ADI17: A Fine-Grained Arabic Dialect Identification Dataset

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

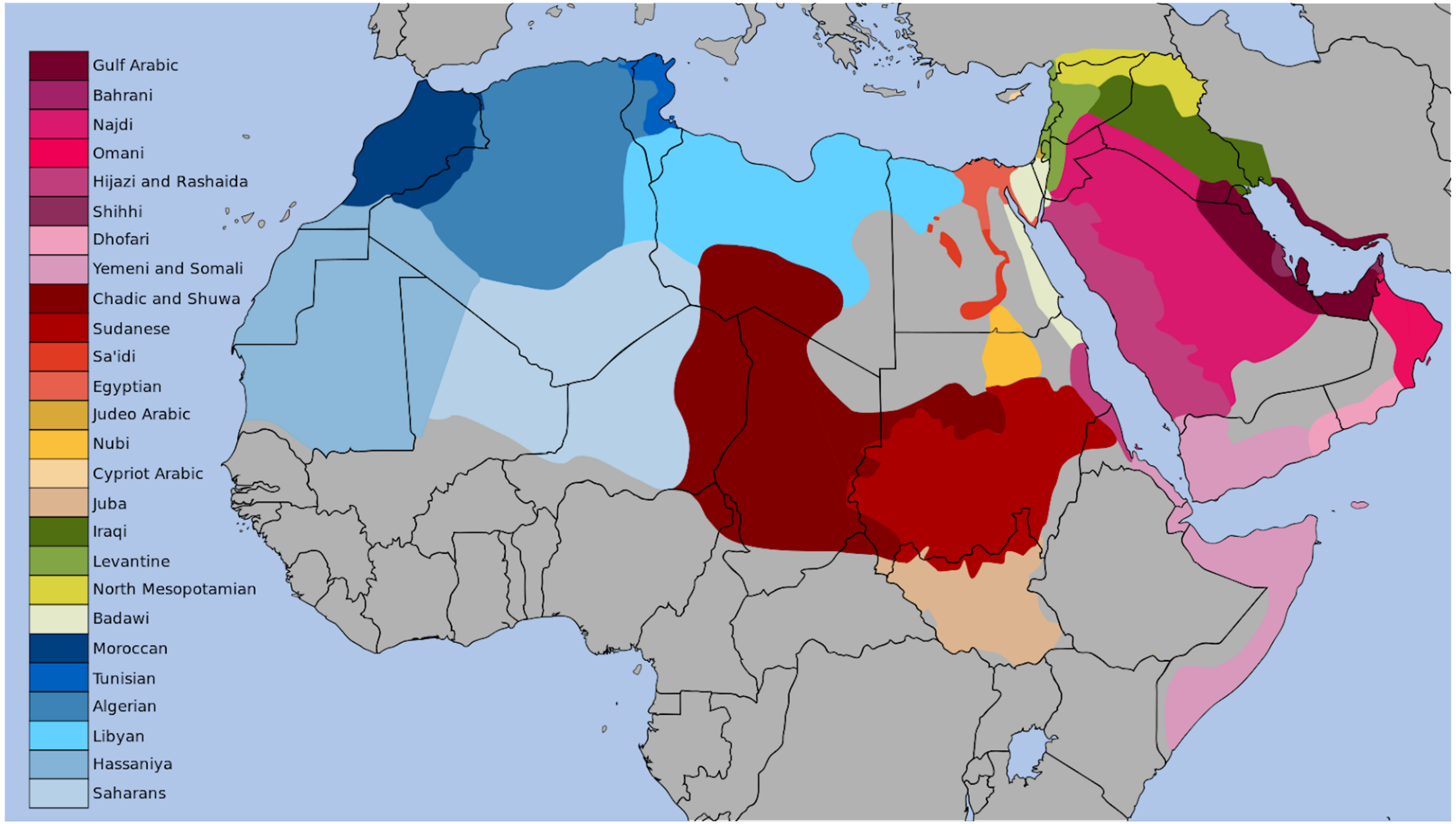
ASAPP Inc, New York, NY, USA[1]
Qatar Computing Research Institute, Doha, Qatar[2]
MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA[3]

