Automatic diacritization of Arabic for Acoustic Modeling in Speech Recognition

Dimitra Vergyri
Speech Technology and Research Lab.,
SRI International,
Menlo Park, CA 94025, USA
dverg@speech.sri.com

Katrin Kirchhoff
Department of Electrical Engineering,
University of Washington,
Seattle, WA 98195, USA
katrin@ee.washington.edu

Abstract

Automatic recognition of Arabic dialectal speech is a challenging task because Arabic dialects are essentially spoken varieties. Only few dialectal resources are available to date; moreover, most available acoustic data collections are transcribed without diacritics. Such a transcription omits essential pronunciation information about a word, such as short vowels. In this paper we investigate various procedures that enable us to use such training data by automatically inserting the missing diacritics into the transcription. These procedures use acoustic information in combination with different levels of morphological and contextual constraints. We evaluate their performance against manually diacritized transcriptions. In addition, we demonstrate the effect of their accuracy on the recognition performance of acoustic models trained on automatically diacritized training data.

1 Introduction

Large-vocabulary automatic speech recognition (ASR) for conversational Arabic poses several challenges for the speech research community. The most difficult problems in developing highly accurate speech recognition systems for Arabic are the predominance of non-diacritized text material, the enormous dialectal variety, and the morphological complexity.

Most available acoustic training material for Arabic ASR is transcribed in the Arabic script form, which does not include short vowels and other diacritics that reflect differences in pronunciation, such as the shadda, tanween, etc. In particular, almost all additional text data that can easily be obtained (e.g. broadcast news corpora) is represented in standard script form. To our knowledge, the only available corpus that does include detailed phonetic information is the CallHome (CH) Egyptian Colloquial Arabic (ECA) corpus distributed by the Linguistic Data Consortium (LDC). This corpus has been transcribed in both the script form and a so-called romanized form, which is an ASCII representation that includes short vowels and other diacritic information and thus has complete pronunciation information. It is quite challenging to create such a transcription: native speakers of Arabic are not used to writing their language in a "romanized" form, or even in fully diacritized script form. Consequently, this task is considered almost as difficult as phonetic transcription. Transcribing a sufficiently large amount of training data in this way is therefore labor-intensive and costly since it involves (re)-training native speakers for this purpose.

The constraint of having mostly non-diacritized texts as recognizer training material leads to problems for both acoustic and language modeling. First, it is difficult to train accurate acoustic models for short vowels if their identity and location in the signal is not known. Second, the absence of diacritics leads to a larger set of linguistic contexts for a given word form; language models trained on non-diacritized material may therefore be less predictive than those trained on diacritized texts. Both of these factors may lead to a loss in recognition accuracy. Previous work (Kirchhoff et al., 2002; Lamel, 2003) has shown that ignoring available vowel information does indeed lead to a significant increase in both language model perplexity and word error rate. Therefore, we are interested in automatically deriving a diacritized transcription from the Arabic script representation when a manual diacritization is not available. Some software companies (Sakhr, Apptek, RDI) have developed commercial products for the automatic diacritization of Arabic. However, these products use only text-based information, such as the syntactic context and possible morphological analyses of words, to predict diacritics. In the context of diacritization for speech recognition, by contrast, acoustic data is available that can be used as an additional knowledge source. Moreover, commer-

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automatic diacritization of Arabic for Acoustic Modeling in Speech Recognition	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

cial products concentrate exclusively on Modern Standard Arabic (MSA), whereas a common objective of Arabic ASR is conversational speech recognition, which is usually dialectal. For this reason, a more flexible set of tools is required in order to diacritize dialectal Arabic prior to speech recognizer training.

In this work we investigate the relative benefits of a variety of knowledge sources (acoustic, morphological, and contextual) to automatically diacritize MSA transcriptions. We evaluate the different approaches in two different ways: (a) by comparing the automatic output against a manual reference diacritization and computing the diacritization error rate, and (b) by using automatically diacritized training data in a cross-dialectal speech recognition application.

