Effect of Gaussian Densities and Amount of Training Data on Grapheme-Based Acoustic Modeling for Arabic

Mohamed ELMAHDY^{1,2}, Rainer GRUHN³, Wolfgang MINKER¹, Slim ABDENNADHER²

¹Faculty of Engineering & Computer Science, University of Ulm, Ulm, Germany

²Faculty of Media Engineering & Technology, German University in Cairo, Cairo, Egypt

³SVOX AG, Ulm, Germany

mohamed.elmahdy@guc.edu.eg, rainer.gruhn@alumni.uni-ulm.de,

wolfgang.minker@uni-ulm.de, slim.abdennadher@guc.edu.eg

Abstract:

Grapheme-based acoustic modeling for Arabic is a demanding research area since high phonetic transcription accuracy is not yet solved completely. In this paper, we are studying the use of a pure grapheme-based approach using Gaussian mixture model to implicitly model missing diacritics and investigating the effect of Gaussian densities and amount of training data on speech recognition accuracy. Two transcription systems were built: a phoneme-based system and a grapheme-based system. Several acoustic models were created with each system by changing the number of Gaussian densities and the amount of training data. Results show that by increasing the number of Gaussian densities or the amount of training data, the improvement rate in the grapheme-based approach was found to be faster than in the phoneme-based approach. Hence the accuracy gap between the two approaches can be compensated by increasing either the number of Gaussian densities or the amount of training data.

Keywords:

Acoustic modeling; Arabic language; Graphemic modeling; Speech recognition

1. Introduction

Arabic language has a strong grapheme-to-phoneme relation but this is only true for fully diacritized script. For diacritized Arabic script, almost all letters and diacritics are unigraphs, and phonetic transcription is a straight forward operation since letters or diacritics can be mapped directly to a corresponding phoneme [6]. Arabic is written in the diacritized form in only sacred books like Koran and in Arabic learning books. However, Arabic is usually written without diacritic marks as in newspapers, books, subtitling, manuals..., and the reader infers missing diacritics from the context. In automatic speech recognition, exact phonetic transcription is required in order to build phoneme-based acoustic

models. Missing diacritics in Arabic poses a problem since it leads to lots of ambiguity about the exact phonetic transcription.

For instance, the word /kataba/ (he wrote) contains three short vowels of the type /a/ that are only estimated from diacritic marks. If the diacritic marks are omitted /ktb/, we will get ambiguity about the exact pronunciation, whether it is /kataba/ (he wrote), /kutiba/ (it was written), /kutub/ (books), /kattaba/ (he dictated), or /kuttiba/ (it was dictated), so we have at least five different pronunciation variants for the word /ktb/ in the grapheme-based form. We use SAMPA notation for phonetic transcription throughout this paper [10].

Lots of research studied how to automatically estimate missing diacritics from the context in order to estimate the exact phonetic transcription as in [2] and [9]. However, still the problem is unsolvable and the accuracy of automatic diacritization systems is in the range of 15–25% WER and manual reviewing is mandatory in order to achieve an accuracy of 99%. Manual reviewing is a very costly task sine the productivity of a well trained linguist is ~1.5K words per work-day [7]. In other words, for a 40 hours speech corpus that consists of 200K words in average, we will need 133 work-days for just reviewing the results of a commercial class automatic diacritization system.

Grapheme-based acoustic modeling (also known by graphemic modeling using all Arabic letters except diacritics) for Arabic was investigated in [5]. Every grapheme was mapped to one phoneme. Phonemes that are estimated from diacritics were ignored relying on that they can be implicitly modeled in the acoustic model. Results showed that the phoneme-based approach performs much better by ~14% increase in accuracy versus the grapheme-based approach. Hence, diacritic marks are mandatory in order to achieve high accuracy recognition rate.

Grapheme-based approach was adopted in [3] and the transcriptions of the used corpora were mainly non-diacritized and it was noticed that the performance is acceptable but they did not compare the results versus

the phoneme-based approach. In order to improve the performance of the grapheme-based approach, in [4], an explicit modeling approach was proposed to model diacritics by using a generic vowel to replace all short vowels and creating all possible pronunciation variants for any given non-diacritized word. However the proposed explicit modeling approach was not compared against the implicit modeling approach.

In this paper, we study implicit modeling of Arabic diacritics, our assumption is: all diacritics can be modeled effectively by using context-dependant acoustic models and Gaussian mixture model, and increasing the number of Gaussian densities can implicitly model all the diacritization possibilities for the same grapheme without the need of full phonetic transcription. Furthermore, the grapheme-based acoustic modeling approach is much more improved by increasing the number of Gaussian densities and the amount of training data. The rate of improvement in the grapheme-based approach should be much higher than the rate of improvement in the phoneme-based approach. In other words, the gap in the accuracy between the two approaches can be decreased by increasing any of the two parameters: the number of Gaussian densities or the amount of training data.

