Exploiting convolution neural networks for phonotactic based dialect identification

Maryam Nafian[1], Sameer Khurana[1], Suwon Shon[1], Ahmed Ali[2], James Glass[1]

MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), Cambridge, MA, USA[1]

Qatar Computing Research Institute, HBKU, Doha, Qatar[2]

Introduction

- One of the challenges of processing real-world spoken content, such as media broadcasts, is the potential presence of different dialects of a language in the material.
- Dialect identification can be a useful capability to identify which dialect is being spoken during a recording.
- Classify additional phone level statistics to model dialect variability

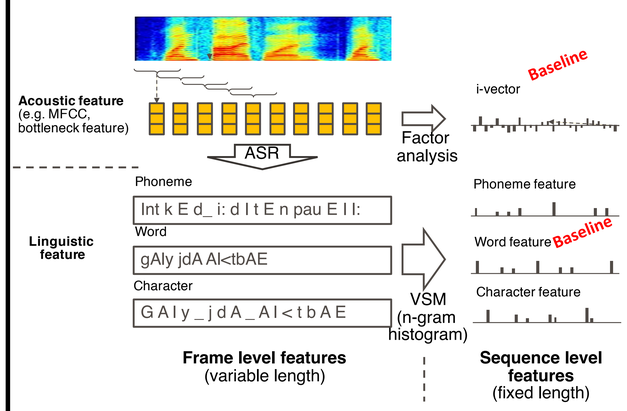
Speech Corpora

- 5 Dialects : Modern Standard Arabic, Egyptian, Levantine, Gulf, North African
- Test dataset domain is different from Training dataset
- Development dataset is relatively small compare to training set, however, it is matched with the test set channel domain

Dataset category	Training (TRN)	Development (DEV)	Test (TST)
Size	53.6 hrs	10 hrs	10.1 hrs
Genre	News Broadcasts		
Channel (recording)	Carried out at 16kHz	Downloaded directly from a high-quality video server	
Availability for system development	O	O	X

Features for ADI

- Acoustic and Linguistic feature



Proposed approach

- **Additional phone level statistics**
  - Using additional phone level statistics such as phone duration and posterior probability
  - This enables Discrimination among different occurrences of the same phone sequences with different phone duration
  - Phone duration representation : classify phoneme into 4 sub-level considering the phone duration

Algorithm 1 Phone representation with phone duration index

```
for c in utterance's phone trasciption do
    if D(c) < M − 0.5S then
        c ← c1
    else {M − 0.5S < D(c) < M}
        c ← c2
    else {M < D(c) < M + 0.5S}
        c ← c3
    else
        c ← c4
    end if
end for
```

  - Phone probability representation: classify phoneme into 4 sub-level considering the occurrence in a utterance

Algorithm 2 Phone representation with phone probability index

```
for c in utterance's phone trasciption do
    if P(c) < M − 0.5S then
        c ← c1
    else {M − 0.5S < P(c) < M}
        c ← c2
    else {M < P(c) < M + 0.5S}
        c ← c3
    else
        c ← c4
    end if
end for
```

  - Fusion of classifier's score from parallel phonotactic DID system

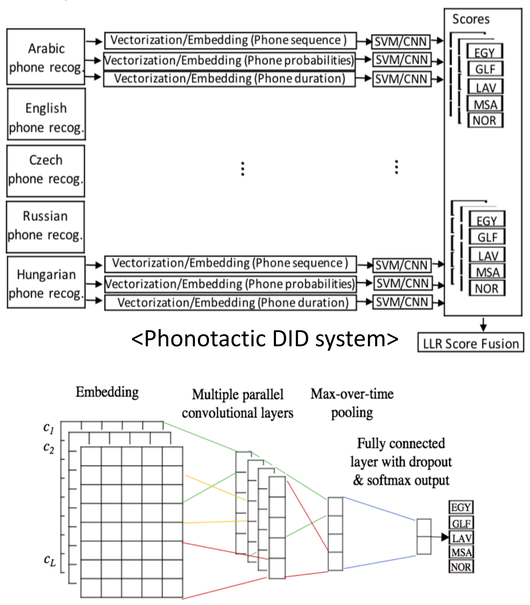


Figure 2: *Architecture of our CNN-based classifier*

Experimental result

- **SVM vs. CNN based classifier**

Language	(%) Phone n-gram seq. Acc. SVM	(%) Phone n-gram seq. Acc. CNN
Arabic	56.82	57.91
English	56.03	56.88
Russian	56.25	57.12
Czech	56.64	57.62
Hungarian	56.71	57.85
Fusion	62.12	64.50

