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

Tuka AlHanai, Wei-Ning Hsu, and James Glass

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{tuka, wnhsu, glass}@mit.edu

ABSTRACT

The Arabic language, with over 300 million speakers, has significant diversity and breadth. This proves challenging when building an automated system to understand what is said. This paper describes an Arabic Automatic Speech Recognition system developed on a 1,200 hour speech corpus that was made available for the 2016 Arabic Multi-genre Broadcast (MGB) Challenge. A range of Deep Neural Network (DNN) topologies were modeled including; Feed-forward, Convolutional, Time-Delay, Recurrent Long Short-Term Memory (LSTM), Highway LSTM (H-LSTM), and Grid LSTM (G-LSTM). The best performance came from a sequence discriminatively trained G-LSTM neural network. The best overall Word Error Rate (WER) was 18.3% ($p < 0.001$) on the development set, after combining hypotheses of 3 and 5 layer sequence discriminatively trained G-LSTM models that had been rescored with a 4-gram language model.

Index Terms— Arabic, Automatic Speech Recognition, MGB Challenge, Deep Neural Networks

1. INTRODUCTION

Increases in computational power and data sizes, along with foundational work on neural networks have motivated a broad range of research adopting, developing, and evaluating such models. These developments have gone beyond the classical use of Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) for Automatic Speech Recognition (ASR) systems. Hinton et al. showed the strength of using feed-forward neural networks to model speech [1, 2, 3]. Further developments to harness temporal context showed the power of Recurrent Neural Networks (RNN) in the flavor of Long Short-Term Memory (LSTM) models [4, 5]. Another powerful topology is the Convolutional Neural Network (CNN) which attempts to model local information in the feature space [6, 7]. However, these have generally found to be more powerful in the domain of vision [8, 9, 10].

Further variations of the LSTM model exist, such as the the Highway LSTM (H-LSTM) and Grid LSTM (G-LSTM). The H-LSTM introduces a directed gated connection between any given memory cell c_t^l in a layer l to that of the corresponding cell c_t^{l+1} in the next layer $l+1$ above [11, 12, 13]. This connection provides a linear dependence between the cells of different layers, in addition to the linear dependence between cells across time that exists in LSTMs. A G-LSTM is a generalized version of the multi-dimensional LSTM, where each grid contains the same number of LSTM blocks as the number of dimensions, which are time and depth in our case. Gated linear dependence is introduced to adjacent cells at each dimension. Both H-LSTM and G-LSTM ease the problem of the vanishing gradient along depth dimension and hence enable the training of deeper neural network models [14].

Another topology being explored, the Time-Delay Neural Network (TDNN), works to capture a wider context of information with respect to time at both the input and at deeper layers of the network [15]. This is managed by splicing together features at different timestamps at some or all of the layers in the network. A recent development is to perform sequence discriminative training without the need of frame-level cross-entropy pre-training. This is done by performing Maximum Mutual Information (MMI) based sequence training at the phone level. This method successfully outperforms CE trained models on datasets of various sizes [16].

Previous work in the domain of Arabic Automatic Speech Recognition has utilized up to 1,800 hours of data [30], with limited use of Deep Neural Networks for acoustic modeling [18, 21, 30]. The largest single Arabic dataset available until now was the 500 hour GALE corpus [31, 32]. Table 1 provides a summary of research in the field as well as the general ASR performance. In comparison, our current system was trained with data on the larger end of developed systems (1,200 hours) with the release of the Arabic MGB dataset, and employed state-of-the-art in acoustic modeling techniques, extending existing work in the field. Specifically, we evaluate the performance of several DNN topologies; Feed-forward, CNN, LSTM, TDNN, H-LSTM, and G-LSTM. Not only did we compare Neural Network topologies of existing toolkits (Feed-forward, CNN, and TDNN), but we also com-

