QCRI-MIT Live Arabic Dialect Identification System

Mohamed Eldesouki[1], Suwon Shon[2], Ahmed Ali[1], and James Glass[2]
[1]QCRI, [2]MIT-CSAIL

Motivation

The task of spoken dialect identification consists of classifying a given spoken utterance into one of the many dialects in a particular language.

Arabic Dialect Identification (ADI) is similar to the more general problem of Language Identification (LID).

ADI is more challenging than LID because of the small and subtle differences between the various dialects of the same language.

A good ADI system can be used to extract dialectal data from the speech database to train dialect specific acoustic models for speech-to-text transcription.
It can also be used for meta-data enrichment.

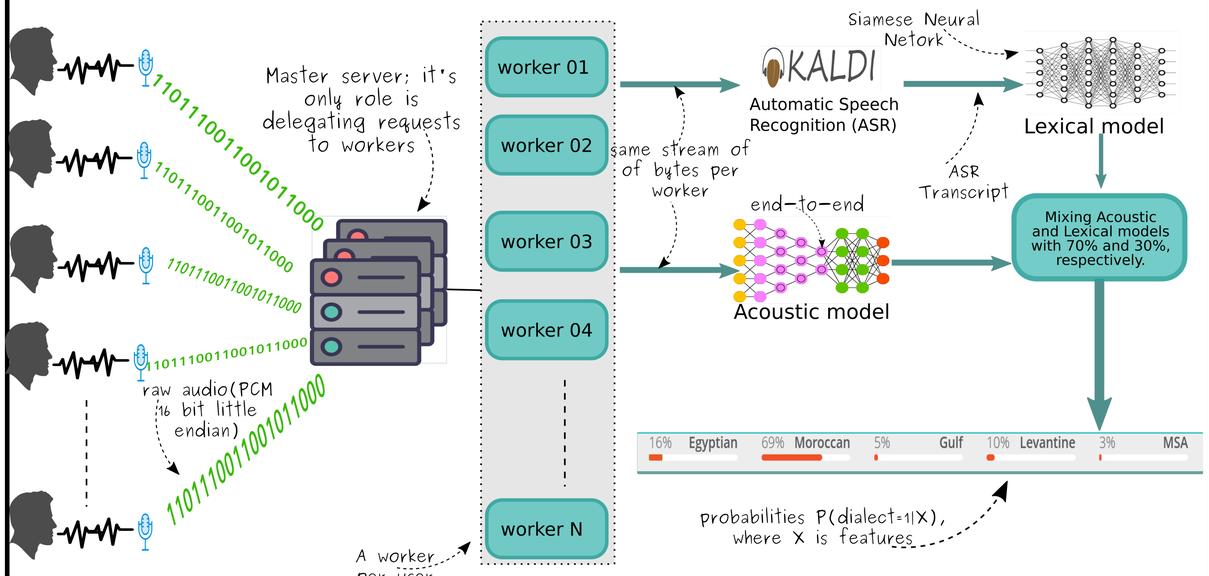
QMDIS

We present our live speech Arabic dialect identification system; QCRI-MIT Advanced Dialect Identification System (QMDIS).

Our demo features modern web technologies to capture live audio, and broadcasts Arabic transcriptions along with the corresponding dialect simultaneously. The detected dialect is visualized using light map, where the intensity of the color reflects the probability of the dialect. We also integrate meter bars to display live the probability for each dialect per sentence.

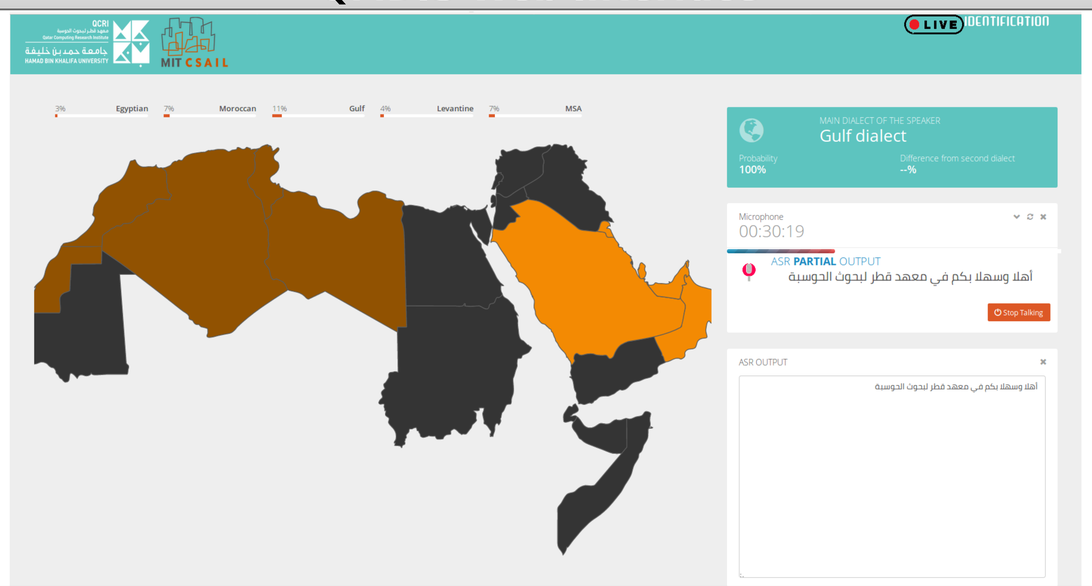
Our demo is publicly available at https:/dialectid.qcri.org

QMDIS Architecture

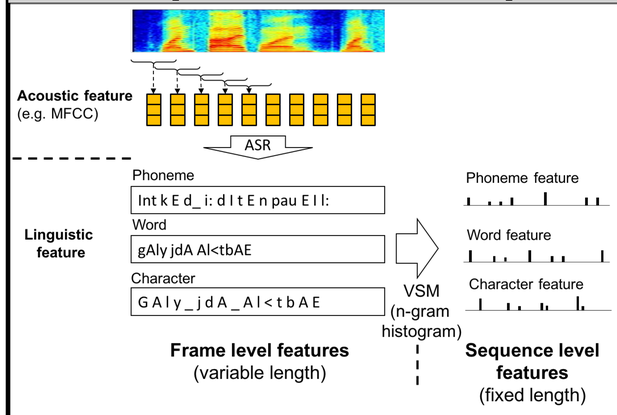


Full duplex communication based on websockets	Very scalable	Uses Gstreamer which supports unlimited set of audio codecs
Supports arbitrarily long speech input (e.g., you can stream live speech into it)	Supports Kaldi's GMM and "online DNN" models	The master server doesn't perform speech recognition itself, it simply delegates client recognition requests to workers.
Python, Java, Javascript clients are available	Can do speech segmentation	The dialect identification task is invoked in a separate process to insure concurrency

QMDIS Web Interface



Experimental Setup



Dataset: Multi-Genre Broadcast 3 (MGB-3)

Acoustic model features; (1)Mel-Freq. Cepstral Coefficients (MFCC), (2)log Mel-scale Filter Bank energies (FBANK), (3)spectrogram energies

Data augmentation through speed perturbation

Siamese neural network models to learn similarity and dissimilarities among Arabic dialects, as well as i-vector post-processing to adapt domain mismatches

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

References

Suwon Shon, Ahmed Ali, James Glass, (2018) Convolutional Neural Networks and Language Embeddings for End-to-End Dialect Recognition, Interspeech

Mohamed Eldesouki, Suwon Shon, Ahmed Ali, Yifan Zhang (2018), QCRI-MIT Live Arabic Dialect Identification System *ICASSP.*