DEVELOPMENT OF A CONVERSATIONAL TELEPHONE SPEECH RECOGNIZER FOR LEVANTINE ARABIC

D. Vergyri[1] *, K. Kirchhoff*[2] *,* *R. Gadde*[1] *, A. Stolcke*[1] *, J. Zheng*[1]

[1]Speech Technology and Research Laboratory
SRI International
Menlo Park, CA

[2]Department of Electrical Engineering
University of Washington
Seattle, WA

ABSTRACT

Many languages, including Arabic, are characterized by a wide variety of different dialects that often differ strongly from each other. When developing speech technology for dialect-rich languages, the portability and reusability of data, algorithms, and system components becomes extremely important. In this paper, we describe the development of a large-vocabulary speech recognition system for Levantine Arabic, which was a new dialectal recognition task for our existing system. We discuss the dialect-specific modeling choices (grapheme vs. phoneme based acoustic models, automatic vowelization techniques, and morphological language models) and investigate to what extent techniques previously tested on other languages are portable to the present task. We present state-of-the-art recognition results on the 2004 Levantine Arabic Rich Transcription evaluation.

1. INTRODUCTION

As speech technology is being applied to a increasingly wider range of languages and dialects, the portability of system components and models becomes very important. Limited resources prohibit lengthy data collection and system development efforts; as a consequence, more attention is being focused on algorithms and techniques that can easily be re-used for novel languages or dialects [1, 2]. In this paper our goal is to assess to what extent acoustic and language modeling techniques adopted for English, Mandarin, or Egyptian Arabic generalize to the recognition of Levantine Arabic, and how much task-specific modeling is required.

Automatic Speech Recognition (ASR) of dialectal Arabic is challenging not only because of the sparseness of the available training data but also because of the rich morphology of the language and the writing system: standard Arabic script lacks short vowels and other diacritics indicating pronunciation differences, which leads to high lexical ambiguity and introduces noise into acoustic model training. Our previous work on dialectal Arabic ASR focused on Egyptian Colloquial Arabic (ECA), whereas the task described here (the NIST 2004 Rich Transcription Evaluation Task) requires the recognition of Levantine Colloquial Arabic (LCA). There are significant grammatical, lexical and pronunciation differences between the two dialects. In addition, the transcription standards used for the two corpora were very different: for ECA, a 'romanized' form was used that was fully vowelized. For LCA, the standard Arabic orthography was used, which is phonetically deficient in that it omits short vowels and other pronunciation information. Due to these differences we treated the new domain as an entirely new language and a separate system was trained for this task.

We used the SRI DECIPHER^{TM} recognition infrastructure for building our system. The setup closely follows that for previous recognition tasks but was modified by modeling choices specific to the task at hand. At the acoustic modeling level we use grapheme-based rather than phoneme-based acoustic models, due to the phonetically deficient training transcriptions. We also experimented with techniques to automatically insert the missing vowels in the transcription and train vowelized acoustic models, that include either a generic vowel or all the short vowels. We found that the ambiguity in the pronunciations affects the efficiency of certain training procedures such as cross-word triphones and discriminative training. At the language modeling level, we use factored language models, which are capable of utilizing morphological and other word class information for more robust probability estimation. In contrast to our previous systems, reliable morphological information was not readily available for this task and had to be inferred automatically.

The following sections describe the data (Section 2), our approaches to acoustic modeling (Section 3), and language modeling (Section 4). Section 5 describes our evaluation system, and how the different acoustic and language model components affected the final performance. Section 6 summarizes our conclusions from this work.

2. TRAINING AND TEST DATA

We used a corpus of LCA data provided by the Linguistic Data Consortium (LDC), consisting of 440 conversations (70 hours of speech with about 500K words). The training corpus vocabulary consists of 37.5K words including 2.5K word fragments and 8 non-speech tokens. The data was transcribed in Arabic script without diacritics. The development (dev04) set consists of 24 conversations (3 hours of speech, about 16K words). The out-of-vocabulary (OOV) token rate for this set based on the training set vocabulary was 5.6%. The test set used for the RT-04 evaluations (eval04) consists of 12 conversations (1.5 hours of speech, 8K words).

