Cross-Dialectal Data Sharing For Acoustic Modeling in Arabic Speech Recognition

Katrin Kirchhoff [a,*], Dimitra Vergyri [b],

[a] *Department of Electrical Engineering, University of Washington, Box 352500, Seattle, WA, 98195-2500, USA*

[b] *SRI International, Menlo Park, California, 94720, USA*

Abstract

Many of the world's languages have a multitude of dialects which differ considerably from each other in their linguistic properties. Dialects are often spoken rather than written varieties; the development of automatic speech recognition systems for dialects therefore requires the collection and transcription of large amounts of dialectal speech. In those cases where sufficient training data is not available, acoustic and/or language models may benefit from additional data from different though related dialects. In this study we investigate the feasibility of cross-dialectal data sharing for acoustic modeling using two different varieties of Arabic, Modern Standard Arabic and Egyptian Colloquial Arabic. An obstacle to this type of data sharing is the Arabic writing system, which lacks short vowels and other phonetic information. We address this problem by developing automatic procedures to restore the missing information based on morphological, contextual and acoustic knowledge. These procedures are evaluated with respect to the relative contributions of different knowledge sources and with respect to their effect on the overall recognition system. We demonstrate that cross-dialectal data sharing leads to significant reductions in word error rate.

Key words: speech recognition, acoustic modeling, Arabic, dialectal variation

1 Introduction

In spite of much recent progress in automatic speech recognition (ASR), dialectal variation is still a very difficult problem. It affects spoken language at

* Corresponding author.
Email addresses: katrin@ee.washington.edu (Katrin Kirchhoff), dverg@speech.sri.com (Dimitra Vergyri).

several levels, ranging from the acoustic realization of phones to differences in vocabulary, morphology and syntax. In many languages, including English, dialectal variation is relatively mild: the number of different dialects is fairly small, dialects are typically mutually intelligible, and dialectal speakers are capable of switching to a more neutral variety. Other languages, e.g. Arabic and Chinese, are characterized by a large number of dialects that differ to such an extent that they are no longer mutually intelligible and could almost be described as different languages. Often, they are also spoken rather than written varieties, i.e. written dialectal material is scarce and does not follow a generally agreed upon writing standard. When developing ASR systems for such dialects, training data is obtained by recording and manually transcribing speech data. This constitutes a major bottleneck for the rapid development of ASR technology for these varieties. It would therefore be desirable to explore methods for sharing training data between different dialects of the same language.

This goal is similar to that of previous work on cross-language acoustic modeling (e.g. (Köhler, 1998; Byrne et al., 1999, 2000; Schultz and Waibel, 2001)), which has addressed the problem of developing recognizers for languages with little or no acoustic training data. In this case, acoustic models need to be bootstrapped from models trained on different languages, possibly followed by adaptation on a small amount of in-language data. Research in cross-language acoustic modeling has concentrated on questions such as how to construct cross-lingual phone mappings, how to develop good seed models, and how much adaptation data is needed. In most of these studies, the data sets used were collected specifically for the purpose of multilingual acoustic modeling and were thus fairly homogeneous (e.g. either clean read speech or broadcast news data).

In our case, the research questions to be addressed are of a different nature: Can the recognition performance on a given dialect be improved by using data from a different dialect even when some amount of in-dialect acoustic training data is already available? In addition, is data-sharing useful when the data sets involved are acoustically and/or stylistically quite different? The latter question reflects real-world conditions of ASR system development: often, constraints on time and resources prevent new data from being collected, and existing data sets need to be re-used.

In this study we investigate the feasibility of using data from one variety of Arabic, Modern Standard Arabic (MSA), to enrich training data for Egyptian Colloquial Arabic (ECA). Our assumption is that a large amount of additional speech data, even when drawn from a heterogeneous source, will provide better coverage of phonetic contexts and help in modeling the resulting acoustic variability. A difficulty encountered in this work is that standard Arabic script lacks diacritics (which indicate short vowels and other pronunciation informa-

tion) and is thus phonetically underspecified. However, some data resources, such as the LDC CallHome corpus used in our experiments, may include diacritics or equivalent information. Therefore, a necessary prerequisite for data sharing between corpora transcribed in Arabic script and those transcribed in a more phonetically oriented way is the automatic diacritization of the script-based data. A third research question thus is how the automatic vowelization of training transcriptions influences data sharing procedures.

The remainder of this paper is structured as follows: in Section 2 we first present the linguistic and dialectal characteristics of Arabic in greater detail, followed by a description of the the data and the baseline recognition system (Sections 3 and 4). We then describe the automatic diacritization procedure and evaluate its output against manually annotated data (Section 5). Section 6 describes the data sharing techniques, recognition experiments and results. Section 7 concludes.

2 Linguistic Properties of Arabic

In this section we describe the linguistic properties of Arabic relevant to this study and their impact on the design of Arabic speech recognizers.

2.1 Dialectal variation

Although the term "Arabic" is commonly used as if referring to a single, homogeneous language, Arabic is more properly described as a collection of different varieties. One of the most widespread varieties is Modern Standard Arabic (MSA), which is a common standard shared by educated speakers throughout the Arabic-speaking world. It is the language used for written and formal oral communication, e.g. broadcast news, courtroom language, lectures, etc. Everyday informal communication, by contrast, is carried out in one of the many regional dialects, of which there are four broad classes: North African, Levantine, Egyptian and Gulf Arabic. There are strong linguistic differences between all of these varieties, affecting pronunciation and phone inventories as well as morphology, word order and vocabulary. Table 1 lists some examples of differences between MSA and Egyptian Colloquial Arabic (ECA), which is one of the more widely known dialects. In some cases, linguistic differences are so pronounced that the dialects in question are no longer intelligible to speakers of other dialects; in fact, linguists sometimes describe the dialects of Arabic as different languages, e.g. Versteegh (2001, p. 189) compares the relation between colloquial dialectal Arabic and MSA to that between Italian and Latin. The dialects are spoken languages and are almost never used in

أُحِبُّ آلسَّفَرَ إِلَىَ آلْقَاهِرَة.
آحب السفر إلى القاهرة.