Table 1. Arabic ASR Approaches in Literature

Reference	Hours	Dataset	Language Model	Acoustic Model	WER (%)
Biadsy et al. [17]	40	T	3gram	GMM	43.1 - 47.3
Cardinal et al. [18]	50	Q	3gram	GMM, DNN	18.0 - 42.6
Billa et al. [19]	60	B	3gram	GMM	15.3 - 31.2
Afify et al. [20]	100	F, T	3gram	GMM	14.2 - 21.9
Thomas et al. [21]	100	EA	3gram	DNN, CNN	31.9 - 40.0
Messaoudi-Lamel et al. [22]	150	F, T, B	3gram	GMM	13.2 - 24.8
Messaoudi-Gauvain et al. [23]	150	F, T, B	3gram	GMM	14.8 - 16.0
Xiang et al. [24]	150	F, T, B	3gram	GMM	17.8 - 31.8
Ali et al. [25]	200	G	3gram	GMM/DNN	15.8 - 43.5
Al-Haj et al. [26]	450	IA	3gram	GMM	33.3 - 37.0
Vergyri et al. [27]	1,100	G	3gram	GMM	8.9 - 36.4
El-Desoky et al. [28]	1,100	G	3gram	GMM	13.9 - 16.3
Ng et al. [29]	1,400	F, T, G, IA	3gram	GMM	10.2 - 18.8
Mangu et al. [30]	1,800	G	3gram	GMM, Bayesian Sensing	7.1 - 12.6
Our System	1,200	MGB	3gram, Rescore: 4gram, RNNLM	GMM, DNN, CNN, TDNN, (H/G-)LSTM	18.3 - 40.3

Dataset: **T** = TDT4, **Q** = QCRI/Aljazeera in-house, **B** = BBN in-house News, **F** = FBIS, **G** = GALE, **EA** = Egyptian Arabic, **IA** = Iraqi Arabic, **MGB** = Multi-Genre Broadcast.

pared with LSTM models developed in-house. We highlight that the G-LSTM model was applied for the first time to an ASR task.

2. METHOD

2.1. Toolkits

Our ASR pipeline employed a number of tools to develop the various components. We used the KALDI speech recognition toolkit to extract features, and to build and evaluate acoustic models [33]. The CNTK toolkit was also used to train acoustic models [34], while the SRILM toolkit was used to build the language models [35].

2.2. Features

We built a baseline recognizer using Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). These were trained using 39-dim Mel Frequency Cepstral Coefficients (MFCC) features that were transformed using Linear Discriminant Analysis (LDA), Maximum Log-Likelihood Transform (MLLT), and feature space Maximum Likelihood Linear Regression (fMLLR). Alignments generated from the GMM-HMM model were used to train a variety of DNN based models. The DNN models were trained using Mel Filterbanks (Fbank) either 30 (for the feed-forward models) or 80 in dimension (rest of the models), all of which were concatenated with 3 pitch features. All DNN models used as input spliced features of width 5, unless stated otherwise.

2.3. DNN Models

Three models were trained using alignments generated by the GMM-HMM model; (1) a feed-forward DNN was trained with the Cross Entropy (CE) criterion, composed of 5 layers and 2048 hidden units in each layer, (2) a Convolutional Neural Network (CNN) with 4 layers and 2000 hidden units in the first layer, and (3) a Time-Delay Neural Network (TDNN) with 6 layers and 3000 hidden units in the first layer.

Alignments from DNN-CE model were then used to train (1) a feed-forward DNN of the same architecture but with the Minimum Phone Error (MPE) criterion, (2) a sequence discriminatively trained ‘chain’ TDNN model (7x625), (3) a 3-layer LSTM model, (4) two H-LSTM models, with 3 layers and 5 layers respectively, and (5) two G-LSTM models, with 3 layers and 5 layers respectively as well.

The **CNN** was composed of 4 layers, the first layer was a 1D convolution component with a maxpooling component, the second layer was a single 1D convolutional component, with the third and fourth layers composed of affine components with ReLU nonlinearities. The first layer had 128 filters, with a patch step size of 1, a dimension of 8, and a pool size of 4. There were 256 filters in the second layer with a patch step of 1, and a patch dimension of 8.

