MIT-QCRI ARABIC DIALECT IDENTIFICATION SYSTEM FOR THE 2017 MULTI-GENRE BROADCAST CHALLENGE

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.
- While MGB-3 Arabic Dialect Identification (ADI) contains 5 Arabic dialect, the evaluation scenario can be viewed as channel and domain mismatched scenario

MGB-3 Challenge

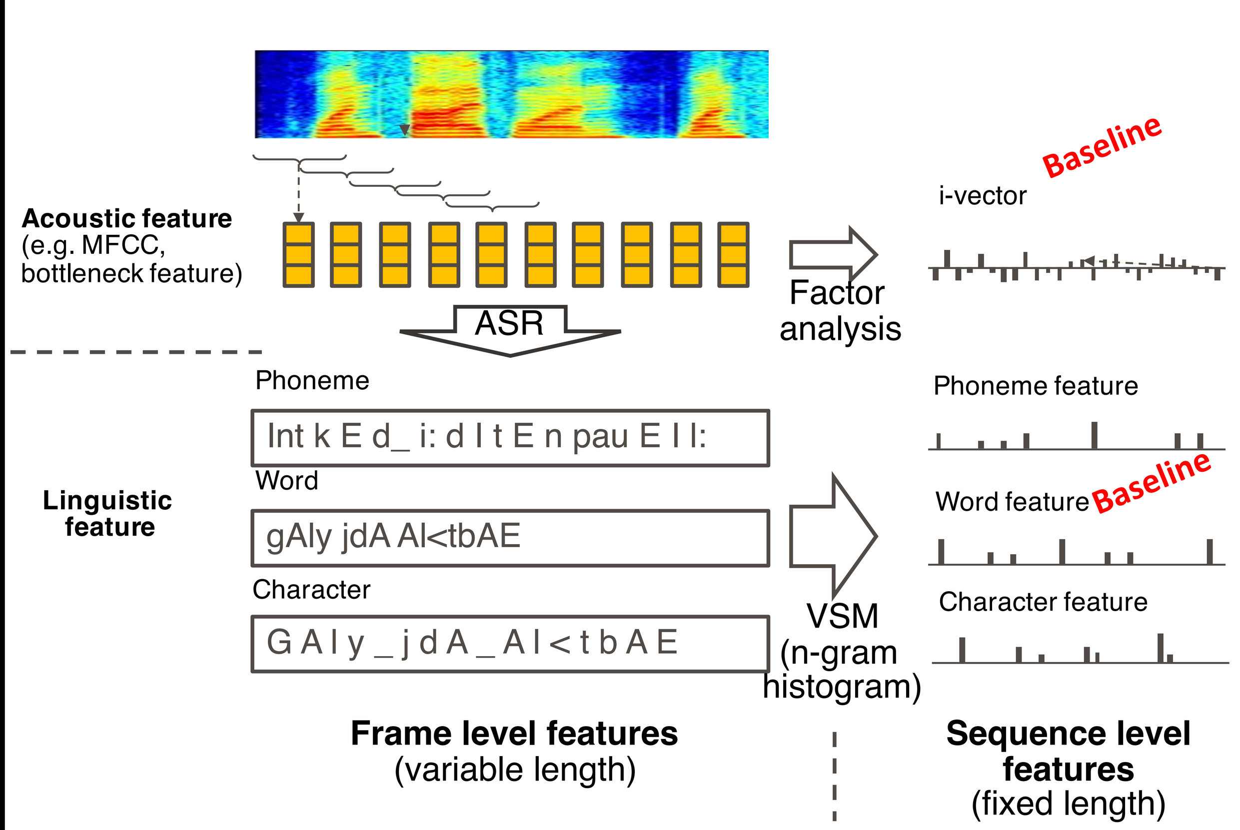
- 5 Dialects : Modern Standard Arabic, Egyptian, Levantine, Gulf, North African
- Test dataset domain is different from Training dataset
- Development dataset is relatively small compare to training set, however, it is matched with the test set channel domain

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

Table 1. MGB-3 ADI Dataset Properties.