**Work done at MIT CSAIL*

Session: HLT-P5: Multilingual Processing of Language
Location: Poster Area A

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	V	Broadcast News	5 (Regional)	74h
VarDial2018 (only test set is available)	V	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)	-
ADI17	V	Multimedia (YouTube)	17 (Arabic countries)	3,091h

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

Lack of fine-grained labeled data

Motivation

- **Previous datasets has 5 regional dialect class**

MSA
Modern Standard Arabic

EGY
Egyptian dialect
- Egyptian
- Sudan

LAV
Levantine dialect
- Lebanon
- Syria
- Palestine
- Jordan

GLF
Gulf dialect
- Iraq
- Kuwait
- UAE
- Qatar
- Oman
- Saudi
- Yemen

NOR
North African dialect
- Morocco
- Algeria
- Libya
- Mauritania

MGB-3

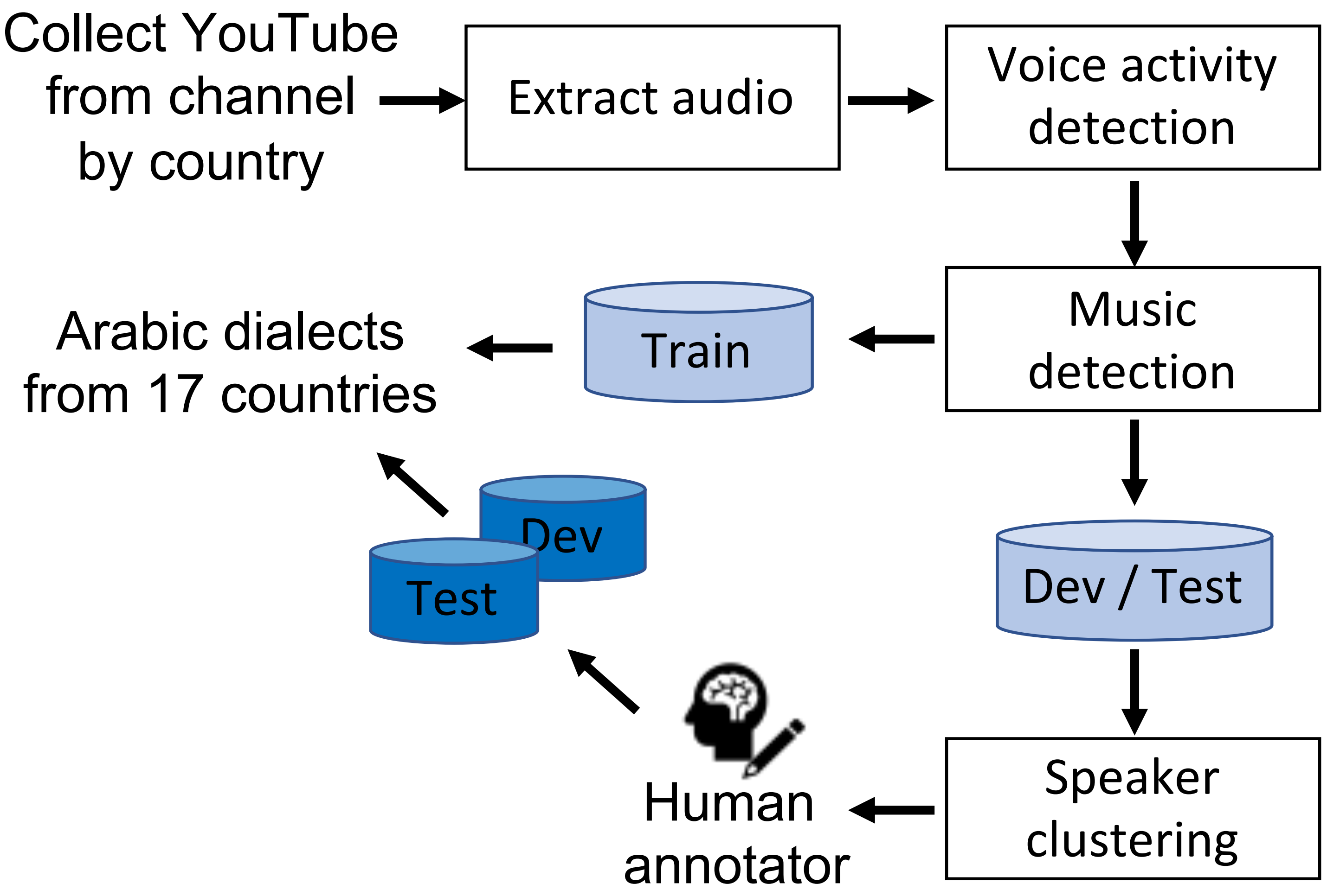
ADI17

-> 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”
  - Segment 1
  - Segment 2
  - Segment 3
  - …
- YouTube ID “b”
- YouTube ID “c”
- …

*** 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**

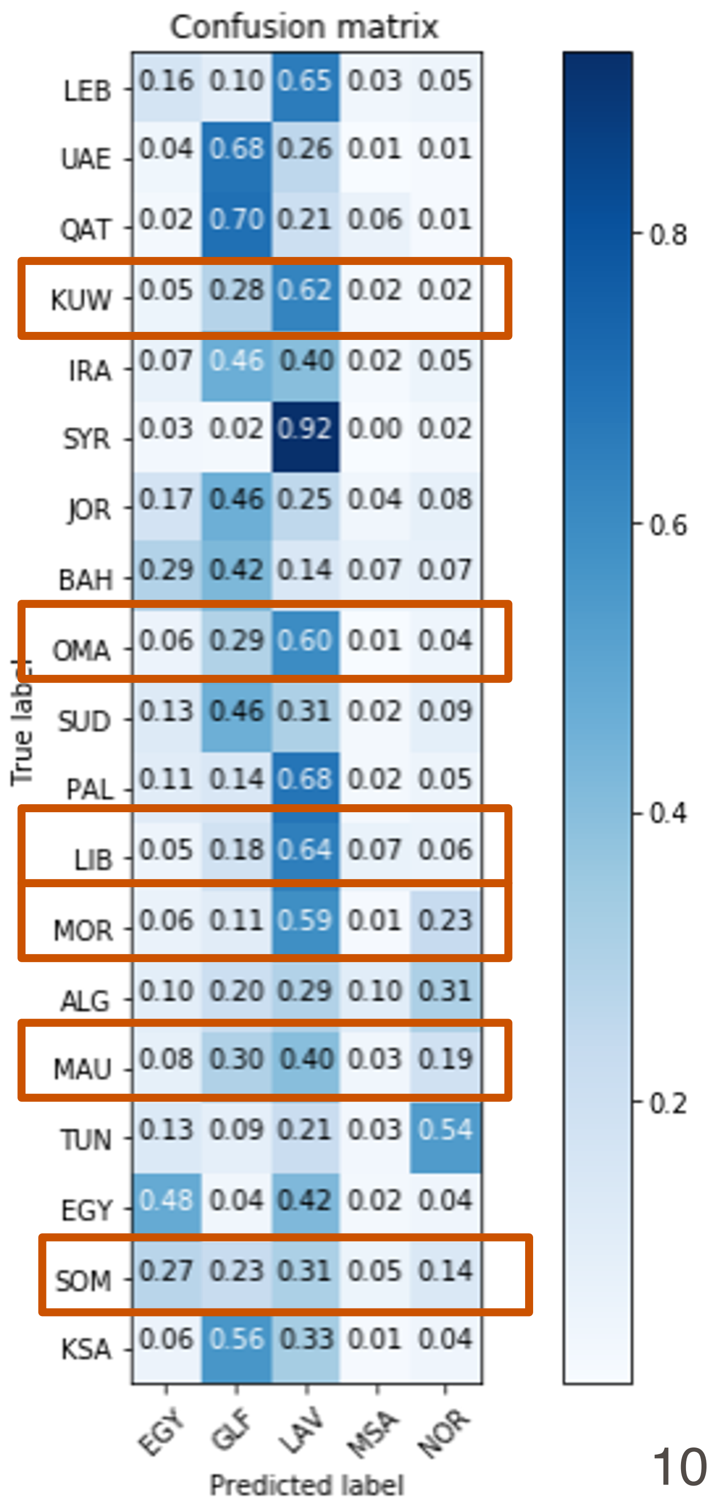
Egypt
YouTube ID “a”
YouTube ID “b”
YouTube ID “c”
…

