Morphological Decomposition in Arabic ASR Systems

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

Department of Engineering, University of Cambridge
Trumpington St., Cambridge, CB2 1PZ, U.K.

Abstract

In recent years, the use of morphological decomposition strategies for Arabic Automatic Speech Recognition (ASR) has become increasingly popular. Systems trained on morphologically decomposed data are often used in combination with standard word-based approaches, and they have been found to yield consistent performance improvements. The present article contributes to this ongoing research endeavour by exploring the use of the 'Morphological Analysis and Disambiguation for Arabic' (MADA) tools for this purpose. System integration issues concerning language modelling and dictionary construction, as well as the estimation of pronunciation probabilities, are discussed. In particular, a novel solution for morpheme-to-word conversion is presented which makes use of an N-gram Statistical Machine Translation (SMT) approach. System performance is investigated within a multi-pass adaptation/combination framework. All the systems described in this paper are evaluated on an Arabic large vocabulary speech recognition

*F. Diehl
Email address: {fd257, mjfg, mt126, pcw}@eng.cam.ac.uk (F. Diehl, M.J.F. Gales, M. Tomalin, and P.C. Woodland)

task which includes both Broadcast News and Broadcast Conversation test data. It is shown that the use of MADA-based systems, in combination with word-based systems, can reduce the Word Error Rates by up to 8.1% relative.

Key words:
Automatic Speech Recognition, Arabic, Morphology, Pronunciation probabilities

1. Introduction

During the past decade, the problems associated with building Automatic Speech Recognition (ASR) systems for Arabic have become a prominent research concern (Kirchhoff et al., 2003; Messaoudi et al., 2006; Gales et al., 2007; Rybach et al., 2007; Vergyri et al., 2008; Soltau et al., 2009; Nguyen et al., 2009; Tomalin et al., 2010; Saon et al., 2010). Though most of the acoustic modelling techniques developed for Indo-European languages (such as English, French, or German) are also well suited for Arabic, standard language modelling techniques are less appropriate.

Arabic is an agglutinative, morphologically complex language. Words are formed from tri-consonantal roots and varying patterns of short vowels (vowelisation). The resulting stems may be expanded by clitics in order to convey information about parts-of-speech, gender, number, case, and so on. In Quranic Arabic, the vowels are marked by diacritics, but in Modern Standard Arabic (MSA), these are usually omitted and the resulting ambiguities are resolved by the syntactic context. This complex morphological process causes non-trivial problems when Arabic speech is modelled in Speech-to-Text (STT) systems: any given stem is likely to have many different morphologically distinct forms, and these must all be handled separately. These distinctive linguistic properties cause data sparsity problems and comparatively high Out-of-Vocabulary (OOV) rates if conventional language modelling approaches developed for non-agglutinative languages are used for Arabic. For instance, for CallHome data the vocabulary growth rate for Arabic is approximately 2.5 times higher than the rate for English (Kirchhoff et al., 2006).

In this work we address these problems by using the 'Morphological Analysis and Disambiguation for Arabic' (MADA) tools[1] (Habash and Rambow, 2005, 2007). The MADA tools implement a morphological decomposition of Arabic words into linguistically motivated prefixes, stems, and suffixes. Input Arabic text, lacking diacritics, can be converted into morphologically decomposed, fully diacritised, vowelised text. In addition, MADA also provides a rank ordered list of all vowelised word-forms associated with the corresponding input lexeme. This list is used by MADA for its disambiguation analysis, but it can also be used for STT-related tasks.

The current investigation focuses on the integration of MADA into the Cambridge University (CU) Arabic Large Vocabulary Continuous Speech Recognition (LVCSR) system. We analyse its impact on system performance and introduce a novel method for mapping the output *morpheme* sequences back into *word* sequences. Both graphemic and phonetic systems are discussed. While graphemic Arabic systems model vowels implicitly, phonetic systems model vowels explicitly in the dictionary – that is, they appear as

[1] MADA version 2.3 was used for this work.

overt letters in the orthography. This dual system design provides insights into the importance of short vowel modelling in Arabic STT. The dictionary design for the phonetic system is discussed in detail, and, in particular, the use of pronunciation probabilities based on the rank ordering of vowelised word-forms provided by MADA is investigated.

