PHONETIC PRONUNCIATIONS FOR ARABIC SPEECH-TO-TEXT SYSTEMS

F. Diehl, M.J.F. Gales, M. Tomalin, & P.C. Woodland

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {fd257, mjfg, mt126, pcw}@eng.cam.ac.uk

ABSTRACT

In this paper two aspects of generating and using phonetic Arabic dictionaries are described. First, the use of single pronunciation acoustic models in the context of Arabic large vocabulary Automatic Speech Recognition (ASR) is investigated. These have been found to be useful for English ASR systems, when combined with standard multiple pronunciation systems. The second area examined is automatically deriving phonetic "pronunciations" for words that standard approaches, such as the Buckwalter Morphological Analyzer, cannot handle. Without pronunciations for these words the OOV rates for various Arabic tasks significantly increase. Here, pronunciations are automatically found by first deriving grapheme-to-phone rules, and associated rule probabilities. These are then used to produce the most likely pronunciation, or pronunciations, for any word. These approaches are evaluated on a large vocabulary Arabic Broadcast News and Broadcast Conversation transcription task. Both schemes are found to yield gains with a multi-pass/combination framework.

Index Terms— Speech Recognition, Arabic, Single Pronunciation Modelling

1. INTRODUCTION

Recently there has been much interest in the problems associated with transcribing Arabic audio [1, 2, 3]. There are a number of issues to be addressed for success due to the nature of the Arabic language. In Arabic texts, short vowels are not normally marked, which means that each "word" in the text may have a large number of pronunciations, with the pronunciations being associated with different, but possibly related, meanings. In addition, Arabic is a highly inflected agglutinative language with lexical items being formed by attaching affixes to triconsonantal roots which requires a very large vocabulary for good coverage of general Arabic audio.

The combination of a large vocabulary and the possibility of a large number of pronunciations per word makes the construction of the phonetic models a challenge for Arabic. The usual approach for Arabic pronunciation generation is to rely on an analysis system such as the Buckwalter Morphological Analyzer and possibly also use some other data which has been explicitly marked for short vowels etc. [1]. The problem is that the coverage of such analysers is limited and hence pronunciations will not be available for a significant portion of the vocabulary: for the most common 350k Arabic words, only 260k have pronunciations with this method.

A simple alternative is to use a *graphemic* representation for the basic acoustic units, and let all pronunciation variability be handled implicitly by standard context-dependent Gaussian mixture hidden Markov models (HMMs). In this case, "pronunciation" generation is trivial. Furthermore, graphemic models have the advantage of reducing the complexity of the recognition search. However they suffer from somewhat increased word error rates (WERs), at least for broadcast news (BN) data, relative to systems that use conventional phonetic acoustic units systems [4].

The aims of this paper are two-fold. First, to investigate whether the simplicity of the graphemic system can be retained by using a single pronunciation entry for each vocabulary item, so that, for example the existence of a short vowel will be explicitly modelled but the range of acoustic realisations will be implicitly modelled by the HMMs used. The technique used here for this single pronunciation (SPron) modelling has previously successfully been used for English [5].

The second aim is to investigate a simple method for automatic generation of missing phonetic dictionary entries by using a simple statistical model, so that phonetic pronunciations, either of SPron style or multiple pronunciations (MPron) for an arbitrary vocabulary size can be used.

The organisation of this paper is as follows. First the SPron method is described and then the technique for augmenting an existing pronunciation dictionary with entries for additional words is discussed. The experimental evaluation investigates the use of graphemic, phonetic MPron and phonetic SPron models with both 350k and 260k word vocabularies for recognition of both Arabic BN and broadcast conversation (BC) data, and demonstrates how the techniques discussed are useful in improving overall system performance.

2. SINGLE PRONUNCIATION DICTIONARY

As described above, the basic idea in a single pronunciation dictionary is to rely on implicit modelling of pronunciation variation and choose, from a corresponding MPron dictionary a set of consistent, representative pronunciations. The approach adopted here is taken from earlier work on English [5], which was particularly effective for conversational data with high pronunciation variability.

SPron dictionaries were constructed using the Arabic acoustic training data, and the main steps in the process are the following:

Pronunciation Variant Frequency: obtain the frequency of each pronunciation in the training dictionary by Viterbi alignment of the acoustic training data.

