Hybrid Phonemic and Graphemic Modeling for Arabic Speech Recognition

Mohamed Elmahdy *mohamed.elmahdy@qu.edu.qa*
Qatar University
Qatar

Mark Hasegawa-Johnson *jhasegaw@illinois.edu*
University of Illinois
USA

Eiman Mustafawi *eimanmust@qu.edu.qa*
Qatar University
Qatar

Abstract

In this research, we propose a hybrid approach for acoustic and pronunciation modeling for Arabic speech recognition. The hybrid approach benefits from both vocalized and non-vocalized Arabic resources, based on the fact that the amount of non-vocalized resources is always higher than vocalized resources. Two speech recognition baseline systems were built: phonemic and graphemic. The two baseline acoustic models were fused together after two independent trainings to create a hybrid acoustic model. Pronunciation modeling was also hybrid by generating graphemic pronunciation variants as well as phonemic variants. Different techniques are proposed for pronunciation modeling to reduce model complexity. Experiments were conducted on large vocabulary news broadcast speech domain. The proposed hybrid approach has shown a relative reduction in WER of 8.8% to 12.6% based on pronunciation modeling settings and the supervision in the baseline systems.

Keywords: Arabic, Acoustic modeling, Pronunciation modeling, Speech recognition.

1. INTRODUCTION

Arabic is a morphologically very rich language that is inflected by gender, definiteness, tense, number, case, humanness, etc. Due to Arabic morphological complexity, a simple lookup table for phonetic transcription -essential for acoustic and pronunciation modeling- is not appropriate because of the high out-of-vocabulary (OOV) rate. For instance, in Arabic, a lexicon of 65K words in the domain of news broadcast leads to an OOV rate in the order of 5% whilst in English, it leads to an OOV rate of less than 1%.

Furthermore, Arabic is usually written without diacritic marks. Text resources without diacritics are known as non-vocalized (or non-diacritized). These diacritics are essential to estimate short vowels, nunation, gemination, and silent letters. The absence of diacritic marks leads to a high degree of ambiguity in pronunciation and meaning [10, 13].

In order to train a phoneme-based acoustic model for Arabic, the training speech corpus should be provided with fully vocalized transcriptions. Then, the mapping from vocalized text to phonetic transcription is almost a one-to-one mapping [10]. State of the art techniques for Arabic vocalization are usually done in several phases. In one phase, orthographic transcriptions are manually written without diacritics. Afterwards, statistical techniques are applied to restore missing diacritic marks. This process is known as “automatic diacritization”. Automatic diacritization techniques can results in diacritization WER of 15%-25% as reported in [10, 12, 16].

In order to avoid automatic or manual diacritization, graphemic acoustic modeling was proposed for Modern Standard Arabic (MSA) in [8] where the phonetic transcription is approximated to be the sequence of word letters while ignoring short vowels. Missing short vowels are assumed to be implicitly modeled in the acoustic mode. It could be noticed that graphemic systems work with an acceptable recognition rate. However the performance is still below the accuracy of phonemic models. Graphemic modeling has an advantage of the straightforward pronunciation modeling approximation, as pronunciation is directly estimated by splitting the word into letters.

Large language models are usually trained with large amounts of non-vocalized text resource. In order to use large language models along with phonemic acoustic model, the pronunciation model should explicitly provide the possible phonemic pronunciations. A morphological analyzer such as the Buckwalter Morphological Analyzer [18] can be used to generate all possible diacritization forms for a given word. This approach was widely used in many Arabic speech recognition systems as in the GALE project and other systems [1, 7, 9]. Actually, this technique results in multiple pronunciation variants for each word. The problem with this approach is that Arabic has a high homograph rate. Hence, the same word has a large number of possible pronunciation variants. In other words, the morphological analyzer provides much more variants than required, and most of them are legacy non-common pronunciations. This large number of variants makes the distance between the different pronunciations becomes very small and hence results in more recognition errors. Another problem with this approach is that pronunciation variants cannot be estimated for words that are not morphologically parsable (e.g. named entities and dialectal words).

