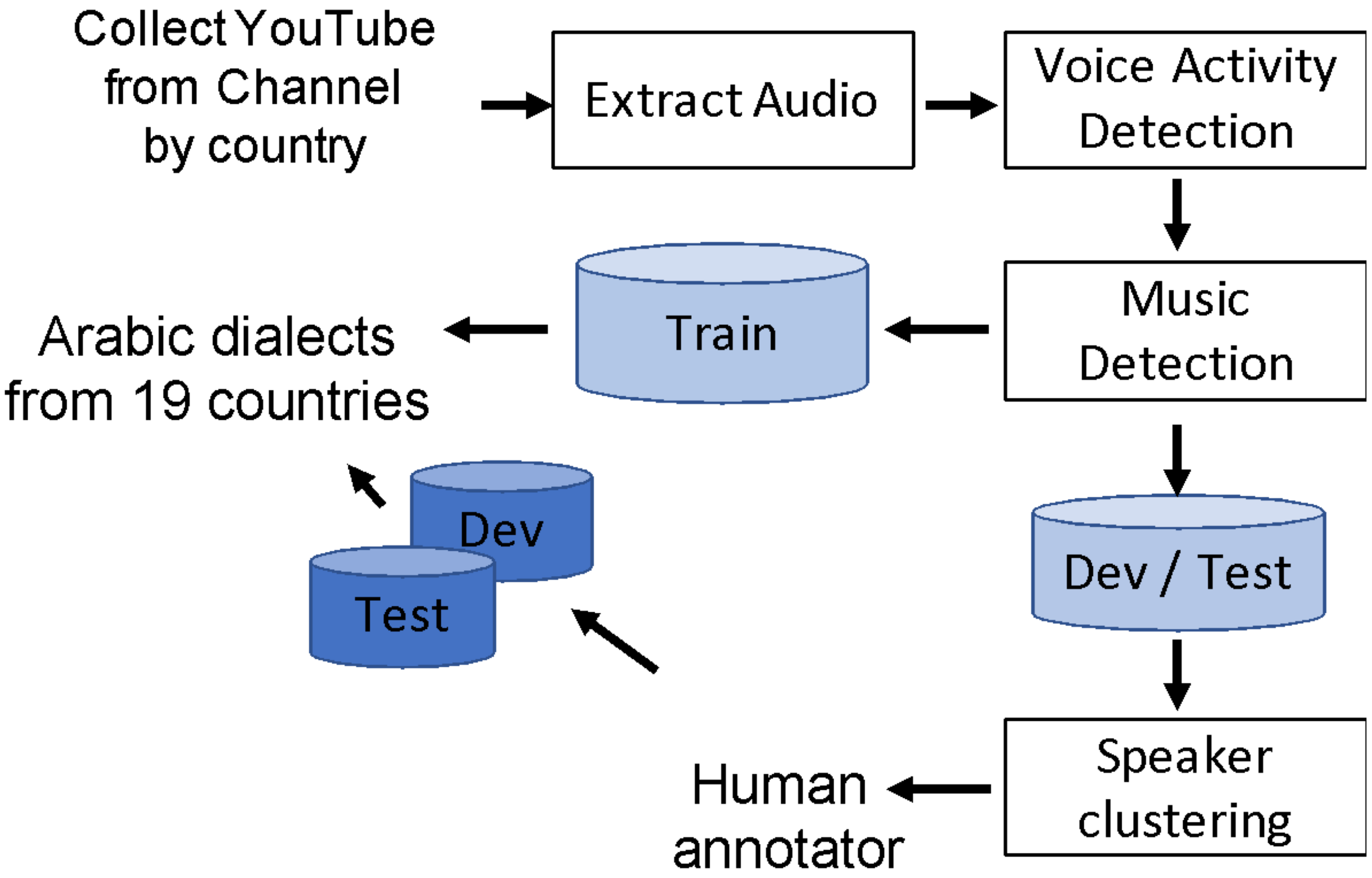
MGB-5

-> Not enough to cover Arab world

Collecting YouTube Speech

- **This year, we focused on speech “in the wild” : YouTube audio**
  - Highly diverse, spanning the whole range of genre
  - Easy to collect dialectal speech
  - Easy to download by anyone without sharing original file

How did we collect dataset?



