Convolutional Neural Networks and Language Embeddings for End-to-End dialect Recognition

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]

Motivation

- 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 (DID) can be a useful capability to identify which dialect is being spoken during a recording.
- The Arabic Multi-Genre Broadcast (MGB) Challenge tasks have provided a valuable resource for re- searchers interested in processing multi-dialectal Arabic speech.
- Investigation of end-to-end DID approach with dataset augmentation for acoustic feature and language embeddings for linguistic feature

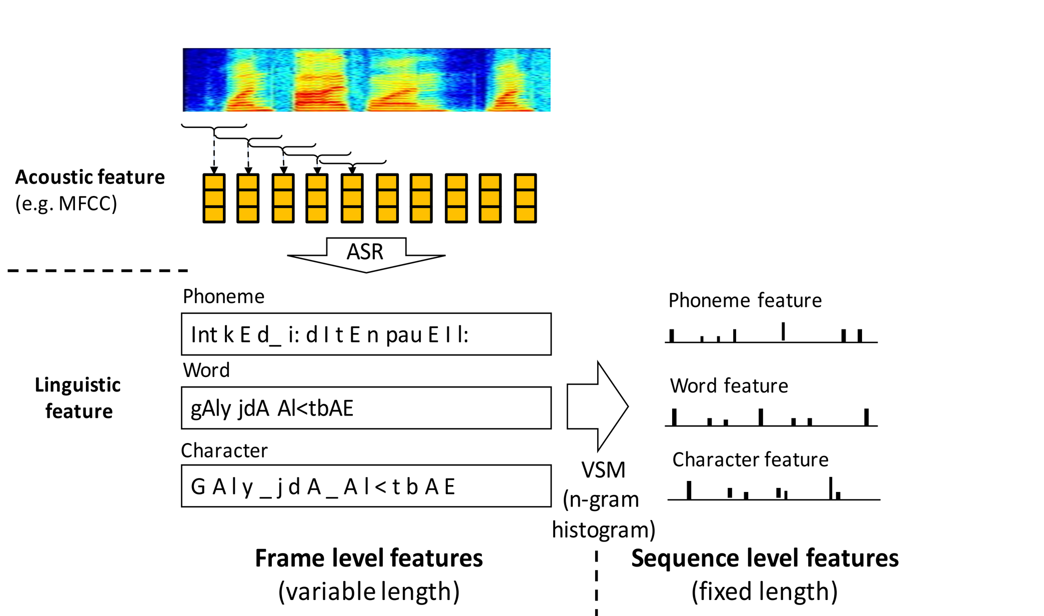
MGB-3 Dataset

- 5 Dialects : Modern Standard Arabic, Egyptian Levantine, Gulf, North African
- Test dataset domain is different from Training dataset

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

<Multi Genre Broadcast (MGB)-3 dataset description>

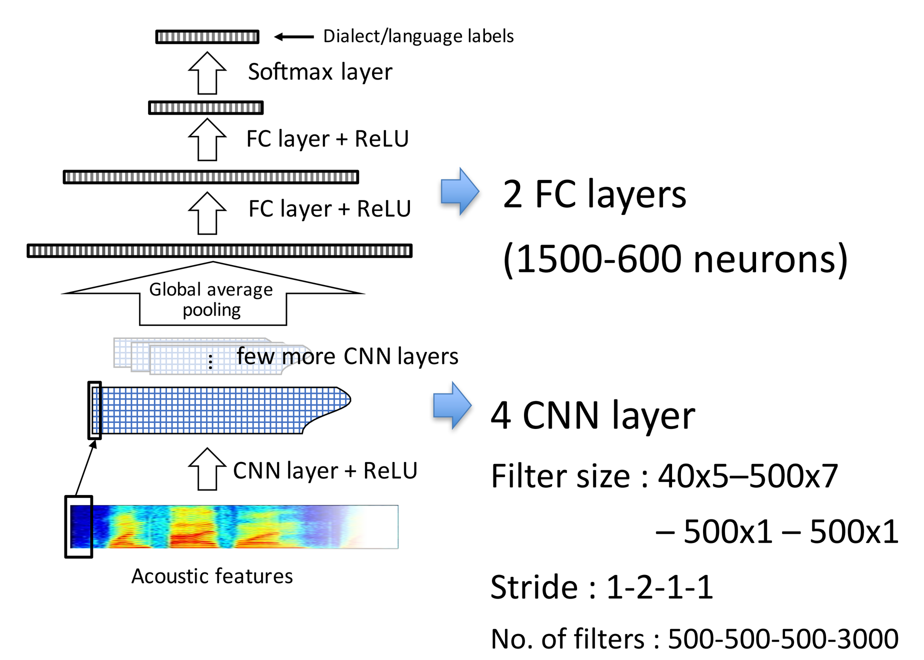
- Feature extraction
  - Acoustic feature : MFCC, FBANK, Spectrogram
  - Linguistic feature : phoneme, word, character



