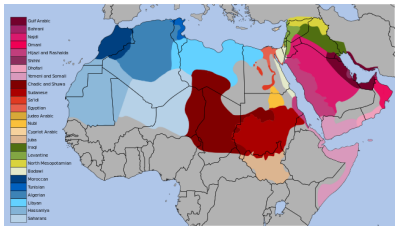
DEVELOPMENT OF THE MIT ASR SYSTEM FOR THE 2016 ARABIC MULTI-GENRE BROADCAST CHALLENGE

Tuka AlHanai, Wei-Ning Hsu, and James Glass
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge MA 02139

1. MOTIVATION

Challenge: The Arabic language has 300 million speakers with significant *diversity* and *breadth*. A large existing and potential user base for Arabic language technologies.



Solution: State-of-the-art techniques for Automatic Speech Recognition (ASR).

Methods: ASR Acoustic Modeling with

- Feed-forward Deep Neural Networks (DNN)
- Time-Dependent Neural Networks (TDNN)
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)
  - Long-Short Term Memory (LSTM)
  - Highway-LSTM (H-LSTM)
  - Grid-LSTM (G-LSTM)
- Various Objective Functions
  - Cross-Entropy (CE), Minimum Phone Error (MPE), Minimum Bayes Risk (MBR), Lattice Free Maximum Mutual Information (LF-MMI)

2. DATASET

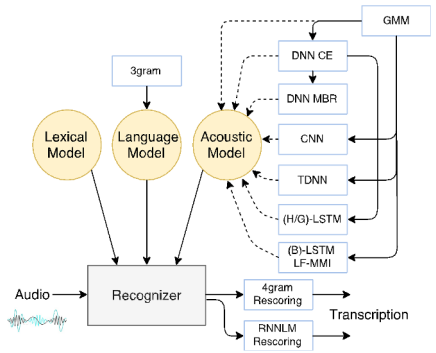
Multi-Genre Arabic Broadcast (MGB) Corpus

- 10 years of Al-Jazeera News Channel programming (2005-2015)
- 1200 hours transcribed audio
- 8M words (200K Vocab)
- 375K utterances
- 10 hour development set

Extra text

- 120M Words (1.4M Vocab)
- 1.75% Out-of-Vocabulary (OOV)

3. SETUP



Experimental Setup of Arabic ASR System

Toolkits

- Kaldi Speech Recognition (Features + Language/Acoustic Models)
- CNTK (Acoustic Models)
- SRILM (Language Models)

Features

- *Baseline GMM-HMM:* 39-dim MFCC + LDA + MLLT + fMLLR.
- *DNN:* 30/80 Mel-Filterbanks + pitch

Language Model

- 3-gram with Kneser-Ney Smoothing
- 4-gram rescoring with MGB + Extra Text
- RNN
  - 1000 hidden units + Hierarchical Softmax
  - 300 hidden units + NCE Criterion

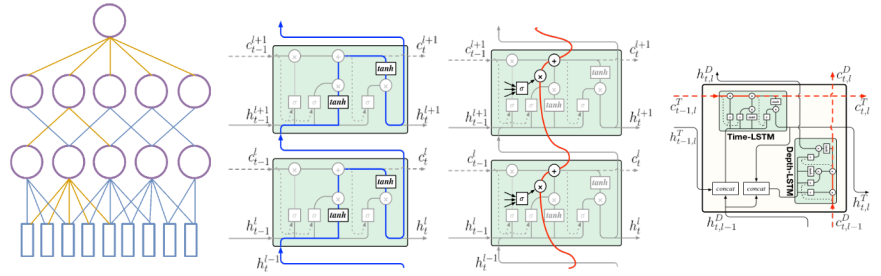
Model Combination

- Lattice combination and hypothesis scoring applying Minimum Bayes Risk (MBR)

Evaluation

- Word Error Rate (WER) and significance testing using Matched Pair Sentence Segment Word Error (MAPSSWE)

4. ACOUSTIC MODELS



5. RESULTS

Model	Topology	Alignment	WER* (%)	$p <$
DNN CE	5x1024	GMM	28.1	-
CNN	4x2000	GMM	28.1	0.734
TDNN	6x3000	GMM	25.8	0.001
DNN MPE	5x1024	CE	24.7	0.001
Chain TDNN	7x625	GMM	23.4	0.001
LSTM	3x1024	CE	22.7	0.001
H-LSTM 3L	3x1024	CE	22.6	0.250
H-LSTM 5L	5x1024	CE	22.4	0.184
G-LSTM 3L	3x1024	CE	21.7	0.001
G-LSTM 5L	5x1024	CE	21.5	0.070
G-LSTM 3L sMBR	3x1024	CE	19.5	0.001
G-LSTM 5L sMBR	5x1024	CE	19.2	0.034
Top 2 Combined	**G-LSTM sMBR (3L + 5L)**	**CE**	**18.3**	**0.001**

*WER = Word Error Rate

- Models that capture temporal context are superior – LSTM
- Could be better ways to leverage CNN
  - e.g. in a hybrid system (CLDNN), for speaker normalization
- TDNN outperformed DNN CE. Captures a wider temporal context.
- Chain TDNN trained on weaker alignments outperformed DNN MPE
- RNN LM hurt performance, but didn't search for optimum parameters
- 8 out of 12 improved performance incrementally by $p < 0.05$