Fig. 1. Arabic sentence with diacritics (upper row) and without (lower row).

writing.

Change	MSA	ECA	Gloss
/θ/ → /s/,/t/	/θala:θa/	/tala:ta/	ثلَاث *three*
/ð/ → /z/,/d/	/ðahab/	/dahab/	ذهب *gold*
/ay/ → /e:/	/saif/	/se:f/	صيف *summer*
inflections	yatakallam(u)	yitkallim	يتكلم 'he speaks'
vocabulary	Tawila	tarabeeza	*table*
word order	VSO	SVO	

Table 1
Linguistic differences between MSA and ECA.

2.2 Writing system

Arabic is written in a script representation consisting of 28 letters representing the consonants and the long vowels /i:/, /a:/, and /u:/. The short vowels, as well as certain other phonemic information such as consonant doubling, are not represented by letters but by diacritics, i.e. short strokes placed above or below the preceding consonant. An example of a sentence transcribed with and without diacritics is shown in Figure 1. Table 2 lists the complete set of diacritics for Arabic.

Most Arabic texts are not diacritized; exceptions are important political or religious texts or texts for beginning students of Arabic. A non-diacritized transcription is the most natural way of writing for native speakers; transcriptions which use a fully diacritized form, or a form where the script is translated into an orthographic form with standard ASCII characters, therefore tend to be more errorful and costly to produce since they require special training. The mapping from graphemes to phonemes is simple compared to languages like English or French: in most cases, there is a one-to-one relationship. The absence of diacritics, however, leads to a high degree of lexical ambiguity. The form كتب, for instance, has 21 possible diacritizations. In Debili et al. (2002) it was shown that a non-diacritized dictionary word has 2.9 corresponding diacritized forms on average. In the same study, an Arabic text of 23,000 script forms was analyzed, which showed a ratio of 11.6 possible diacritizations for

Diacritic	Name	Pronunciation/Meaning
اَ	fatHa (on alif)	/a/
اِ	kasra (under alif)	/i/
اُ	Damma (on alif)	/u/
رّ	shadda (on raa)	consonant doubling
رْ	sukuun (on raa)	vowel absence
اً	tanween al-fatHa (on alif)	/an/
اٍ	tanween al-kasr (under alif)	/in/
اٌ	tanween aD-Damm (on alif)	/un/

Table 2
Arabic diacritics.

every non-diacritized word.

2.3 Consequences for ASR

Since the dialects of Arabic are essentially spoken languages, the only way to collect large amounts of appropriate training data is to record and transcribe conversations, which is a slow and cost-intensive procedure. Existing collections of dialectal Arabic are small, and systems trained on these data sets typically do not achieve a state-of-the-art performance. Although several new collection and transcription efforts are underway (Zitouni et al., 2002; Siemund et al., 2002; Maamouri et al., 2004), the data requirements of large-vocabulary speech recognition will still exceed available resources for some time to come. This naturally leads to the question whether cross-dialectal data sharing techniques can be used to improve a dialect-specific system trained on sparse data.

We have previously investigated the use of MSA text data to enhance a language model for ECA (Kirchhoff et al., 2002); however, no improvements were obtained due to the different linguistic properties of the two varieties and the different topic structure of the corpora (broadcast news vs. informal conversations between friends and family members). More advanced techniques (e.g. data transformation prior to data pooling) could be applied to this problem but have to date not been explored.

In this study we investigate the possibility of sharing acoustic rather than language model training data across different varieties of Arabic. Our assumption is that additional data from a different dialect can increase the amount of training data for context-dependent phones (e.g. triphones) when the existing

training corpus is small. Even if the inventory of context-dependent units in different corpora is influenced by differences in pronunciation, vocabulary and topic structure, the number of training samples for context-dependent units might still be higher, which could result in better acoustic models. In the experiments reported here, we test this hypothesis by using data from MSA to improve acoustic modeling for ECA.

3 Data

Two different corpora were used for this study: the LDC CallHome corpus of Egyptian Colloquial Arabic, and the FBIS Modern Standard Arabic corpus.

The CallHome (CH) corpus is a collection of telephone conversations between family members or friends, with one speaker being located in the U.S. and the other in Egypt (mostly in the Cairene dialect region). We use a training set of 120 conversations (approximately 20 hours, 170K words), a development set of 20 conversations (32K words) and an evaluation set (eval03) of 10 conversations (11K words). Two different sets of transcriptions are available for this corpus: Arabic script-based transcriptions without diacritics, and "romanized" transcriptions, i.e. an ASCII-based representations which includes short vowels and other pronunciation information and thus resembles a phonetic transcription. An example is shown below:

Romanized: ilHamdulilla kuwayyisaB wi inti izzayyik
Script: الحمد لله كويسة و انتي ازيك

The recognition lexicon contains about 18K entries. The corpus has a high rate of disfluencies (about 9%), and an out-of-vocabulary rate of 5%.

The FBIS corpus is a collection of broadcast news shows from various radio stations in the Arabic-speaking world (e.g. Cairo, Baghdad, Riyadh, Damascus). The total size of the data set is 40 hours (240K words). The nature of the recordings differs widely, and includes clean studio-style speech, voice-overs with background music, and field reports recorded outdoors. This corpus was transcribed in Arabic script without diacritics. The size of the script-based lexicon is approximately 51K words.