Train set
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

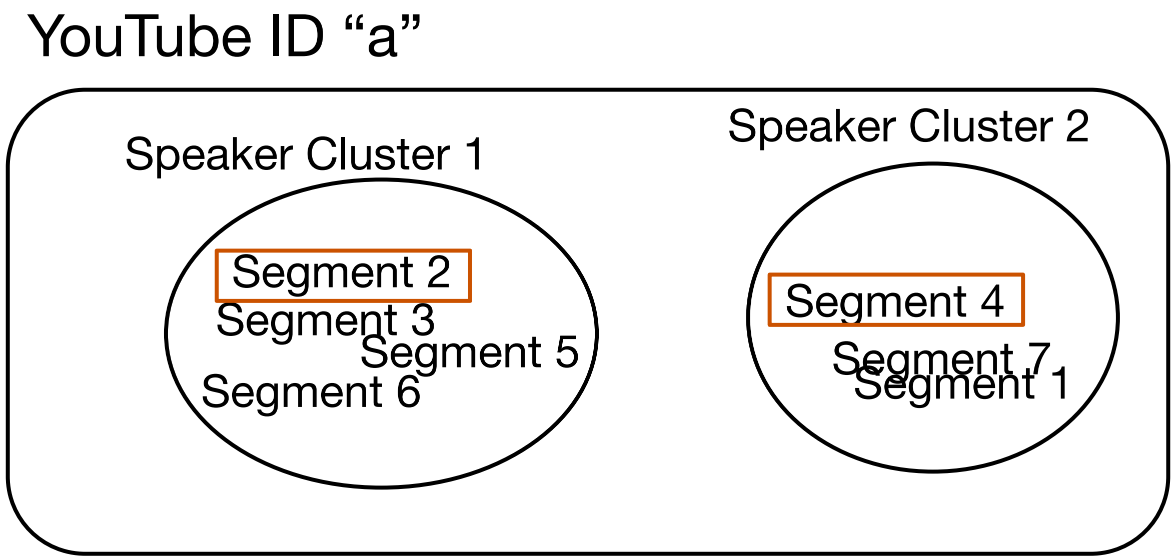


*Suwon Shon, Ahmed Ali, and James Glass. "Convolutional Neural Network and Language Embeddings for End-to-End Dialect Recognition." In Proc. Odyssey: The Speaker and Language Recognition Workshop, pp. 98-104. 2018.