3. ACOUSTIC MODELING

Since no phonetic lexicon was provided for the LCA corpus, we used grapheme-based rather than phoneme-based acoustic models. Due to the lack of short vowels in the grapheme-based representation, each acoustic model implicitly models either a long vowel or a consonant with optional adjacent short vowels. We used both PLP and MFCC front ends, each with 13 coefficients with 1st, 2nd, and 3rd derivatives. HLDA was used to reduce the feature vector to 39 dimensions, and mean, variance, and vocal tract length

	MFCC		PLP	
	non-cw	cw	non-cw	cw
MLE	54.3	53.6	52.1	53.0
MPFE-iter1	53.4	53.1	51.0	52.0
MPFE-iter2	53.8	53.4	51.0	52.3

Table 1. WER results (%) on the dev04 test set, comparing MLE and discriminative MPFE training for unadapted within-word (non-cw) models and SAT cross-word (cw) models

	unadapted		MLLR adapted	
	MFCC	PLP	MFCC	PLP
cw	53.1	52.0	48.0	47.9
cw+wdbd	49.7	48.3	47.2	47.1

Table 2. WER results (%) on the dev04 test set, comparing plain cross-word (cw) and cw with word-boundary attributes (cw+wdbd) models for MFCC and PLP MPFE trained models, before and after MLLR adaptation. All models also include SAT.

(VTL) normalization was performed per conversation side. We trained gender-independent "tri-grapheme" models in combination with decision-tree based (top-down) state clustering [3], resulting in 650 state-clusters with 128 gaussians for each cluster. In the absence of phone-level transcriptions we used the acoustic models from our 2003 ECA system to initialize the LCA models and re-estimated the model parameters by iterative embedded maximum-likelihood training.

The following sections address the problems associated with the grapheme-based modeling approach in more detail.

3.1. Grapheme-based Pronunciation Lexicon

The pronunciation lexicon was obtained by directly mapping the graphemes to phones and applying the following pronunciation rules, some of which added certain short vowels at specific word locations.
- taa marbuta was converted to /i/ plus an optional /t/.
- hamza was pronounced as glottal stop except after "Al".
- hamza over alif inserted short /a/ before the glottal stop.
- hamza over waaw inserted the short /u/ before the glottal stop.
- hamza under alif and hamza over yaa inserted the short /i/ before the glottal stop.
- tanween was converted to /an/ at the end of the word.
- assimilation of the "sun" letters was incorporated.

All the rest of the short vowels were missing from the resulting pronunciations.

3.2. Discriminative Training on Grapheme LCA Models

Discriminative training using the minimum phone frame error (MPFE) approach ([4]) was applied in addition to maximum likelihood estimation (MLE). However, we found that the effect of the discriminative training procedure was not as significant as in other languages. In our comparable English and Mandarin ASR system, MPFE has yielded relative improvements over MLE of 10% and 6-9%, respectively, using multiple MPFE iterations. In Table 1 we show that for the grapheme based models in this task MPFE training only produced a 2% relative improvement in the first iteration, while subsequent iterations increased WER. It is likely that grapheme models cannot substantially benefit from the discriminative training procedure since each grapheme represents a class of heterogeneous acoustic models rather than one single model. Also the high WER and the numerous inconsistencies in the transcriptions can limit the effect of the MPFE procedure, especially since it relies on accurate phone alignments for discrimination.