The **TDNN** was composed of 6 layers, with connections of [{-4,-3,-2,-1,0,1,2,3,4};{0};{-2,2};{0};{-4,4};{0}]. The input layer had 3000 hidden units, with the input feature a splice of width 4 (window of +/- 4 frames). The second layer was fully connected to the layer below, the third layer concatenated the input from activations at timestamps only at minus and plus 2 with respect to the node being considered, the third layer was fully connected to the layer below, and so on.

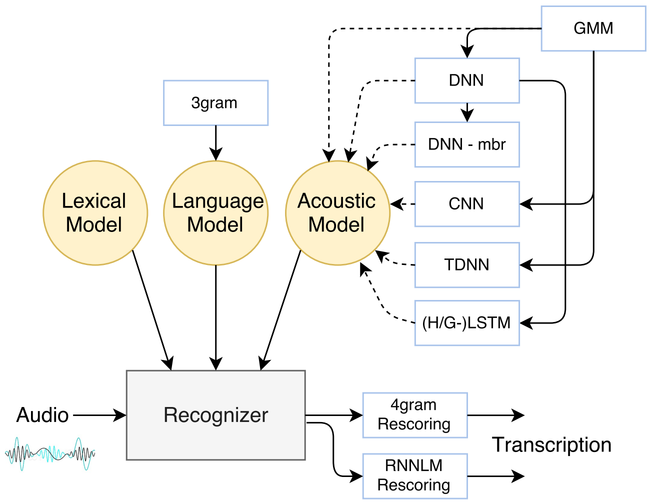

Fig. 1. Experimental Setup of Arabic ASR System.

The **chain TDNN** model was composed of 7 layers with 625 Rectified Linear units (ReLU) at the input layer. The spliced indices at the different layers were [{-1,0,1};{-1,0,1,2};{-3,0,3};{-3,0,3};{-3,0,3};{-6,-3,0};{0}] with LDA applied to the input features. We used the default parameters as defined in the Kaldi recipe [1].

For the **LSTM** and **H-LSTM**, each layer was comprised of 1024 memory cells, and the cell output was fed into a 512-unit linear projection layer [12]. For the **G-LSTM**, each layer contained 1024 memory cells for the time dimension and 1024 memory cells for the depth dimension, of which the outputs were also projected respectively through a linear layer to 512 dimensions. Furthermore, we refined two G-LSTM models by sequence discriminative training with state-level minimum Bayes risk (sMBR) criterion, using the alignments and denominator lattices generated by each model, respectively.

2.4. Model Combination

We also assessed how models complement each other in generating a hypothesis transcript. This was done using lattice combination and hypothesis scoring method presented in [36], which applies Minimum Bayes Risk to minimize the expected WER.

2.5. Lexicon

We use the provided lexicon composed of graphemes, which is a one-to-one mapping between character and acoustic unit, containing a total of 960,000 word entries, and 38 acoustic units. This lexicon had an Out-Of-Vocabulary (OOV) rate of 1.76% on the development set.

[1]kaldi/egs/swbd/s5c/local/chain/run_tdnn_7b.sh

2.6. Language Model

We decoded with a 3-gram model trained on the training data only (8 million words). A 4-gram model was trained on the larger provided text using Knesser-Ney Discounting with a pruning threshold of 1e-10. We also experimented with a Recurrent Neural Network (RNN) language model, using the faster-rnnlm tool of the Kaldi toolkit. We trained two RNN language models on the larger text, one with 1000 hidden units and a hierarchical softmax, while the second was composed of 300 hidden units using the Noice Contrastive Error Criterion set to 20.

2.7. Evaluation

In addition to the standard Word Error Rate (WER) metric for evaluating ASR performance, we present the statistical significance of these values compared to (1) the GMM-HMM baseline, and (2) compared to its closest and lesser performing model in order to gauge the incremental significance of WER improvements. The Matched Pair Sentence Segment Word Error (MAPSSWE) significance test was used [37].