In order to assess the degree of variation present in these corpora, we computed the vocabulary overlap as the percentage of shared unigrams, bigrams and trigrams (shown in Table 3). Prior to computing overlap, different orthographic conventions (e.g. the use of hamza) were normalized. As can be seen, the inventory of words (unigrams) only overlaps by 10%, and there is hardly any overlap in bigrams and trigrams. The percentage of shared triphones (determined according to lexicon pronunciations) was found to be 40% (a detailed

Varieties	ECA-MSA	BE-AmE
shared unigrams	10.3%	44.5%
shared bigrams	1%	19.2%
shared trigrams	$< 1\%$	5.3%

Table 3
Percentage of shared unigrams, bigrams and trigrams in the LDC ECA CallHome corpus (ECA) and the MSA FBIS corpus (MSA), and for the conversational British English (BE) CHRISTINE corpus and American English (AME) Broadcast News data.

description follows below in Section 6). It is interesting to compare these numbers to equivalent statistics for two varieties of English: conversational British English as represented by a subset (75K words) of the CHRISTINE corpus (Sampson, 2004), and an equivalent amount of American English Broadcast News data from the NIST 2004 Rich Transcription evaluations. Although topics and domains are similarly divergent, the number of shared n-grams is much higher. The triphone overlap for English was 85%. These results confirm that, at least for the corpora under investigation, that the difference between ECA and MSA are much stronger than for varieties of English.

4 Recognition System

Our baseline recognition system for the CallHome ECA corpus was developed using the SRI DECIPHER^{TM} multipass recognition engine. The system was trained on the 120-conversation CallHome training set. In previous work it was shown that conversational Arabic recognizers that utilize short vowel information yield lower word error rates than those trained on non-diacritized script-based representations (Kirchhoff et al., 2002; Messaoudi et al., 2004). We therefore use the romanized transcriptions for training and include acoustic models for short vowels. [1]

The front-end of our recognition system consists of 39 features (13 mel-frequency cepstral coefficients with first and second derivatives). Cepstral mean and variance normalization as well as vocal tract length normalization are applied. The acoustic models are non-crossword continuous-density genonic hidden Markov models (HMMs) (Digalakis and Murveit, 1994). 250 genones with 128 Gaussians per genone are used. Recognition proceeds in several stages: first, phone-loop adaptation with two Maximum Likelihood Linear Regression (MLLR) transforms is applied. A bigram language model is then used to

[1] We do not include separate acoustic models for doubled consonants (shadda) since did not have a significant effect on the performance.

generate the first-pass recognition hypotheses. These are used to adapt the acoustic models by constrained MLLR with six adaptation transforms. Next, bigram lattices are generated and expanded with a trigram language model. N-best lists are generated from the trigram lattices, and the final one-best hypothesis is obtained by N-best ROVER (Stolcke et al., 2000). This system is a simplified version of our best evaluation system for this data. For faster experimental turn-around we decided not to use maximum-mutual information training (all models are maximum-likelihood trained), and to use a single front-end (MFCC) only. Moreover, we do not perform lattice rescoring with cross-word triphones and nbest rescoring with more complex morphological language models.

The baseline word error rates obtained on the development and evaluation sets, are 57.3% and 42.7%, respectively. This is about 3% worse than the performance obtained by our best system (and within 5% of the best reported result of 37.5% on the evaluation set) while greatly facilitating experimentation.

5 Automatic diacritization

The information conveyed by Arabic diacritics is of a lexico-grammatical nature. Word-medial short vowels often indicate the part-of-speech of the word (e.g. noun vs. verb) as well as secondary grammatical categories, such as tense, voice, etc. Word-final vowels and tanweens indicate grammatical case and/or adverbial status. Therefore, knowledge of the syntactic context and morphological composition of the word form may help in determining the correct diacritization. If acoustic data is available, the acoustic signal can be used directly as an additional knowledge source.

The few previous studies on automatic diacritization of Arabic have exclusively concentrated on text-based information. An HMM-based approach to diacritization was presented in (Gal, 2002). This approach uses a bigram HMM decoder to derive diacritized sentences from non-diacritized sentences. The technique was evaluated on the Qur'an and achieved a word error rate (i.e. the percentage of incorrectly diacritized words) of 14%. An attempt at developing an automatic diacritizer for dialectal speech was reported in (Kirchhoff et al., 2002). The basic approach of this study was to use a small set of parallel script and diacritized data (obtained from the ECA CallHome corpus) in order to learn diacritization rules in an example-based fashion. This approach achieved a 16.6% word error rate on the ECA CallHome development set. Finally, various commercial automatic diacritization products for Arabic have been developed. These typically use a combination of statistical, rule-based and lexical information (e.g. proper names look-up tables). All of these

products are targeted towards MSA; to our knowledge, there are no products for dialectal Arabic. In a previous study (Kirchhoff et al., 2002) a commercial diacritization package 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. On the ECA text, the diacritization software obtained an error rate of 48%. Other studies (El-Imam, 2004) 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.

In contrast to the methods described above, we include acoustic information in our diacritization procedure, which consists of the following steps:

(1) First, all possible diacritized forms of the word are generated, together with the corresponding morphological analyses. This is accomplished by applying an automatic morphological analyzer. The analyzer produces all possible morphological analyses of the word, along with the appropriate diacritizations.
(2) A statistical tagger is trained on the morphological tags produced by the analyzer. The purpose of this step is to obtain probabilities for transitions between morphological tags (and thereby probabilities for transitions between diacritizations). The tagger exploits contextual information in the form of a trigram model.
(3) All possible sequences of morphological tags for a given utterance are scored by the trained tagger and are output in the form of weighted transition networks of diacritized forms.
(4) The decoder of our recognition system is used in combination with the the transition networks, the acoustic waveforms and previously trained acoustic models in order to identify the most likely paths. Each best path identifies the diacritized word sequence which is then used as the "true" training transcription.