3.3. Cross-Word Grapheme-Based Acoustic Models

As shown in Table 1, the performance of cross-word models is either worse or only slightly better than that of within-word models, even though speaker adaptive transforms (SAT) were applied only to cross-word models. Without SAT, cross-word models actually performed worse than within-word models. This was again contrary to the behavior of cross-word models in our in-house ASR systems for other languages (English, Mandarin, ECA). The fact that short vowels are not explicitly included in the acoustic model context can cause problems in cross-word models if the nature of the hidden short vowels is different at word boundaries compared to the within-word location. To test this assumption we built models by letting the decision trees use word boundary markers as explicit attributes. This technique has been used before [5], but our experience in English ASR systems was that it yielded only very small improvement over standard cross-word models. The comparison in WER between the plain and the word boundary cross-word models is shown in Table 2. We see that before adapting with maximum likelihood linear regression (MLLR) the models with the boundary information gain about 3.5% absolute. After adaptation the improvement is reduced to 0.8%, which, however, is still significant.

3.4. Modeling of Short Vowels

Work on the ECA corpus [6] has compared the WER in recognizers with phone-based vs. grapheme-based acoustic models. It was shown that the relative loss in performance due to grapheme-based models was close to 10%. We therefore explored techniques for automatic vowelization of the training transcripts.

In our first effort to use vowels in the LCA system we generated word pronunciation networks that included one optional generic vowel phone in all possible positions in the pronunciation. The possible positions were determined by applying a morphological analysis tool for Modern Standard Arabic (MSA), the LDC Buckwalter stemmer. The stemmer produced diacritized variants for a 13K subset of the 37K vocabulary. For about 150 words, the possible vowel positions was annotated manually. For all other words, we added an optional vowel between every consonant pair in the written form. This system was still using a non-vowelized orthography for LM purposes.

In our second approach we manually added the vowels on a small subset of the training data (about 40K words), which was selected to have a high vocabulary coverage (covering 43% of the vocabulary and 80% of the word tokens in the training data). We trained a 4-gram character-based language model on this data, which was used as a hidden tag model to predict the missing vowels on the whole training data transcriptions. On a held-out subset of about 6K words, we estimated this procedure to have a character error rate of about 7%, with about 30% of the words having at least one wrong character. From the automatically vowelized data we obtained pronunciations that included short vowels (/a/, /u/, /i/), and applied the rules described in Section 3.1. We also had the option to use a grapheme-based language model or train a new model based on the automatically vowelized transcripts. For

	WER (%)
grapheme AM + grapheme LM	54.3
generic vowel AM + grapheme LM	54.9
auto-vowelized AM + grapheme LM	54.5
auto-vowelized AM + auto-vowelized LM	54.0

Table 3. WER comparison of the grapheme based and vowelized models on the dev04 testset. In all experiments the acoustic model (AM) is using within-word MLE trained MFCC models.

computing the WER for that system we stripped the vowels from the output words and compared it against the script references.

In Table 3 we show the effect of the different vowel modeling approaches after MLE training. The generic vowel system performs worse than the grapheme based one, but in Section 5 we show that it improves the final result when combined with other system components since it is sufficiently different from the grapheme models. The auto-vowelized system with all short vowels performs better than the generic vowel system. The use of a vowelized LM yields an additional improvement even though the automatic vowelization procedure increased the LM vocabulary and added diacritization errors to the transcriptions. We found that MPFE training on the vowelized models had the same effect as on the grapheme models (about 1% absolute WER reduction). The effect is still lower than in other ASR tasks, probably due to the noisy transcriptions and the low accuracy of the MLE model.

4. LANGUAGE MODELING

For language model training we used the LCA training transcriptions provided by the LDC. The vocabulary was reduced to 17,638 words by removing singletons from the training data; these were mapped to a generic reject model. In our development experiments, this resulted in a slight (0.5% absolute) improvement in WER, possibly because of the elimination of spelling errors and word fragments. All noise events except laughter were removed as well. We used two different types of language models in our system: standard word-based language models and factored language models (FLMs). The baseline word-based bigram and trigram models were trained using modified Kneser-Ney smoothing and interpolation of higher-order with lower-order n-grams. The bigram was used for generating the initial lattices, whereas the trigram was used for lattice expansion and N-best list rescoring.