Morphological decomposition schemes for Arabic LVCSR have recently become an active research area (Vergyri et al., 2004; Afify et al., 2006; Nguyen et al., 2009; Ng et al., 2009; Xiang et al., 2006; Choueiter et al., 2006; Lamel et al., 2008; Kuo et al., 2009, 2010). However, although MADA has already been widely used both for Arabic-to-English Statistical Machine Translation (SMT) systems (Habash and Sadat, 2006; Bender et al., 2007; de Gispert et al., 2008) and for Arabic dictionary generation (Vergyri et al., 2008), work on MADA-based morphological decomposition for Arabic STT has only recently started to receive attention (Tomalin et al., 2010; El-Desoky et al., 2009). In addition to providing decompositions, vowelisations, and disambiguation analysis for Arabic, one advantage of the MADA tools is that they are publicly available for research use.[2] This contrasts with some of the commercially available alternative software.

This article is structured as follows: section 2 introduces the MADA tools; section 3 discusses the integration of MADA into the CU Arabic LVCSR system; section 4 describes the experimental setup; section 5 provides information concerning experiments and results, while, finally, section 6 summarises the main conclusions.

[2] Available from http://www1.ccls.columbia.edu/∼cadim/

2. The MADA Tools

Arabic is a highly inflected, morphologically rich language in which grammatical features are often indicated by the attachment of clitics and affixes to lexical roots. The attachment of conjunction ($CONJ$) and particle ($PART$) proclitics, the definite article Al, and pronominal ($PRON$) enclitics is largely rule-governed, and the basic patterns are indicated in the following schema (Sadat and Habash, 2006):

$$[CONJ+ \quad [PART+ \quad [Al+ \quad ROOT \quad +PRON]]]$$

As mentioned in section 1, all the MADA-based systems described in this article were trained on data that had been processed using the MADA tools, and specifically the MADA D2 configuration. The tools implement a tokenisation and tagging stage which is then followed by a morphological disambiguation stage. The final output contains decomposed morpheme sequences (Habash and Rambow, 2005).

The MADA tools support various levels of morphological detail. In the D2 configuration, four particle proclitics (specifically, l+ (to/for), b+ ($by/with$), k+ ($as/such$), f+ (in)) and one conjunction proclitic, w+ (and), are identified and separated from their associated word roots (Habash and Sadat, 2006).[3] If required, the basic D2 processing can be modified. For instance, in Diehl et al. (2009) a processing referred to as D2+Al was investigated.

[3] The Buckwalter romanised Arabic transliteration conventions are used throughout this paper.

The D2+Al scheme separates the same morphemes as D2, but, in addition, it separates all definite articles (*Al*) which do not precede one of the so-called 'solar' consonants.[4] However, since the D2+Al scheme degraded system performance when compared to the D2 scheme, the latter was used in the work reported in the current article.

The following two schemes are discussed in this work:[5]

- **Word**: a word-based system, with no morphological decomposition (baseline)

- **D2**: the prefixes *l*+, *b*+, *k*+, *f*+, and *w*+ are separated from the word stem using MADA tools

An example of D2 decomposition is given in Table 1. In addition to illustrating the decomposition itself, this example also indicates a further feature which MADA provides – namely, stem normalisation. The second *A* (*alif*) in *AlAyrAnY* becomes *M* (*alif with hamza below*), and the *Y* (*alif maksura*) in *AlAyrAnY* becomes *y* (*yeh*).

The stem normalisation is based on an analysis of both the lexical token and its immediate grammatical context (Habash and Rambow, 2005). Typically, approximately 20% of the tokens are affected by normalisation.

[4] In Buckwalter notation the solar consonants are *v t S $ s z r d D Z C l n T*.

[5] The MADA tools can also split off pronominal suffixes. However, such decomposition schemes were not explored in the work described in this article. We also explored the possibility of merging affixes to create 'compound' affixes (e.g., $l + w + ROOT \rightarrow lw + ROOT$). However, compared to the original decomposition without merging affixes we found that this kind of decomposition does not improve system performance.