The remainder of this paper is structured as follows: Section 2 gives a detailed description of the motivation as well as prior work. Section 3 describes the corpora used for the experiments reported in this paper. The automatic diacritization procedures and results are explained in Section 4. The speech recognition experiments and results are reported in Section 5. Section 6 presents our conclusions.

2 Motivation and Prior Work

We first describe the Arabic writing system and its inherent problems for speech recognizer training, and then discuss previous attempts at automatic diacritization.

2.1 The Arabic Writing System

The Arabic alphabet consists of twenty-eight letters, twenty-five of which represent consonants and three of which represent the long vowels (/i:/,/a:/,/u:/). A distinguishing feature of Arabic-script based writing systems is that short vowels are not represented by the letters of the alphabet. Instead, they are marked by so-called *diacritics*, short strokes placed either above or below the preceding consonant. Several other pronunciation phenomena are marked by diacritics, such as consonant doubling (phonemic in Arabic), which is indicated by the "shadda" sign, and the "tanween", i.e. word-final adverbial markers that add /n/ to the pronunciation of the word. These diacritics are listed in Table 1. Arabic texts are almost never fully diacritized; normally, diacritics are used sparingly and only to prevent misunderstandings. Exceptions are important religious and/or political texts or beginners' texts for students of Arabic. The lack of diacritics may lead to considerable lexical ambiguity that must be resolved by contextual information, which in turn presupposes knowledge of the language. It was observed in (Debili et al., 2002) that a non-diacritized dictionary word form has 2.9 possible diacritized forms on average and that an Arabic text containing 23,000 word forms showed an average ratio of 1:11.6. The form كتب, for instance, has 21 possible diacritizations. The correspondence between graphemes and phonemes is relatively transparent compared to other languages like English or French: apart from certain special graphemes (e.g. laam alif), the relationship is one to one. Finally, it is worth noting that the writing system described above is that of MSA. Arabic dialects are primarily oral varieties in that they do not have generally agreed-upon writing standards. Whenever there is the need to write down dialectal speech, speakers will try to approximate the standard system as far as possible and use a phonetic spelling for non-MSA or foreign words.

MSA	Symbol Name	Meaning
◌َ	fatHa	/a/
◌ِ	kasra	/i/
◌ُ	Damma	/u/
◌ّ	shadda	consonant doubling
دْرس	sukuun	vowel absence
◌ً	tanween al-fatHa	/an/
◌ٍ	tanween al-kasr	/in/
◌ٌ	tanween aD-Damm	/un/

Table 1: Arabic diacritics

The lack of diacritics in standard Arabic texts makes it difficult to use non-diacritized text for training since the location and identity of short vowels and other phonetic segments are unknown. One possible approach is to use acoustic models for long vowels and consonants only, where the acoustic signal portions corresponding to unwritten segments are implicitly incorporated into the acoustic models for consonants (Billa et al, 2002). However, this leads to less discriminative acoustic and language models. Previous work (Kirchhoff et al., 2002; Lamel, 2003) has compared the word error rates of two CH ECA recognizers: one trained on script transcriptions and another trained on romanized transcriptions. It was shown that the loss in information due to training on script forms

results in significantly worse performance: a relative increase in word error rate of almost 10% was observed.

It seems clear that diacritized data should be used for training Arabic ASR systems whenever possible. As explained above, however, it is very expensive to obtain manually transcribed data in a diacritized form. Therefore, the corpora that do include detailed transcriptions are fairly small and any dialectal data that might become available in the future will also very likely be of limited size. By contrast, it is much easier to collect publicly available data (e.g. broadcast news data) and to transcribe it in script form. In order to be able to take advantage of such resources, we need to restore short vowels and other missing diacritics in the transcription.

2.2 Prior Work

Various software companies have developed automatic diacritization products for Arabic. However, all of these are targeted towards MSA; to our knowledge, there are no products for dialectal Arabic. In a previous study (Kirchhoff et al., 2002) one of these products was tested on three different texts, two MSA texts and one ECA text. It was found that the diacritization error rate (percentage of missing and wrongly identified or inserted diacritics) on MSA ranged between 9% and 28%, depending on whether or not case vowel endings were counted. However, on the ECA text, the diacritization software obtained an error rate of 48%.