Each of these steps is described in detail in the following sections.

5.1 Generating candidate diacritizations

Each Arabic script form in our MSA text is passed through a morphological analysis tool, the Buckwalter stemmer available from the LDC. The stemmer is based on a lexical lookup table and morphophonemic rules. It produces all possible morphological analyses of the script form; as a by-product, the corresponding diacritizations are output as well. Figure 2 shows an example. In the version of the analysis software that was used for this study, case vowel

Number of tags	763
Number of word tokens	547, 752
Number of word types	51,085
Average number of tags per word	10.7
Maximum number of tags per word	13

Table 4
Statistics of morphological analysis of FBIS corpus.

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>

Fig. 2. 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).

endings were not produced; these are added later in the form of pronunciation variants in the dictionary used by the aligner (see below Section 5.1.2). Table 4 shows the total number of different morphological tags produced by the stemmer, the number of word types and tokens in the corpus, and the average number of diacritized forms per script form. A small percentage of words (1.5%) could not be analyzed, due to spelling/transcription errors, proper names, or other gaps in the lexicon used by the stemmer. These words were replaced by a generic REJECT tag which is mapped to an acoustic garbage model during acoustic modeling.

5.1.1 Morphological tagging

The next step is to find probabilities for transitioning from one tag to the next (and thereby from one diacritized form to the next), in order to model contextual constraints on diacritization. To this end we trained a statistical trigram tagger, according to a standard HMM-based tagging model (see e.g. (Brants,

2000)):

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

where t is a tag, w is a word, and n is the total number of words in the sentence. In general, taggers can be trained in a supervised or an unsupervised way (e.g. (Schütze, 1993)). In the first case, a training set with annotated true tags is available; in the second case, only the possible tags for each word are known and the probabilities in the tagger (the lexical probabilities $P(t_i|w_i)$ and the contextual probabilities $P(t_i|t_{i-1}, t_{i-2})$) are estimated iteratively using an Expectation-Maximization (EM) (Dempster et al., 1977) update procedure. Since manual tag annotations were not available for the FBIS corpus we used unsupervised training. During testing, only the word sequence is observed, and the tags (and thereby the corresponding diacritizations) are found by identifying the sequence that maximizes the probability in Equation 1.

The tag sets commonly used for standard tagging tasks usually consist of 30-50 tags. By contrast, the tag set produced by the morphological analyzer generates a much larger set of 763 tags. The reason for this is that many Arabic words are morphologically complex and express information equivalent to entire phrases in languages like English, e.g.

وباستطاعتها (wbAstTAEthA):

CONJ+PREP+NOUN+NSUFF_FEM_SG+POSS_PRON_3FS
and + by/with + capability/possibility + its/their/her
"and with its capability"

In order to train the tagger robustly while preserving the morphosyntactic information relevant for diacritization, the original tag set was reduced to approximately half its size by mapping it to the tags used in the LDC Arabic Treebank project, which was also developed based on the Buckwalter morphological analyzer. Tag correspondences were identified by longest common substring matching. The final number of tags was 392.

We used the graphical modeling toolkit GMTK (Bilmes and Zweig, 2002) to train the tagger. Training was done using Expectation-Maximization. The probability distributions in Equation 1 were not smoothed. Since the tagger is ultimately applied to the same set as it is trained on (the ultimate goal is to derive a diacritized transcription of the training data), using smoothing to avoid zero probabilities for unseen tag sequences or word/tag combinations was not a major concern. The tagger was used to assign probabilities to all possible trigram sequences of morphological tags (and their associated diacritizations) for each utterance in the FBIS corpus, thus creating probability-weighted transition networks of diacritized forms. An example is shown in Figure 3.

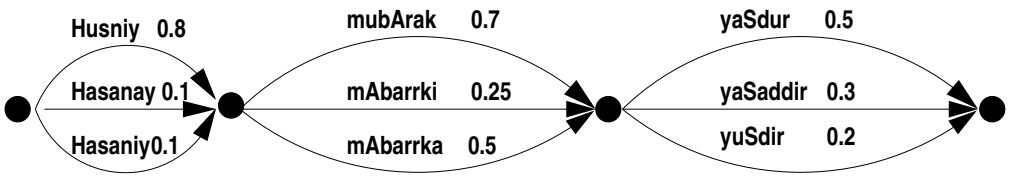

Fig. 3. Transition network showing possible diacritizations of حسني مبّارك يصدر "Hsny mbArk ySdr".

5.1.2 Acoustic scoring of transition networks

The final step in our diacritization procedure is the alignment of the weighted transition networks with the acoustic waveforms. To this end we use the acoustic models from our baseline ECA recognition system in combination with the DECIPHER decoder and a pronunciation dictionary. The pronunciation dictionary lists word variants with case vowel endings (since these were not produced by our version of the Buckwalter stemmer), as well as some additional pronunciation rules, in particular the /a/ - /at/ alternation of so-called *taa marbuta* endings. Words rejected by the Buckwalter stemmer were paired with a generic acoustic reject model. After performing an initial alignment of the acoustic data, approximately 10% of the data were discarded due to alignment failures. In most cases, this was due to segmentation problems caused by background noise (e.g. music). The remaining 90% were used for the experiments reported below in Section 6.

5.2 Evaluation

In order to evaluate our diacritization method, and to assess the relative contributions of the morphological, contextual and acoustic knowledge sources, the automatically diacritized transcriptions were compared against manual diacritizations on a subset of the data.

