AUTOMATIC SPEECH RECOGNITION OF ARABIC MULTI-GENRE BROADCAST MEDIA

Maryam Najafian [1]*, Wei-Ning Hsu* [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]
{najafian, wnhsu, glass}@mit.edu [1], amali@qf.org.qa [2]

ABSTRACT

This paper describes an Arabic Automatic Speech Recognition system developed on 15 hours of Multi-Genre Broadcast (MGB-3) data from YouTube, plus 1,200 hours of Multi-Dialect and Multi-Genre MGB-2 data recorded from the Aljazeera Arabic TV channel. In this paper, we report our investigations of a range of signal pre-processing, data augmentation, topic-specific language model adaptation, accent specific re-training, and deep learning based acoustic modeling topologies, such as feed-forward Deep Neural Networks (DNNs), Time-delay Neural Networks (TDNNs), Long Short-term Memory (LSTM) networks, Bidirectional LSTMs (BLSTMs), and a Bidirectional version of the Prioritized Grid LSTM (BPGLSTM) model. We propose a system combination for three purely sequence trained recognition systems based on lattice-free maximum mutual information, 4-gram language model re-scoring, and system combination using the minimum Bayes risk decoding criterion. The best word error rate we obtained on the MGB-3 Arabic development set using a 4-gram re-scoring strategy is 42.25% for a chain BLSTM system, compared to 65.44% baseline for a DNN system.

Index Terms— Speech recognition, RNNs, Acoustic mis-match, multi-dialect, multi-genre

1. INTRODUCTION

There are a number of major challenges associated with automatic speech recognition of conversational multi-genre broadcasts, including background noise variation, cross-talk, talker dialects, and transcriber inconsistency of reference transcripts [1, 2]. The goal of the Arabic MGB-3 challenge for the multi-genre speech transcription task is to further the state-of-the-art in Arabic speech recognition, which is a challenging task considering that Arabic dialectal variation is inherent in most real life speech transcription applications [3].

In this paper, we present a speech transcription system we developed for the Arabic MGB-3 ASR task that investigated a wide range of techniques. For acoustic modeling, we examined a feed-forward deep neural network (DNN), a 'chain' Time-Delay Neural Network (TDNN) [4], a 'chain' Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN) [5], a 'chain' Bi-directional LSTM (BLSTM) model using the lattice-free maximum mutual information (LF-MMI) framework [4]. We also examined the Prioritized Grid LSTM (PGLSTM) model that has achieved good results on other conversational speech tasks compared to other LSTM model variations [2, 6, 7]. Specifically, we investigate a Bi-directional version of the PGLSTM (BPGLSTM). We also report the effect of a range of signal preprocessing, data augmentation, accent-specific retraining, topic specific language model training, acoustic model combination using Minimum Bayes Risk (MBR) criterion [8], and language model re-scoring. Our best system achieves a 42.25% word error rate (WER) on the MGB-3 Arabic development set, which is a more than 20% absolute error reduction over a DNN baseline.

2. DATA DESCRIPTION

The five hours of MGB-3 training data (which was called adaptation data) is not enough by itself to build a robust Arabic speech recognition system. Therefore, we created models using both the MGB-3 5 hours and 1,200 hours of MGB-2 data. The MGB-2 data was recorded from Aljazeera Arabic TV channels, and contains a total of 1,200 hours of audio from 19 different programs. The MGB-2 content can be split into three broad categories: conversational (63%), where a presenter talks with more than one guest discussing current affairs; interview (19%), where a presenter speaks with one guest; and report (18%), such as news or documentary. The MGB-2 recordings originate from TV programs with Modern Standard Arabic (MSA) dominating most of them. It is estimated that more than 70% of the MGB-2 speech is MSA, and the rest is spoken in different Dialectal Arabic (DA) namely: Egyptian (EGY), Gulf (GLF), Levantine (LEV), and North African (NOR). The output of a speech recognizer was aligned with the original transcription to generate small speech segments on average between five and 30 seconds per segment that were suitable for building a speech recognizer.