A fully automatic approach to diacritization was presented in (Gal, 2002), where an HMM-based bigram model was used for decoding diacritized sentences from non-diacritized sentences. The technique was applied to the Quran and achieved 14% word error (incorrectly diacritized words).

A first attempt at developing an automatic diacritizer for dialectal speech was reported in (Kirchhoff et al., 2002). The basic approach was to use a small set of parallel script and diacritized data (obtained from the ECA CallHome corpus) and to derive diacritization rules in an example-based way. This entirely knowledge-free approach achieved a 16.6% word error rate.

Other studies (El-Imam, 2003) have addressed problems of grapheme-to-phoneme conversion in Arabic, e.g. for the purpose of speech synthesis, but have assumed that a fully diacritized version of the text is already available.

Several knowledge sources are available for determining the most appropriate diacritization of a script form: analysis of the morphological structure of the word (including segmentation into stems, prefixes, roots and patterns), consideration of the syntactic context in which the word form appears, and, in the context of speech recognition, the acoustic data that accompanies the transcription. Specific dictionary information could in principle be added (such as information about proper names), but this knowledge source is ignored for the purpose of this study. All of the approaches described above make use of text-based information only and do not attempt to use acoustic information.

3 Data

For the present study we used two different corpora, the FBIS corpus of MSA speech and the LDC CallHome ECA corpus.

The FBIS corpus is a collection of radio newscasts from various radio stations in the Arabic speaking world (Cairo, Damascus, Baghdad) totaling approximately 40 hours of speech (roughly 240K words). The transcription of the FBIS corpus was done in Arabic script only and does not contain any diacritic information. There were a total of 54K different script forms, with an average of 2.5 different diacritizations per word.

The CallHome corpus, made available by LDC, consists of informal telephone conversations between native speakers (friends and family members) of Egyptian Arabic, mostly from the Cairene dialect region. The corpus consists of about 20 hours of training data (roughly 160K words) and 6 hours of test data. It is transcribed in two different ways: (a) using standard Arabic script, and (b) using a romanization scheme developed at LDC and distributed with the corpus. The romanized transcription contains short vowels and phonetic segments corresponding to other diacritics. It is not entirely equivalent to a diacritized Arabic script representation since it includes additional information. For instance, symbols particular to Egyptian Arabic were used (e.g. "g" for /g/, the ECA pronunciation of the MSA letter ج), whereas the script transcriptions contain MSA letters only. In general, the romanized transcription provides more information about actual pronunciation and is thus closer to a broad phonetic transcription.

4 Automatic Diacritization

We describe three techniques for the automatic diacritization of Arabic text data. The first combines acoustic, morphological and contextual information to predict the correct form, the second ignores contextual information, and the third is fully acoustics based. The latter technique uses no morphological or syntactic constraints, and allows for all possible items to be inserted at every possible position.

4.1 Combination of Acoustic, Morphological and Contextual Information

Most Arabic script forms can have a number of possible morphological interpretations, which often correspond to different diacritized forms. Our goal is to combine morphological knowledge with contextual information in order to identify possible diacritizations and assign probabilities to them. Our procedure is as follows:

1. Generate all possible diacritized variants for each word, along with their morphological analyses (tags).

2. Train an unsupervised tagger to assign probabilities to sequences of these morphological tags.

3. Use the trained tagger to assign probabilities to all possible diacritizations for a given utterance.

For the first step we used the Buckwalter stemmer, which is an Arabic morphological analysis tool available from the LDC. The stemmer produces all possible morphological analyses of a given Arabic script form; as a by-product it also outputs the concomitant diacritized word forms. An example of the output is shown in Figure 1. The next step was to train an unsupervised tagger on the output to obtain tag n-gram probabilities. The number of different morphological tags generated by applying the stemmer to the FBIS text was 763. In order to obtain a smaller tag set and to be able to estimate probabilities for tag sequences more robustly, this initial tag needed to be conflated to a smaller set. We adopted the set used in the LDC Arabic TreeBank project, which was also developed based on the Buckwalter morphological analysis scheme. The FBIS tags were mapped to TreeBank tags using longest common substring matching; this resulted in 392 tags. Further possible reductions of the tag set were investigated but it was found that too much clustering (e.g. of verb subclasses into a single verb class) could result in the loss of important information. For instance, the tense and voice features of verbs are strong predictors of the short vowel patterns and should therefore be preserved in the tagset.

```
LOOK-UP WORD: قبل (qbl)
SOLUTION 1: (qabola) qabola/PREP
  (GLOSS): + before +
SOLUTION 2: (qaboli) qaboli/PREP
  (GLOSS): + before +
SOLUTION 3: (qabolu) qabolu/ADV
  (GLOSS): + before/prior +
SOLUTION 4:(qibal) qibal/NOUN
  (GLOSS): + (on the) part of +
SOLUTION 5:(qabila)
  qabil/VERB_PERFECT+a/PVSUFF_SUBJ:3MS
  (GLOSS): + accept/receive/approve + he/it <verb>
SOLUTION 6: (qab˜ala)
  qab al/VERB_PERFECT+a/PVSUFF_SUBJ:3MS
  (GLOSS): + kiss + he/it <verb>
```

Figure 1: Sample output of Buckwalter stemmer showing the possible diacritizations and morphological analyses of the script form قبل (*qbl*). Lower-case *o* stands for sukuun (lack of vowel).

We adopted a standard statistical trigram tagging model:

$$P(t_0, \ldots, t_n | w_0, \ldots, w_n) = \prod_{i=0}^{n} P(w_i|t_i) P(t_i|t_{i-1}, t_{i-2}) \quad (1)$$

where t is a tag, w is a word, and n is the total number of words in the sentence. In this model, words (i.e. non-diacritized script forms) and morphological tags are treated as observed random variables during training. Training is done in an unsupervised way, i.e. the correct morphological tag assignment for each word is not known. Instead, all possible assignments are initially considered and the Expectation-Maximization (EM) training procedure iteratively trains the probability distributions in the above model (the probability of word given tag, $P(w_i|t_i)$, and the tag sequence probability, $P(t_i|t_{i-1}, t_{i-2})$) until convergence. During testing, only the word sequence is known and the best tag assignment is found by maximizing the probability in Equation 1. We used the graphical modeling toolkit GMTK (Bilmes and Zweig, 2002) to train the tagger. The trained tagger was then used to assign probabilities to all possible sequences of three successive mor-

phological tags and their associated diacritizations to all utterances in the FBIS corpus.

Using the resulting possible diacritizations for each utterance we constructed a word-pronunciation network with the probability scores assigned by the tagger acting as transition weights. These word networks were used as constraining recognition networks with the acoustic models trained on the CallHome corpus to find the most likely word sequence (a process called alignment). We performed this procedure with different weights on the tagger probabilities to see how much this information should be weighted compared to the acoustic scores. Results for weights 1 and 5 are reported below.

Since the Buckwalter stemmer does not produce case endings, the word forms obtained by adding case endings were included as variants in the pronunciation dictionary used by the aligner. Additional variants listed in the dictionary are the taa marbuta alternations /a/ and /at/. In some cases (approximately 1.5% of all words) the Buckwalter stemmer was not able to produce an analysis of the word form due to misspellings or novel words. These were mapped to a generic reject model.

4.2 Combination of Acoustic and Morphological Constraints

We were interested in separately evaluating the usefulness of the probabilistic contextual knowledge provided by the tagger, and the morphological knowledge contributed by the Buckwalter tool. To that end we used the word networks produced by the method described above but stripped the tagger probabilities, thus assigning uniform probability to all diacritized forms produced by the morphological analyzer. We used the same acoustic models to find the most likely alignment from the word networks.

4.3 Using only Acoustic Information

Similarly, we wanted to evaluate the importance of using morphological information versus only acoustic information to constrain the possible diacritizations. This is particularly interesting since, as new dialectal speech data become available, the acoustics may be the only information source. As explained above, existing morphological analysis tools such as the Buckwalter stemmer have been developed for MSA only.

