A COMPLETE KALDI RECIPE FOR BUILDING ARABIC SPEECH RECOGNITION SYSTEMS

Ahmed Ali[1], *Yifan Zhang*[1], *Patrick Cardinal*[2], *Najim Dahak*[2],

Stephan Vogel[1], *James Glass*[2]

[1] Qatar Computing Research Institute
[2] MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, Massachusetts 02139, USA

amali@qf.org.qa, yzhang@qf.org.qa, patrick.cardinal@csail.mit.edu
najim@csail.mit.edu, svogel@qf.org.qa, glass@mit.edu

ABSTRACT

In this paper we present a recipe and language resources for training and testing Arabic speech recognition systems using the KALDI toolkit. We built a prototype broadcast news system using 200 hours GALE data that is publicly available through LDC. We describe in detail the decisions made in building the system: using the MADA toolkit for text normalization and vowelization; why we use 36 phonemes; how we generate pronunciations; how we build the language model. We report results using state-of-the-art modeling and decoding techniques. The scripts are released through KALDI and resources are made available on QCRI's language resources web portal. This is the first effort to share reproducible sizable training and testing results on MSA system.

Index Terms: Arabic, ASR system, lexicon, KALDI, GALE

1. Introduction

Arabic Automatic Speech Recognition (ASR) is challenging because of the lexical variety and data sparseness of the language. Arabic can be considered as one of the most morphologically complex languages. Reducing the entry barrier to build robust Automatic Speech Recognition (ASR) for Arabic has been a research concern over the past decade[1]–[4].Unlike American English, for example, which has CMU dictionary, standard KALDI scripts available, Arabic language has no freely available resource for researchers to start working on ASR systems. To build an Arabic ASR system, a researcher will need not only to understand the technical details, but also to have the language expertise, which is a barrier for many people. This has been the main motivation for us to release, and share with the community, all the needed bits and pieces, including code, experiential results as well as the required resources to get an Arabic ASR system with reasonable WER in short time. Researchers who are interested in building a baseline Arabic ASR system can use it as reference. This work was developed in parallel to Al-Jazeera system [5].

The main motivation to use KALDI Speech Recognition toolkit[6] is that, it has attracted speech researchers, and it has been very actively developed over the past few years. Furthermost, most of the state-of-the-art techniques have already been implemented, and heavily used by the research community. KALDI is released under the Apache license v2.0, which is flexible and fairly open license. Recipes for training ASR systems with many speech corpuses have been made available and frequently updated with the latest techniques, such as Bottle-Neck Features (BNF), Deep Neural Networks (DNN), etc.

In this paper, we describe our ASR system for Modern Standard Arabic (MSA) built using the 200 hours broadcast news database GALE phase 2 [7] released by LDC.

The scripts as well as the lexicon used to reproduce the reported results are made available on QCRI's language resource web portal[1]. This includes all the intermediate results to reach the best reported system. In the following sections, we describe the main design characteristics of our system:

- Acoustic Modeling (AM): we used the state-of-the-art modeling techniques, starting by building GMM systems for the first pass with fMLLR adaption, and for the second pass, we explored various technologies; MMI, MPE, SGMM, DNN and we report the gain obtained in each one of them.

- Data and text pre-processing: We have used the 200 hours data training, which contains two types of speech: Broadcast Conversations (BC) and Broadcast Reports (BR). We use all data from both BC and BR for training, and we report results on each of the BR and BC as well as the combined WER on both of them, we also built two systems using both the original text and the text after being processed with the Morphological Analysis and Disambiguation for Arabic (MADA) [8] toolkit.

[1] http://alt.qcri.org//resources/speech/

- Lexicon: we use the QCRI lexicon, which has 526K unique words, with 2M pronunciation variants, .i.e on average 4 pronunciations per word.

- Language Model (LM): We built standard trigram LMs using the training data transcripts with Kneser-Ney smoothing. The same LM hasbeen throughout the experiments reported in this paper.

The following section describes acoustic data and models; Section 3 describes language data and; Section four discusses the experimental results; Section five concludes the paper with findings and future work.

2. Acoustic Model

This section will describe the acoustic modeling data, the details of our acoustic training and models.