In this research, we propose a hybrid modeling approach that can benefit simultaneously from both the grapheme-based and the phoneme-based techniques. Our assumption is that for a relatively small vocalized text corpus, high frequency words always exist (e.g. في , من , على , etc). On the other hand, for larger non-vocalized corpora, we have better lexical coverage for low frequency words. By combining a phonemic acoustic model along with a graphemic model, we can benefit from little amounts of vocalized text for accurate pronunciation modeling of high frequency words. Moreover, for low frequency words, graphemic modeling is still possible and the pronunciation model will not fail.

2. SPEECH CORPORA

Three speech corpora have been chosen in our work. All of them are from the domain of news broadcast. Two corpora were sourced from the European Language Resources Association (ELRA) [6] and the third one was sourced from the Linguistic Data Consortium (LDC) [11]. All resources were recorded in linear PCM format, 16 kHz, and 16 bit. The ELRA speech resources were:

- The NEMLAR Broadcast News Speech Corpus: consists of ~40 hours from different radio stations: Medi1, Radio Orient, Radio Monte Carlo, and Radio Television Maroc. The recordings were carried out at three different periods between 30 June 2002 and 18 July 2005. The corpus is provided with fully vocalized transcriptions [17].
- The NetDC Arabic BNSC (Broadcast News Speech Corpus): contains ~22.5 hours of broadcast news speech recorded from Radio Orient during a three month period between November 2001 and January 2002. The orthographic transcriptions are fully vocalized with the same guidelines as the Nemlar corpus. Detailed composition of the ELRA databases is shown in Table 1.

The LDC resource was the Arabic Broadcast News Speech (ABNS) corpus [11]. The corpus consists of ~10 hours recorded from the Voice of America satellite radio news broadcasts. The recordings were made at time of transmission between June 2000 and January 2001. The orthographic transcription provided with this corpus is partially vocalized.

The two ELRA resources have been taken as the training set (~62 hours) and the LDC corpus has been taken as the testing set (~10 hours) as shown in Table 1. This way we can guarantee complete independence between the training and the testing sets including recording setups, speakers, channel noise, time span, etc.

Training set		
Corpus	**Source**	**Duration (hours)**
NetDC	Radio Orient	22.5
Nemlar	Radio Orient	12.1
	Medi1	9.5
	Radio Monte Carlo	9.0
	Radio Tele. Maroc	9.3
Testing set		
Corpus	**Source**	**Duration (hours)**
ABNS	Voice of America	10.0

TABLE 1: Composition of the Arabic speech broadcast news resources.

3. LANGUAGE MODELING

Two language models have been trained: a small language model (Small-LM-380K) and a large language (Large-LM-800M). The two models are backoff tri-gram models with Kneser-Ney smoothing. The Small-LM-380K model has been trained with the transcriptions of the speech training set (~380K words) that consists of 43K unique words after eliminating all diacritic marks. The evaluation of the Small-LM-380K model against the transcriptions of the speech testing set resulted in an OOV rate of 10.1%, tri-grams hit of 19.6%, and perplexity of 767.5 (entropy of 9.6 bits) as shown in Table 2.

The Large-LM-800M has been trained with the LDC Arabic Gigaword fourth edition corpus [15] that consists of ~800M words. Vocabulary was chosen to be the top 250K unique words. The evaluation of the Large-LM-800M model against the transcriptions of the speech testing set resulted in an OOV rate of 3.1%, tri-grams hit of 41.8%, and perplexity of 464.6 (entropy of 8.9 bits) as shown in Table 2.

Language modeling in this research was carried out using the CMU-Cambridge Statistical Language Modeling Toolkit [2, 14].

Language Model	Small-LM-380K	Large-LM-800M
Training words	380K	800M
Vocabulary	43K	Top 250K
OOV	**10.1%**	**3.1%**
Perplexity	767.5	464.6
Tri-grams hit	19.6%	41.8%

TABLE 2: Language models properties and evaluation against the transcriptions of the testing set.