<Feature extraction flow >

End-to-end DID with Acoustic features*

- CNN based End-to-end model structure



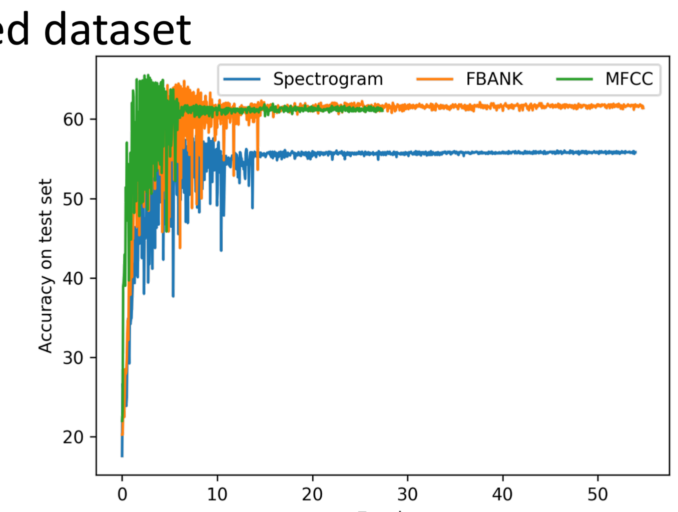
<Network structure>

- Performance by input feature

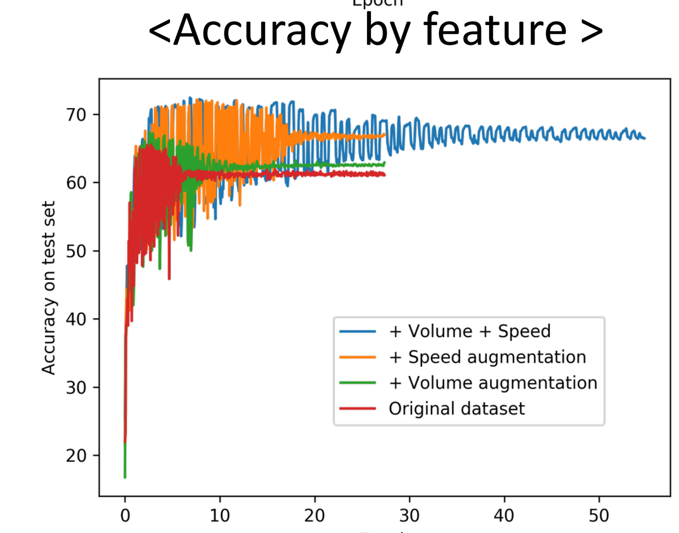
Feature	Maximum			Converged		
	Accuracy	EER	C_{avg}	Accuracy	EER	C_{avg}
MFCC	65.55	20.24	19.92	61.33	21.95	21.53
FBANK	64.81	20.22	19.91	61.26	22.12	21.79
Spectrogram	57.57	24.48	24.49	54.22	25.90	25.09

<Performance evaluation by features>

- The maximum condition: the network achieves the best accuracy
- The converged condition: the average loss of 100 mini-batches < 1e-5.
- Theoretically, spectrograms have more information than MFCC or FBANKS, but it seems hard to optimize the network using the limited dataset



<Accuracy by feature >



<Accuracy by augmentation(speed and volume) >

- Dataset augmentation

Augmentation method (feature = MFCC)	Maximum			Converged		
	Accuracy	EER	C_{avg}	Accuracy	EER	C_{avg}
Volume	67.49	20.37	20.00	62.47	21.55	21.08
Speed	70.51	17.54	17.39	65.42	19.87	19.19
Volume and speed	70.91	17.79	17.93	67.02	19.37	19.01

<Performance evaluation by augmentation method>

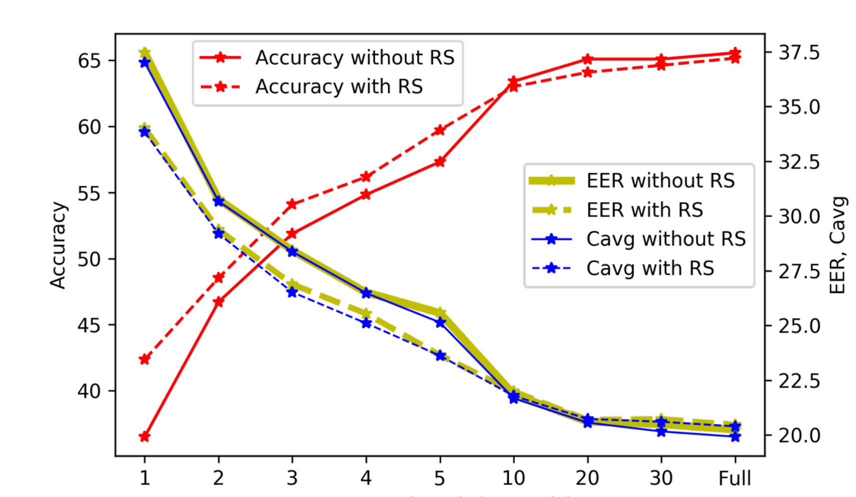
- Perturb slightly original dataset attributes
- Speed factor of 0.9 and 1.1, Volume factor of 0.25 and 2.0