2.1. Data

The LDC released GALE Arabic Broadcast News dataset used in our system consists of 100K speech segments recorded at 16kHz across nine different TV channels, a total of 284 episodes. Overall, there are 203 hours speech data. The dataset is a mix of 76 hours BR, and 127 hours BC. We split the data by episodes: 273 episodes for training (194 hours, 95k segments), and 11 episodes for testing (9 hours, 5k segments). The testset consists of three hours BR and six hours BC data. We split the data this way to make sure that episodes appear in test data will not appear in training, to reduce the chance of having speakers overlapping between training and testing. We report three Word Error Rate (WER) results from the test data: BR, BC, and combined for both of them.

2.2. Acoustic Modeling

[1] has shown that the MADA vowelization based phoneme system is superior comparing to grapheme-based system. For this reason, we built our system to be phoneme based. We experimented with different numbers of phonemes to see how this affects the performance of the system. More details can be found in Section 3.3.

Our models are trained with the standard 13-dimensional cepstral mean-variance normalized (CMVN) Mel-Frequency Cesptral Coefficients (MFCC) features without energy, and its first and second derivatives. For each frame, we also include its neighboring ±4 frames and apply Linear Discriminative Analysis (LDA) transformation to project the concatenated frames to 40 dimensions, followed by Maximum Likelihood Linear Transform (MLLT) [6]. We use this setting of feature extraction for all models trained in our system. Speaker adaptation is also applied with feature-space Maximum Likelihood Linear Regression (fMLLR) [9].

Our system includes all conventional models supported by KALDI: diagonal Gaussian Mixture Models (GMM), subspace GMM (SGMM) and DNN models. Training techniques including discriminative training such as boosted Maximum Mutual Information (bMMI), Minimum Phone Error (MPE), and Sequential Training for DNN are also employed to obtain the best number.

Figure 1 shows the workflow for acoustic training: MFCC features are extracted from speech frames; MFCC+LDA+MLLT are then used to train the Speaker-Independent (SI) GMM model; Two discriminative training method MPE and MMI are used to train two individual models for GMM; fMLLR are estimated based on each training utterance with SI GMM; fMLLR transformed features are used for SGMM training and DNN training separately; and after that SGMM and DNN models are discriminatively trained further with bMMI and Sequential Training techniques respectively.

In the end, we obtain three different sets of models: GMM-HMM based models, SGMM-HMM based models and DNN-HMM based models. The system will use the intermediate basic GMM model for first pass decode to obtain fMLLR transformation, and the second pass decoding with one of the more advanced final models.

These models are all standard 3-states context-dependent triphone models. The GMM-HMM model has about 512K Gaussians for 8K states; the SGMM-HMM model has 5K states and 40K total substates. The DNN-HMM model is trained with 5 layers; each layer has 2K nodes. The DNN training was done using one GPU on a single machine. 10K samples from the training data were held out for cross-validation.

3. Language Model, and Lexicon

The LM was built using GALE training data transcripts with a total of 1.4M words. In this section we explain how we do text normalization, vowelization, pronunciation and language modeling.

3.1. Normalization

Since Arabic is a morphologically rich language, it is common to find discrepancies in written text. We integrate a preprocessing phase which does auto-correction of the raw input text, targeting Common Arabic Mistakes (CAM) [10]. A study of spelling mistakes in Arabic text has been carried out over one thousand articles picked randomly from public Arabic web news sites. A semi-manually tagging procedure for detecting spelling mistakes shows an error rate of 6%,

which is considered very high compared to English [11]. In the case of speech transcription such as the GALE transcripts, the Arabic text could be linguistically wrong if the transcriber is more faithful to the speech than to the grammar. We use MADA for text normalization. When we run MADA to disambiguate words based on their contexts, we notice that it has modified the text as following:

original text	هذا الإلتزام الأخلاقي القوي (h*A Al<ltzAm Al>xlAqy Alqwy)
MADA text	هذا الالتزام الأخلاقي القوى (h*A AlAltzAm Al>xlAqy AlqwY)