Initial Dictionary: sort the pronunciations for each word in the baseline MPron dictionary according to frequency of occurrence in the training data. If a word is observed in the training data, delete any associated unseen pronunciation variants.

Merging of Phoneme Substitutions: for a given word, align each pair of pronunciation variants. If phonemes are only substituted, the variant with the higher frequency of occurrence is retained and the

This work was in part supported by DARPA under the GALE programme via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

frequency of the second variant is added. If the variants occurred equally often the selection is random.

After these stages, two variant types remain: (i) variants observed in the training data but which cannot be solely described by phoneme substitutions, (ii) variants for words not observed in the training data. To handle both these cases a simple statistical model is developed which, given an alignment between the remaining pronunciations under consideration, determines the most probable variant according to the model. The case of insertions/deletions is simply handled at this stage by the use of extra symbols in the alignment of the dictionary entries [5].

3. PHONETIC PRONUNCIATIONS

A 39-phone set used for all phonetic systems in this paper. This comprises the 36 consonants, in contrast to the system described in [4] alif and ya and wa variants are modelled separately[1]. In addition the three short vowels are modelled. In Arabic these short vowels (*fatha* `/a/`, *kasra* `/i/` and *damma* `/u/`) are commonly not marked in texts. Additionally, *nunation* can result in a word-final *nun* (`/n/`) being added to nouns and adjectives in order to indicate that they are unmarked for definiteness. Thus for any word in Arabic there may be multiple valid phonetic pronunciations, compared to the graphemic system where only the consonants are modelled. Note in the phonetic systems used here, the diacritics *shadda* and *sukun* are not considered and are required to be implicitly modelled.

This section describes the two approaches to obtaining the phonetic pronunciations, which form the basis of the SPron and MPron systems, used in this paper. The first is based on the Buckwalter Morphological Analyzer (version 2.0)[2], referred to as Buckwalter in this paper. The second is the procedure adopted to automatically derive phonetic pronunciations for any Arabic word.

3.1. Baseline Pronunciations

The baseline dictionary used in this work is generated by Buckwalter. All initial recognition dictionaries were based on this analysis. However for training data Buckwalter was used in combination with the Treebank and the FBIS pronunciations (similar to the procedure described in [1]). Here the following strategy is used:

$$\text{Buckwalter} \rightarrow \text{Treebank} \rightarrow \text{FBIS pron.}$$

where $\rightarrow$ means if the word is not found in the left dictionary search in the the right dictionary. This expands the coverage for the training data and is not felt to be a major issue as inconsistencies in the dictionaries will minimally impact other words as training is an alignment process. In contrast for decoding, an inconsistent dictionary may affect both the word in question and the surrounding words. Thus only the Buckwalter pronunciations were used during decoding. This procedure yields an average of about 4.3 pronunciations per word.

Word-list	bcad06	bnad06	dev07
350K	1.8	1.0	1.6
260K	4.1	2.2	5.0

Table 1. OOV rates for the bcad06, bnad06, dev07 using the 350k and 260k wordlists.

[1]This was found to give small, but consistent gains, in the final phonetic graphemic system combination.

[2]Available at http://www.qamus.org/index.html.

The starting decoding word-list used here was the 350K word-list derived using frequency counts derived from the language model sources, see section 4. Table 1 shows the OOV rates for this word-list on the three test sets, bcad06, bnad06 and dev07 evaluated in section 4. As expected the OOV rates on the BC data is larger than on the BN-style data. Buckwalter was able to obtain pronunciations for about 260K word in this word-list. The OOV rates for this 260K word-list is also shown in table 1. As expected the OOV rates increase on all test sets. However the increases on the BC-style bcad06 test set and the dev07 data which contains some BC-style data is larger than on the BN-style bnad06 data.

