An Investigation in Speech Recognition for Colloquial Arabic

Sarah Al-Shareef and Thomas Hain

Department of Computer Science, University of Sheffield, Sheffield, UK

s.al-shareef@dcs.shef.ac.uk , t.hain@dcs.shef.ac.uk

Abstract

This paper describes a study of grapheme-based speech recognition for colloquial Arabic. An investigation of language and acoustic model configurations is carried out to illustrate the differences between colloquial and modern standard Arabic (MSA) on the example of Levantine telephone conversations. The study defines extensive and carefully crafted data sets for different dialects and studies their overlap with MSA sources. The use of grapheme models is re-investigated, and alternative configuration for acoustic models to correct obvious shortcomings are tested. The recognition performance was analyzed on two levels: corpus-level and dialect-level. In addition modifications of dictionaries to allow better specification of sound patterns is explored. Overall the experiments highlight the need for higher level information on acoustic model selection.

Index Terms: Automatic speech recognition, conversational telephone speech recognition, colloquial Arabic

1. Introduction

Work on colloquial Arabic automatic speech recognition is difficult for a variety of reasons. First, colloquial Arabic (CA) normally exists in spoken form only; it is not considered a written language. Therefore, in comparison with the volume of Modern Standard Arabic (MSA) text available, only limited textual data exists. Secondly, no standard convention is agreed on how the various dialects should be transcribed [1]. CA transcribers typically use the MSA undiacritised orthographic system, which lacks some important phonemic information such as short vowels and gemination [2]. Also, transcribers are usually influenced by their MSA knowledge and sometimes substitute a spoken word with its original MSA form, even if the latter does not describe the spoken word phonetically [1]. Finally, CA still inherits the complex morphological form of MSA. Moreover, additional affixes are introduced informally for each dialect, thereby increasing cross-dialectical differences.

In this work a comparative study of grapheme-based speech recognition for CA conversational telephone speech is presented. Several aspects of the representation of CA in acoustic and language modelling are described and investigated. As outlined above, the pronunciation of an undiacritised word can only be clarified by the context. Experiments were conducted to investigate the impact of the phonotactic rules.

The organisation of this paper is as follows. Section 2 describes the data selected for study in this paper. Sections 3 and 4 describe the experiments in language and acoustic modelling. In section 5 the interaction of these components is illustrated, together with final performance figures. An exploration of phonotactic rules in dictionary creation for decoding is detailed in section 6. Finally, section 7 summarises the work and concludes the paper with a discussion of possible future directions.

2. Training and Testing Data

The data used in this study was drawn from the Fisher Levantine colloquial Arabic (LCA) corpus. This corpus has a regional constraint, containing a dialect spoken mainly in Jordan, Lebanon, Palestine and Syria. The corpus is distributed by the Linguistic Data Consortium (LDC) and consists in total of more than 175 hours of conversational telephone speech (CTS) recordings. The data represents conversations by more than 2000 native Levantine Arabic speakers talking to their friends and families, as well as unrelated individuals, about topics suggested by the corpus collectors [3].

The orthographic transcripts were generated by LDC in a semi-diaccriticised form using standard MSA scripts. Each side in the conversation is labelled with the speaker’s identity, gender and sub-dialect. To maintain homogeneous and balanced recording conditions, a test set was constructed by the random selection of conversation sides from the corpus with the objective to have good coverage of all sub-dialects and speakers. A selection on a side-level instead of conversation-level was preferred for better speaker coverage, while allowing speaker set separation between training and test sets. The test set, l06, consists of 5.10 hours of speech and specifically excludes sides with speaker changes, foreign speakers or overlapped speech. The training set, referred to as LCA, consists of 170.96 hours of speech covering 2355 speakers. Table 1 illustrates the distribution of gender and major dialects: Lebanese (LEB), Jordanian (JOR), Palestinian (PAL) and Syrian (SYR).

The sets were also used for language modelling purposes. The l06.train set contains 1.54M tokens with 69.4K unique words. The test set consists of 44.2K tokens with 7.6K unique words. In addition, MSA data was included in the language model development. A part of the Arabic Gigaword [4] newswire resources, known as An Nahar, was chosen. It consists of 232.4M words and 1.9M unique words. Table 2 shows the coverage of vocabulary across the four major dialects within the corpus and the MSA vocabulary derived from Arabic Gigaword [4]. For instance, 29.39% of words from LEB conversa-
The complex morphology of Arabic results in multiple forms for a single word. Indeed, a relatively large number of distinct words can be found in a fixed amount of text. Morphological decomposition has been found to be effective in reducing the word-list size, such as in [5, 2, 6]. However, this work emphasises the impact of conversational cues and disfluencies on LM performance. In addition, it investigates the within and cross-dialect coverage. Thus, no morphological decomposition was applied.