A subset of 500 words was annotated by transcribers experienced in orthographic transcription of Arabic. The annotation was in Arabic script with diacritics, i.e. not in romanized form. It was subsequently converted to the ASCII representation (Buckwalter transliteration) used in our recognition and automatic diacritization system, and it was aligned to the automatic output. The alignment was first performed by a standard dynamic programming procedure using a Levenshtein distance measure; it was then manually verified to ensure that diacritization slots were lined up with each other. Performance was measured in two different ways:

- character-level error rate: this is the percentage of diacritization errors rel-

ative to the number of possible diacritization slots:

$$\frac{\#\text{ of diacritization errors}}{\#\text{ of diacritization slots}} \times 100 \quad (2)$$

Diacritization errors occur under the following circumstances:
- no diacritic is present in the manual reference annotation but a diacritic was hypothesized by the automatic diacritizer (insertion)
- a diacritic is present in the hand annotation but was not hypothesized by the diacritizer (deletion);
- a wrong diacritic was substituted for the correct one

In the case of tanween endings, which consist of a vowel followed by /n/, one error is assigned if the vowel is recognized correctly but the nasal is not; two errors are assigned if neither of them is recognized.

- word-level error rate, i.e. the percentage of incorrectly diacritized words:

$$\frac{\#\text{ of incorrectly diacritized words}}{\#\text{ of words}} \times 100.0 \quad (3)$$

In order to be counted as incorrect, at least one diacritization error must have occured.

The system was run in several different configurations, in order to determine the relative contributions of the different knowledge sources:

(1) *morphological, contextual, and acoustic information*: This configuration corresponds to the full diacritization procedure described above. Different weights (1 and 5) for the probabilities obtained from the tagger were tested in order to evaluate the importance of the tagger information.

(2) *morphological and acoustic information*: In this configuration, the probabilities from the tagger were weighted by 0, which corresponds to the elimination of the contextual information. Thus, diacritization is constrained only by acoustic and morphological information.

(3) *acoustic information*: In order to assess the feasibility of diacritization using only acoustic information, we created word transition networks that allow all possible diacritics to be inserted at all possible positions. In principle, every consonant can be followed by three vowels, by a duplication of itself (shadda), or it can transition directly to the next consonant. Combinations of shadda and a short vowels are also possible. Word endings can have the three tanweens and three case vowel endings. Allowing for all these possibilities creates a very large number of possible word forms. In order to reduce the number of variants, the shadda insertion was omitted since our recognizer does not use separate acoustic models for doubled consonants. Furthermore, we only allow the three case vowels at word endings since the tanween are comparatively rare. In spite of these simplifications, possible word forms are drastically overgenerated and the size of the networks drastically decreases the speed of the decoder.

(4) *morphological information*: in this case, a random path through the diacritization networks was chosen, without using tagger probabilities (weight = 0) and without using acoustic alignment. Thus, the result represents the diacritization accuracy that can be obtained by only using morphological information, without use of contextual or acoustic information. Since the word-final case endings are only represented in the pronunciation dictionary used by the acoustic aligner, they can by definition not be recognized in this case. For this reason, we consider two different evaluation conditions: in the first case, we count all diacritization slots (including word-final case endings), in the second case, we only evaluate word-internal vowels.

Table 5 lists the diacritization results. First, we observe that error rates are considerably higher throughout when word-final vowels are included. The reason may be that word-final vowels are often coarticulated with the following word, such that both the recognition of these vowels and the reference annotation may be less reliable in these cases. Second, the use of contextual information seems to hurt rather than help. A possible reason is the large tag set that was used in the tagger; as a consequence, the contextual probabilities may not be robust enough. A further reduction of the tag set to a typical size of 20-30 may be advisable here. The most striking result is obtained when using only acoustic information. In this case, both the diacritization error rate and the word error rate increase significantly. It was found that most errors result from spurious vowel insertions (e.g. بَغدَاد → بَغُدَاد (Baghdad → Baghudad), which may be due to the quality of certain neighbouring consonants. The possibilities for vowel insertions are greatly constrained when using lexical/morphological knowledge in addition to acoustic information. The diacritizations created by the morphological analyzer essentially define permissible words, which reduces error rates considerably. However, the final row of results shows that acoustic information also plays a significant role in improving diacritization accuracy. As explained above, this condition could only be evaluated for word-internal vowels.

6 Recognition Experiments

Our goal is to use the automatically diacritized MSA data corpus to improve the acoustic models in our ECA recognizer. To this end we use the diacritized transcriptions obtained by the best of the diacritization methods described above (Method 1) as the training transcriptions for the MSA corpus. The phone inventories for ECA and MSA are fairly similar; differences are shown in Table 6. For the purpose of aligning the training data, the MSA-specific phones were replaced with the corresponding ECA phones. An arbitrary choice

Information used	with word-final vowels		w/o word-final vowels	
	Word level	Character level	Word level	Character level
1a) acoustic + morphological + contextual (tagger prob. weight=1)	27.3	11.54	11.8	5.2
1b) acoustic + morphological + contextual (tagger prob. weight=5)	27.3	13.24	14.6	7.5
2) acoustic + morphological (tagger prob. weight=0)	27.3	11.94	9.9	3.4
3) acoustic only	50.0	23.08	37.3	16.2
4) morphological only	–	–	22.9	12.7

Table 5
Automatic diacritization error rates (in %), with and without counting word-final case endings. Condition (4) could only be evaluated for word-internal vowels since word-final vowels are produced by the acoustic aligner.

	phones					
MSA	θ	ð	q	ai	au	ʤ
ECA	s,t	z,d	ʔ	e	o	g

Table 6
Differences in MSA and ECA phone inventories (in IPA notation).

was made in cases where several ECA phones may correspond to the original MSA phone since the choice is usually dependent on the individual word and on the speaker.