3.2. Automatic Pronunciation Generation

As part of the training process it is necessary to obtain pronunciations for words that can not be handled by Buckwalter [4, 3]. A series of rules were automatically generated from a 250K Buckwalter derived phonetic dictionary. Though this derives many of the standard expert rules, it ensures that they were consistent with pronunciations from Buckwalter. The pronunciations were derived in a "right-associative" fashion and the start (`_S`) and end (`_E`) of word pronunciations were kept distinct from standard variations (`_V`) (this also allows inter-word silence to be correctly added to the pronunciations). For example, the pronunciation and derived rules for the Arabic word *ktAb* are

```
ktAb    /k/ /i/ /t/ /A/ /b/
k_S     /k/
t_V     /i/ /t/
A_V     /A/
b_E     /b/
```

This procedure yielded 1215 derived pronunciations and was guaranteed to yield a pronunciation for each word. The vast majority resulted from *nunation* at the end of words. Segments containing one or more words with no Buckwalter pronunciations were then forced aligned using the rules to give allowable pronunciations. This resulting phone sequences can be used for acoustic model training. In addition to using these rules for the training data, pronunciations for 734 words that were felt to be reliable (occurred greater than 5 times) were added to the test vocabulary in [4].

In this paper, these automatically derived rules are used to find pronunciations for any word, not just those seen in the training data. A simple procedure was used.

1. Using all derived rules, force align the training data to obtain the 1-best phone sequence.
2. Derive "pronunciation" probabilities for all pronunciation rules using the statistics obtained from the forced alignment.
3. For all words requiring pronunciations generate the top N most likely pronunciations using the rules and "pronunciation" probabilities.

This process can be used to derive pronunciations for both single and multiple-pronunciation systems. For the MPron system, $N = 5$ was used as this approximately matches the average number of pronunciations per word. For the single-pronunciation system, $N = 1$ was used.

It is worth noting that there is a natural bias in the above procedure to producing shorter pronunciations. Despite this bias, short-vowels occurred frequently in the most likely derived pronunciation, which were thus different from the graphemic "pronunciation".

4. EXPERIMENTS

4.1. Acoustic and language models

The performance of the single pronunciation system and automatic pronunciation generation was evaluated using the HTK Arabic broadcast news transcription system [4]. All acoustic models were trained on about 1000 hours of acoustic training data. Three sources of acoustic data were used: FBIS data, TDT4 Arabic data, and data released by LDC under the DARPA GALE programme. The majority of this data was used with supervised transcriptions, but lightly supervised and unsupervised approaches were also used. For more details of the acoustic training sources see [4]. Four forms of acoustic model were built. A graphemic system (G0) was built. This allows reliable pronunciations to be obtained for all possible words. Three phonetic systems were built. These require pronunciations to be found. As previously discussed the Buckwalter Morphological Analyzer (version 2.0) was used to find the baseline pronunciations. All phonetic systems used 39 phones (compared to the 36 phones used for the graphemic system). The three additional phones, the short vowels, were modelled using either three emitting state HMMs, as used for all other phones, (V1 system), or using 2 states for short vowels (*nun* was also modelled using 2 states) in the V3 system, as suggested by [6]. Both the V1 and V3 systems used the standard multiple pronunciations obtained from Buckwalter. An SPron system was then built using the same configuration as the V3 system. This is denoted the V5 system. Each of these acoustic models used state-clustered triphone HMMs with approximately 9000 distinct states and an average of 36 components per state. Minimum Phone Error (MPE) discriminative training was used to train all the acoustic models. Pronunciation probabilities were used for the phonetic systems[3].

The N-gram language models were trained on data from 22 Arabic sources, including the acoustic transcriptions. As discussed previously two word list sizes were used, the first consisting of 350K words, derived by extracting the most frequently occurring words from the language model training data. The 260K word-list was the subset of the 350K word-list for which phonetic pronunciations could be automatically derived.

4.2. Multi-pass combination framework

Figure 1 shows the multi-pass combination framework used in this paper. P1 is a fast decoding pass used to generate hypotheses for least-squares linear regression (LSLR) and variance adaptation for the second, P2, stage. The P2 stage generates lattices using a trigram word-based language model. This is then expanded using a 4-gram language model to yield the lattices that are passed to the P3 stage. Here 1-best CMLLR and lattice-based MLLR adaptation is applied. The lattices are then rescored and confusion networks generated. In contrast to the systems described in [4] only a single acoustic segmentation was used, the CU1 segmentation. It is expected that cross-segmentation gains could also be obtained. Two separate branches were run using this segmentation, allowing different acoustic models, or language models, to be used in each of the branches. The outputs from the individual branches were combined using confusion network combination (CNC).