~~Step 5: Annotation by Human~~

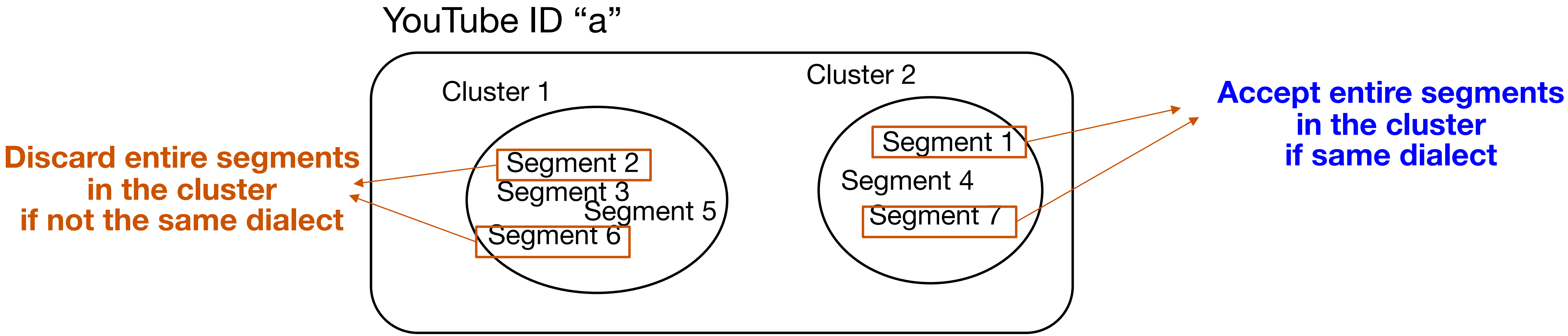
Step 5: Speaker Clustering

- **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)**

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		$\boxed{3033.4h}$	1,043,269	$\boxed{24.9h}$	8,955	3,242	4,857	856	$\boxed{33.1h}$	12,615	4,940	6,667	1,008

ADI 17 Baseline

- i-vector
- X-vector
- E2E(x-vector)
- E2E(softmax)*
- E2E(Tuplemax)
- E2E(AM-Softmax)

*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 17 Evaluation conditions

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 1: Evaluation conditions for ADI17 dataset.

ADI 17 evaluation result

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 $C_{avg} * 100$.