### 3.1. Handling backchannels and hesitations

Backchannels indicate acknowledgment and the speaker’s engagement in the conversation discourse and thus are important for the recognition of conversational speech. Hesitations do not bear syntactical or discourse information and hence are typically excluded from the recognition output. However, the ability to recognise hesitations adds robustness to the ASR system. The sound patterns of the backchannels in the targeted corpora are similar to each other; thus, using one backchannel word with different pronunciation may improve the LM predictions and overall recognition performance. When using standard NIST output scoring, Arabic hesitations are mapped into a single word. The same strategy was applied to both backchannels and hesitations.

In order to investigate the impact of mapping backchannels and hesitations, different trigram, bigram and unigram LMs were estimated using the SRILM toolkit [7]. The standard n-gram models were trained, using modified Kneser-Ney discounting and backoff. All LMs were tested on a modified of the training and testing sets.

### 3.2. Using MSA training data

As indicated in Table 2, the MSA vocabulary derived from the Arabic Gigaword corpus covers between 62-77% of the words in the LCA conversations. Thus, one would expect limited success for the interpolation of MSA language models with conversational ones. Three different vocabulary lists consisting of 100k words each were chosen. The first list (S\textsubscript{top}) included the most frequent 100K words from Gigaword set of words. These cover only 33.4% of the LCA vocabulary. The second list (S\textsubscript{c}) includes all words common between LCA and the Gigaword corpus (a total of 43.9k words) and is padded with the most frequent Gigaword words. The third list (S\textsubscript{p}) includes the complete set of 69k LCA words and is again padded with the most frequent MSA words.

Table 5 shows the perplexity results of three language models for each of the vocabulary lists: trained on MSA text only, trained on LCA text, and interpolation of the two component LMs. The very high perplexity result with the LM trained on the MSA text clearly shows that MSA data alone does not reflect CA text very well. As expected, interpolation (using optimised weights) gives hardly an improvement over the LCA component LM. Table 6 presents the perplexity in the dialect-level for LMs built using the three vocabulary lists: S\textsubscript{top}, S\textsubscript{c} and L and trained on LCA only. Despite the very high OOV rate (30-35%), S\textsubscript{top} computes lower perplexity for each dialect. This is an indication that the 33.4% of MSA found in the LCA top 100k word-list is frequent in LCA as well, which translated into better perplexity. Introducing more of the LCA words (up to 63.1%) in S\textsubscript{c} negatively affects the perplexity. Indeed, the perplexity reaches its worst level when including all LCA word-list in L, where more uncertainty is added to the LM.

### 4. Acoustic Modelling

Most of the available Arabic acoustic transcriptions lack short vowels and gemination information. Therefore, several studies of Arabic speech recognition have used grapheme-based modelling [2, 8, 9]. Here, each grapheme will represent up to eight phonetic classes: the phoneme represented by that let-
4.1. Long-span grapheme model
In all known previous studies, e.g. [2, 8], a standard 3-state HMM was used. Using a 3-state HMM suggests that most of the grapheme models will have the same minimum duration as a single phoneme. Apart from long vowels, the majority of consonants occur as vowelised phonemes, i.e. consonant+short vowel. In addition, a geminated phoneme is considered to be double the length of a normal phoneme [1]. This suggests that using more states should be more suitable. As an alternative, a 5-state HMM (as shown in Figure 1.a) was implemented. Training parameters were modified to keep the same number of allo-grapheme states (G5a), showing a log likelihood improvement of 0.9% relative to G3. However, the minimum number of states occupying a cluster was lowered by 33.3%. If clustering parameters were kept fixed instead (G5b), the training set log likelihood would improve by 0.7% relative to G3, with an increase in the number of states (and thus parameters) of 19%.

4.2. Generic vowel model
The main issue in acoustic modelling is the absence of short vowels. One way to address this is by including all possible vowelised versions of a word; however, this will expand the pronunciation variations exponentially with the number of consonants. For instance, a 3-consonant word could have up to 64 variations. [10] solves this by using one symbol to represent 3 short vowels. Using such a generic vowel model (GV) might capture some of the essence of short vowels in the language. In their work, this GV was initially trained on a small amount of diacritised data then was used in unsupervised training on undiacritised data. In contract to their work, no prior training was performed for the GV in this study; in addition, the GV topology provides an option to skip the short vowels. However, in comparison to G3 and G5, this model captures much less contextual information, which might degrade its performance.