System	**Example Sentence**		
Word		*wktb*	*AlAyrAnY*
D2	*w+*	*ktb*	*AlMyrAny*
Translation	*and*	*he-writes*	*the Iranian*

Table 1: Examples of D2 morphological decomposition.

From the perspective of morpheme-to-word conversion, stem normalisation introduces an additional level of complexity. Before reattaching the prefixes to their corresponding stems, the normalised words must be mapped back to their original forms. However, due to the context dependent nature of the stem-normalisation process, there is no unique bidirectional mapping from the MADA domain back to the word domain. Instead, a context dependent morpheme-to-word back-mapping procedure is required.

Prior to MADA processing, a normalisation step is applied which regularises a range of common inconsistencies of the Arabic orthography. This involves

1. mapping non-word-final *alif maksura* and all other *alif*s to plain *alif*
2. mapping word-final *yeh* to *alif maksura*
3. deleting possible short vowel markers *fatha*, *damma*, *kasra*, *fathatan*, *dammatan*, *kasratan*
4. deleting the vowel omission marker *sukun* and the consonant gemination marker *shadda*

The MADA vowelisation and stem normalisation incorporates these into the text, though in a consistent way.

2.1. Dictionary Generation

In Arabic, word-forms are based on tri-consonantal roots. Although these roots sometimes convey the core semantic information, it is by means of vowelisation and affixation that they become full lexical items. The content-dependent addition of short vowels (i.e., *fatha* /a/, *damma* /u/ and *kasra* /i/) and affixes determines the semantic and grammatical features of the word, as well as its pronunciation. Table 2 gives some examples for possible word variations for the tri-consonantal root *ktb* (write).

Vowelised Form	**Meaning**
kataba	he writes
kutiba	it is written/fated/destined
kitab	book (indefinite)
kutubun	books (indefinite)
kutubu	books (definite)

Table 2: Several vowelised forms for the tri-consonantal root *ktb*

The short vowels and the word-final nunations *fathatan* /an/, *dammatan* /un/ and *kasratan* /in/, are indicated by diacritic markers placed over the preceding consonant.[6] However, these diacritic markers are not normally written in MSA, and the reader is expected to infer the diacritised form of

[6] In Buckwalter notation the nunations /an/, /un/, and /in/ are transcribed as ‘F’, ‘N’, and ‘K’, respectively. However, phonologically, they correspond to a short vowel followed by the nasal /n/, which is reflected by the transcriptions ‘an’, ‘un’, and ‘in’ used in this work.

the word from the semantic and grammatical context. Since there exists a roughly one-to-one mapping between the fully diacritised orthographic word-forms and the associated phoneme sequences (Habash, 2010), the diacritised word-forms are used as phonetic transcriptions in the STT dictionary.

From Table 2, it is clear that many possible pronunciations exist for a given tri-consonantal root. The absence of the short vowel markers in MSA Arabic text complicates the process of dictionary generation for STT. The usual approach for Arabic pronunciation generation is to rely on an analysis system such as Buckwalter (Buckwalter, 2002). For instance, Buckwalter expands the grapheme sequence '*ktb*' ('book') in 8 distinct ways (some of which are given in Table 2). These are generally verbal and nominal forms.

The Buckwalter analyser forms an integral part of the MADA tools, and the generation of diacritised morphemes is controlled by setting the 'DIAC' option. For the example given in Table 1, this setting produces the diacritised morpheme sequence shown in Table 3.

System	**Example Sentence**		
Word		*wktb*	*AlAyrAnY*
D2+DIAC	*wa+*	*kataba*	*AlMiyrAniyu*

Table 3: Examples of D2+DIAC morphological decomposition.

The difference between the morpheme sequences generated by the D2+DIAC and the D2 options consists in the additional annotation of the three short vowels, as well as the use of *shadda* /∼/ and *sukun* /o/. Removing the corresponding markers from the D2+DIAC output produces the same morpheme sequence as that generated by the D2 option. The D2+DIAC output in Table 3 constitutes the most likely diacritised morpheme sequence for the input word sequence. It results from the morphological analysis and disambiguation performed by MADA. However, for all words in the input text, MADA also provides ranked listings of all possible diacritised morpheme forms (Habash, 2010), as shown in table 4.