The MGB-3 data contains Egyptian broadcast data collected from 80 programs from different YouTube channels. All programs were transcribed by four different annotators to explore the non-orthographic nature of dialectal Arabic. MGB-3 content can be split into seven broad genres; namely comedy, fashion, sports, cooking, family, movies, and sci-

ence. Using a total of 15 hours of Arabic speech recorded from YouTube, MGB-3 data is divided into adaptation (12 minutes * 24 programs), development (12 minutes * 24 programs), and evaluation (12 minutes * 31 programs) sets. A transcription was provided for the MGB-3 adaptation and development sets, while no transcription is available for the evaluation data.

3. SYSTEM DESCRIPTION

Previous studies showed that neural network models capturing temporal context. RNNs incorporate feedback cycles into the network architecture, which leads to better modeling of sequences. There are many implementations of RNNs, such as LSTMs, and Gated Recurrent Units (GRUs). LSTM success in acoustic modeling can be explained by their strength in memorizing sequences with long range temporal dependencies. LSTMs are easy to train, and do not suffer from the exploding gradient problems when performing back-propagation-through-time. The LSTM blocks in the hidden layers of RNNs consist of input, output, and forget gates that control the flow of input information from the previous hidden layer and the output information to be passed on to the next layer. Sequence training of neural networks using the Connectionist Temporal Classification (CTC) training objective is a common trend in ASR. In this work, we train chain LSTM and BLSTM acoustic models with the Lattice Free version of the Maximum Mutual Information training criterion (LF-MMI) [9] modeling framework.

3.1. Data Pre-processing

The SOX toolkit is applied to trim and normalize the MGB-2 and MGB-3 audio recordings that have been recorded using different microphones, and with differing background noise. Initially, a high-pass filter with a cutoff frequency of 100Hz is applied to the signal to remove any DC offset. Then we apply a companding procedure during the signal pre-processing. The algorithm mimics tone-to-tone suppression and masking in the auditory system to improve automatic speech recognition performance in noise using the following setup with the following SOX companding option $[$ *compand* $0.05, 0.26$: $-54, -90, -36, -36, -24, -24, 0, -120 - 900.1]$. The attack and decay parameters (in seconds) determine the time over which the instantaneous level of the input signal is averaged to determine its volume; attacks refer to increases in volume and decays refer to decreases. For most situations, the attack time (response to the music getting louder) should be shorter than the decay time because the human ear is more sensitive to sudden loud music than sudden soft music. Our input channel is companded separately with values of 0.05 and 0.2 seconds. We defined a list of points on the companders transfer function specified in dB relative to the maximum possible signal amplitude. The input values must be in a strictly increasing order but the transfer function does not have to be monotonically rising. Then we preceded by a soft-knee-dB value, and the points at where adjacent line segments on the transfer function meet will be rounded by the amount given. We applied an additional gain in dB which is applied at all points on the transfer function which allows easy adjustment of the overall gain. Our values for the transfer function are $[6 : -54, -90, -36, -36, -24, -24, 0, -12]$. The '6:' selects 6dB soft-knee companding. The 0 (dB) output gain is needed to avoid clipping (the number is inexact, and was derived by experimentation). The input signal is analyzed immediately to control the compander, but it is delayed before being fed to the volume adjuster. Specifying a delay approximately equal to the attack/decay times allows the compander to effectively operate in a 'predictive' rather than a reactive mode. We selected a value of 0.2 seconds. The 90 (dB) for the initial volume will work fine for a clip that starts with near silence, and the delay of 0.1 (seconds) has the effect of causing the compander to react a bit more quickly to sudden volume changes.

3.2. Data Augmentation

Previous studies showed great improvement in accuracy using audio augmentation and perturbation [10]. We performed audio speed and volume perturbation with speed factors that are uniformly sampled from the interval [0.85 , 1.3]. The speed perturbed data is followed by volume perturbation with volume factors that are uniformly sampled from the interval [0.1 , 3.0]. We apply 40 rounds of data augmentation to the MGB-3 data in the aforementioned intervals. This gives us 40 times the original speech utterances from MGB3.