In order to verify our assumption that the use of additional data (even when drawn from a different source) results in a better coverage of context-dependent phones, we collected triphone statistics from the two different corpora, which are shown in Table 7. The first three rows show the statistics for the training set (excluding foreign words and hesitations). We see that out of the 8780 triphones in the CallHome training corpus, about 40% also occur in the FBIS corpus. The number of training samples for these triphones is increased by a factor of 2.5. The next three rows show the same statistics for the development set. Here, too, a larger number (more than 50%) of CallHome triphones obtain a larger number of training samples through the added FBIS data. The FBIS data also contains triphones which occur in the CallHome development set but not in the CallHome training set. However, we currently cannot benefit from those since our recognition lexicon is limited to in-vocabulary words.

	in CH	in FBIS
# unique triphones in train set	8780	6211
# shared training set triphones	3476	
# samples for shared training set tripones	888,875	2,261,266
# unique triphones in dev set	4167	–
# shared dev set triphones	2141	
# samples for shared dev set triphones	875,264	1,802,778

Table 7
Triphone statistics for CH and FBIS. The upper three rows show the number of triphones found in each of the training corpora, the number of shared triphones (found in both), and the number of samples for the shared triphones. The lower part of the table repeats these statistics for the triphones in the CH dev set (note that the FBIS corpus only provides a training set, while system development is done on the CH dev set)

6.1 Basic combination experiments

The automatically diacritized FBIS data was downsampled from 16 to 8 kHz and was added to the CH training set. Vocal tract length normalization was performed by speaker clustering and applying individual normalization parameters (i.e. spectral scaling factors) to each cluster. This has the effect of eliminating spectral shifts due to speakers' different vocal tract lengths. Cepstral mean and variance normalization, which help eliminate additive and convolutive noise and channel variation, are standard features of our front-end. Acoustic models were trained by pooling the CH and FBIS training data, using a 2:1 weighting ratio of CH data to FBIS data. The number of genones was increased from 250 to 300, in order to be able to take advantage of the larger training set. After training, a final MAP adaptation pass (Digalakis et al., 1995) was performed on the CallHome data. This was done in order to counter effects caused by the different acoustic conditions of the FBIS corpus, which might otherwise result in large model variances and large shifts in model means. As a final step in the recognition procedure, we use N-best ROVER combination of the CH-only and CH+FBIS N-best lists. Results for the different systems, as well as for the combined system, are shown in Table 8.

We observe that the performance of the CH-only and the CH+FBIS system is approximately the same; combining both systems, however, improves the word error rate moderately by 1% absolute, demonstrating that the acoustic models in the two systems learn different types of information. An analysis of the recognition hypotheses produced by the two systems clearly showed differences in errors, though no systematic pattern could be observed. Errors

System	dev	eval03
CH-only	56.1	42.7
CH+FBIS	56.3	42.6
combined	55.3	41.7

Table 8
Word error rates (%) obtained by the baseline system (CH-only), by the system trained on the pooled CH+FBIS data (CH+FBIS), and by combining N-best lists from both systems.

were not correlated with speaker, conversation side, or similar variables. A drawback of this data sharing procedure is the acoustically different nature of the FBIS corpus: the variety of acoustic background conditions in this corpus has the effect of adding noise to the training data. Although mean, variance, and Vocal Tract Length normalization were applied [2] and models were adapted on the CH data, the increased acoustic variability may still affect the model parameters considerably, and may override the benefits of better modeling of phonetic variability to some extent.
This problem might be addressed by selecting subsets of the MSA corpus rather than adding all of the data to the ECA training set. This is described in the next section.

6.2 Data filtering

Additional experiments were performed in order to determine whether data filtering prior to combination might result in better performance. The FBIS utterances were filtered in two different ways: (a) according to the number of additional triphone samples provided, and (b) according to their acoustic background condition. In the first case, triphones were extracted from the reference transcription of the utterance, and utterances were ranked according to the amount by which the counts for existing CH triphones were increased by these additional training samples. In the second case, phone backtrace information obtained from forced alignments during automatic diacritization were used to identify the acoustic scores for the silence model (which had been trained on the CH corpus). Utterances from the FBIS corpus were ranked according to the duration-normalized acoustic scores of the silence model. High scores were assumed to indicate a good match of the acoustic background condition of the utterance to the acoustic properties of the CH corpus. The top 75% of all utterances from the FBIS corpus were selected for the combined CH+FBIS training set. Table 9 shows the results. Although the performance of

[2] For the FBIS corpus the normalization parameters were estimated on automatically obtained speaker clusters, while for CH the conversational side was used.

Filtering method	CH-only	CH+FBIS	combined
no filtering	57.3	58.4	56.5
by acoustic score	57.4	57.8	56.5
by triphone score	57.4	57.2	56.5

Table 9
Word error rates (%) on the CH dev set for the CH+FBIS and combined system after data filtering. The results for the CH-only system are given for comparison. Results are based on first-pass recognition without adaptation.

the CH+FBIS system was improved by data filtering, the combination results did not improve. This can be explained by the fact that the filtering procedure renders the system more similar to the CH system. As a consequence, errors become more similar and the beneficial effect of system combination disappears.

6.3 Weighted acoustic score combination

In a third experiment, we combined the acoustic scores from the CH-only and the CH+FBIS systems with different weights for different sets of triphones.

The weighted combination of acoustic or language model scores has consistently shown advantages over unweighted combination methods (Byrne et al., 2000; Vergyri, 2000; Glotin et al., 2001). We use the framework of weighted log-linear score combination (Beyerlein, 1998; Ostendorf et al., 1991). The weights of the log-linear combination are trained discriminatively; they are optimized to directly minimize the word error rate on the development set. The optimization method is a simplex downhill method known as amoeba search (Nelder and Mead, 1965).