As part of our previous work on Arabic ASR we have developed a so-called factored language model approach [7], which is based on a representation of words as feature vectors and a generalized parallel backoff scheme which utilizes the word features for more robust probability estimation. The word features represent the morphological properties of the word. Complex Arabic words may consist of affixes and a stem, which can be further subdivided into roots and patterns. In previous work [8] we found that the most useful of these for the purpose of language modeling were the stem, the root, and a tag indicating the morphosyntactic properties of the affixes. A factored language model exploiting these features can assign more robust probabilities to unseen combinations of words in the test data when the combination of corresponding morphological features has been observed in the training data. The structure of the model , i.e. the set of features to be used and the combination of partial probabilities estimates from features, is optimized using a genetic algorithm [9]. In our previous ECA system, factored language models yielded an improvement of up to 2% absolute, for a baseline with approximately 40% word error rate [8]. For this system, however, information about the morphological features of each word was available in the form of a lexicon distributed with the ECA corpus. For our present LCA system, this information was not available and had to be inferred by other means. Since automatic morphological analyzers do not currently exist for dialectal Arabic, we used a simple script and knowledge of Levantine Arabic morphology to identify affixes and a subsets of the parts-of-speech from the surface script forms. We also applied a morphological analyzer developed for MSA [10] to obtain the roots of the script forms. Those forms that could not be analyzed retained the original script form as factors. It was found that this type of decomposition, although error-prone, yielded better results than using data-driven word classes. On the development set the perplexity was reduced from 222.7 to 211.8.

	dev04	eval04
RT-04	43.0	47.4
post-eval-04	42.5 (-0.5)	46.9 (-0.5)

Table 4. WER (%) of the RT-04 evaluation system. The submitted system is compared to a post-eval system where the original cross-word models were replaced with those that used word-boundary information.

5. EVALUATION SYSTEM

The processing stages of the full system submitted for the RT04 evaluation follow the setup of the SRI RT04 20xRT English CTS system. The system consists of two stages, both of which include lattice generation with within-word models, lattice acoustic rescoring with multiple models to obtain different sets of N-best lists, and final N-best combination (implemented as N-best-Rover) to obtain the final consensus hypotheses [11].

At the first stage, phone-loop adapted PLP acoustic models (AMs) and a bigram language model (LM) are used to generate the lattices. After trigram LM rescoring on the lattices, word confusion networks are constructed in order to obtain the best posterior word hypotheses. These hypotheses are used as reference to estimate SAT transforms for each of the models used in the following passes. The following models are then used to perform acoustic rescoring of the lattice and generate N-best lists:
(a) PLP cross-word models adapted on the best word posterior hypotheses from the lattices
(b) Phone-loop adapted MFCC within-word models
(c) MFCC cross-word models adapted on the hypothesis from (b).

The final hypotheses from the first stage (after N-best-Rover of (a)-(c)) are used as adaptation references to compute new MLLR transforms for the PLP within-word model. The new model is used to generate the lattices for the second stage of the system. Then new N-best lists are generated using
(d) the newly adapted PLP within-word models
(e) PLP cross-word models adapted on the hypotheses from (c)
(f) MFCC cross-word models adapted on the hypotheses from (a)
(g) MFCC within-word model that include short-vowels, adapted on the hypotheses from (d).
All the N-best lists at this stage are rescored with FLMs. The best posterior hypotheses are obtained after optimized 4-way N-best-Rover combination.

	dev04	eval04
grapheme	43.1	47.3
+ generic-vowel non-cw MFCC	42.5 (-0.6)	46.9 (-0.4)
+ auto-vowel MFCC models	42.1 (-1.0)	46.5 (-0.8)

Table 5. Effect of the vowelized models on the final system WER. The post-eval improved cross-word acoustic models are used in both cases and a bigram FLM is used as described in Section 5.

For the system submitted for the evaluations (RT-04 system), the cross-word models did not include the word boundary information; this was added in the post-eval-04 system shown in Table 4 resulting in 0.5% absolute improvement of the final system result.

