CROSS-DIALECTAL ACOUSTIC DATA SHARING FOR ARABIC SPEECH RECOGNITION

Katrin Kirchhoff

Department of Electrical Engineering
University of Washington, Seattle, WA, USA

Dimitra Vergyri

SRI International
Menlo Park, CA, USA

ABSTRACT

The automatic recognition of Arabic dialectal speech is a challenging task since Arabic dialects are essentially spoken varieties, for which only sparse resources (transcriptions and standardized acoustic data) are available to date. In this paper we describe the use of acoustic data from Modern Standard Arabic (MSA) to improve the recognition of Egyptian Conversational Arabic (ECA). The cross-dialectal use of data is complicated by the fact that MSA is written without short vowels and other diacritics and thus has incomplete phonetic information. This problem is addressed by automatically vowelizing MSA data before combining it with ECA data. We described the vowelization procedure as well as speech recognition experiments and show that our technique yields improvements over our baseline system.

1. INTRODUCTION

Recent research in large-vocabulary conversational speech recognition has expanded to accommodate a wider range of languages (e.g. Mandarin and Arabic) in addition to more "mainstream" languages like English or Spanish. Often these new languages present problems that are not encountered in mainstream languages, such as extreme dialectal variation and non-standardized speech representations. Arabic in particular is characterized by its multitude of dialects. While one variety, Modern Standard Arabic (MSA) is used in writing, TV and radio broadcasts and for formal communication, all informal communication is typically carried out in one of the regional dialects of Arabic. The linguistic differences between different dialects, and between dialects and MSA, are considerable and affect pronunciation, phonology, morphology, syntax, and the lexicon. Moreover, the regional dialects of Arabic are spoken languages; very little written dialectal material exists. This is a serious problem for the automatic recognition of Arabic dialectal speech since large-vocabulary recognizers rely on large amounts of training material. Previous attempts at utilizing MSA data to improve language modeling for Egyptian Colloquial Arabic (ECA) [1] were largely unsuccessful and demonstrated that the two varieties do indeed behave like two different languages.

In this paper we attempt to use MSA data to improve not the language model but the acoustic models in a large-vocabulary conversational speech recognizer for ECA. Acoustic differences between these two varieties are smaller than the differences at the language level, and since only a small amount of acoustic data is currently available for ECA, acoustic models might benefit from a larger amount of similar data that provides more training instances of context-dependent phones. Moreover, the difference between dialectal and MSA speech is not necessarily clear-cut; it is a continuum, with speakers varying between the two ends of the continuum depending on the situational context. Cross-dialectal data sharing may be helpful in modeling this type of mixed speech.

Our approach is similar to sharing acoustic training data across different languages. Previous studies in this field [2, 3] have mainly addressed the problem of building a recognition system for a new target language given several source languages with sufficient acoustic data. Data sharing was implemented through the use of multilingual acoustic models trained on pooled data from the source languages. Extensive experiments were reported in [3], where cross-language transfer of acoustic models (without any data from the target language) was compared to multilingual training followed by adaptation, bootstrapping and retraining, and to training on a large set of data from the target language. It was found that multilingual training plus adaptation performed only slightly worse than training on the target language.

In contrast to these studies, Arabic presents an additional problem: whereas ECA is available in a romanized form with almost phonetic spelling, MSA is usually written without short vowels. Thus our goal is to combine two different data sets for acoustic modeling, one of which has incomplete phonetic information. This is done by automatically vowelizing the MSA data before combining it with the ECA data.

The following describes the linguistic differences between MSA and ECA in greater detail. Section 3 provides a description of the data used for this study. In Section 4 the process of automatic vowelization of the MSA data is explained. Section 5 describes the baseline systems and ex-

periments and Section 6 discusses the results.

2. LINGUISTIC DIFFERENCES BETWEEN MSA AND ECA

MSA has a system of 28 phones, which largely overlaps with the phone system of ECA. Differences in the phone inventory include the substitution of ECA [g], [d] and [t] for MSA /ʤ/, /ð/ and /θ/, respectively, modification of /a ɪ/ and /au/ in MSA to [ɛ] and [o] in ECA, and the replacement of the uvular stop /q/ in MCA by glottal stop [ʔ] in ECA. A further pronunciation differences is the insertion of a so-called linking vowel into consonant clusters spanning word boundaries in ECA but not in MSA. Morphological differences are e.g. the lack of case endings on nouns in ECA; and different inflectional morphs. All of these factors influence the acoustic structure (triphone statistics) of the two varieties; another important factor is linguistic usage: MSA is typically used for formal communication and in the media. Existing corpora of MSA data are mostly broadcast news corpora, such that the range of topics influences the lexical choice and thereby the triphone structure. However, many of triphones can still be expected to be shared between the two varieties. Whereas ECA is a spoken variety and is almost never written, MSA is the language of writing. However, it is mostly written without diacritics (short vowel markers and special markers for consonant doubling etc.), consisting of consonant and long vowels only. In order to be able to use an MSA text for ASR it either needs to be diacritized, or the recognizer must use grapheme-based models [4]. Since our goal is to combine MSA data with romanized ECA data which does include vowel information, automatic diacritization of the MSA data was a prerequisite for our experiments.