3.3. Language Modelling

The organizers provided a trigram and 4-gram language model trained on MGB-2 data, and some additional data. In order to familiarize the language model with the MGB-3 data we developed a trigram and 4-gram language model (LM) trained using the MGB-2 and MGB-3 transcripts (excluding the development and test data). The trigram is used for decoding to generate decode lattices. A 4-gram LM is then used for re-scoring the lattices. We use interpolated Kneser-Ney smoothing on both the LMs, which are built using the SRILM toolkit [11]. A previous study on the MGB challenge showed that using an RNN based language model [12] using limited data will not lead to an improvement in accuracy [13]. Therefore, we did not explore that option.

3.4. Lexicon

Arabic is a phonologically complex language [14], and using a grapheme-based lexicon to reduce out-of-vocabulary (OOV) words can be an effective strategy, especially for multi-dialect Arabic speech. In our experiments we used the

grapheme-based lexicon in BuckWalter format provided by the organizer, and added the extra vocabulary from MGB-3 for our experiments. The grapheme based lexicon has a one-to-one word-to-grapheme mapping, which means that the vocabulary size is the same as the lexicon size.

3.5. Feature Extraction and Acoustic Modelling

This section presents the details of the neural network acoustic modelling approaches, the architectures, the hyperparameter settings, and the input features used for developing these models.

3.5.1. Alignment Generation

We built a baseline recognizer using Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). These were trained using 39-dimension Mel Frequency Cepstral Coefficients (MFCC) derived features (i.e., 13 static MFCCs with delta and delta-delta MFCCs) that were transformed using Linear Discriminant Analysis (LDA), a Maximum Log-Likelihood Transform (MLLT) [15], and feature space Maximum Likelihood Linear Regression (fMLLR). Senone-based frame-level alignments generated from the GMM-HMM model were used to train a variety of neural network based models. Five models were trained using alignments generated by the GMM-HMM model, namely a feed forward DNN, BPGLSTM (5x1024), a sequence discriminatively trained chain Time-Delay Neural Network (TDNN) model (7x625), and a 7 layer LSTM and BLSTM.

3.5.2. Feature Extraction

For the BPGLSTM model, we use a 30 dimensional filter bank features without splicing. For the remaining models we found that the high resolution MFCC features outperform the filter bank features. All the remaining models are trained using concatenated 40 dimensional high resolution MFCC features, and 100 dimensional i-vectors for each frame, unless mentioned otherwise [16]. The i-vector extractor is trained on top of features that are not mean-normalized, so that the mean offset information can be encoded in the i-vectors [17]. This eliminates the need for adaptation or normalization of the MFCC input features to our systems [18]. Once the i-vector extractor is trained, i-vectors for the training and test data are extracted in an on-line fashion. During this stage, only prior frames to the current frame are used along with the prior utterances from the same speaker to extract the i-vectors. The i-vector extraction framework consists of a GMM Universal Background Model (UBM) trained on LDA+MLLT-transformed MFCCs, that consist of 512 GMM components, and that makes use of 300k feature frames. The UBM first-order statistics are then modeled using a factor analysis model known as total variability subspace model[17]. The parameters of the model are learned in an unsupervised manner. In this work, variability subspace parameter was set to 100 i-vector dimensions. All neural network based models used spliced features of width 5 (window of +/- 5 frames) as input, unless stated otherwise.

3.5.3. BPGLSTM Structure

The BPGLSTM models arrange LSTM blocks into multidimensional grids such that each grid contains one set of LSTM blocks for each dimension, including the depth dimension. This architecture introduces per-dimension gated linear dependencies between adjacent cell states, which mitigates the vanishing gradient problem along all dimensions. The best PGLSTM architecture reported in [6] has 5 layers, and each layer contains 1,024 memory cells along with a 512-node linear projection layer. To keep the number of parameters comparable, the BPGLSTM model we consider here also has 5 layers, but each LSTM now contains 512 memory cells along with a 300-node linear projection layer added on top of each LSTM output.