In Table 5 we demonstrate the contribution of the vowelized models in the post-eval system. In the system without the vowelized model, we replaced the model in (d) with a grapheme-based within-word MFCC model. We see that using the generic-vowel model in step (d) reduces the error rate by 0.6%-0.4% absolute, even though this system was worse in isolation (see Table 3). For the auto-vowel model we modified the structure of the system by replacing the MFCC models at the second stage of the system, with vowelized models. We generated a third set of lattices using the vowelized LM, which were used to obtain the vowelized MFCC N-bests with within-word and cross-word models. These models improve the final performance by 0.8-1.0% absolute over the grapheme based system.

Table 6 shows the contribution of the FLM. The system without FLM uses only the standard word 3-gram for all rescoring steps. We see that including the eval04 FLM only for final N-best ((d)-(g)) rescoring improves the result by 0.2% and 0.1% on each test set. This is a much smaller contribution than that observed previously in our ECA system [8]. In a newer post-eval system, we replaced the bigram FLM with a more optimized trigram FLM and used it at every step in the system, for rescoring the lattices and N-best lists, replacing the standard word 3-gram in steps (4), (6) and (9). This resulted in improvements of 0.6% and 0.3% absolute.

6. CONCLUSIONS

We found that most of the techniques developed for English or ECA ASR could be ported to the development of a LCA system. However, due to use of script-based training transcriptions, some techniques did not have the expected effect.

Discriminative training (MPFE) yielded a much smaller win compared to previous tasks, probably because of the ambiguity in the pronunciations, the noisy training transcriptions and the low accuracy of the MLE models.

Adding extra information improves the grapheme-based models. Word-boundary information proved more valuable for cross-word models in this task than in e.g. our phone-based system for English. The vowelized models were beneficial for system combination, even when they did not improve over a purely grapheme-based system in isolation. The system that uses vowelization for both acoustic and language modeling performs the best, even though the automatic vowelization approach induced errors in the training transcriptions Techniques that would improve the accuracy of the automatic vowelization deserve further investigation.

Compared to our 2003 evaluation results on ECA, the FLM was not as helpful as before. The larger size of the 2004 training set was ruled out as a possible cause; FLMs were observed to yield consistent improvement over the word-based language model regardless of the training data size. The most likely causes are the lack of accurate morphological information and the increased lexical ambiguity due to the use of non-vowelized script forms for language modeling.

	dev04	eval04
No-FLM	42.7	47.0
eval04 FLM	42.5	46.9
Post-eval04 FLM	42.1 (-0.6)	46.7 (-0.3)

Table 6. Effect of the FLM on the WER in the post-eval-04 system (cross word models with word-boundary attributes, and vowelized system included). The eval04 FLM is the bigram FLM used in the submitted eval04 system, applied only to the rescoring of the final N-best hypotheses. The Post-eval FLM is an improved trigram model which is applied at all steps of the system, except for lattice generation (the word-bigram is still used for this step).

Acknowledgments

This work was funded by DARPA under the EARS Project, contract No. MDA972-02-C-0038. Distribution is unlimited.

7. REFERENCES

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

[2] Lori Lamel, Fabrice Lefevre, Jean-Luc Gauvain, and Gilles Adda, "Portability issues for speech recognition technologies," in *Proc. of HLT*, 2001.

[3] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings ARPA Workshop on Human Language*, 1994, pp. 307–312.

[4] J. Zheng and A. Stolcke, "Improved discriminative training using phone lattices," in *submitted to EUROSPEECH*, 2005.

[5] P. Woodland, C. Leggetter, J. Odell, V. Valtchev, and S. Young, "The 1994 HTK large vocabulary speech recognition system," in *Proc. of ICASSP*, Detroit, 1995.

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

[7] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proceedings of HLT/NACCL*, 2003, pp. 4–6.

[8] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," in *Proceedings of ICSLP*, 2004.

[9] K. Duh and K. Kirchhoff, "Automatic learning of language model structure," in *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 2004.

[10] K. Darwish, "Building a shallow Arabic morphological analyser in one day," in *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA, 2002.

[11] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. of Eurospeech*, 1999.