The performance of the systems was evaluated on three test sets. The first two `bnad06` and `bcad06` were defined by BBN technology and consist of about 3 hours of BN and BC data respectively (collected during Dec05-Jan06). These test sets comprise complete shows collected between December 2005 and January 2006. The `dev07` test set contains "snippets" from 55 shows recorded in November 2006. The data is a mixture of BC and BN data.

[3]For the SPron system this only distinguished between a silence model being at the end of a word or not.

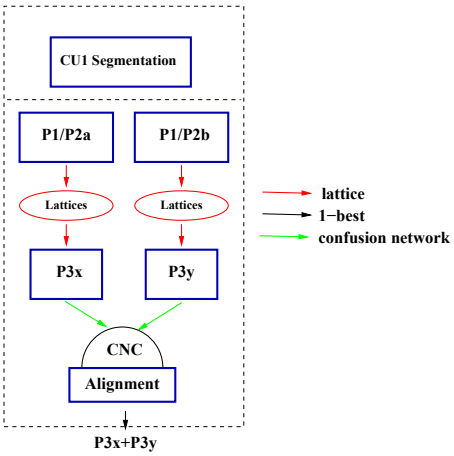

Fig. 1. Multi-pass combination framework

4.3. Single Pronunciation Modelling

All the initial experiments used the Graphemic, G0, system for the P1 and P2 stages. In contrast to the 2007 GALE CUED evaluation system, only gender independent models were used. The use of gender-dependent models was found to only give a small performance gain in the current configuration. The 260K word-list was used in this section so that the effects of using larger approximate dictionaries, could be separated from the acoustic modelling effects.

System		bcad06	bnad06	dev07
G0	Graph	25.2	19.9	15.7
V1	MPron	24.6	18.4	14.3
V3	MPron	24.2	18.3	14.3
V5	SPron	25.1	19.2	14.8
G0+V3		23.1	17.7	13.9
G0+V5	CNC	24.0	18.4	14.1
V3+V5		23.7	17.7	13.8

Table 2. P3 and CNC WER(%) using Graphemic (G0), MPron models with (V3) and without (V1) 2-state short vowel models, and a SPron (V5) system with 260K word-list LM, G0 models in P1/P2.

Table 2 shows the P3 performance for the graphemic (G0) and three phonetic systems. Note the phonetic systems have an advantage due to cross-adaptation effects from the G0 P2 stage. All the phonetic systems outperformed the graphemic system. Comparing the V1 and V3 performance, it can be seen that the use of short-vowel modelling with two-states rather than three gives gains on the BC data, but not on the BN data. This slight gain was found to be reduced when the V1 system was combined with the graphemic system output. The V3 system was used as the basis for the single pronunciation, V5, system.

The performance of the V5 SPron system on the BC data, bcad06, was only slightly better than the graphemic system and significantly worse than the other phonetic systems. This is partly

because the use of pronunciation probabilities gives gains of around 0.5% for the MPron systems, and no gain for the SPron system. For bnad06 and dev07 the performance was approximately half way between the graphemic system and the MPron systems.

The second half of table 2 shows CNC combination results. As expected combining any of the phonetic systems with the graphemic system gave large gains. Interestingly the V5 SPron system when combined with the V3 MPron system also gave good gains. This shows that the SPron is complementary to the MPron system as observed for English systems in, for example, [7]. However on bcad06 using the standard G0+V3 system outperforms the V3+V5 system.

4.4. Dictionary Expansion

System		LM	WER%		
P2	P3	Vocab	bcad06	bnad06	dev07
G0	G0 (Graph)	260K	25.2	19.9	15.7
		350K	24.1	18.5	14.6
G0	V3 (MPron)	260K	24.2	18.3	14.3
		350K	23.6	17.9	13.9
G0	V5 (SPron)	260K	25.1	19.2	14.8
		350K	24.5	18.7	14.5
V5	V5 (SPron)	260K	25.6	19.3	14.6
		350K	25.0	18.9	14.5

Table 3. Pass 3 results for the graphemic G0 models, and the phonetic V3 MPron and V5 SPron models using the 260K and 350K word-list language models. P1/P2 used either G0 or V5 systems.