The Computational Network Toolkit (CNTK) [19] is used for BPGLSTM training. As [20] suggests, all weights are randomly initialized from the uniform distribution with range $[-0.05, 0.05]$, and all biases are initialized to 0, without generative or discriminative pre-training [21]. The model is trained with a cross-entropy (CE) criterion, using latency-controlled back-propagation-through-time (BPTT) [22] for optimization, where each BPTT segment contains 80 frames, with additional 22 future frames used to provide the right context. 40 utterances are parallelized in each mini-batch. No momentum is used for the first epoch, and a momentum of 0.9 is used for subsequent epochs [23]. $L2$ constraint regularization [24] with weight 10^{-5} is applied.

3.5.4. Chain LSTM and BLSTM Structures

The chain LSTM model is composed of a total of 7 recurrent and non-recurrent projection layers, as described previously [5]. The spliced indexes at the different layers were [{-3,-2,-1,0,1,2,3}; {0}; {0}; {0}; {0}; {0}; {0}], and the delay at each layer is chosen to be -3, and the output label delay is 5. The time delay introduces a delay between the inputs and the targets. Providing the network with a few time-steps of future context can have a positive impact on the robust training process, since it provides short distortions, especially when it is used with LSTMs. Both the recurrent, and the non-recurrent projection dimensions are set to 256, and the training process is repeated for 6 epochs. Purely sequence trained models are trained using a sequence objective, without the need for Cross Entropy training.

The acoustic model architecture for the chain bidirectional LSTM (BLSTM) is the same as that of the chain LSTM, except that the training occurs in both forward and backward directions.

3.5.5. Chain TDNN Structure

The chain TDNN model is composed of 7 layers with 725 Rectified Linear units (ReLUs) 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. It require less training time than sequence models such as LSTMs, while attempting to capture the long-term temporal dependencies that a sequence model is capable of doing.

3.5.6. DNN Structure

The FDNN model has 7 layers, each layer having 1024 sigmoidal neurons. Input to the FDNN is a 40 dimensional transformed MFCC feature vector. 9 frames of 13 dimensional MFCC feature vectors are spliced together, mean normalized, and reduced to a 40 dimensional representation using LDA, followed by Maximum Likelihood Linear Transform (MLLT) [15], and Maximum Likelihood Linear Regression (fMLLR) for feature level speaker adaptation and normalization. The fMLLR transform is obtained from a baseline GMM-HMM system with speaker adaptive training (SAT) [25]. The output of the FDNN is a softmax layer, whose units correspond to triphone states. A baseline GMM-HMM system provides frame-level HMM-state alignments that are used as training examples in a multi-class classification setting. The FDNN is trained to minimize the Cross Entropy loss function using Stochastic Gradient Descent (SGD). We use a learning rate of 0.008 for SGD for the first epoch and for later epochs, the learning rate is decided using the "new-bob" algorithm [26]. Mini-batches of size 256 are used during the training stage.

3.5.7. Model Combination

Different models complement each other in generating a hypothesis transcript. Therefore, we exploit a round of model combination on our most successful systems. This was done using lattice combination and a hypothesis scoring method using Minimum Bayes Risk (MBR) to minimize the expected WER [8].

3.6. Measurement of Word Error Rate (WER)

There is varying qualities across the transcriptions provided by multiple transcribers. The word error rates (WERs) reported in this work average the % accuracy across the transcriptions provided by all four transcribers. The official results by the organizers is reported using multi-Reference Word Error Rate (MR-WER) [27]. They Also reported the average WER (AV-WER) across the four transcribers for the test set.

4. EXPERIMENTAL RESULTS

Previous studies showed that neural network models capturing temporal context at acoustic and phonemic level outperformed all other models in multi-dialect ASR [2, 28, 29, 30], dialect Identification [31, 32, 33], speaker diarization [34, 35], phone classification [36], and acoustic-physiological measurements [37]. In this section we explore their strength in speech recognition in presence of dialectal speech, genre and channel mismatch.

We explore different approaches to address the dialect, acoustic background, and topic mismatch between the training and the test set at the pre-processing, feature extraction, acoustic modelling, and language modelling stages of the speech recognition process. The Computational Network Toolkit (CNTK) [19], Kaldi speech recognition [38], and SRILM language modelling [11] toolkits were used for this research. Our recipe will be made available for the MGB-3 challenge via a GitHub repository, and the databases are accessible through the challenge website.