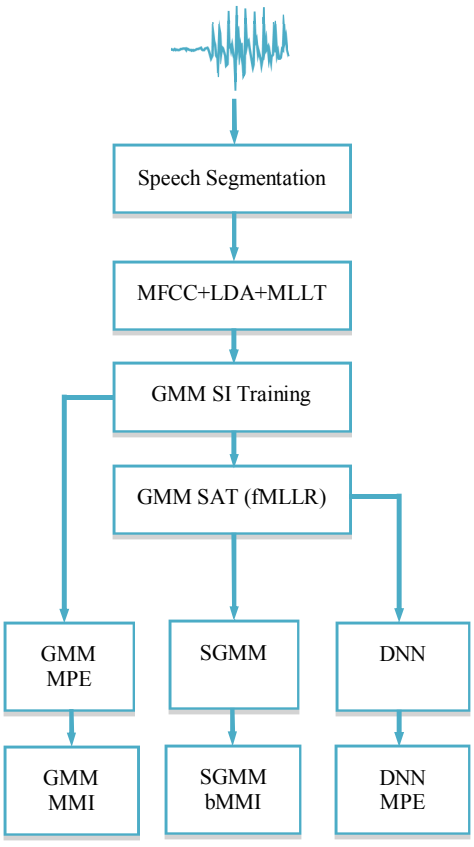

Figure 1: Acoustic training flow-chart

3.2. Vowelization

The Arabic alphabet contains two types of representations; characters which are always written, and diacritics (most importantly indicating short vowels), which are not written in most cases. This makes it difficult to distinguish between different pronunciations, and most importantly, different possible meanings for the same spelling. However, native Arabic speakers can understand and pronounce words without seeing those short vowels. The challenge of missing diacritics has been studied by many researchers [11][12][13][14][15]. For example the word (علم/Elm); can mean science (عِلْمEilm), or flag (عَلَمEalam), or teach (عَلَّمEal~am) or knew (عَلِمEalim), the only difference being the diacritics, which will not be written in most of the cases. The diacritics can be classified into three classes:

a- Three short vowels /ُ/, /َ /, /ِ/ (/u/,/a/,/i/).
b- Three nunation diacritics, which appear at the last character /ٌ /, /ً /, /ٍ / - (/un/, /an/, /in/).
c- Shadda /ّ/ (/~/) which is the doubling diacritic, and can be combined with short vowels from class a.

MADA generates all the possible fully vowelized representation for each word ordered by confidence score. We normalize the confidence sore with respect to the top score for each word, and choose the top three candidates as long as it has 80% or higher confidence relative to the top one.

The output of this is to be used to build a vowelization dictionary. The approach we used is to keep the grapheme representation for each word with the corresponding vowelization candidates. The QCRI vowelization dictionary has about 526K unique grapheme words, with 1.8M vowelization, with an average of 3.43 vowelization for each grapheme word. This dictionary is the input to the V2P to generate ASR lexicon. In a situation where MADA is not able to do automatic vowelization, which typically happens for words with no context, we back off to the grapheme level for the input text word.

3.3. Vowelize to Phones V2P

The mapping between the vowelized Modern Standard Arabic text and phonetic representation is straightforward process, and almost one-to-one mapping. Our V2P rules have been developed following [16].In addition we have an extra rule, for each word ending with vowels, we add an extra entry by removing the last vowel. We found that it happens very often especially in conversational speech, that speakers tend to eliminate the last vowel. We built a preliminary system with this extra rule, and we gained 1.8% relative reduction in WER.

We apply the pronunciation rules on top of the vowelization candidates from MADA as explained in the previous section. In the case of those words, which could not be vowelized by MADA, we apply the V2P rules to the grapheme representation. We also experimented with G2P that has been trained on an initial lexicon of 400K words generated by MADA for the same task; however the result was worse by 2% relative in WER. One possible explanation is that G2P make decisions using only word internal information, while the pipeline of MADA followed by V2P considers

the word context to generate phoneme sequences; especially the last vowel depends on the context.

We investigated different phoneme set for the lexicon and observed no significant difference in WER, when using larger phoneme sets, eg. By distinguishing between long and short vowels, or using different forms of the hamzas. We therefore settled with a condensed phonetic representation using 36 phonemes; 35 phonemes for speech, and one phoneme for silence.

3.4. Lexicon

In QCRI, we have collected a news archive from many news websites, and through our collaboration with Aljazeera we had access to the past five years of news articles from the Arabic news website Aljazeera.net, and processed the text using MADA. The collected text is mostly MSA, but we do find some colloquial words every now and then. We selected all words that occurred more than once in our news archive, and created QCRI ASR lexicon. The lexicon has 526K unique grapheme words, with 2M pronunciations, with an average of 3.84 pronunciations for each grapheme word.