4. SYSTEM DESCRIPTION

Our system is a GMM-HMM architecture based on the CMU Sphinx engine [3, 4]. Acoustic models are all fully continuous density context-dependent tri-phones with 3 states per HMM trained with MLE. The feature vector consists of the standard 39 MFCC coefficients. During acoustic model training, linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) were applied to reduce dimensionality to 29 dimensions. This was found to improve accuracy as well as recognition speed. Decoding is performed in multi-pass, a fast forward Viterbi search using a lexical tree, followed by a flat-lexicon search and a best-path search over the resulting word lattice.

5. PHONEMIC BASELINE SYSTEM

5.1 Phonemic Acoustic Modeling

The 62 hours training set was used to train the phonemic acoustic model. A grapheme-to-phoneme module was developed to convert the vocalized transcriptions to phonetic ones. The phoneme set consists of 28 consonants, 3 short vowels, and 3 long vowels. Diphthongs were treated as two consecutive phonemes (a vowel followed by a semi-vowel). The acoustic model consists of both context-independent (CI) and context-dependent (CD) phones. During decoding, CI models were used to compute the likelihood for tri-phones that have never been seen in the training set. The CI models consist of 102 states with 32 Gaussians each. The total number of CD tied-states is 3000 with 32 Gaussians each.

5.2 Phonemic Pronunciation Modeling

Phonemic pronunciation modeling is done through a lookup table lexicon, where each entry word is associated with one or more pronunciation variants. The lexicon was built using the phonetic transcription of the training data set (380K words), resulting in 43K unique words with an average of ~1.6 variants per word. Each word is also associated with a rank based on its frequency in the vocalized text resource we have used.

6. Graphemic Baseline System

6.1 Graphemic Acoustic Modeling

All diacritic marks have been removed from the transcriptions of the training set. A graphemic acoustic model was trained by approximating the pronunciation to be the word letters rather than the actual pronunciation. Each letter was mapped to a unique model resulting in a total number of 36 base units (letters in the Arabic alphabet). The graphemic acoustic model consists of 108 CI states and 3000 CD tied-states with 32 Gaussians each.

6.2 Graphemic Pronunciation Modeling

In this case, pronunciation modeling is a straightforward process. For any given word, pronunciation modeling is done by splitting the word into letters. Each word is associated with only one graphemic pronunciation. The major advantage of this modeling technique is the ability to generate a model for any word. However, it is still just an approximation to overcome the problem of out of lexicon words in phonemic modeling.

7. Hybrid System

7.1 Hybrid Acoustic Modeling

Hybrid acoustic modeling is performed by combining the phonemic and the graphemic models into one hybrid model after two independent trainings. This results in having all the phonemic and graphemic HMMs into the same model. The final model consists of 70 base units (34 phonemic and 36 graphemic) with 210 CI states. The total number of CD tied-states is 6000 (3000 phonemic and 3000 graphemic).

One limitation of hybrid acoustic modeling is that the acoustic model does not contain cross tri-phones between graphemic and phonemic units. These types of tri-phones can appear between two words, one with grapheme-based pronunciation and the other one with phoneme-based pronunciation. In this case, the decoder in our system backs off to CI units to compute acoustic likelihood.

7.2 Hybrid Pronunciation Modeling

In hybrid modeling, pronunciation is estimated from a phonemic lexicon in conjunction with the previously discussed graphemic approach. The variants in this case can be phonemic and/or graphemic. The decoder selects the appropriate acoustic phone models (either phonemic or graphemic), based on the pronunciation(s) generated by the pronunciation model. The phonemic lexicon is the same 43K lexicon used in the phonemic baseline. For any given non-vocalized word, three different hybrid pronunciation modeling techniques are proposed: Hybrid-Or, Hybrid-And, and Hybrid-Top(n).