3. DATA

We used two different corpora for this study: the LDC CallHome (CH) corpus of Egyptian Colloquial Arabic and the FBIS corpus of MSA. The CallHome corpus consists of telephone conversations between native speakers of Egyptian Arabic. The training set used for the experiments described in this paper consists of the "train", "h5_new" and "eval96" subsets and contains approximately 200K words. The FBIS corpus is a collection of radio news casts from various radio stations in the Arabic speaking world (Cairo, Damascus, Baghdad) totalling approximately 40 hrs of speech (roughly 240K words). The CallHome corpus was 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 includes information about short vowels and other diacrictics. The transcription of the FBIS corpus was done in Arabic script only and does not contain any diacritic information. The state-of-the-art word error rate obtained on the CH eval03 set in the 2003 EARS evaluations was 37.5%; the authors' system obtained 39.7%.

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

4. AUTOMATIC DIACRITIZATION

Since the FBIS transcriptions did not contain diacritics we applied the following automatic diacritization procedure:

1. Generate all possible diacritized variants for each word, along with their morphological analyses.
2. Train an unsupervised tagger to assign probabilities to sequences of morphological tags.
3. Use the trained tagger to assign probabilities to all possible diacritizations for a given utterance.
4. Use the weighted diacritizations as pronunciation networks and use trained acoustic models from a different corpus to find the most likely diacritization.

For the first step we used the Buckwalter stemmer, which is an Arabic morphological analysis tool available from the LDC. The stemmer produces all 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, which was conflated into smaller set of 392 consistent with the tag set used in the LDC Arabic TreeBank corpus.

This set was also developed based on the Buckwalter morphological analysis and can be matched to the FBIS tags by finding the longest common substring. We adopted a standard trigram tagging model:

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

Since the true tag assignment was not known, only the set of possible tags for each word were available during training. The probabilities for $P(w_i|t_i)$ and for the tag sequence model $P(t_i|t_{i-1}, t_{i-2})$ were updated iteratively using the Expectation-Maximization algorithm. We used the graphical modeling toolkit GMTK [5] to train the tagger. The trained tagger was then used to assign probabilities to all possible sequences of three successive tags and their associated diacritizations, for all utterances in the FBIS corpus. The resulting word networks were used in combination with the acoustic models trained on CallHome to find the most likely word sequence. Since the Buckwalter stemmer does not produce case endings for nouns, the word forms obtained by adding case endings were included as variants in the pronunciation dictionary used by the aligner. In some cases (1.5%) 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. 10% of the FBIS data was discarded since no alignment could be found. The remaining 90% were used for our experiments. Since a manual diacritization was not available as a reference standard we cannot give an accurate assessment of the diacritization error rate. However, listening to a small set of audio files and comparing the corresponding automatically diacritized word strings yielded an estimated diacritization accuracy of approximately 95%.

5. BASELINE SYSTEMS AND EXPERIMENTS

One motivation for using cross-dialectal data was the assumption that infrequent triphones in the CallHome corpus might have more training samples in the larger MSA corpus. In order to verify this assumption we collected triphone statistics on the diacritized MSA corpus and compared them to the CallHome triphone statistics. Since we did not use cross-word triphones in our recognizer, only within-word triphones were considered. The total number of unique triphones in the CallHome training set is 8780 (computed on Arabic words only, excluding foreign words, hesitations, etc.), compared to 6211 unique triphones in the FBIS corpus. 40% of the CallHome triphones also occur in the FBIS corpus and can thus potentially profit from additional training data. The overall number of training samples for these triphones in the FBIS corpus is 2.5 times larger than in the CallHome corpus alone. About 50% percent of the triphones have more occurrences in FBIS than in CallHome, 14% have more samples in CallHome than FBIS, and 36% have an approximately equal number of samples ($+/-$ 50) in both corpora. We computed the same statistics for the triphones in the development set and obtained similar results: of the 4167 unique triphones in the development set, 2141 occur in both training corpora and have in total about twice as many training samples in FBIS as in CallHome. The FBIS corpus also contains additional triphones which might occur in the CallHome development and test sets but not in the training set - however, since the recognition lexicon only contains the words in the training set, we currently do not have a way of utilizing these.