For that purpose, we generated word networks that include all possible short vowels at each allowed position in the word and allowed all possible case endings. This means that after every consonant there are at least 5 different choices: no vowel (corresponding to the sukuun diacritic), /i/, /a/, /u/, or consonant doubling caused by a shadda sign. Combinations of shadda and a short vowel are also possible. Since we do not use acoustic models for doubled consonants in our speech recognizer, we ignore the variants involving shadda and allow only four possibilities after every word-medial consonant: the three short vowels or absence of a vowel. Finally, we include the three tanween endings in addition to these four possibilities in word-final position. As before, the taa marbuta variants are also included.

In this way, many more possible "pronunciations" are generated for a script form than could ever occur. The number of possible variants increases exponentially with the number of possible vowel slots in the word. For instance, for a longer word with 7 possible positions, more than 16K diacritized forms are possible, not even counting the possible word endings. As before, we use these large pronunciation networks to constrain our alignment with acoustic models trained on CallHome data and choose the most likely path as the output diacritization.

In principle it would also be possible to determine diacritization performance in the absence of acoustic information, using only morphological and contextual knowledge. This can be done by selecting the best path from the weighted word transition networks without rescoring the network with acoustic models. However, this would not lead to a valid comparison in our case because case endings are only represented in the pronunciation dictionary used by the acoustic aligner; they are not present in the weighted transition network and thus cannot be hypothesized unless the acoustic aligner is used.

4.4 Autodiacritization Error Rates

We measured the performance of all three methods by comparing the output against hand transcribed references on a 500 word subset of the FBIS corpus. These references were fully diacritized script transcriptions created by a native speaker of Arabic who was trained in orthographic transcription but not in phonetic transcription. The diacritization error rate was measured as the percentage of wrong diacritization decisions out of all possible decisions. In particular, an error occurs when:

- a vowel is inserted although the reference

transcription shows either sukuun or no diacritic mark at the corresponding position (insertion).

• no vowel is produced by the automatic procedure but the reference contains a vowel mark at the corresponding position (deletion).

• the short vowel inserted does not match the vowel at the corresponding position (substitution).

• in the case of tanween and taa marbuta endings, either the required consonants or vowels are missing or wrongly inserted. Thus, in the case of a taa marbuwta ending with a following case vowel /i/, for instance, both the /t/ and the /i/ need to be present. If either is missing, one error is assigned; if both are missing, two errors are assigned.

Results are listed in Table 2. The first column reports the error rate at the word level, i.e. the percentage of words that contained at least one diacritization mistake. The second column lists the diacritization error computed as explained above. The first three methods have a very similar performance with respect to diacritization error rate. The use of contextual information (the tagger probabilities) gives a slight advantage, although the difference is not statistically significant. Despite these small differences, the word error rate is the same for all three methods; this is because a word that contains at least one mistake is counted as a word error, regardless of the total number of mistakes in the word, which may vary from system to system. Using only acoustic information doubles the diacritization error rate and increases the word error rate to 50%. Errors result mostly from incorrect insertions of vowels (e.g. بَغدَاد → بَغُدَاد). Many of these insertions may stem from acoustic effects created by neighbouring consonants, that give a vowel-like quality to transitions between consonants. The main benefit of using morphological knowledge lies in the prevention of such spurious vowel insertions, since only those insertions are permitted which result in valid words. Even without the use of morphological information, the vast majority of the missing vowels are still identified correctly. Thus, this method might be of use when diacritizing a variety of Arabic for which morphological analysis tools are not available. Note that the results obtained here are not directly comparable to any of the works described in Section 2.2 since we used a data set with a much larger vocabulary size.

Information used	Word level	Character level
acoustic + morphological + contextual (tagger prob. weight=5)	27.3	13.24
acoustic + morphological + contextual (tagger prob. weight=1)	27.3	11.54
acoustic + morphological (tagger prob. weight=0)	27.3	11.94
acoustic only	50.0	23.08

Table 2: Automatic diacritization error rates (%).

5 ASR Experiments

Our overall goal is to use large amounts of MSA acoustic data to enrich training material for a speech recognizer for conversational Egyptian Arabic. The ECA recognizer was trained on the romanized transcription of the CallHome corpus described above and uses short vowel models. In order to be able to use the phonetically deficient MSA transcriptions, we first need to convert them to a diacritized form. In addition to measuring autodiacritization error rates, as above, we would like to evaluate the different diacritization procedures by investigating how acoustic models trained on the different outputs affect ASR performance.