4.1. Effect of Data Pre-processing and Augmentation

The baseline WER on the MGB-3 development set using a DNN based system trained on 5 hours of the MGB-3 adaptation set, and 1,200 hours of the MGB-2 data is 79.8%. After applying data pre-processing to the MGB-2 and the MGB-3 data, and 40 rounds of speed followed by volume perturbation to the 5 hours of data from MGB-3 adaptation set, this WER is reduced to 72.2%.

4.2. Adapting Lexicon and Language Model to MGB-3

The original language model provided by the organizers was created with the MGB-2 data and some additional data. When we augmented the language model and lexicon with the MGB-3 adaptation data, the DNN baseline WER was reduced from 72.2% to 65.4%. We consider the 65.4% WER to be our baseline result, as we explore other modeling methods.

4.3. Alternative Neural Network Acoustic Models

In this section we show how using more complex deep learning strategies has improved our results. In these experiments we used the data augmentation and pre-processing procedures, and we used the MGB-3 modified lexicon and language model. The WER results of the DNN, TDNN, LSTM, BLSTM, and BPGLSTM are shown in Table 1.

Model	DNN	TDNN	LSTM	BLSTM	BPGLSTM
WER (%)	65.44	53.53	51.90	44.89	42.95

Table 1: *WERs of DNN, TDNN, LSTM, and BLSTM models.*

4.4. Dialect Specific Retraining

The MGB-3 data consists of Egyptian dialect data while the MGB-2 data consists of multiple Arabic dialects. In this experiment we applied a dialect identification system presented in [39] to the MGB-2 data, and selected the utterances considered to be the Egyptian dialect, in addition to the MGB-3 adaptation data for accent specific retraining in our systems (In total 40 hours). Unfortunately, as shown in Table 2, dialect specific training degraded WER performance for all models. We suspect that the reason might be due to the small amount of dialect specific data, or acoustic mismatch between the MGB-2 and MGB-3 data.

System	DNN	TDNN	LSTM	BLSTM
Baseline (No re-training)	65.44	53.53	51.90	44.89
Dialect specific retraining	66.28	56.00	54.61	49.53

Table 2: *WERs after the dialect specific retraining.*

4.5. Topic Specific Language Models

MGB-3 content can be partitioned into seven broad genres; namely comedy, fashion, sports, cooking, family, movies, and science. We attempted to create topic specific language models by interpolating the language models created using MGB-3 plus MGB-2 data, with MGB-3 topic specific data. As shown in Tables 3 and 4, topic specific language models did not improve the WERs across a majority of topics. We suspect that this is due to the small amount of topic specific material in the MGB-3 adaptations set (approximately 5 hours).

Topics	TDNN	LSTM
Comedy	55.61	53.04
Cooking	54.29	51.62
Family	42.69	40.42
Fashion	68.66	67.74
Movies	66.06	63.31
Science	51.68	49.53
Sport	51.14	49.93

Table 3: *WERs with topic specific language models.*

4.6. Language Model Re-scoring

Up to this stage we have used 3-gram language models in all our experiments. This section reports the WER after applying 4-gram re-scoring to the development set, which leads a WER improvement across all acoustic models, as shown in Table 5.

4.7. Acoustic Model Combination with LF-MMI

In this section we report the result after combining the chain LSTM and BLSTM acoustic models using the Lattice free Maximum Mutual Information (LF-MMI) approach. More details on the LF-MMI training objective can be found in [9]. Unfortunately, we were unable to improve the WER using model combination. We believe this result might be because there was not enough difference between the two acoustic models, although this is clearly an area that needs further investigation.