5.1. Baseline Systems

We trained two different systems, one trained only on CallHome data (CH-only), and one trained on pooled data from both corpora (CH+FBIS). Both systems used the same front end and modeling techniques. The front-end consisted of 39 mel-frequency cepstral coefficients (13 base coefficients + first and second differences). Mean and variance as well as VTL normalization were performed per conversation side for CH and per speaker cluster (obtained automatically) for FBIS. We trained continuous-density, genonic hidden Markov models (HMMs) [6], with 128 gaussians per genone (only non-crossword models). For the CH-only system we used 250 genones, while for the CH+FBIS trained system we used 300, in order to take advantage of the extra training data. To obtain the CH+FBIS models we first trained using the whole data with weight 2 for CH and 1 for FBIS. Then we adapted the final models on the CH only data using MAP adaptation [7]. Recognition was done by SRI's DECIPHER^{TM} engine in a multipass approach: in the first pass, phone-loop adaptation with two MLLR transforms was applied. A recognition lexicon with 18K words and a bigram language model were used to generate the first pass hypothesis. In the second pass the acoustic models were adapted using constrained MLLR (with 6 transformations) based on the previous hypotheses. Bigram lattices were generated which were then expanded using a 3gram language model. Finally Nbest lists were generated using the adapted models and the 3gram lattices. The final best hypothesis was found using N-best ROVER. The CH+FBIS trained system before MAP adaptation was actually about 1% worse than the CH-only system, which suggests that we may have used too high a weight for the FBIS data in the combination. After MAP the two systems become very close in performance. Nevertheless we will demonstrate in the next paragraph that the two systems were actually different enough to help in combination. It should also be noted here that, in contrast to the evaluation system mentioned in Section 3 the present system was only a single front-

System	dev96	eval03
CH-only	57.3	42.7
CH+FBIS	57.2	43.0

Table 1. Baseline word error rates (%) for first-pass systems trained on CallHome only and pooled CH+FBIS data.

System	dev96	eval03
CH-only	56.1	42.7
CH+FBIS	56.3	42.6
combined	55.3	41.7

Table 2. Word error rates (%) obtained after the final recognition pass and with ROVER combination.

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 that we used for the evaluation. The lack of these features resulted in a higher error rate but our goal here was to explore exclusively the effect of the extra training data.

5.2. Combination Experiments

We wanted to explore the assumption that even though the FBIS data did not help improving the models, they actually resulted in a sufficiently different system to be able to help in combination with the CH-only baseline. Both models (MLLR adapted) were used as described above in the third pass of our system, to obtain N-best hypothesis from the trigram lattices. We combined these results using a 2-way N-best ROVER, whose parameters were optimized on the dev96 set. The results are shown in Table 2. We obtained an 0.8% absolute improvement on the development set and a 1.0% improvement on the evaluation set; the latter is significant at the 0.1 level using a difference of proportions significance test. Although the improvement seems modest, it is higher than that obtained by simply combining systems based on different front-ends (0.5% absolute on the eval03 set), and it amounts to one third of the combined improvement gained by different front ends, a more complex language model, MMIE training and HLDA.

6. DISCUSSION

This study represents - to our knowledge - the first attempt at cross-dialectal data sharing for Arabic speech recognition. We have shown how the problem of incomplete phonetic information in the MSA data can be circumvented by automatic diacritization using a combination of syntactic, morphological and acoustic information sources. One likely reason why our experiments did not show more gain may be that the FBIS corpus is acoustically very different from the CallHome corpus, essentially adding noise to the training data. Some acoustic normalization procedure in addition to mean and variance normalization might need to be performed. Furthermore, different weighting schemes for the initial training of the CH+FBIS system might prove beneficial. Furthermore, a much larger data set of MSA might have to be added before stronger effects can be observed: for instance, a related study on using multilingual acoustic data for cross-language adaptation [2] showed that 72hrs of English data added to 1h of Czech training data only resulted in a 1% absolute reduction in error rate (for a baseline for 30% WER) of the Czech ASR system trained on the native data only. Future work will include different adaptation techniques as well as different weighting schemes for different dialect regions represented in the FBIS corpus.

Acknowledgements
This word was funded by DARPA under contract No. MDA-972-02-C-0038. We are grateful to Kathleen Egan for making the FBIS corpus available to us, and to Andreas Stolcke for valuable advice on several aspects of this work.

7. REFERENCES

[1] K. Kirchhoff et al., "Novel approaches to Arabic speech recognition - final report from the JHU summer workshop 2002," Tech. Rep., John-Hopkins University, 2002.

[2] B. Byrne et al., "Towards language-independent acoustic modeling," in *Proceedings of ICASSP*, 2000.

[3] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[4] J. Billa et al., "Audio indexing of Broadcast News," in *Proceedings of ICASSP*, 2002.

[5] J. Bilmes & G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series proessing," in *Proceedings of ICASSP*, 2002.

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

[7] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Transactions SAP*, vol. 3, pp. 357–366, 1995.