2. Data sets

We have chosen the Nemlar broadcast news speech corpus [8] for training and testing in our research. The corpus consists of 40 hours of Modern Standard Arabic (MSA) news broadcast speech. The broadcasts were recorded from four different radio stations. All files were recorded in linear PCM format, 16 kHz, and 16 bit. The total number of speakers is 259 and the lexicon size is 62K words.

This corpus was mainly selected because: the transcription is fully diacritized and manually reviewed -i.e. contains all types of diacritic marks- and the high number of speakers is required for a better speaker independent acoustic modeling.

We have processed the Nemlar corpus to exclude speech segments with music or noise in the background and also excluded cross-talks, segments for non-native speakers, and segments containing truncated words. After filtration the 40 hours were been reduced to 32 hours. An amount of 28 hours (~85%) of the filtered data was taken as the training set and 4 hours (~15%) was taken as the testing set.

3. Transcription systems

We have prepared two transcription systems: phoneme-based transcription and grapheme-based transcription.

3.1. Phoneme-based transcription system

This system is a full phonetic transcription system. The transcription contains all diacritic marks and hence the exact phonetic transcription is available. This system will be used in building the phoneme-based acoustic models. The total number of phonemes is 34 phonemes (28 consonants, 3 long vowels, and 3 short vowels). Foreign and rare phonemes were ignored: /p/, /v/, /g/, and /ʔ/, and we mapped them to the closest standard Arabic phonemes. The lexicon size in this system is 62K words and in average 1.6 variants per word as shown in Table 1.

3.2. Grapheme-based transcription system

This system is a grapheme only transcription (letters without diacritics). We have removed all diacritic marks from the original transcription. So this system contains only letters in the common Arabic writing system. Every letter is mapped to one phoneme and this system will be used in building the grapheme-based acoustic models. In this system the total number of unique phonemes that can be estimated from the text is 31 phonemes, the short vowels /a/, /i/, and /u/ are not included in this system since they can be only estimated from diacritics.

There are nine types of diacritic marks in Arabic, and every written letter can be followed by one of them and sometimes two. We have calculated the frequency of diacritics in the corpus and we have found that 44.9% of the corpus consists of diacritic marks as shown in Table 2. So in the grapheme-based transcription we are missing 44.9% of the information about the exact phonetic transcription. The frequencies of all diacritic marks are shown in Table 3. The lexicon size in this system has been decreased by 37% and this is due to that all the pronunciation variants for a word (different diacritic combinations in the phoneme-based system) has been reduced here to only one pronunciation because of the absence of diacritics.

For example: the word /taʔʔallama/ (he learned) contains five phonemes estimated from diacritics (four Fatha /a/ and one Shadda: doubling the consonant /l/), after removing the diacritics, the word will become /tʔʔlm/ (grapheme-based transcription). In this specific example, we are missing 55% of the original phonetic transcription.

4. System description

Training and decoding are based on CMU Sphinx engine. The number of states per HMM is 3 without skip state topology (in our experiment, we found that increasing the number of states per HMM or using skip state topology did not improve the accuracy). All the acoustic models are context-dependant tri-phone models

with a total number of 2000 tied-states and 13 MFCC coefficients with 40 Mel frequency bands were taken. The sampling rate is 16 kHz as in the original data.

One of the common problems in Arabic ASR is the high out-of-vocabulary (OOV) rate, for example: a typical 65K lexicon in the domain of news broadcast, the OOV rate is 4% while in English it is less than 1% [3]. In order to avoid this high OOV rate in Arabic and since we are concerned in our work with acoustic modeling, we decided to work on the phoneme level recognition. We have built a closed vocabulary 7-gram statistical language model with Good Turning discounting using CMU SLM toolkit [1] on the phoneme level by considering every phoneme as a word.

5. Evaluation and results

5.1. Effect of Gaussian densities

The whole 28 hours of the training set were used to train several acoustic models. We fixed all training and decoding parameters except the number of Gaussian densities. We started by using one Gaussian till reaching 32 Gaussians. The training was performed for all Gaussians using the grapheme-based transcription then it was repeated using the phoneme-based transcription. Overall we created 6 acoustic models using the grapheme-based transcription and another 6 acoustic models using the phoneme-based transcription. We used the whole 4 hours of the testing set in decoding, and we repeated the decoding test using every acoustic model and calculated the phoneme error rate (PER) each time. The decoding results are shown in Table 4 and Figure 1.