3. DATASET

We trained on 1,200 hours of transcribed audio provided by the 2016 Arabic MGB Challenge. [2] The dataset is composed of 4,000 programs broadcast on the Aljazeera News Channel, spanning 10 years of programming from 2005 to 2015. The transcription is generated from a lightly supervised system, with varying levels of manual annotation. The data is organized into 375,000 utterances containing over 8 million words, and a vocabulary of 200,000 words.[3] Development and evaluation sets were partitioned from the larger set and are 10 hours in duration each. A larger text corpus was also provided, containing over 120 million words, and a vocabulary of 1.4 million words. In addition to the audio, transcriptions, and text, a lexicon was provided. Further details on the dataset can be found here [38].

4. RESULTS

A summary of our results are in Table 2. We found that training with DNNs provided at least a 10% absolute gain in performance when compared with the classical GMM-HMM baseline. LSTM based models provided the best performance (23.6% WER) compared to the feed-forward DNN (25.6%), TDNN (27.1%), and CNN models (29.5%), while the best performing system was the discriminatively trained 5 layer G-LSTM with a WER of 20.1%. The chain TDNN performed as

[2]http://www.mgb-challenge.org/arabic.html
[3]http://alt.qcri.org/MGB_challenge_Arabic_Track_2016/MGB_Arabic_description_2016.pdf

Table 2. Development Set Results of Models

Model	Topology	Features	Alignments	WER (%) $p <$ (prev/base)	WER (%) 4gram $p <$ (prev/base)
GMM-HMM	-	MFCC+LDA+MLLT+FMLLR	-	40.3 (-/-)	-
DNN CE	5x1024	30 Fbank + Pitch	GMM	29.7 (0.001/0.001)	28.1 (0.001/0.001)
CNN	4x2000	80 Fbank + Pitch	GMM	29.5 (0.472/0.001)	28.1 (0.734/0.001)
TDNN	6x3000	80 Fbank + Pitch	GMM	27.1 (0.001/0.001)	25.8 (0.001/0.001)
DNN MPE	5x1024	30 Fbank + Pitch	CE	25.6 (0.001/0.001)	24.7 (0.001/0.001)
Chain TDNN	7x625	80 Fbank + Pitch	GMM	23.6 (0.001/0.001)	23.4 (0.001/0.001)
LSTM	3x1024	80 Fbank + Pitch	CE	23.6 (0.936/0.001)	22.7 (0.001/0.001)
H-LSTM 3L	3x1024	80 Fbank + Pitch	CE	23.3 (0.027/0.001)	22.6 (0.250/0.001)
H-LSTM 5L	5x1024	80 Fbank + Pitch	CE	23.1 (0.055/0.001)	22.4 (0.184/0.001)
G-LSTM 3L	3x1024	80 Fbank + Pitch	CE	22.4 (0.001/0.001)	21.7 (0.001/0.001)
G-LSTM 5L	5x1024	80 Fbank + Pitch	CE	22.2 (0.110/0.001)	21.5 (0.070/0.001)
G-LSTM 3L sMBR	3x1024	80 Fbank + Pitch	CE	20.4 (0.001/0.001)	19.5 (0.001/0.001)
G-LSTM 5L sMBR	5x1024	80 Fbank + Pitch	CE	20.1 (0.009/0.001)	19.2 (0.034/0.001)
Top 2 Combined	**G-LSTM sMBR (3L+ 5L)**	**80 Fbank + Pitch**	**CE**	**-**	**18.3 (0.001/0.001)**

well as the LSTM model (23.6%). We also found that rescoring with a 4-gram model improves performance by 0.2% to 1.6% absolute WER. Although we do not report the numbers, we note that when rescoring with the RNN language models there were no observable improvements in performance. Finally, combining the hypotheses of the top two systems - the 3 and 5 layer G-LSTM sMBR after 4-gram rescoring - yielded the best results with a WER of 18.3%. All results were found to be significant at $p < 0.001$ with respect to the GMM-HMM baseline, while 8 out of the 13 results differed significantly from their lower neighbors at $p < 0.05$.