Topics	TDNN	LSTM
Comedy	52.65	49.15
Cooking	51.28	50.93
Family	40.51	37.45
Fashion	68.27	63.95
Movies	63.90	60.22
Science	48.87	46.56
Sport	49.87	47.23

Table 4: *WERs without topic specific language models.*

Language Model	TDNN	LSTM	BLSTM	BPGLSTM
Baseline (3-gram)	53.53	51.90	44.89	42.95
4-gram re-scoring	52.13	48.53	42.25	42.64

Table 5: *WERs after 4-gram re-scoring of acoustic models.*

Language Model	TDNN	LSTM	BLSTM	Combination
3-gram	53.53	51.90	44.89	50.58
4-gram	52.13	48.53	42.25	47.18

Table 6: *Effect of LF-MMI system combination.*

4.8. Topic-Specific Performance of Best Overall System

In previous sections we explored different strategies to reduce the multi-conditional data problem and language mismatch between the development set and the training data. This section reports the topic specific WERs of our best overall system, which was achieved after applying data signal preprocessing, data augmentation, and 4-gram re-scoring using the chain BLSTM structure. As shown in Table 7, this led to an overall 42.25% WER on the development set data. We can also see that the WERs varied considerably depending on the genre of the broadcast, ranging from a low of 31.90% WER for family shows, to a high of 53.89% WER for fashion shows.

5. CONCLUSIONS

In this paper we describe the ASR systems we have investigated as part of the MGB-3 Arabic ASR speech transcription challenge. In particular, we examined the recently introduced LF-MMI modeling framework, and achieved the our best overall WER on the development set of 42.25%. In order to

Topic	BLSTM WER (%)	BPGLSTM WER (%)
Comedy	42.70	42.58
Cooking	42.30	41.86
Family	31.90	29.50
Fashion	53.89	54.40
Movies	52.48	53.39
Science	38.29	38.35
Sports	39.05	39.38
All	**42.25**	42.64

Table 7: *Topic-specific WER of best overall ASR system.*

be able to apply a range of complex deep learning algorithms, and address the limited data problem we decided to exploit the 1,200 hours of MGB-2 data from the Aljazeera Arabic TV channel, in addition to the original 5 hours of adaptation data and 5 hours of development data from YouTube. To address the acoustic mis-match between the MGB-2 and MGB-3 we applied multiple rounds of speed and volume perturbation to the MGB-3 data. Next, to address the multi-genre problem we created a genre-specific language model created as a result of interpolation of the MGB-2, and MGB-3 genre specific language model that was created for each topic. We used a dialect identification system to enable dialect-dependent acoustic model retraining to address dialectal mis-match between the training and test data. We used a grapheme-based lexicon provided for this task and added missing words from MGB-3 transcripts.

The best WER was achieved using a chain BLSTM. This system is trained using in total 1,400 hours of data comprising 1,200 hours of MGB-2 data and 5 hours of MGB-3 adaptation data, after a round of data pre-processing and 40 rounds of speed and volume perturbation. This model is re-scored using a 4-gram language model. We did not manage to gain any improvement in accuracy through exploiting dialect specific re-training, and topic specific language modeling, so these remain issues for future investigations. The official results for the MGB-3 are 36.8% MR-WER and 44.9% average WER.

6. REFERENCES

[1] Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," in *SLT*, 2016, pp. 279–284.

[2] Tuka AlHanai, Wei-Ning Hsu, and James Glass, "Development of the MIT ASR system for the 2016 Arabic multi-genre broadcast challenge," in *SLT*. IEEE, 2016, pp. 299–304.

[3] Ahmed Ali, Stephan Vogel, and Steve Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *ASRU*. IEEE, 2017.

[4] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free MMI.," in *INTERSPEECH*, 2016, pp. 2751–2755.

[5] Hasim Sak, Andrew W Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling.," in *INTERSPEECH*, 2014, pp. 338–342.

[6] Wei-Ning Hsu, Yu Zhang, and James Glass, "A prioritized grid long short-term memory RNN for speech recognition," in *SLT*, 2016, pp. 467–473.

[7] 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.

[8] Shankar Kumar and William Byrne, "Minimum bayes-risk decoding for statistical machine translation," Tech. Rep., CLSP, 2004.

[9] Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks.," in *INTERSPEECH*, 2013, pp. 2345–2349.

[10] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[11] Andreas Stolcke et al., "SRILM-an extensible language modeling toolkit.," in *INTERSPEECH*, 2002.

[12] Mikolov Tomas, Deoras Anoop, Kombrink Stefan, Burget Lukas, and H Gernocky Jan, "RNNLM-recurrent neural network language modeling toolkit," in *ASRU*, 2011.