Features for ADI

- Acoustic and Linguistic feature



Proposed approach - Acoustic

- Siamese Neural Network based dimension reduction*
  - To learn similarity and dissimilarities among Arabic dialects
  - Using two parallel convolutional network that shares parameters

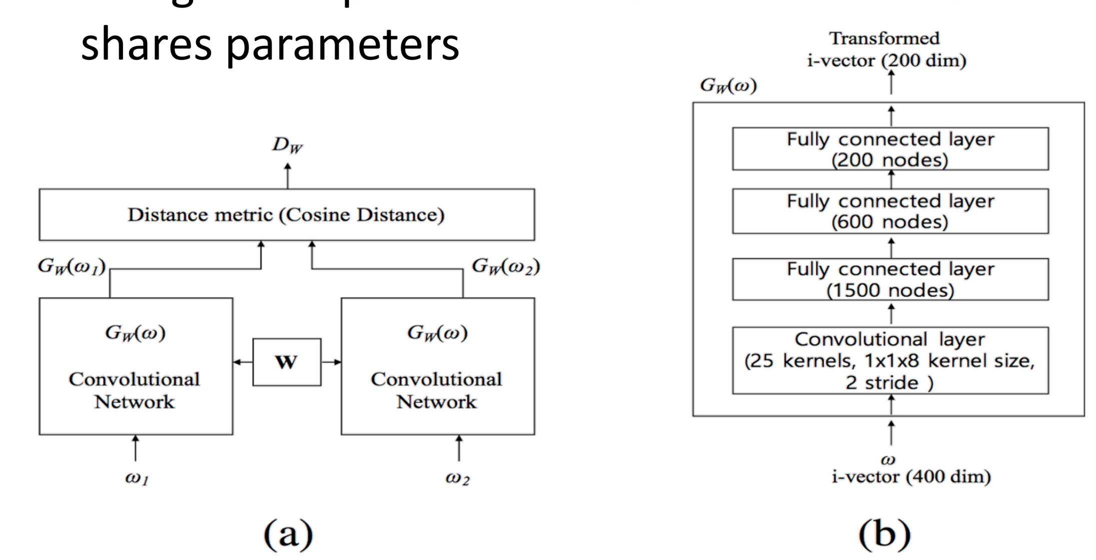


Fig. 1. (a) Siamese network for i-vector (b) Architecture of convolutional neural network G_W

  - To optimize the network, Euclidean distance loss function L is used between the label and cosine distance

$$L(\omega_i, \omega_j, Y_{ij} = ||Y_{ij} - D_W(\omega_i, \omega_j)||_2^2$$

- Interpolated i-vector dialect model
  - Since we have two dialect dataset for training and development, we can use interpolation approach with parameter γ

$$\overline{\omega_d^{\text{Inter}}} = (1-\gamma)\overline{\omega_d^{\text{TRN}}} + \gamma\overline{\omega_d^{\text{DEV}}}$$

  - Figure 2 shows performance heavily depend on parameter gamma, and shows max 15% improvement on test dataset

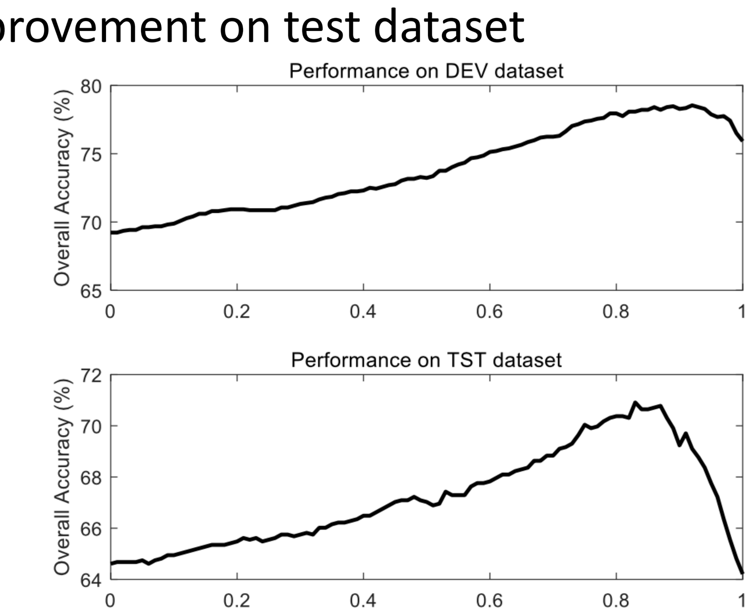


Fig. 2 Overall accuracy on DEV and TST sets by gamma: The DEV set shows the best performance at gamma = 0.91, while the TST set shows the best result at gamma=0.83. For our experiments, we used gamma = 0.91