Figure 1.b illustrates the topology of the generic vowel model. The addition of the model appears to add more confusion, since in order to obtain a similar number of states as in G3, the likelihood split threshold had to be lowered by 85%. When leaving the clustering parameters to be the same, the number of clustered states is reduced by 34%.

Table 7 illustrates the log-likelihood differences for all the acoustic model configurations tested. The highest per frame log-likelihood is obtained with the 5-state HMM. All GV model configurations show poorer matches to the training data. This would indicate that using temporally longer acoustic models represents the graphemic model better.

5. Recognition Experiments
Initially, a language model (LL0) similar to the LL2a described in section 3.1 was constructed, but with a smaller vocabulary size of only 46k words, based on the highest frequency in the training set. Table 8 shows the impact on recognition performance of an increase in vocabulary size from 46k to 69K. Introducing more vocabulary to the system improves G3 performance by 8.3% absolute, whereas no improvement was observed for G5 or GV. In addition, using an MSA vocabulary with less LCA data improves G3 and G5 performance by 5.8-7.1% absolute to the system using LL0; this improvement agrees with the interpretation in section 3.2. Again, the GV performs 7% absolute worse when introducing MSA vocabulary with less LCA data. Although G5 models acoustically match the data, based on the likelihood presented previously in Table 7, their performance is significantly worse than G3 by 10-17% absolute over all systems. On the dialect-level, it is noticeable that the WER is correlated with training data size: more training data introduces more confusion. This effect begins to smear with the likelihood of the model getting worse, such as in the case of GV.

6. Phonotactic Rules
In addition to the initial baseform in the pronunciation dictionary described in section 4, multiple pronunciations were generated using linguistic and phoneme-to-sound rules by employ-
Table 8: Recognition performance for different acoustic model designs with different LM. Both overall and sub-dialect level WER are provided when using LLa.

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_L</td>
<td>LL0</td>
</tr>
<tr>
<td>G3</td>
<td>60.5</td>
</tr>
<tr>
<td>G5a</td>
<td>73.1</td>
</tr>
<tr>
<td>G5b</td>
<td>71.7</td>
</tr>
<tr>
<td>GVa</td>
<td>86.3</td>
</tr>
<tr>
<td>Gvb</td>
<td>86.1</td>
</tr>
</tbody>
</table>

Table 9: Pronunciation rule coverage within the vocabulary word list and training data text.

<table>
<thead>
<tr>
<th>rule</th>
<th>%wlist</th>
<th>%text</th>
<th>word:prons</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>18.89</td>
<td>11.73</td>
<td>1.2</td>
</tr>
<tr>
<td>W</td>
<td>4.64</td>
<td>1.73</td>
<td>1.1</td>
</tr>
<tr>
<td>N</td>
<td>0.55</td>
<td>1.11</td>
<td>1.1</td>
</tr>
<tr>
<td>T</td>
<td>12.82</td>
<td>8.13</td>
<td>1.3</td>
</tr>
<tr>
<td>A</td>
<td>11.57</td>
<td>16.52</td>
<td>1.1</td>
</tr>
<tr>
<td>Y</td>
<td>1.47</td>
<td>1.13</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Table 10 shows the effect on training data of such changes for each rule together with the number of pronunciations generated. As it can be seen, W, N and A rules are just mapping rules and do not add pronunciation variants. In order to assess the effect of each rule on recognition performance, multiple recognition dictionaries were generated employing each rule independently. A final dictionary is generated where all rules are applied jointly.

Table 10 shows WER results for the associated dictionaries. None outperform the raw grapheme dictionary. The largest degradation is observed with a mapping rule for Alif, followed by the rule that generated the most alternative pronunciation variants. The phonotactic rules as outlined above are highly context-dependent and simply adding the variants will not help models to automatically select the accurate variant.

7. Conclusion

In this paper, a number of issues and design decisions for developing a grapheme-based ASR for colloquial Arabic CTS were addressed. First, the impact of hesitation and backchannel mapping in a language modelling was shown. In addition, different schemes for training language models were investigated, including the use of MSA background material. Second, different acoustic model configurations were reviewed and examined, in particular the use of longer HMMs and an insertion model for generic vowels. None of the two designs outperformed the standard 3-state HMM in recognition experiments, although better acoustic matches were observed. Finally, the impact of applying phonotactic rules during the decoding stage was reviewed. Most of the experiments have a common theme. Simply adding the variation capability does not help; instead it degrades performance. Future work must address this issue with wider span acoustic model selection.

8. References