Step 1: Channel collection

- **Compiled an average of 30 YouTube channels per country**
- **The list was reviewed by a native speaker from each country**
- **Tried to diversify the channels across multiple genres per country**
- **We can get the low-quality, noisy label to help annotator,**
  -> because labeling dialect is *difficult*.

Egypt
YouTube ID “a”
YouTube ID “b”
YouTube ID “c”
…

Qatar
YouTube ID “d”
YouTube ID “e”
…

Step 2: Extract Audio

- **Download:** extract audio in 16kHz
- **Voice Activity Detection*:** to remove non-speech
- **Music detection**:** to remove music segment

Egypt
YouTube ID “a”
YouTube ID “b”
YouTube ID “c”
…
Segment 1
Segment 2
Segment 3
…

*** Google WebRTC Voice Activity Detector**

****David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. "An open-source speaker gender detection framework for monitoring gender equality." IEEE ICASSP, pp. 5214-5218. 2018.**

Step 3: Divide into Train / Eval set

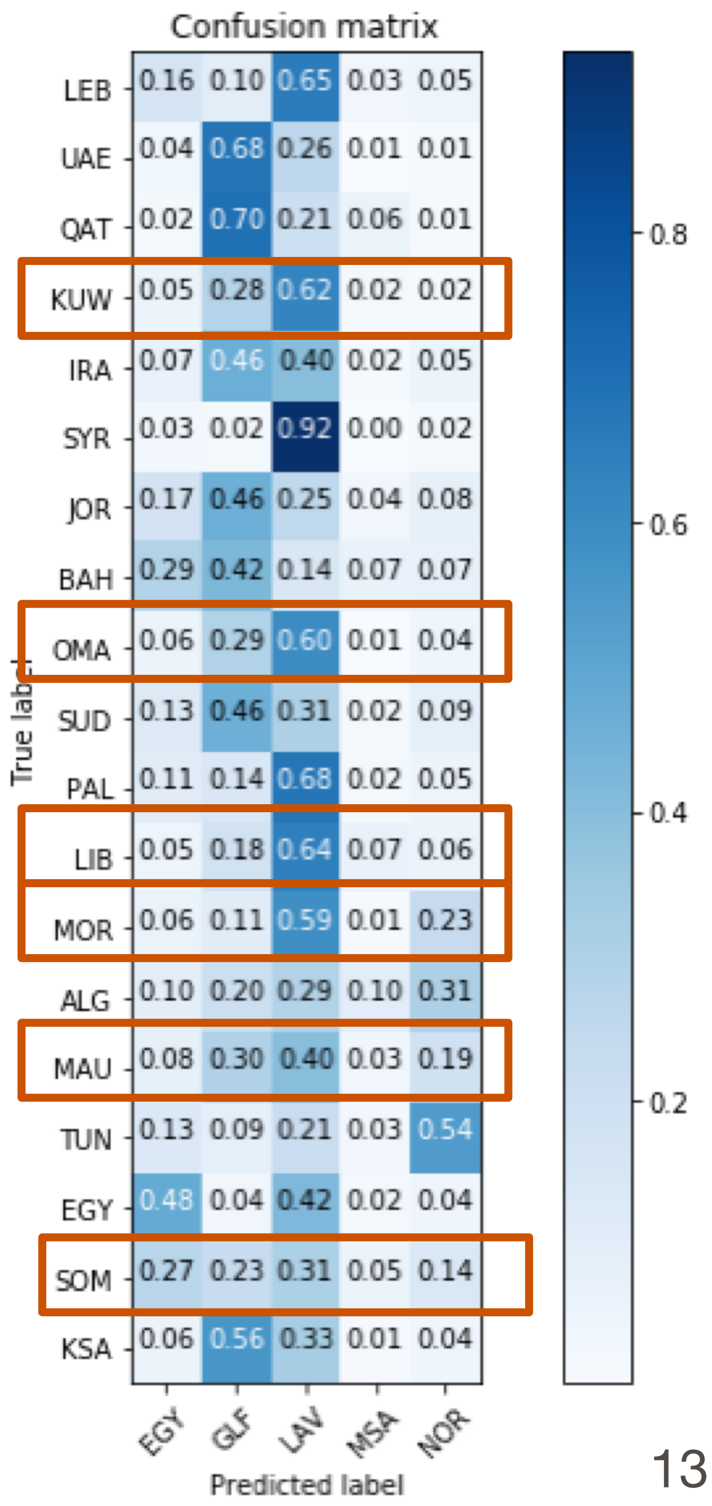
- We randomly picked YouTube IDs to have an average 15 hours for each dialect