One motivation for using cross-dialectal data is the assumption that infrequent triphones in the CallHome corpus might have more training samples in the larger MSA corpus. In (Kirchhoff and Vergyri, 2004) we demonstrated that it is possible to get a small improvement in this task by combining the scores of models trained strictly on CallHome (CH) with models trained on the combined FBIS+CH data, where the FBIS data was diacritized using the method described in Section 4.1. Here we compare that experiment with the experiments where the methods described in Sections 4.2 and 4.3 were used for diacritizing the FBIS corpus.

5.1 Baseline System

The baseline system was trained with only CallHome data (CH-only). For these experiments we used a single front-end (13 mel-frequency cepstral coefficients with first and second differences). Mean and variance as well as Vocal Tract Length (VTL) normalization were performed per conversation side for CH and per speaker cluster (obtained automatically) for FBIS. We trained non-crossword,

System	dev96	eval03
simple CH-only	56.1	42.7
RT-2003 CH-only	52.6	39.7

Table 3: CH-only baseline WER (%)

continuous-density, genonic hidden Markov models (HMMs) (Digalakis and Murveit, 1994), with 128 gaussians per genone and 250 genones. Recognition was done by SRI's DECIPHER^{TM} engine in a multipass approach: in the first pass, phone-loop adaptation with two Maximum Likelihood Linear Regression (MLLR) transforms was applied. A recognition lexicon with 18K words and a bigram language model were used to generate the first pass recognition hypothesis. In the second pass the acoustic models were adapted using constrained MLLR (with 6 transformations) based on the previous hypothesis. Bigram lattices were generated and then expanded using a trigram language model. Finally, N-best lists were generated using the adapted models and the trigram lattices. The final best hypothesis was obtained using N-best ROVER (**?**). This system is simpler than our best current recognition system (submitted for the NIST RT-2003 benchmark evaluations) (Stolcke et al., 2003) since we used a single front end (instead of a combination of systems based on different front ends) and did not include HLDA, cross-word triphones, MMIE training or a more complex language model. The lack of these features resulted in a higher error rate but our goal here was to explore exclusively the effect of the additional MSA training data using different diacritization approaches. Table 3 shows the word error rates of the system used for these experiments and the full system used for the NIST RT-03 evaluations. Our full system was about 2% absolute worse than the best system submitted for that task. This shows that even though the system is simpler we are not operating far from the state-of-the-art performance for this task.

5.2 ASR Systems Using FBIS Data

In order to investigate the effect of additional MSA training data, we trained a system similar to the baseline but used training data pooled from both corpora (CH+FBIS). After performing alignment of the FBIS data with the networks described in Section 4.1, 10% of the data was discarded since no alignments could be found. This could be due to segmentation problems or noise in the acoustic files. The remaining 90% were used for our experiments. In order to account for the fact that we had much more data, and also more dissimilar data, we increased the model size to 300 genones.

For training the CH+FBIS acoustic models, we first used the whole data set with weight 2 for CH utterances and 1 for FBIS utterances. Models were then MAP adapted on the CH-only data (Digalakis et al., 1995). Since training involves several EM iterations, we did not want to keep the diacritization fixed from the first pass, which used CH-only models. At every iteration, we obtain better acoustic models which can be used to re-align the data. Thus, for the first two approaches, where the size of the pronunciation networks is limited due to the use of morphological information, the EM forward-backward counts were collected using the whole diacritization network and the best diacritization path was allowed to change at every iteration. In the last case, where only acoustic information was used, the pronunciation networks were too large to be run efficiently. For this reason, we updated the diacritized references once during training by realigning the networks with the newer models after the first training iteration. As reported in (Kirchhoff and Vergyri, 2004) the CH+FBIS trained system by itself did not improve much over the baseline (we only found a small improvement on the eval03 testset) but it provided sufficiently different information, so that ROVER combination (Fiscus, 1997) with the baseline yielded an improvement. As we can see in Table 4, all diacritization procedures performed practically the same: there was no significant difference in the word error rates obtained after the combination with the CH-only baseline. This suggests that we may be able to obtain improvements with automatically diacritized data even when using inaccurate diacritization, produced without the use of morphological constraints.