7.2.1 Hybrid-Or

In the Hybrid-Or approach, either graphemic modeling or phonemic modeling is applied for any given word. The approach is adopted as follows:

1. Check the existence of the word in the phonemic lexicon.
2. If the word does not exist in the lexicon, only one graphemic variant is generated.
3. If the word exists in the lexicon, all phonemic pronunciation variants associated with that entry word are extracted from the lexicon.

According to our assumption that high frequency words always exist in the lexicon, this means that for a high frequency word like من , only phonemic variants are generated: /m i n/, /m i n a/, and /m a n/.

The drawback of this approach is that for low frequency words that might appear in the lexicon, the generated phonemic variant cannot model all possible variations. That is because low frequency words have always a low rank in the lexicon lacking the coverage for all possible variants.

7.2.2 Hybrid-And

In the Hybrid-And approach, a graphemic pronunciation is always generated for any given word in addition to the existing phonemic pronunciations in the lexicon as follows:

1. Check the existence of the word in the phonemic lexicon.
2. If the word does not exist in the lexicon, only one graphemic variant is generated.
3. If the word exists in the lexicon, one graphemic variant is generated in addition to all existing variants in the lexicon.

In this approach, we are trying to compensate low ranked words in the lexicon, with one generic graphemic variant to model the missing variants. The drawback is that we also generate a redundant graphemic variant for high frequency words as well, and this might decrease recognition rate. For instance, the word من will have one redundant graphemic variant /م ن/ and the phonemic variants: /m i n/, /m i n a/, and /m a n/.

7.2.3 Hybrid-Top(n)

The Hybrid-Top(n) approach is a mixture of Hybrid-Or and Hybrid-And. For n=N, pronunciation modeling is performed as follows:

1. Check the existence of the word in the phonemic lexicon.
2. If the word does not exist in the lexicon, only one graphemic variant is generated.
3. If the word exists in the lexicon, check the word's rank in the phonemic lexicon.
   - If the word exists among the Top(N) high frequently used words, only the phonemic variants associated with that entry word are generated.
   - If the word's rank is below the Top(N) words, one graphemic variant is generated in addition to all existing variants in the lexicon.

In the Hybrid-Top(n) approach, we are trying to keep only phonemic pronunciations for high frequency words. On the other hand, for low frequency words, a generic graphemic model is added to compensate missing variants.

8. Recognition Results

8.1 Phonemic Modeling Results

Performance was evaluated against the 10 hours testing set. The phonemic acoustic model along with the Small-LM-380K language model resulted in a WER of 47.8% as shown in Table 3. A significant percentage of the errors was due to the high OOV rate. Moreover, phonemic modeling for low frequency was less accurate than high frequency words. The large-LM-800K resulted in a WER of 41.1%, as shown in Table 4, with a relative reduction in WER of 14.0% compared to the case of Small-LM-380K. The relative reduction in WER shows only the effect of a larger language model. In the case of the large language model, pronunciation modeling suffers from the inability of generating the appropriate pronunciation for words that do not exist in the lexicon.

8.2 Graphemic Modeling Results

The graphemic acoustic model along with the Small-LM-380K language model resulted in a WER of 53.4%, as shown in Table 3, with a relative increase in WER of 11.7% compared to phonemic modeling. This relative increase in WER shows the difference in performance between the phonemic and the graphemic approach when each word in the language model has a phonemic pronunciation.

The Large-LM-800K resulted in a WER of 42.1%, as shown in Table 4, with a relative increase of 2.4% compared to the case of phonemic modeling. The small relative difference is mainly interpreted because of the lack in pronunciation modeling for the high number of words in the Large-LM-800K model.

Acoustic model	WER	Relative
Phonemic AM	47.8%	-
Graphemic AM	53.4%	+11.7%

TABLE 3: WERs on the 10 hours testing set using the Small-LM-380K language model and the conventional acoustic modeling techniques: phonemic and graphemic.

8.3 Hybrid Modeling Results

Hybrid modeling was first tested with the Small-LM-380K language model. However, no improvement was observed compared to the phonemic baseline system. That was expected since the entire vocabulary of the Small-LM-380K model already exist in the phonemic lexicon.