Table 2: *Employing language-dependent parallel PRLMS in a conventional versus an attention-based context for DID*

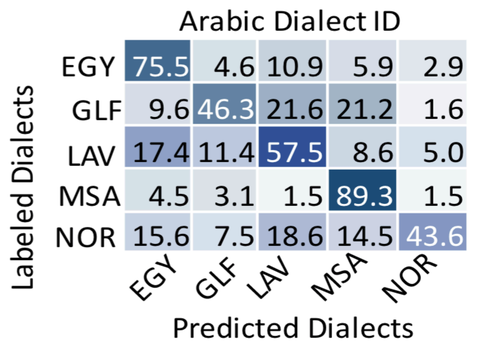
- **Proposed Multi-lingual phonotactic system**

Language	System	(%) Acc.
Arabic	Phone n-gram sequence with CNN	57.91
	Phone n-gram (duration relabeled) with CNN	59.55
	Phone n-gram (probability relabeled) with CNN	59.72
	LLR fusion of 3 systems	68.95
English	Phone n-gram sequence with CNN	56.88
	Phone n-gram (duration relabeled) with CNN	56.30
	Phone n-gram (probability relabeled) with CNN	56.24
	LLR fusion of 3 systems	63.70
Russian	Phone n-gram sequence with CNN	57.12
	Phone n-gram (duration relabeled) with CNN	57.59
	Phone n-gram (probability relabeled) with CNN	57.29
	LLR fusion of 3 systems	65.10
Czech	Phone n-gram sequence with CNN	57.62
	Phone n-gram (duration relabeled) with CNN	57.71
	Phone n-gram (probability relabeled) with CNN	57.37
	LLR fusion of 3 systems	67.85
Hungarian	Phone n-gram sequence with CNN	57.85
	Phone n-gram (duration relabeled) with CNN	58.74
	Phone n-gram (probability relabeled) with CNN	58.90
	LLR fusion of 3 systems	68.31
Fusion	LLR fusion of all systems	71.60
	LLR fusion of Arabic, Hungarian, and Czech systems	73.27

Table 3: *Employing language-dependent parallel PRLMS in a conventional versus an attention-based context for DID*

Experimental result

- **Confusion matrix of final fusion system**



Conclusion

- Arabic dialect identification system using phonotactic feature
- Direct mapping of acoustic and phonotactic feature to one of five dialects
- New phone level statistics based phonotactic feature based dialect identification with 73% accuracy
- For future work, we would explore long short-term memory RNN using raw acoustic waveform to make dialect prediction per frame

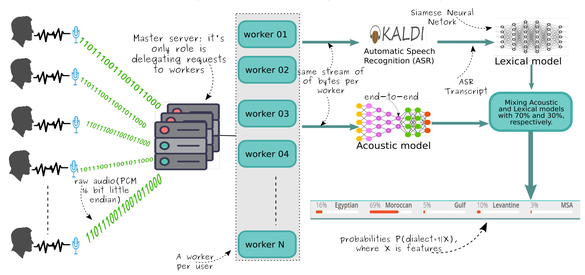
Real-time Arabic Dialect Identification (Demo session, Friday 13:30)

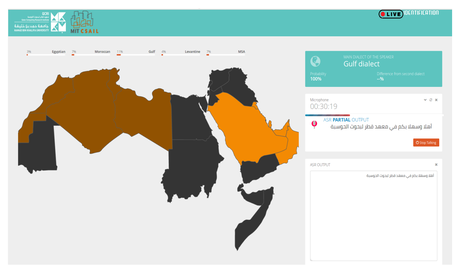
- Real-time online Arabic dialect identification and recognition system*

April 20 (Friday) 13:30 – 15:30 @Exhibit Hall Foyer

Demo-4.1: QCRI-MIT LIVE ARABIC DIALECT IDENTIFICATION SYSTEM

Mohamed Eldesouki, Suwon Shon, Ahmed Ali, James Glass





Our demo is also publicly available at https://dialectid.qcri.org

* Applied algorithms are based on the paper below

S. Shon, A. Ali, and J. Glass, "Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition," *to be appeared on Odyssey 2018*

Acknowledgements: This research was carried out in collaboration between the HBKU Qatar Computing Research Institute (QCRI) and the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).