Egypt
YouTube ID “a”
YouTube ID “b”
Train set
YouTube ID “c”
Eval set
…

Step 4: Dataset Pre-validation

- **MGB-3 system to validation***
  - identified 20 dialect into 5 regional class

- **Misclassification on few dialects**
  - MGB-3 dataset cannot cover entire dialects in each regional class
  - Channel mismatch

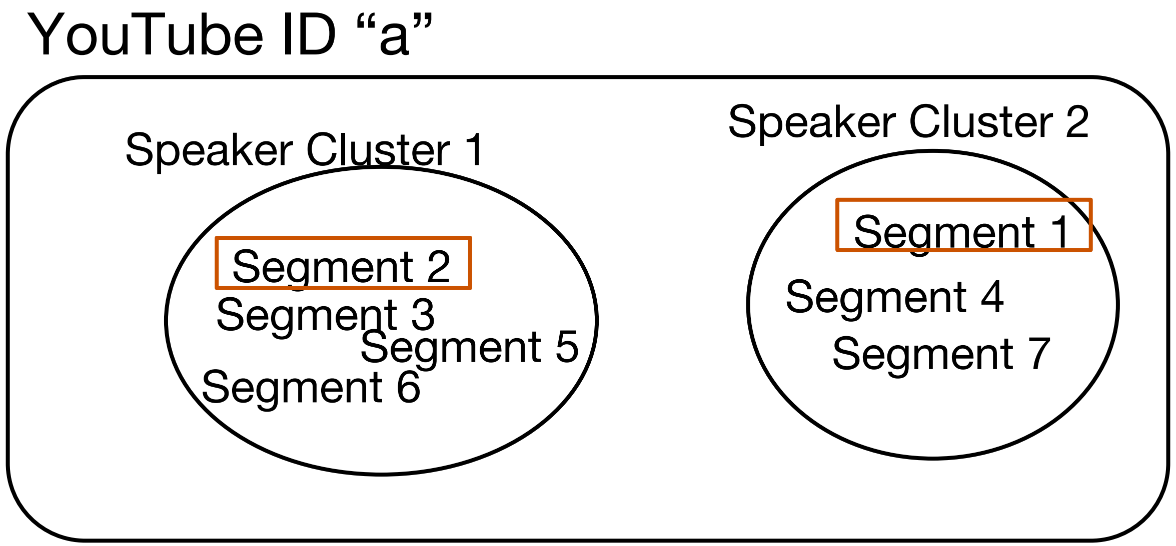


*Suwon Shon, Ahmed Ali, and James Glass. "Convolutional Neural Network and Language Embeddings for End-to-End Dialect Recognition." In Proc. Odyssey: The Speaker and Language Recognition Workshop, pp. 98-104. 2018.

~~Step 5: Annotation by Human~~

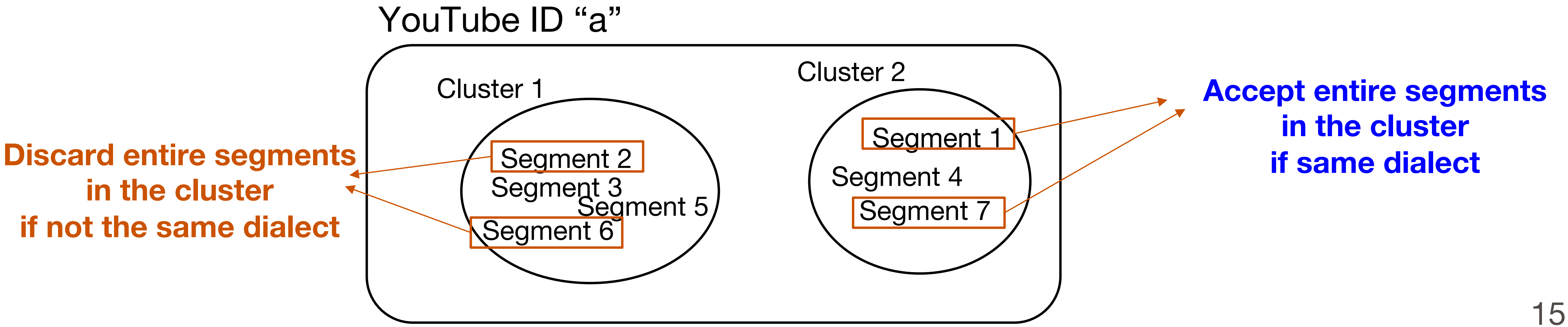
Step 5: Speaker Clustering

- For cost efficiency
- Assumption: same speaker speaks same dialect
- Similar to speaker diarization