[13] Sameer Khurana and Ahmed Ali, "QCRI advanced transcription system (QATS) for the arabic multi-dialect broadcast media recognition: MGB-2 challenge," in *SLT*, 2016, pp. 292–298.

[14] Arfath Pasha, Mohamed Al-Badrashiny, Mona T Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic.," in *LREC*, 2014, vol. 14, pp. 1094–1101.

[15] Ramesh A Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *IEEE transactions on acoustics, speech, and signal processing*, 1998, vol. 2, pp. 661–664.

[16] Steven Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[17] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[18] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *INTERSPEECH*, 2014.

[19] 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., Tech. Rep. MSR, Microsoft Research, 2014, http://codebox/cntk, 2014.

[20] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *ICASSP*, 2015, pp. 4580–4584.

[21] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011, pp. 24–29.

[22] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass, "Highway long short-term memory RNNs for distant speech recognition," in *ICASSP*, 2016, pp. 5755–5759.

[23] Yu Zhang, Dong Yu, Michael L Seltzer, and Jasha Droppo, "Speech recognition with prediction-adaptation-correction recurrent neural networks," in *ICASSP*, 2015, pp. 5004–5008.

[24] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[25] Spyros Matsoukas, Rich Schwartz, Hubert Jin, and Long Nguyen, "Practical implementations of speaker-adaptive training," in *DARPA Speech Recognition Workshop*, 1997.

[26] Shakti P Rath, Daniel Povey, Karel Veselỳ, and Jan Cernockỳ, "Improved feature processing for deep neural networks.," in *INTERSPEECH*, 2013, pp. 109–113.

[27] Ahmed Ali, Walid Magdy, Peter Bell, and Steve Renals, "Multi-reference WER for evaluating ASR for languages with no orthographic rules," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 576–580.

[28] Maryam Najafian, Andrea DeMarco, Stephen J. Cox, and Martin J. Russell, "Unsupervised model selection for recognition of regional accented speech," in *INTERSPEECH*, 2014, pp. 2967–2971.

[29] Maryam Najafian, Saeid Safavi, John HL Hansen, and Martin Russell, "Improving speech recognition using limited accent diverse british english training data with deep neural networks," in *MLSP*. IEEE, 2016, pp. 1–6.

[30] Maryam Najafian, Saeid Safavi, Abualsoud Hanani, and Martin J. Russell, "Acoustic model selection using limited data for accent robust speech recognition," in *EUSIPCO*, 2014, pp. 1786–1790.

[31] Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor, "Identifying dialects with textual and acoustic cues," *VarDial 2017*, p. 93, 2017.

[32] Maryam Najafian, Saeid Safavi, Philip Weber, and Martin J. Russell, "Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems," in *ODYSSEY*, 2016, pp. 1–6.

[33] Abualsoud Hanani, Aziz Qaroush, and Stephen Taylor, "Classifying asr transcriptions according to arabic dialect," *VarDial 3*, p. 126, 2016.

[34] Maryam Najafian and John HL Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *SLT*. IEEE, 2016, pp. 114–120.

[35] Maryam Najafian and John HL Hansen, "Environment aware speaker diarization for moving targets using parallel dnn-based recognizers," in *ICASSP*. IEEE, 2017, pp. 5450–5454.

[36] Linxue Bai, Peter Jančovič, Martin Russell, Philip Weber, and Steve Houghton, "Phone classification using a non-linear manifold with broad phone class dependent dnns," *INTERSPEECH*, pp. 319–323, 2017.

[37] Mehdi Shokoueinejad, Chris Fernandez, Carroll, et al., "Sleep apnea: a review of diagnostic sensors, algorithms, and therapies," *Physiological Measurement*, vol. 38, no. 9, pp. R204–R252, 2017.

[38] 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 *ASRU*, 2011.

[39] Sameer Khurana, Maryam Najafian, Tuka Ali Ahmed and, Al Hanai, Yonatan Belinkov, and Jim Glass, "QMDIS: QCRI-MIT advanced dialect identification system," in *INTERSPEECH*, 2017.