Feature (on augmented dataset)	Accuracy	EER	C_{avg}
MFCC	70.91	17.79	17.93
FBANK	71.92	18.01	17.63
Spectrogram	68.83	18.70	18.69

<Performance evaluation by features on augmented dataset >

- Spectrogram is worst, but gain from increasing dataset size is much higher than MFCC, FBANK

- Random Segmentation (RS)



- Segmentation of the training dataset into small chunks randomly between 2 to 10 seconds
- Since random segmentation provides diversity given a limited dataset, the performance is improved on short utterance

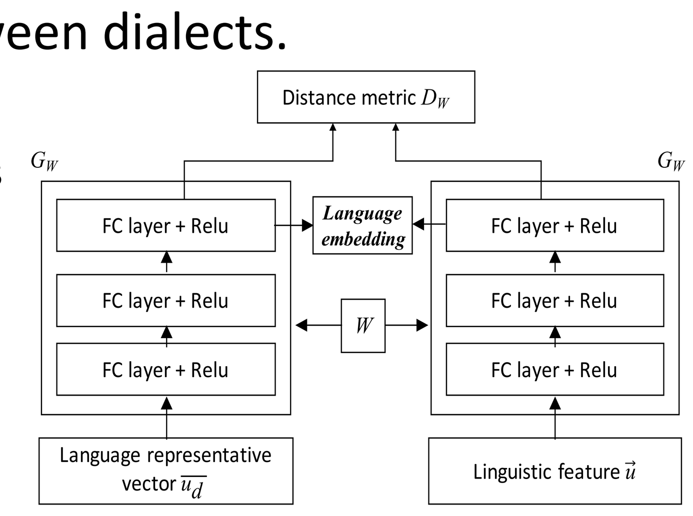
- Final result with augmented dataset
  - End-to-end system outperforms other conventional i-vector approaches.

System	Accuracy	EER	C_{avg}
i-vector	60.32	26.98	26.35
i-vector-LDA	62.60	21.05	20.12
End-to-End (MFCC)	71.05	18.01	17.97
End-to-End (FBANK)	**73.39**	**16.30**	**15.96**
End-to-End (Spectrogram)	70.17	17.64	17.27

<Performance comparison with conventional i-vector approach >

Language embedding**

- Siamese neural network to learn dis-similarity and similarity between dialects.
  - FC layer has 1500-600-200 neurons
  - Distance metric is Cosine similarity



<Network structure>

- Result
  - Words feature shows best improvement among three features
  - Another benefit is that the linguistic feature dimension can be significantly reduced

Phoneme Recognizer	System	Accuracy	EER	C_{avg}
Hungarian	Baseline	48.86	29.94	29.16
	Embedding	**54.49**	**28.69**	**27.77**

<Phoneme feature>

Feature	System	Accuracy	EER	C_{avg}
Character	Baseline	51.34	30.03	30.17
	Embedding	**58.18**	**25.48**	**25.68**
Word	Baseline	50.00	30.73	30.41
	Embedding	**58.51**	**24.87**	**24.99**

<Character and word feature>

Fusion result

- Fusion between end-to-end system and language embeddings shows better efficiency than between end-to-end system such as MFCC and FBANK
- Spectrograms achieve slightly better results than MFCCs

Fusion system (**Bold** : end-to-end system *italic* : language embedding)	Accuracy(%)	EER(%)	C_{avg}
FBANK + *word*	76.94	13.66	13.57
FBANK + *char*	76.61	13.89	13.87
FBANK + *phoneme*	75.13	14.95	14.79
FBANK + **MFCC**	74.40	15.63	15.50
MFCC + *word* + *char* + *phoneme*	77.48	14.02	14.00
FBANK + *word* + *char* + *phoneme*	**78.15**	**12.77**	**12.51**
Spectrogram + *word* + *char* + *phoneme*	77.88	13.34	13.24
i-vector + **FBANK** + *word* + *char* + *phoneme*	**81.36**	**11.03**	**10.90**

<Performance of score fusion systems with end-to-end system and language embeddings>

Systems	Accuracy(%)	
	Single System	Fusion System
Khurana et al. [5]	67	73
Shon et al. [9]	69.97	75.00
Najafian et al. [7]	59.72	73.27
Bulut et al. [10]	-	79.76
Our approach	**73.39**	**81.36**

<Comparison with recent studies>

Discussion

- Data augmentation by perturbing speed gives impressive gain on performance
- If we have large dataset, we can use raw signals as input features
- At the same time, however, it is difficult to determine how much training data is required for training raw features
- Fusion with different features such as between acoustic and linguistic gives great effectiveness

Conclusion

- We present end-to-end dialect identification system using acoustic and linguistic features
- We investigated several techniques for end-to-end DID on acoustic features and language embeddings of linguistic features
- Using a limited dataset, we can increase diversity by perturbing the attribute of speech audio and random segmentation
- The end-to-end DID system has a simplified topology and training methodology compared to conventional bottleneck feature based i-vector extraction

* https://github.com/swshon/dialectID_e2e
**https://github.com/swshon/dialectID_siam

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).