In section 3.2 a simple scheme for allowing phonetic pronunciations for any word was described. The impact of using additional automatic pronunciations along with the full 350K word-list is shown in table 3. For the graphemic system (G0), where consistent dictionary entries can be derived for all words, large reductions in WER were achieved using the full 350k vocabulary. For example on bcad06 the error rate was reduced by 1.1% absolute.

Taking the results of the P3 passes using the G0 P2 pass, both the MPron and SPron systems yield WER reductions using the larger word-list. However, the gains are smaller than those for the graphemic G0 system. For example the MPron system gained 0.6% absolute on bcad06. Thus the automatic pronunciations are, as expected, not as consistent as the graphemic ones but are still useful.

As an additional contrast, the SPron system was used in the P1-P2 lattice generation stage. to investigate whether the graphemic lattices were limiting performance changes due to poor pronunciations. From the results this is not the case, with the larger word-list still showing consistent gains in this configuration.

4.5. System Combination

Table 4 shows a range of individual P3 outputs and the results of CNC combination of multiple branches. The best performing P3 branches both used the V3 MPron system with either the SPron V5 P2 stage, or the graphemic G0 stage. For the easier dev07 data the V5/V3 (P3c) configuration was 0.3% absolute better than the G0/V3 (P3b) configuration. Both these systems were evaluated in combination with the G0 system at the CNC stage. As expected large gains were obtained. The P3a+P3c configuration was at least as good, and 0.7% absolute better on dev07, than the P3a+P3b system. Both were better than the P3d SPron system in combination with the P3a graphemic G0 system. Thus using the SPron system to give additional cross-phonetic adaptation is useful in this multi-pass combination framework.

System		P2/P3	WER%		
			bcad06	bnad06	dev07
P3a	Graph	G0/G0	24.1	18.5	14.6
P3b	MPron	G0/V3	23.6	17.9	13.9
P3c	MPron	V5/V3	23.8	17.8	13.6
P3d	SPron	V5/V5	25.0	18.9	14.5
P3a+P3b		CNC	22.5	17.2	13.7
P3a+P3c			22.5	17.0	13.1
P3a+P3d			23.0	17.6	13.6

Table 4. Single branch and combination numbers using the MPron and SPron phonetic systems, V3 and V5, and the graphemic G0 system. The 350K word-list LM was used at all stages.

5. CONCLUSION

In this paper two aspects of generating and using phonetic Arabic dictionaries have been investigated. First the use of single pronunciation systems was evaluated. These single pronunciation phonetic systems are simpler, similar to graphemic systems. With accurate pronunciations, this SPron system outperforms a graphemic system on both BN and BC-style data. The second scheme investigated handling words for which standard approaches do not yield pronunciations, graphemic systems do not suffer from this problem. Here, a series of grapheme-to-rules, and probabilities, are found and used to generate pronunciations. Though these automatically derived pronunciations do not give the same level of gains as, for example, those seen for the graphemic system, consistent gains are still observed. By combining these approaches, a consistent reduction in WERs within a multi-pass combination framework was obtained.

6. REFERENCES

[1] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent progress in Arabic broadcast news transcription at BBN," in *Proc. InterSpeech*, 2005.

[2] A. Messaoudi, J.-L. Gauvain, and L. Lamel, "Arabic transcription using a one million word vocalized vocabulary," in *Proc. ICASSP*, 2006.

[3] D. Rybach, S. Hahn, C. Gollan, R. Schlüter, and H. Ney, "Advances in arabic broadcast news transcription at RWTH," in *Proc. ASRU*, December, 2007.

[4] M.J.F. Gales, F. Diehl, C.K. Raut, M. Tomalin, P.C. Woodland, and K Yu, "Development of a phonetic system for large vocabulary Arabic speech recognition," in *Proc. ASRU Workshop*, December 2007.

[5] T. Hain, "Implicit modeling of pronunciation variation in automatic speech recognition," *Speech Communication*, June 2005.

[6] H. Soltau, G. Saon, B. Kingsbury, H.-K. Kuo, L. Mangu, D. Povey, and G. Zweig, "The IBM 2006 GALE Arabic ASR system," in *Proc. ICASSP*, 2007.

[7] M.J.F. Gales, D.Y Kim, P.C. Woodland, D Mrva, R Sinha, and S.E. Tranter, "Progress in the CU-HTK broadcast news transcription system," *IEEE Transactions Speech and Audio Processing*, September 2006.