A range of different weighting schemes were applied. First, triphones were divided into subsets according to their frequency of occurrence in FBIS vs. CallHome. Three different clusters were established, depending on whether the triphone was more frequent in FBIS, more frequent in CH, or equally frequent in both corpora. Our assumption was that phones with different frequency characteristics would be weighted differently. An additional cluster was used for pause models. The results are shown in Table 10. We observe an additional 0.3%-0.5% improvement over the previous results (Table 10); however, the same improvement is observed in the baseline system when using a phone-class based weighting procedure.

An analysis of the combination weights showed that the highest weights were given to the phone clusters having more samples in CH or more samples in

	dev	eval03
CH only	55.6	42.3
CH+FBIS	55.8	42.5
combined	54.8	41.4

Table 10
Word error rates (%) for CH-only, CH+FBIS and combined system after phone-class dependent combination of acoustic model scores.

FBIS (as opposed to phones having an equal number of samples in both). This demonstrates that the weight optimization procedure favours models whose parameters preserve the characteristics of the distinct corpora, whereas models built from equal contributions of both corpora seem to provide less information.

6.4 Effect of different diacritization procedures on joint training

In the above experiments, the best-performing diacritization method was used. An additional goal of this study was to determine how different diacritization procedures with varying degrees of accuracy influence acoustic model training and data sharing. Table 11 shows recognition results for systems trained on all different diacritization procedures described in Section 5, except for the morphology-only condition since it produces only word-internal diacritics. These results were obtained without data filtering or phone-class dependent weighting. We found that the different procedures produce virtually the same results. Even Method 3, which uses only acoustic information and has a high diacritization error rate, does not have a detrimental effect on the recognition performance.

The reasons for this are threefold: first, we only select those triphones from the automatically diacritized FBIS data that occur in the CallHome lexicon. Since our final goal is recognition of the CallHome test data, we use only the words in the CallHome lexicon and therefore have a fixed, constrained list of triphones we can use. Those triphones that do not occur in this lexicon are discarded; thus, many of the triphones resulting from erroneously introduced vowel segments will not have any effect on the system. In particular, the CallHome triphone list does not include word-final case ending vowels since case endings are not normally used in dialectal Arabic. Therefore, many of the wrong diacritizations are eliminated. Note that this use of automatic diacritization (diacritization for data sharing) is quite different from an application where diacritized data is used in its entirety to train a recognition system from the beginning. In that case, a much stronger effect of diacritization errors might be observed. Second, EM training allows model parameters to be adjusted

during joint training on both corpora; the influence from the manual diacritic information in the CallHome data might be sufficient to keep parameters from being tuned too much to the errorful FBIS data. Finally, the acoustic models undergo a final adaptation pass to the (accurately transcribed) CallHome data, which helps eliminate noise introduced by the diacritization procedure.

	dev		eval03	
System	alone	combined	alone	combined
CH-only	56.1		42.7	
CH+FBIS 1a(weight 1)	56.3	55.3	42.2	41.6
CH+FBIS 1b(weight 5)	56.1	55.2	42.2	41.8
CH+FBIS 2	56.2	55.3	42.4	41.6
CH+FBIS 3	56.6	55.7	42.1	41.6

Table 11
Word error rates (%) obtained after the final recognition pass; in isolation (alone) and with ROVER combination with the baseline system (combined). FBIS1, FBIS2 and FBIS3 correspond to the diacritization procedures described in Table 5. FBIS1a and FBIS1b systems use acoustic, morphological and contextual knowledge with different weights (1 vs 5) for the contextual tagger probabilities. FBIS2 uses acoustic and morphological knowledge; FBIS3 uses only acoustic information. No data filtering or phone-class dependent weighting was used.

It should be observed, however, that erroneous diacritization might have a much more pronounced effect in other applications, e.g. when using the resulting transcriptions for training a completely new system (i.e. where the set of triphones to be used is not constrained by a manually constructed dictionary), or for natural language processing applications, such as speech-to-speech translation.

7 Conclusions

We have presented a study on data sharing across different dialects of Arabic (MSA and ECA) to improve the recognition performance of a recognizer for conversational Arabic. An equal amount of MSA data was added to a 20-hour corpus of ECA to increase phonetic coverage. A prerequisite to this procedure was the automatic diacritization of MSA data. We tested various diacritization procedures based on a combination of knowledge sources: morphological, contextual and acoustic. The diacritization techniques were evaluated both by comparison with manually diacritized data and in the context of the recognition system trained on both data sources.

We found that high diacritization accuracy (12% character-level error rate for all diacritics, 3.4% for word-internal diacritics) can be achieved by our proposed techniques. Morphological information was found to be the most important knowledge source in determining the correct diacritization since it provides constraints on possible Arabic words.

The combination of systems trained on pooled ECA and automatically diacritized MSA data, and on ECA data only, resulted in significant word error reductions (from 56.1% to 54.8% on the development set, and from 42.7% to 41.4% on the evaluation set)

The observed gains could potentially be larger when using a larger MSA data set. Another way of improving the current data sharing approach might be to develop techniques for reducing purely acoustic variability across corpora (that are caused by background noise or different recording channels) while retaining the benefits of additional training data for variability caused by phonetic context and coarticulation. Various ways of data filtering prior to training and combining the individual systems were investigated. While some improvements to the baseline systems were obtained, their combination did not improve the overall word error rate. We also tested the effect of using different combination weights for phone classes established on the basis of phone frequency counts in the different corpora. This led to improvements in both individual baseline word error rates and in the combination result; however, the relative improvement obtained by the combination was not greater than before. A technique not investigated here is the modification of the triphone clustering procedure, e.g. by varying the number of genones. Finally, our experiments could be repeated with only the data from the Egyptian part of the FBIS corpus. Since some regional dialectal influences are present even in MSA (especially in the pronunciation), limiting the additional MSA data to this specific region might be preferable. On the other hand, this would severely limit the amount of available data, resulting in less than the in-domain data from the CallHome corpus.