TABLE 1
LEXICON SIZE AND AVERAGE VARIANTS PER WORD IN THE GRAPHEME-BASED AND THE PHONEME-BASED TRANSCRIPTION SYSTEMS (NEMLAR NEWS BROADCAST CORPUS)

System	Lexicon size	Variants per word
Grapheme-based	39.2K	1.0
Phoneme-based	62.7K	1.6

TABLE 2
THE FREQUENCY OF GRAPHEMES AND DIACRITICS IN THE DIACRITIZED TRANSCRIPTION OF A 32 HOURS SPEECH DATA (NEMLAR NEWS BROADCAST CORPUS)

Type	Frequency
Graphemes	1,179,623 (55.1%)
Diacritics	963,008 (44.9%)
Total	2,142,631 (100%)

TABLE 3
ARABIC DIACRITICS AND THEIR FREQUENCY OF OCCURRENCE

Type	Frequency
Fatha /a/	17.88%
Kasra /i/	9.98%
Damma /u/	3.53%
Shadda (consonant doubling)	3.40%
Sukun (no vowel)	9.49%
Tanween Fatha /an/**	0.29%
Tanween Kasra /in/	0.33%
Tanween Damma /un/	0.07%

**Tanween (Nunation) may only appears on the last letter of a word

The results show that the accuracy is improved by increasing the number of Gaussian densities in both the grapheme-based and phoneme-based approaches, but the rate of improvement is not the same. The rate of improvement in the grapheme-based approach was found to be higher than the rate of improvement in the phoneme-based approach. In the case of one Gaussian, the difference (Delta) in the accuracy between the two approaches was 13.43. The delta in accuracy was found to decrease by increasing the number of Gaussians till reaching a delta of ~2% in the case of 32 Gaussians (see Figure 2). In our experiment we found that by doubling the number of Gaussians, delta is reduced in average by 30% (using 28 hours of training data).

5.2. Effect of the amount of training data

We have fixed the number of Gaussian densities to 32. We fixed all training and decoding parameters except the amount of training data. Using the grapheme-based transcription system, we created four acoustic models with training data amount of 7, 14, 21, and 28 hours. Then, we repeated the training using the phoneme-based transcription system. Overall we created 4 acoustic models using the grapheme-based transcription system and another 4 acoustic models using the phoneme-based transcription system. We used the whole 4 hours of the testing set in decoding. We repeated the decoding test using every acoustic model and calculated PER each time. The decoding results are shown in Table 5 and Figure 3.

The results show that the accuracy is improved by increasing the amount of training data in both approaches but the rate of improvement is not the same. The rate of improvement in the grapheme-based approach was found to be higher than the rate of improvement in the phoneme-based approach. By using 7 hours of training data, delta between the accuracy in the two approaches was 9.25%. The delta in accuracy was found to decrease by increasing the amount of

training data till reaching a delta of 2.08% by using the whole 28 hours of the training set (see Figure 4). In our experiment we found that delta is reduced in average by 53% by doubling the amount of training data (using 32 Gaussians). This rate was not observed using 8 Gaussians and below. The interpretation of that is: sufficient number of Gaussians should exist in order to notice a convergence between the accuracy of two approaches.

6. Discussion

6.1. Toward multi-accent acoustic modeling

Grapheme-based acoustic modeling can be thought as a multi-accent approach for Arabic varieties, because of its ability to implicitly model different pronunciations for the same grapheme or letter, in other words different possible diacritics without the need to explicitly include different pronunciation variants for the same word in the lexicon. For instance, the word /jal?'ab/ (he plays) in MSA is transformed to the word /jil?'ab/ in Egyptian Colloquial Arabic (ECA) and the only difference between the two words is: the vowel /a/ is transformed to the vowel /i/. This transformation is found in almost all present tense verbs in ECA. In the case of the grapheme-based approach, no changes are needed in order to deal with that word in the ECA accent. On the other hand, in the phoneme-based approach, the lexicon should be modified to add the new word /jil?'ab/ as it is considered as a different pronunciation than the existing one /jal?'ab/.

7. Conclusions and future work

The major advantages of the grapheme-based acoustic modeling approach in Arabic are: the fast transcription development since it is the normal Arabic writing system, and there is no need for automatic diacritization or manual reviewing. Furthermore, the acoustic model is capable to model implicitly all pronunciation variants for the same grapheme and hence reducing the lexicon size and the number of variants per word. In the case of the grapheme-based approach, our research shows that we miss 44.9% of the information about the exact phonetic transcription compared to the phoneme-based approach.