The lexicon (Pronouncing Dictionary) can be downloaded from QCRI language resource website [2]

3.5. Language Model

We used the MADA toolkit to pre-process the text used to build the LMs. Three language models were built using different types of text; 1) original text, 2) MADA normalized text, and 3) MADA normalized and vowelized text.

Table 1. *LM text pre-processing*

	Word	Phone
Original text	A l < l t z A m A l q w y	A l A i l t i z A m A l q a w i y
MADA +Diacritics	A l A l t z A m A l q w Y	A l A i l t i z A m A l q a w a
MADA +Diacritics	A l Ai l ti za A m A l qa wa Y	A l A i l t i z A m A l q a w a

Table 1 shows an example for the LM corpus and lexicon representations. In the experiments using the vowelized LM, the recognition results were devowelized to make the recognition result compatible with the original references.

4. Results

The KALDI system produces lattices as recognition result. To obtain the best path, we followed the standard KALDI procedures and report the best WER based on evaluation on a set of language model scaling factors. Although this is not the ideal setup, the differences between the results from these parameters are rather small and do not change any conclusion we drew from the results.

Table 2. *Language Model Comparison using the* SGMM+bMMI *AM*

LM	Unigram (vocab)	OOV	PPL	WER
Original	105k	9.4%	746.44	32.38%
+MADA	102k	9.3%	728.31	31.72%
++Diac	133k	10.4%	911.48	32.74%

Table 2 shows the performance using the different language models setup as mentioned in Section 3.4. While the same data has been used, but the unigram count is not the same due to the text pre-processing. The first row shows the unigram count as seen in the input text data. The second row shows the unigram count for the same text after MADA normalization. The three thousand words difference in vocabulary,105K words in original text and 102K words in MADA text can be justified due to text normalization. For example the original text has two formats for the word "Egyptian" /مصري/ /mSrY/ and /مصرى/ /mSry/, but after preprocessing the text with MADA only format survived /مصري/ /mSrY/. This increases the count for the first format, and consequently reduces the unigram. This has positive impact on the overall system by slight reduction in the OOV from 9.4% to 9.3% and nice reduction in the WER of 2% relative 32.38% to 31.7%.

Adding diacritics as shown in the third row (++Diacritics) to MADA text actually increased the WER by one percent absolute 32.74% compared to 31.72%. While diacritics add information, which should help the recognition system, it also increases the OOV rate, despite increasing the vocabulary size, and in parallel increases the perplexity of the LM. We will end up with fewer matching n-grams.

Final system has been built using QCRI lexicon which has 526K unique grapheme words with about 2M pronunciation entries. The vocabulary helps in reducing the OOV, by using QCRI lexicon the OOV has gone to 3.9% with relative reduction in the OOV of 56%. The WER has 3.5% relative reduction to be 30.6% instead of 31.72% and the perplexity 997.2. This is the LM setting which we used in our scripts, and which we used for subsequent experiments.

All the WER numbers reported in the LM comparison section was using the SGMM+bMMI Acoustic Model as shown in the next table.

[2] http://alt.qcri.org/resources/speech/dictionary/ar-ar_lexicon_2014-03-17.txt.bz2

Table 3. *WER for our models*

	Report	Conversational	Combined
GMM	22.32%	43.53%	36.74%
GMM+fMLLR	20.98%	41.07%	34.63%
GMM+MPE	19.54%	39.07%	32.84%
GMM+bMMI(0.1)	19.42%	38.88%	32.63%
SGMM+fMLLR	19.9%	39.08%	32.94%
SGMM+bMMI	18.86%	36.34%	30.73%
DNN	17.36%	35.7%	29.81%
DNN+MPE	15.81%	32. 21%	26.95%