On the other hand, hybrid modeling along with the Large-LM-800M language model resulted in significant accuracy improvement compared to both the phonemic and the graphemic baselines. The Large-LM-800M along with the hybrid acoustic model were used in decoding the testing set. The three hybrid pronunciation modeling approaches have been evaluated as follows:

8.3.1 Hybrid-Or

In the Hybrid-Or approach, the absolute WER was 37.5% outperforming the phonemic and the graphemic baseline systems by 8.8% and 10.9% relative reduction in WER respectively as shown Table 4.

8.3.2 Hybrid-And

In Hybrid-And settings, the absolute WER was reduced to 36.9% absolute, outperforming the phonemic and the graphemic approaches by 10.2% and 12.4% relative reduction. The improvement in Hybrid-And compared to Hybrid-Or was mainly interpreted as the lexicon does not have enough variants for low frequency words. That is why by generating a graphemic variant along with available phonemic variants, missing variants can be modeled by the generic graphemic model.

8.3.3 Hybrid-Top(N)

The Hybrid-Top(n) approach was evaluated by taking n=100 (i.e. the top 100 most frequently used words), the WER was slightly improved achieving 36.8% absolute WER, outperforming the phonemic and the graphemic approaches by 10.5% and 12.6% relative reduction in WER respectively.

The slight improvement is interpreted since high frequency words in the lexicon are already associated with almost all possible pronunciation variants, and by adding an extra graphemic variant, the distance between them become smaller and more recognition errors are expected. That is why eliminating graphemic variants from the top words may slightly improve recognition accuracy.

Actually, the slight accuracy improvement of Hybrid-Top(n) compared to Hybrid-And is not significant. Thus, accuracy wise, Hybrid-And and Hybrid-Top(n) are in fact equivalent. However, the advantage of the Hybrid-Top(n) is that it can reduce system complexity by eliminating the redundant graphemic variant associated with the top high frequency words, and hence can improve real time factor.

Approach	WER	Relative to Phonemic	Relative to Graphemic
Phonemic	41.1%	-	-2.4%
Graphemic	42.1%	+2.4%	-
Hybrid-Or	37.5%	-8.8%	-10.9%
Hybrid-And	36.9%	-10.2%	-12.4%
Hybrid-Top(n), n=100	36.8%	-10.5%	-12.6%

TABLE 4: WERs on the 10 hours testing set using the Large-LM-800M language model and the different acoustic and pronunciation modeling techniques.

9. Conclusions and Future Work

Arabic is a morphologically very rich language. This morphological complexity results in high OOV rate compared to other languages like English. In large vocabulary speech recognition, the high OOV rate can significantly reduce speech recognition accuracy due to the limitation of pronunciation modeling.

In this paper, we have proposed a hybrid approach for Arabic large vocabulary speech recognition. The proposed approach benefits from both phonemic and graphemic modeling techniques where two acoustic models are fused together after two independent trainings.

First, two baseline systems were built: phonemic and graphemic. With the Small-LM-380K language model where all words in the vocabulary are associated with phonemic pronunciation, the phonemic baseline has outperformed the graphemic baseline by -11.7% relative decrease in WER. With the Large-LM-800M language model, the gap was decreased where the phonemic system has outperformed the graphemic system by -2.4% relative reduction in WER.

In order to create the hybrid acoustic model, the phonemic and the graphemic models were combined together. Three different approaches have been proposed for hybrid pronunciation modeling: Hybrid-Or, Hybrid-And, and Hybrid-Top(n).

The Hybrid-Or technique has resulted in 37.5% absolute WER outperforming the phonemic and the graphemic baselines by -8.8% and -10.9% relative reduction in WER respectively. The hybrid-And technique has resulted in 36.9% absolute WER outperforming the phonemic and the graphemic baselines by -10.2% and -12.4% relative.