C_{avg} : defined in NIST LRE 2017 with P_{target}=0.5

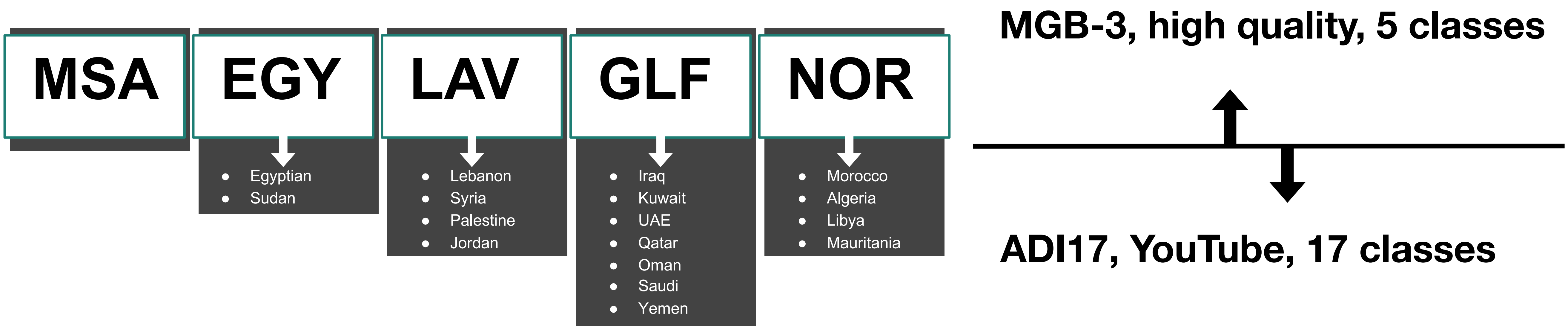
Limitations of the ADI-17

- ➢ **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

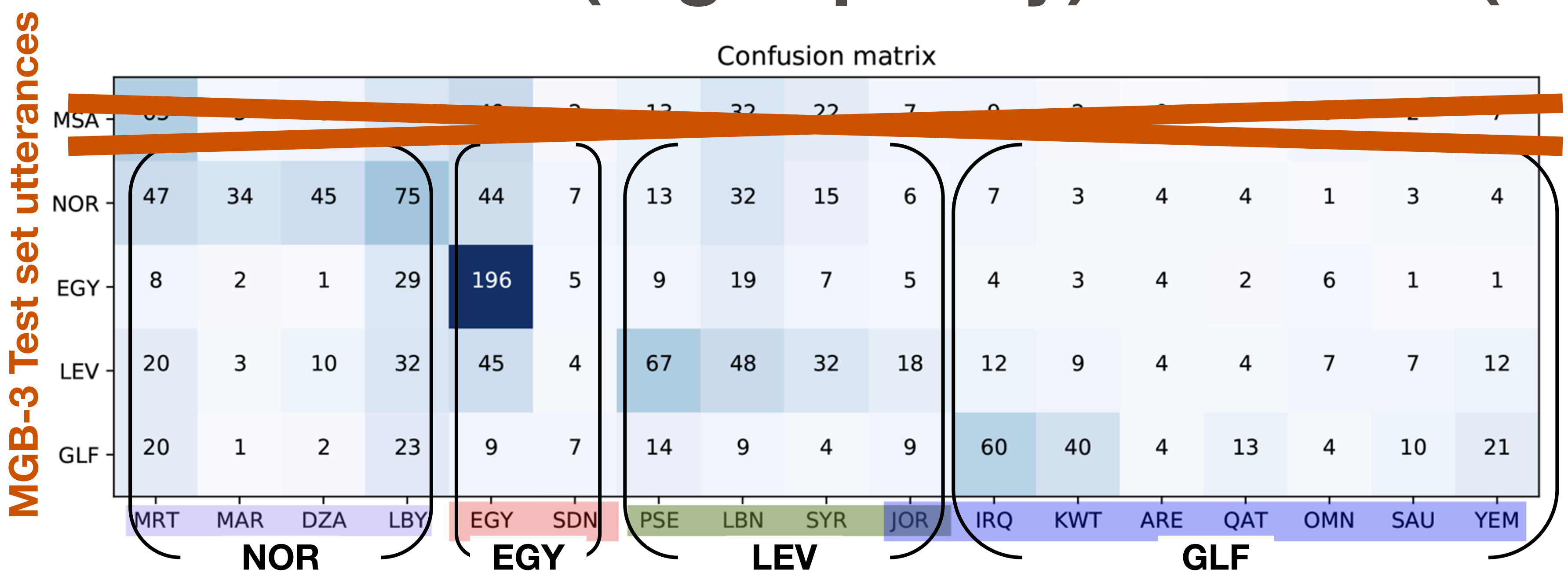
Further analysis

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

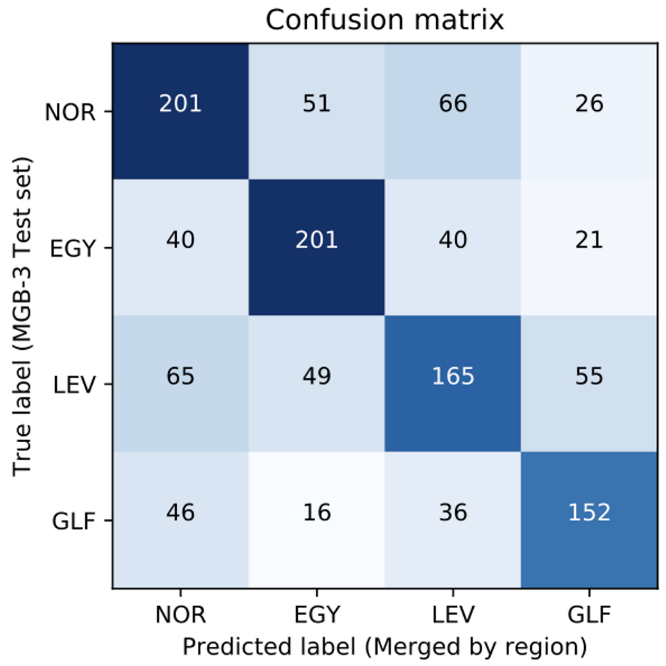


Further analysis

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



ADI17 system ID result



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
    - What information is learned on the network
    - Channel mismatch problem
    - Effective use of noisy labeled train set
- **Supplement on Dataset**
  - **Annotate the MGB-3 to map country level information**
  - **Cover the 22 Arab countries**
  - **Reach 1,000 hours per country**

Thank you

Email: swshon@csail.mit.edu, amali@hbku.edu.qa

ADI17 dataset

- **Download :** `https://goups.csail.mit.edu/sls/downloads/adi17`
- **Github :** `https://github.com/swshon/arabic-dialect-identification`
- **Arabic speech website:** `https://arabicspeech.org/`
- **MGB-challenge infomation :** `https://mgb-challenge.org/`