Step 6: Annotation by Human

- **Gave two binary task**
  - Speech? or not
  - **IF** speech, target dialect? or not
- **First/last segments of each clusters are labeled**
- **Avoid 17 dialect classification task**



Label noise

- 3 dialects was discarded based on the annotation result
- Average 75% is properly labeled

	Dialect (%)	Other (%)
Palestine	91	9
Lebanon	85	15
Qatar	85	15
Egyptian	85	15
Iraq	83	17
Saudi	82	18
Libya	79	21
Oman	78	22
Kuwait	77	23
Syria	77	23
Jordan	75	25
UAE	73	27
Moroccan	66	34
Mauritania	63	37
Yemen	63	37
Algeria	57	43
Sudan	54	46
Tunisia	44	56
Bahrain	32	68
Somalia	-	-

17 dialects survived

discard

Step 7: Final dataset

- **Total 17 Arabic dialects**
  - Discarded 3 dialects based on the annotation result
- **Divide annotated data into Dev / Test set**
- **Balancing Test set**
  - Duration per dialects
  - Number of utterances in Sub-categories per dialects
    - **Short (<5 s)**
    - **Mid (5s~20s)**
    - **Long (> 20s)**

Challenge plan

- Release Train and Dev set : April 25, 2019
- Release Test set : June 10, 2019
- Submission Deadline : June 24, 2019

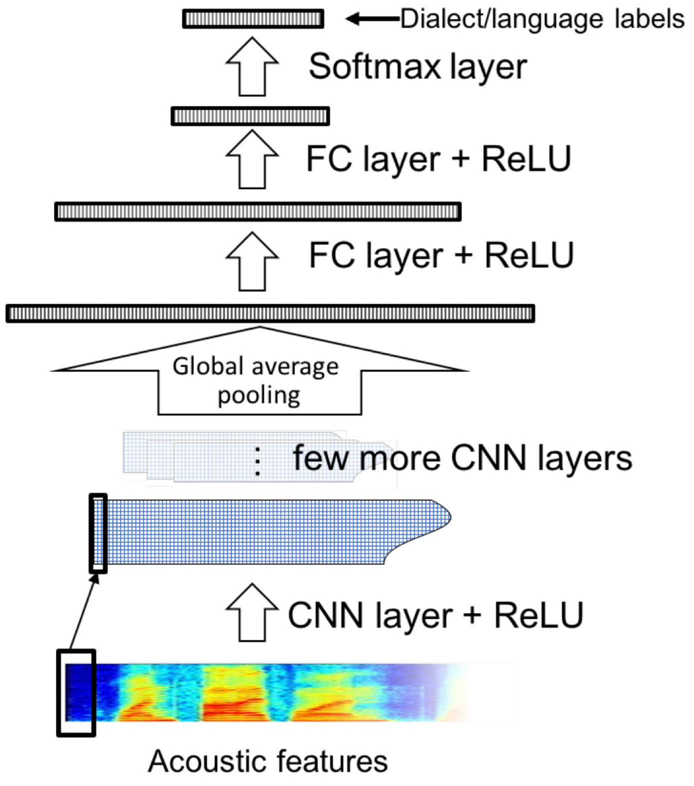
Dataset for ADI task

- **Arabic Dialect Identification for 17 countries (ADI17) Dataset**

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

ADI Baseline : E2E Dialect ID

- **CNN structure***



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 ($C_{avg} * 100$)

*Suwon Shon, Ahmed Ali, and James Glass. "Convolutional Neural Network and Language Embeddings for End-to-End Dialect Recognition." In Proc. Odyssey: The Speaker and Language Recognition Workshop, pp. 98-104. 2018.