6 Conclusions

In this study we have investigated different options for automatically diacritizing Arabic text for use in acoustic model training for ASR. A comparison of the different approaches showed that more linguistic information (morphology and syntactic context) in combination with the acoustics provides lower diacritization error rates. However, there is no significant difference among the word error rates of ASR sys-

	dev96		eval03	
System	alone	Rover with CH-only	alone	Rover with CH-only
CH-only	56.1		42.7	
CH+FBIS1(weight 1)	56.3	55.3	42.2	41.6
CH+FBIS1(weight 5)	56.1	55.2	42.2	41.8
CH+FBIS2	56.2	55.3	42.4	41.6
CH+FBIS3	56.6	55.7	42.1	41.6

Table 4: Word error rates (%) obtained after the final recognition pass and with ROVER combination with the baseline system. FBIS1, FBIS2 and FBIS3 correspond to the diacritization procedures described in Sections 4.1, 4.2 and 4.3 respectively. For the first approach we report results using the tagger probabilities with weights 1 and 5.

tems trained on data resulting from the different methods. This result suggests that it is possible to use automatically diacritized training data for acoustic modeling, even if the data has a comparatively high diacritization error rate (23% in our case). Note, however, that one reason for this may be that the acoustic models are finally adapted to the accurately transcribed CH-only data. In the future, we plan to apply knowledge-poor diacritization procedures to other dialects of Arabic, for which morphological analyzers do not exist.

7 Acknowledgments

This work was funded by DARPA under contract No. MDA972-02-C-0038. We are grateful to Kathleen Egan for making the FBIS corpus available to us, and to Andreas Stolcke and Jing Zheng for valuable advice on several aspects of this work.

References

J. Billa et al. 2002. Audio indexing of Broadcast News. In *Proceedings of ICASSP*.

J. Bilmes and G. Zweig. 2002. The Graphical Models Toolkit: An open source software system for speech and time-series processing. In *Proceedings of ICASSP*.

F. Debili, H. Achour, and E Souissi. 2002. De l'étiquetage grammatical à la voyellation automatique de l'arabe. Technical report, Correspondances de l'Institut de Recherche sur le Maghreb Contemporain.

V. Digalakis and H. Murveit. 1994. GENONES: Optimizing the degree of mixture tying in a large vocabulary hidden markov model based speech recognizer. In *Proceeding of ICASSP*, pages I–537–540.

V.V. Digalakis, D. Rtischev, and L. G. Neumeyer. 1995. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions SAP*, 3:357–366.

Yousif A. El-Imam. 2003. Phonetization of Arabic: rules and algorithms. *Computer, Speech and Language*, in press, preprint available online at www.sciencedirect.com.

J. G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, pages 347–352, Santa Barbara, CA.

Ya'akov Gal. 2002. An HMM approach to vowel restoration in Arabic and Hebrew. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pages 27–33, Philadelphia, July. Association for Computational Linguistics.

K. Kirchhoff and D. Vergyri. 2004. Cross-dialectal acoustic data sharing for Arabic speech recognition. In *Proceedings of ICASSP*.

K. Kirchhoff, J. Bilmes, J. Henderson, R. Schwartz, M. Noamany, P. Schone, G. Ji, S. Das, M. Egan, F. He, D. Vergyri, D. Liu, and N. Duta. 2002. Novel approaches to Arabic speech recognition - final report from the JHU summer workshop 2002. Technical report, Johns Hopkins University.

L. Lamel. 2003. Personal communication.

A. Stolcke, Y. Konig, and M. Weintraub. 1997. Explicit word error minimization in N-best list rescoring. In *Proceedings of Eurospeech*, volume 1, pages 163–166.

A. Stolcke et al. 2003. Speech-to-text research at sri-icsi-uw. Technical report, NIST RT-03 Spring Workshop. availble online http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/sri+-rt03-stt.pdf.