5. DISCUSSION

We found that models capturing context (LSTMs) with respect to time were superior to other neural network topologies. Although the CNN model only performed as well as the DNN-CE model, there could be other ways to use this model to leverage its strengths. CNNs have been found to be good at extracting feature representations and reducing variance in the frequency domain [39], therefore, it may be better utilized if piped within a hybrid-like DNN topology, as a feature extraction step [40]. The TDNN performed better than the DNN-CE which may be due to the way it captures a wider temporal context at both the input and at deeper layers of the network [15]. Interestingly, the chain TDNN model outperformed the sequence discriminatively trained DNN and performed as well as the LSTM even though it trained on weaker alignments (GMM-HMM versus CE). This highlights the strength and feasibility of sequence discriminative training with a phone level MMI objective function of a neural network with a TDNN topology. Although rescoring with an RNN language model did not help performance, gain from the RNN language model may be achieved with more optimized training parameters, a space which we did not extensively search. The significance of the results (8/13 results with $p < 0.05$ compared to next increase in WER) highlights that each incremental improvement in WER introduced by a different network topology is a significant increase, even if it is a difference of only 0.3% absolute.

6. CONCLUSIONS

We have described the MIT system for Arabic ASR developed on the 1,200 hour dataset of the 2016 Multi-Genre Broadcast Challenge. We evaluated several DNN topologies; Feed-forward, CNN, TDNN, LSTM, H-LSTM, and G-LSTM. We found that models capturing temporal context (LSTMs) out-performed all other models, with sequence discriminative training (chain, sMBR) showing strength. A discriminatively trained 5 layer G-LSTM was the best performing acoustic model, with a WER of 19.2% ($p < 0.05$) after 4-gram language model rescoring. The absolute best performance achieved was 18.3% ($p < 0.001$) WER with a system combination of the top two hypotheses from the sequence trained G-LSTM models.

7. ACKNOWLEDGEMENTS

Tuka AlHanai thanks AlNokhba Scholarship and the Abu Dhabi Education Council for sponsoring her studies. This research was also sponsored by the Qatar Computing Research Institute (QCRI). The authors thank Sameer Khurana and Ahmed Ali of QCRI for their input and suggestions during this work.

8. REFERENCES

[1] Abdel-rahman Mohamed, George Dahl, and Geoffrey Hinton, "Deep belief networks for phone recognition," in *NIPS workshop on deep learning for speech recognition and related applications*, 2009, vol. 1, p. 39.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Li Deng, Geoffrey Hinton, and Brian Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8599–8603.

[4] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] Alex Graves, Douglas Eck, Nicole Beringer, and Juergen Schmidhuber, "Biologically plausible speech recognition with LSTM neural nets," in *Biologically Inspired Approaches to Advanced Information Technology*, pp. 127–136. Springer, 2004.

[6] Li Deng, Ossama Abdel-Hamid, and Dong Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6669–6673.

[7] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back, "Face recognition: A convolutional neural-network approach," *Neural Networks, IEEE Transactions on*, vol. 8, no. 1, pp. 98–113, 1997.

[10] Dan C Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011, vol. 22, p. 1237.

[11] Kaisheng Yao, Trevor Cohn, Katerina Vylomova, Kevin Duh, and Chris Dyer, "Depth-gated recurrent neural networks," *arXiv preprint arXiv:1508.03790*, 2015.

[12] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass, "Highway long short-term memory RNNs for distant speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5755–5759.

[13] Wei-Ning Hsu, Yu Zhang, Ann Lee, and James Glass, "Exploiting depth and highway connections inconvolutional recurrent deep neural networks for speech recognition.," in *Interspeech*, 2016, pp. 395–399.

[14] Nal Kalchbrenner, Ivo Danihelka, and Alex Graves, "Grid long short-term memory," *arXiv preprint arXiv:1507.01526*, 2015.

[15] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," .

[16] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahrmani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free mmi," .

[17] Fadi Biadsy, Nizar Habash, and Julia Hirschberg, "Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2009, NAACL '09, pp. 397–405, Association for Computational Linguistics.

[18] Patrick Cardinal, Ahmed Ali, Najim Dehak, Yu Zhang, Tuka Al Hanai, Yifan Zhang, James Glass, and Stephan Vogel, "Recent advances in ASR applied to an arabic transcription system for al-jazeera," 2014.