ADI Result

- **Total 15 teams registered, 15 submissions from 6 teams**

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**

<Result of primary submission on ADI task>

* ***DKU (Duke Kunshan University) - Weicheng Cai, Haiwei Wu, Ming Li***
** ***UKent (The University of Kent) - Xiaoxiao Miao, Ian McLoughlin***

Dialectal ASR: Moroccan

- **93 YouTube videos distributed**
- **12 minutes from each program selected for transcription**
- **7 genres collected from YouTube**
  - Comedy
  - Cooking
  - Family/children
  - Fashion
  - Drama
  - Sports
  - Science (TEDx)
- **NO strict guidelines to ensure a standardized orthography.**

14h genre labeled and transcribed	48 hours genre labeled with no transcription (in-domain and genre adaptation)

Dataset for ASR

- **Moroccan ASR**

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

Dataset for ASR

- **Inter annotator disagreement**

	Ref2	Ref3	Ref4
Ref1	44	49	48
Ref2	--	47	47
Ref3	--	--	47

The inter annotator disagreement:
word-level word error rate

Dataset for ASR

- **Inter annotator disagreement**

	Ref2	Ref3	Ref4
Ref1	44/43	49/48	48/47
Ref2	--	47/46	47/46
Ref3	--	--	47/45

The inter annotator disagreement:
word-level word error rate
normalized text word-level error rate*

*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.

Dataset for ASR

- **Inter annotator disagreement**

	Ref2	Ref3	Ref4
Ref1	44/43/15	49/48/17	48/47/17
Ref2	--	47/46/16	47/46/17
Ref3	--	--	47/45/17

The inter annotator disagreement:
word-level word error rate
normalized text word-level error rate*
character-level error rate

*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.

ASR Baseline

Train TDNN system using transcribed data

1. Augment the data by four-multiple transcriptions (*4)
2. Speed and volume perturbation (*4*3)= 170h
3. Evaluate WER[1-4], average WER and multi-reference WER

	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**

ASR Result

	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

Result of ASR task

1. **RDI-CU:**
   a. Combine i-vector and x-vector for speaker adaptation
   b. Apply semi-supervised genre adaption
2. **DARTS:**
   a. Mix MGB-2 with MGB-5 data and train single system
3. **ZXIAT:**
   a. Train end-to-end transformer based model

Limitations of the MGB-5 challenge

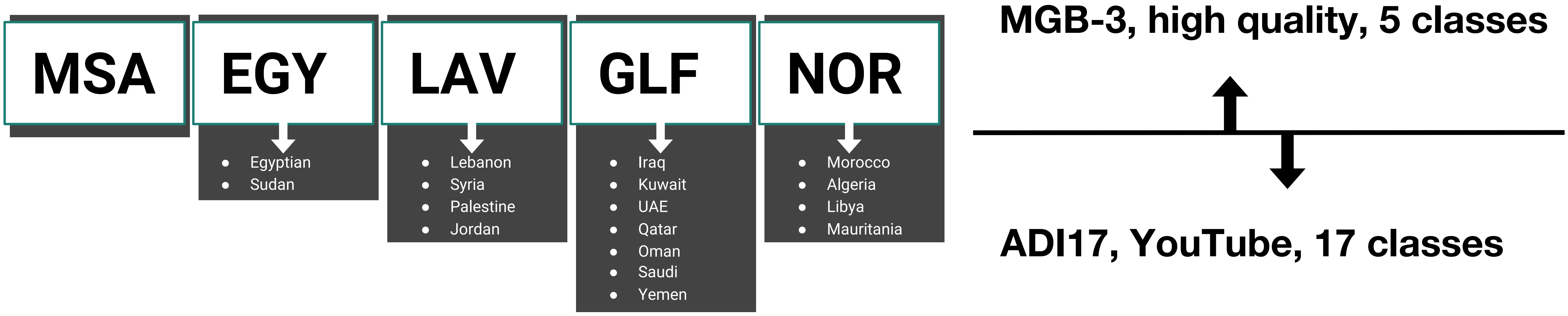
- **Too tight schedule (only 1.5 months are given)**

- **Dividing set only considering YouTube id**
  - Same speaker could appear across the sets
  - Same broadcast program could appear across the sets
  - Duplicated content might exist

- **Channel domain of the train and test was matched**
  - Very high accuracy by over-fitted system