- Recursive whitening transformation
  - To remove un-whitened residual components in the dataset associated with i-vector length normalization

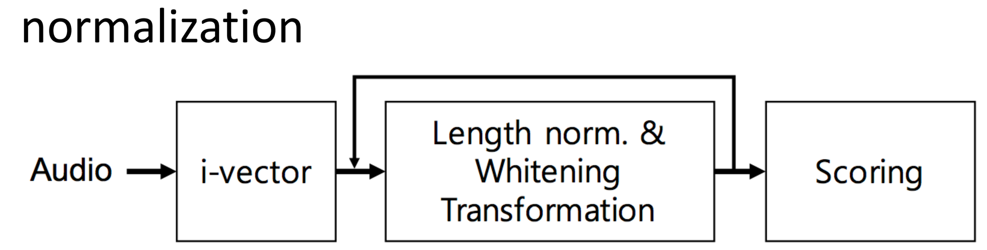


Fig. 3. Flowchart of recursive whitening transformation.

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

Proposed approach – Linguistic

- Phoneme feature
  - Extracting the phone sequence, and phone duration statistics using four different speech recognizers
  - Table 2 shows the Hungarian phoneme recognition obtained the best results

System	Accuracy(%)	Precision(%)	Recall(%)
Czech	45	45.2	45.8
Hungarian	**47**	**47.3**	**48.1**
Russian	46	47	46.8
English	33.3	33	34

Table 2. Evaluating four phoneme recognition systems.

- Character feature
  - Word sequences are extracted using a state-of-the-art Arabic speech-to-text transcription system built as part of the MGB-2 : Combination of TDNN, LSTM and BLSTM acoustic models, followed by 4-gram and Recurrent Neural Network (RNN) language model rescoring using grapheme lexicon during both training and decoding

System (scoring method)	Accuracy (%)	Precision (%)	Recall (%)
Baseline word(SVM)	48.43	50.99	49.25
Character	**57.28**	**60.83**	**58.03**
Phoneme	47.18	47.66	48.23

Table 4. Linguistic feature evaluation on DEV set: TRN and DEV sets were used for training.

Discussion

- Siamese network can learn similarity and dissimilarity effectively even without target domain information
- Interpolated i-vector dialect model shows significant performance improvements by leveraging target domain information
- Fusion rule from system 1 prevented overfitting on the target domain
- As the linguistic features is not affected by the domain mismatch, linguistic features show useful contributions for all systems.

Conclusion

- Arabic dialect identification system using both audio and linguistic features
- Several approaches to address dialect variability and domain mismatches between the training and test sets.
- On both conditions, fusion of audio and linguistic feature guarantees substantial improvements on dialect identification.
- In future, we will explore their utility on other speaker and language recognition problems in the future.

Experiments

- Final submitted system (***marked 1st on MGB-3 Challenge ADI task***)

System		TST Accuracy	TST Precision	TST Recall	DEV Accuracy
TRN	i-vector - baseline	55.29	59.27	56.44	57.28
	Siamese i-vector	60.99	60.88	61.72	63.65
	+ fusion w. linguistic feature	**67.76**	**68.00**	**67.88**	**66.60**
TRN + DEV	i-vector - baseline	65.82	65.80	66.35	64.79
	i-vector	60.86	61.87	61.49	62.07
	+ interpolated	68.23	68.95	68.56	75.52
	+ recursive	69.97	70.37	70.37	78.54
	+ fusion w. linguistic feature (primary)	**75.00**	**75.46**	**75.03**	**76.38**
	Siamese i-vector	62.47	62.28	63.32	65.81
	+ interpolated	68.23	68.75	68.63	76.05
	+ recursive	68.30	68.81	68.69	76.31
	+ fusion w. linguistic feature (contrastive)	**72.72**	**73.02**	**72.99**	**73.43**

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).