Table 3 shows the results of different models generated with our scripts. GMM model with LDA and MLLT is 36.74%. MPE gave an impressive gain of almost 4% absolute, 10.6% relative. GMM+bMMI which have been trained with 0.1 boosting factor reduced the WER further by 0.2% absolute. SGMM+fMLLR gave another 1.7% absolute gain on top of the GMM+fMLLR. Since SGMM+fMLLR start from GMM, the contribution is almost 4% absolute. SGMM+bMMI models give us a nice gain of another 2.2% absolute compared to SGMM+fMLLR. The Deep Neural Network system has 29.81% WER, with 3% relative gain compared to the best SGMM models. The best results are coming from the sequential training for DNN, with an overall WER of 26.95% which is almost 10% relative improvement to the DNN models. The reports data have a relative gain of 8.9% with final WER of 15.81%, and the conversational data have a relative gain of 9.7% with final WER of 32.21%. This summarizes that the sequential training for DNN gave about 12.3% relative reduction to the best SGMM system. The DNN models have been trained with five layers; each layer has 2K nodes, and learning rate 0.008. The DNN training was done using one GPU machine, and it took nearly 90 hours to finish training.

5. Conclusion

In this paper, we present our work on establishing KALDI recipes to build Arabic broadcast news speech recognition systems. Using this recipe and the language resources we provide, researcher can build a broadcast news system with 15.81% WER on Broadcast Report (BR) and 32.21% WER on Broadcast Conversation (BC), with a combined WER of 26.95%.

We provided the rationale behind the decisions made in text processing and generation of the pronunciation dictionary. We also demonstrated the effect of vowelization – for our rather low-resource scenario we did not see any benefits - and that LM word coverage can be improved substantially by adding the dictionary vocabulary into the LM. For future work, we will continue to update the scripts to incorporate new KALDI developments, and maybe introduce bottleneck features and other recently published techniques to further improve the performance of our systems, particularly on handling Arabic dialects.

6. References

[1] F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, "Morphological decomposition in Arabic ASR systems," *Comput. Speech Lang.*, vol. 26, no. 4, pp. 229–243, Aug. 2012.

[2] D. Rybach, S. Hahn, C. Gollan, R. Schluter, and H. Ney, "Advances in Arabic broadcast news transcription at RWTH," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 449–454.

[3] L. Mangu, H.-K. Kuo, S. Chu, B. Kingsbury, G. Saon, H. Soltau, and F. Biadsy, "The IBM 2011 GALE Arabic speech transcription system," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2011, pp. 272–277.

[4] B. Kingsbury, H. Soltau, G. Saon, S. Chu, H.-K. Kuo, L. Mangu, S. Ravuri, N. Morgan, and A. Janin, "The IBM 2009 GALE Arabic speech transcription system," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4672–4675.

[5] P. Cardinal, A. Ali, N. Dahak, T. Al Hanai, Y. Zhang, J. Glass, and V. Stephan, "Recent Advances in ASR Applied to an Arabic Transcription System for Al-Jazeera," in *To appear in Proceedings of 15th Annual Conference of the International Speech Communi- cation Association (Interspeech)*, 2014.

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *The Kaldi Speech Recognition Toolkit*, 2011.

[7] LDC, "GALE Phase 2 Arabic Broadcast Conversation Speech," 2013.

[8] N. Habash, O. Rambow, and R. Roth, "MADA+ TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization," in *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*, 2009, pp. 102–109.

[9] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians.," in *INTERSPEECH*, 2006.

[10] T. Buckwalter, "Issues in Arabic orthography and morphology analysis," pp. 31–34, Aug. 2004.

[11] A. Said, M. El-Sharqwi, A. Chalabi, and E. Kamal, "A Hybrid Approach for Arabic Diacritization," in *Natural Language Processing and Information Systems*, Springer, 2013, pp. 53–64.

[12] N. Habash and O. Rambow, "Arabic diacritization through full morphological tagging," pp. 53–56, Apr. 2007.

[13] I. Zitouni, J. S. Sorensen, and R. Sarikaya, "Maximum entropy based restoration of Arabic diacritics," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, 2006, pp. 577–584.

[14] K. K. Dimitra Vergyri, "Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition."

[15] M. Rashwan, M. Attia, S. Abdou, M. Al Badrashiny, and A. Rafea, "Stochastic Arabic hybrid diacritizer," in *2009 International Conference on Natural Language Processing and Knowledge Engineering*, 2009, pp. 1–8.

[16] F. Biadsy, J. B. Hirschberg, and N. Y. Habash, "Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-Based Pronunciation Rules," 2009.