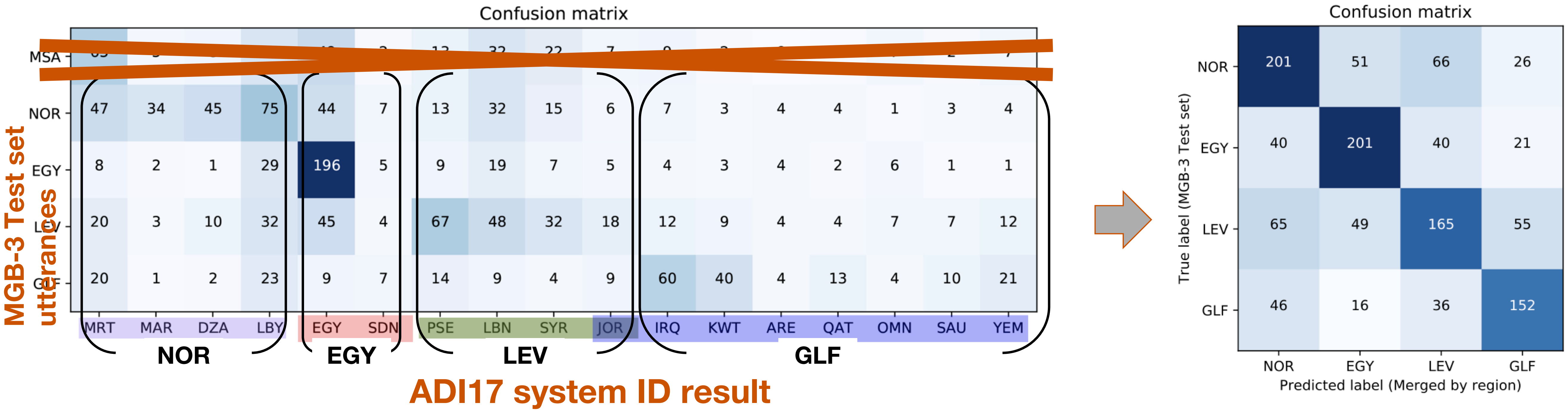
Further analysis

- **More objective evaluation protocol**
  - Train using ADI17, test on MGB-3
    - **Mismatched channel to prevent overfitted system**
  - Classes are mismatched
    - **Use hierarchical relationship**



Further analysis

➢ MGB-3 Test(high-quality) on ADI17(YouTube) system



Accuracy = **58%**

Previous result*
Train with matched dataset (5 class, 63h, high-quality) : **65%**
Train with mismatched data (5 class, 1,000h, YouTube) : **51%**

*** Suwon Shon,, Ahmed Ali, and James Glass. "Domain Attentive Fusion for End-to-end Dialect Identification with Unknown Target Domain." In *IEEE ICASSP*, pp. 5951-5955, 2019.**

Ongoing and Future Work

- **Further investigation on the new evaluation**
  - Use MGB-3 Test set for more objective evaluation
    - Annotate MGB-3 test set into country-level dialect
  - To explore
    - Channel mismatch problem
    - Effective use of noisy labeled train set
- **Supplement on Dataset**
  - ASR

14h genre labeled and transcribed	48 hours genre labeled with no transcription (in-domain and genre adaptation)	+	10,000 hours unsupervised
↓ **100h**	↓ **1,000h**		

  - ADI
    - * **Annotate the MGB-3 to map country level information**
    - * **Cover the 22 Arab countries**
    - * **Reach 1,000 hours per country using distant supervision**

ArabicSpeech

- **Website:** `https://arabicspeech.org/`

- **Call for Posters**
  - ArabicSpeech 2020 Meeting: April 20,21 QCRI, Qatar

- **Focus:**
  - Dialectal speech processing: Arabic as an example

- **Contacts:**
  - You can also email us at: info@arabicspeech.org

Thank you

- Challenge Website: www.mgb-challenge.org
- ADI17 dataset (just type "adi17" on google)
  groups.csail.mit.edu/sls/downloads/adi17
  - Baseline
    github.com/swshon/arabic-dialect-identification
- Moroccan ASR
  - Kaldi/egs/mgb5 github.com/kaldi-asr/kaldi/tree/master/egs/mgb5
  - Kaldi/egs/mgb2_arabic is also available
  - http://www.islrn.org/resources/938-639-614-524-5/