Our results show that the grapheme-based acoustic modeling approach is improved by increasing the number of Gaussian densities and the amount of training data. The improvement rate in the grapheme-based approach was found to be higher than the improvement rate in the phoneme-based approach and we were able to notice a clear convergence between the accuracy of the grapheme-based approach and the phoneme-based approach. In our research, we were able to decrease the

TABLE 4
EFFECT OF GAUSSIAN DENSITIES ON PER IN THE GRAPHEME-BASED APPROACH (GBA) AND THE PHONEME-BASED APPROACH (PBA) USING 28 HOURS OF TRAINING DATA

Gaussian densities	PER (GBA)	PER (PBA)	Delta
1	49.76%	36.33%	13.43%
2	43.58%	32.37%	11.21%
4	37.37%	29.85%	7.52%
8	33.83%	27.97%	5.86%
16	30.63%	27.31%	3.32%
32	28.41%	26.33%	2.08%

Figure 1. PER (%) versus the number of Gaussian densities in the grapheme-based approach (GBA) and the phoneme-based approach (PBA) using 28 hours of training data.

Figure 2. Delta PER(GBA-PBA) (%) versus the number of Gaussian densities using 28 hours of training data.

difference in the accuracy between the grapheme-based approach and the phoneme-based approach to ~2%. By examining the improvement trend from the graphs, this difference is expected to decrease more by using more training data or by using more Gaussian densities. In our

experiment, we were limited to 28 hours of training data and that amount is not suitable to train more than 32 Gaussians (e.g. 64 or 128) (in our work, max. total Gaussians was 64K using 32 Gaussians per state and 2000 tied-states). Finally, we found that by adding more training data, there is a noticeable convergence between the accuracy of the two approaches if sufficient Gaussian densities are available (16 or more).

For future work, the grapheme-based acoustic modeling approach will be studied in more depth with dialectal Arabic (dialectal Arabic is only spoken and rarely utilized in the written form). Phonetic transcription for dialectal Arabic is much more difficult than MSA because available dialectal resources are very limited and still there is no commonly accepted standard for the phonetic transcription of dialectal Arabic. Hence the grapheme-based approach represents a better solution.

References

- [1] Carnegie Mellon Statistical Language Modeling (CMU SLM) Toolkit, <http://www.speech.cs.cmu.edu/SLM/toolkit.html>.
- [2] Dimitra Vergyri and Katrin Kirchhoff, "Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition", In COLING workshop on Arabic-script based languages, 66-73, 2004.
- [3] J. Billa, M. Noamany, A. Srivastava, D. Liu, R. Stone, J. Xu, J. Makhoul, and F. Kubala, "Audio Indexing of Arabic Broadcast News", In ICASSP, 1:5-8, 2002.
- [4] Lori Lamel, Abdel. Messaoudi, and Jean-Luc Gauvain, "Improved Acoustic Modeling for Transcribing Arabic Broadcast Data", In INTERSPEECH, 2077-2080, 2007.
- [5] Mohamed Afify, Long Nguyen, Bing Xiang, Sherif Abdou, and John Makhoul, "Recent Progress in Arabic Broadcast News Transcription at BBN", In INTERSPEECH, 1637-1640, 2005.
- [6] Mohamed Elmahdy, Rainer Gruhn, Wolfgang Minker, and Slim Abdennadher, "Survey on Common Arabic Language Forms from a Speech Recognition Point of View", In NAG-DAGA, 63-66, 2009.
- [7] Muhammad Atiyya, Khalid Choukri, and Mustafa Yaseen, "Specifications of the Arabic Written Corpus", Nemlar project, 2005.
- [8] Nemlar project, <http://www.nemlar.org>.
- [9] Ruhi Sarikaya, Ossama Emam, Imed Zitouni, and Yuqing Gao, "Maximum Entropy Modeling for Diacritization of Arabic Text", In INTERSPEECH, 145-148, 2006.
- [10] Speech Assessment Methods Phonetic Alphabet (SAMPA) for Arabic, <http://www.phon.ucl.ac.uk/home/sampa/arabic.htm>.

TABLE 5
EFFECT OF THE TRAINING DATA AMOUNT ON PER IN THE PHONEME-BASED APPROACH (PBA) AND THE GRAPHEME-BASED APPROACH (GBA) USING 32 GAUSSIANS

Training amount(hours)	PER (GBA)	PER (PBA)	Delta
7	49.45%	40.20%	9.25%
14	35.96%	28.82%	7.14%
21	32.18%	28.53%	3.65%
28	28.41%	26.33%	2.08%

Figure 3. PER (%) versus the amount of training data in the grapheme-based approach (GBA) and the phoneme-based approach (PBA) using 32 Gaussians.

Figure 4. Delta PER(GBA-PBA) (%) versus the amount of training data using 32 Gaussians.