[19] Jayadev Billa, Mohamed Noamany, Amit Srivastava, John Makhoul, and Francis Kubala, "Arabic speech and text in TIDES OnTAP," in *Proceedings of the second international conference on Human Language Technology Research*, San Francisco, CA, USA, 2002, HLT '02, pp. 7–11, Morgan Kaufmann Publishers Inc.

[20] Mohamed Afify, Long Nguyen, Bing Xiang, Sherif Abdou, and John Makhoul, "Recent progress in Arabic broadcast news transcription at BBN," in *INTERSPEECH'05*, 2005, pp. 1637–1640.

[21] Samuel Thomas, George Saon, Hong-Kwang Kuo, and Lidia Mangu, "The IBM BOLT speech transcription system," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] Abdelkhalek Messaoudi, Lori Lamel, and Jean-Luc Gauvain, "Modeling vowels for Arabic BN transcription.," in *INTERSPEECH*, 2005, pp. 1633–1636.

[23] A. Messaoudi, J.-L. Gauvain, and L. Lamel, "Arabic Broadcast News Transcription Using a One Million Word Vocalized Vocabulary," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, may 2006, vol. 1, p. I.

[24] Bing Xiang, Kham Nguyen, Long Nguyen, Richard Schwartz, and John Makhoul, "Morphological decomposition for Arabic broadcast news transcription," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[25] Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and James Glass, "A complete kaldi recipe for building arabic speech recognition systems," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 525–529.

[26] Hassan Al-Haj, Roger Hsiao, Ian Lane, Alan W Black, and Alex Waibel, "Pronunciation modeling for dialectal Arabic speech recognition," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 525–528.

[27] Dimitra Vergyri, Arindam Mandal, Wen Wang, Andreas Stolcke, Jing Zheng, Martin Graciarena, David Rybach, Christian Gollan, Ralf Schlüter, Katrin Kirchhoff, Arlo Faria, and Nelson Morgan, "Development of the SRI/nightingale Arabic ASR system," in *INTERSPEECH*, 2008, pp. 1437–1440.

[28] Amr El-Desoky Mousa, Ralf Schlüter, and Hermann Ney, "Investigations on the use of morpheme level features in Language Models for Arabic LVCSR," in *ICASSP*, 2012, pp. 5021–5024.

[29] Tim Ng, Kham Nguyen, Rabih Zbib, and Long Nguyen, "Improved morphological decomposition for Arabic broadcast news transcription," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, april 2009, pp. 4309 –4312.

[30] L. Mangu, Hong-Kwang Kuo, S. Chu, B. Kingsbury, G. Saon, Hagen Soltau, and F. Biadsy, "The IBM 2011 GALE Arabic speech transcription system," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, 2011, pp. 272–277.

[31] "LDC2013S02, GALE Phase 2 Arabic Broadcast Conversation Speech Part 1 Linguistic Data Consortium," 2013.

[32] "LDC2013S07, GALE Phase 2 Arabic Broadcast Conversation Speech Part 2 Linguistic Data Consortium," 2013.

[33] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[34] Dong Yu, Adam Eversole, Mike Seltzer, Kaisheng Yao, Zhiheng Huang, Brian Guenter, Oleksii Kuchaiev, Yu Zhang, Frank Seide, Huaming Wang, et al., "An introduction to computational networks and the computational network toolkit," Tech. Rep.

[35] Andreas Stolcke et al., "Srilm-an extensible language modeling toolkit.," .

[36] Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

[37] David S Pallet, William M Fisher, and Jonathan G Fiscus, "Tools for the analysis of benchmark speech recognition tests," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. IEEE, 1990, pp. 97–100.

[38] Ahmed Ali, Peter bell, James Glass, Yacine Messaouie, Hamdy Mubarak, Steve Renals, and Yifan Zhang, "The MGB-2 Challenge: Arabic Dialect Multi-Broadcast Media Recognition," 2016.

[39] Tara N Sainath, Brian Kingsbury, Abdel-rahman Mohamed, and Bhuvana Ramabhadran, "Learning filter banks within a deep neural network framework," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 297–302.

[40] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4580–4584.