Finally, a comparison of different systems trained on the differently diacritized data showed that the diacritization procedures did not have a significant effect on the acoustic models. This could be due to the pre-selection of triphones according to a manually constructed dictionary, the use of EM training, or the adaptation to accurately transcribed ECA data. This is a promising result since it indicates that even inaccurately diacritized data can be used in cross-dialectal data sharing.

Acknowledgements
This material is based upon work funded by DARPA under contract No. MDA-972-02-C-0038. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of these

agencies.

References

Beyerlein, P., 1998. Discriminative model combination. In: Proceedings of the International Conference on Acoustic, Speech and Signal Processing. pp. 481–484.

Bilmes, J., Zweig, G., 2002. The Graphical Models Toolkit: An open source software system for speech and time-series proessing. In: Proceedings of the International Conference on Acoustic, Speech and Signal Processing. pp. 3916–3919.

Brants, T., 2000. TnT - a statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference. pp. 224–231.

Byrne, W., Beyerlein, P., Huerta, J., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D., Wang, W., 1999. Towards language independent acoustic modeling. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

Byrne, W., Beyerlein, P., Huerta, J., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri, D., Wang, W., 2000. Towards language independent acoustic modeling. In: Proceedings of the International Conference on Acoustic, Speech and Signal Processing. pp. 1029–1032.

Debili, F., Achour, H., Souissi, E., 2002. De l'étiquetage grammatical à la voyellation automatique de l'arabe. Correspondances de l'Institut de Recherche sur le Maghreb Contemporain 17.

Dempster, A., Laird, H., Rubin, D., 1977. Maximum-likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. Ser. B 39.

Digalakis, V., Murveit, H., 1994. GENONES: Optimizing the degree of mixture tying in a large vocabulary hidden markov model based speech recognizer. In: Proceedings of the International Conference on Acoustic, Speech and Signal Processing. pp. I–537–540.

Digalakis, V. V., Rtischev, D., Neumeyer, L. G., 1995. Speaker adaptation using constrained estimation of Gaussian mixtures. IEEE Transactions on Speech and Audio Processing 3, 357–366.

El-Imam, Y. A., 2004. Phonetization of Arabic: rules and algorithms. Computer, Speech and Language 18, 339–373.

Gal, Y., 2002. An HMM approach to vowel restoration in Arabic and Hebrew. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages. Association for Computational Linguistics, Philadelphia, pp. 27–33.

Glotin, H., Vergyri, D., Neti, C., Potamianos, G., Luettin, J., 2001. Weighting schemes for audio-visual fusion in speech recognition. In: Proceedings of the International Conference on Acoustic, Speech and Signal Processing. pp. 173–176.

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

Köhler, J., 1998. Language adaptation of multilingual phone models for vocabulary-independent speech recognition tasks. In: Proceedings of the International Conference on Acoustic, Speech and Signal Processing. pp. 417–420.

Maamouri, M., Buckwalter, T., Cieri, C., 2004. Dialectal Arabic telephone speech corpus: principles, tool design, and transcription conventions. In: Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools. Cairo, Egypt.

Messaoudi, A., Lamel, L., Gauvain, J., 2004. The LIMSI RT-04 BN Arabic system. In: Proceedings of the 2004 DARPA Rich Transcription Workshop. Pacific Palisades, New York.

Nelder, J., Mead, R., 1965. A simplex method for function minimization. Computing Journal 7(4), 308–313.

Ostendorf, M., Kannan, A., Austin, S., Kimball, O., Schwartz, R., Rohlicek, J. R., 1991. Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In: Proceedings of the DARPA Speech and Language Workshop. pp. 83–87.

Sampson, G., 2004. www.grsampson.net/ChrisDoc.html.

Schultz, T., Waibel, A., 2001. Language independent and language adaptive acoustic modeling for speech recognition. Speech Communication 35, 31–51.

Schütze, H., 1993. Part-of-speech induction from scratch. In: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. pp. 251–258.

Siemund, R., Heuft, B., Choukri, K., Emam, O., Tropf, H., Edge, O., Shammass, S., Moreno, A., Rodriguez, A. N., Zitouni, I., Iskra, D., 2002. Orientel: Arabic speech resources for the IT market. In: Proceedings of the LREC Workshop on Arabic Language Resources and Evaluation: Status and Prospects.

Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Gadde, V. R. R., Plauche, M., Richey, C., Shriberg, E., Sonmez, K., Weng, F., Zheng, J., 2000. The SRI March 2000 Hub-5 conversational speech transcription system. In: Proceedings of the NIST Speech Transcription Workshop. College Park, MD.

Vergyri, D., 2000. Integration of multiple knowledge sources in speech recognition using minimum error training. Ph.D. thesis, Johns-Hopkins University.

Versteegh, K., 2001. The Arabic Language. Edinburgh University Press.

Zitouni, I., Olive, J., Iskra, D., Choukri, K., Emam, O., Gedge, O., Maragoudakis, E., Tropf, H., Moreno, A., Rodriguez, A. N., Heuft, B., Siemund, R., 2002. ORIENTEL: speech-based interactive communication applications for the Mediterranean and the Middle Eeast. In: Proceedings of ICSLP - Interspeech. pp. 325–328.