Best hybrid modeling approach was found to be the Hybrid-Top(n), where a lexicon-based pronunciation modeling has been used for the top n words and we apply the Hybrid-And approach on the non-top n words. Hybrid-Top(n) results show that hybrid modeling outperforms phonemic and graphemic modeling by -10.5% and -12.6% relative reduction in WER respectively.

In large vocabulary speech domains, acoustic and pronunciation modeling is a common problem among all the Arabic varieties and not only limited to standard Arabic form. Thus, for future work, the proposed approach will be extended and evaluated with the different Arabic colloquials (e.g. Egyptian, Levantine, Gulf, etc.). Moreover, the proposed technique can be also applied on others morphological rich languages like Turkish, Finnish, Korean, etc.

10. Acknowledgements

This publication was made possible by a grant from the Qatar National Research Fund under its National Priorities Research Program (NPRP) award number NPRP 09-410-1-069. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Qatar National Research Fund.

We would like also to acknowledge the Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) for providing us with the required speech and text resources to conduct this research.

11. REFERENCES

[1] A. Messaoudi, L. Lamel, and J. Gauvain, “Transcription of Arabic Broadcast News”. *In International Conference on Spoken Language Processing (INTERSPEECH)*, pp. 1701-1704, 2004.

[2] Carnegie Mellon University-Cambridge, CMU-Cambridge Statistical Language Modeling toolkit, http://www.speech.cs.cmu.edu/SLM/toolkit.html

[3] Carnegie Mellon University Sphinx, Speech Recognition Toolkit, http://cmusphinx.sourceforge.net/

[4] D. Huggins-Daines, M. Kumar, A. Chan, A. W Black, M. Ravishankar, and A I. Rudnicky, “Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices”, *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 185-188, 2006.

[5] D. Vergyri and K. Kirchhoff, “Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition”, *In proceedings of COLING Computational Approaches to Arabic Script-based Language*s, pp. 66-73, 2004.

[6] ELRA: European Language Resources Association, http://www.elra.info/

[7] H. Kuo, S. Chu, B. Kingsbury, G. Saon, H. Soltau, F. Biadsy, “The IBM 2011 GALE Arabic speech transcription system”, *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 272- 277, 2011.

[8] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, “Audio indexing of Arabic broadcast news”, *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 5–8, 2002.

[9] L. Lamel, A. Messaoudi, and J. Gauvain, "Automatic Speech-to-Text Transcription in Arabic", *ACM Transactions on Asian Language Information Processing (TALIP)*, 8(4), 2009

[10] M. Elmahdy, R. Gruhn, and W. Minker, "Novel Techniques for Dialectal Arabic Speech Recognition", *Springer*, 2012.

[11] M. Maamouri, D. Graff, C. Cieri, "Arabic Broadcast News Speech", *Linguistic Data Consortium(LDC)*, LDC Catalog No.: LDC2006S46, 2006.

[12] N. Habash and O. Rambow, "Arabic Diacritization through Full Morphological Tagging", *Proceedings of NAACL HLT 2007*, pp. 53-56, 2007.

[13] N. Habash, "Introduction to Arabic Natural Language Processing", Morgan and Claypool Publishers, 2010.

[14] P. Clarkson, and R. Rosenfeld, "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *In Proceedings of ISCA Eurospeech*, 1997.

[15] R. Parker, D. Graff, K. Chen, J. Kong, and K. Maeda, "Arabic Gigaword Fourth Edition", *Linguistic Data Consortium(LDC)*, LDC Catalog No.: LDC2009T30, 2009.

[16] R. Sarikaya, O. Emam, I. Zitouni, and Y. Gao, "Maximum Entropy Modeling for Diacritization of Arabic Text", *In Proceedings of International Conference on Speech and Language Processing INTERSPEECH*, pp. 145–148, 2006.

[17] The Nemlar project, http://www.nemlar.org/

[18] T. Buckwalter, "Buckwalter Arabic Morphological Analyzer Version 1.0", *Linguistic Data Consortium(LDC)*, LDC Catalog No.:LDC2002L49, 2002.