Rank	**Weight**	**Transliteration**
1.	0.740871	wa+ kataba
2.	0.710576	wa+ kutiba
3.	0.535714	wa+ kutubun
=3.	0.535714	wa+ kutubu
=3.	0.535714	wa+ kutub
=3.	0.535714	wa+ kutubin
=3.	0.535714	wa+ kutubi
=3.	0.535714	wa+ kutuba

Table 4: Ranked MADA transliterations for the word *wktb*. The higher the weight, the higher the likelihood for the corresponding transliteration.

Table 4 shows the ranking for the complex *wktb*. Besides providing all possible diacritised morpheme sequences, weights are also provided and these define a rank ordering of all transliterations. The weights are generated by the morphological analysis and disambiguation performed by MADA and they indicate which transliteration is the most likely variant (in the given syntactic context). The most likely morpheme sequence for the word *wktb* in Table 3 is the highest ranked entry in of Table 4. Therefore, this sequence constitutes the 1-best MADA output.

2.2. Pronunciation Probabilities

MADA offers an elegant, self-contained way to obtain pronunciation probabilities for the various vowelisation alternatives it assigns to the surface form of a word. According to Table 4, for each token processed by MADA, a list of vowelised word-forms is generated. These forms are ordered by a weight indicating respective vowelisation likelihood. Given the vowelisation frequency counts, 'pseudo' alignment counts can be generated by multiplying the number of vowelisation occurrences by the corresponding weight. These pseudo counts can be normalised to obtain estimates for the pronunciation probabilities.

The MADA-derived pseudo counts have the advantage that they do not depend on the availability of acoustic alignment data (unlike the alignment-based pronunciation probabilities used in (Gales et al., 2007)). Alignment-based pronunciation probabilities are based on the alignment of acoustic training data, and therefore they suffer from the fact that counts for different pronunciation alternatives can only be obtained for words seen in the acoustic training data. By contrast, MADA-derived pseudo counts are based on the Language Model (LM) training data which contains over 50 times more tokens. Consequently, greater dictionary coverage is obtained. In fact, pronunciation probabilities can be generated for all words for which MADA provides vowelised word-forms.

3. System Integration and SMT-style Domain Conversion

The MADA decomposition of a given Arabic text, and the corresponding dictionary generation, is initiated by applying MADA using the D2+DIAC option. To obtain the final graphemic morpheme sequence, the short vowels, as well as *shadda* and *sukun*, are removed from the 1-best output. The dictionary generation is based on the rank ordered listing of the possible diacritisations of a word. Final pronunciations (the 'phonetic' forms), are obtained by removing *shadda* and *sukun* from the transliterations.[7] As this description indicates, graphemic forms and their corresponding phonetic forms are distinguished only by the use of the short vowels and the related word-final nunation forms. Consequently, in summary, the main processing steps in the word-to-morpheme sequence conversion and dictionary generation are as follows:

- Apply MADA with the option D2+DIAC to the Arabic text.
- For word-to-morpheme sequence conversion the short vowels, nunations, *shadda*, and *sukun* are stripped-off from the 1-best MADA transliteration. The prefixes l+, b+, k+, f+, and w+ (with their prefix marker) are left unchanged. The resulting morpheme sequences provide the transcriptions for LM and acoustic training.
- For dictionary generation, all word-morpheme pairs are collected from the rank ordered transliterations. As for morpheme sequence generation, the graphemic forms are obtained by stripping-off the short vowels, *shadda*, and *sukun*. However, for the phonetic forms, only *shadda* and *sukun* are removed.

[7] So, *shadda* and *sukun* are not modelled as acoustic base units.

3.1. MADA-to-Word Back-mapping

As mentioned above, the stem normalisation introduced by MADA complicates the morpheme-to-word conversion. Crucially, the prefixes cannot simply be reattached to their stems. Instead, an additional stem back-mapping is required. The stem normalisation performed by MADA is a context-sensitive process. There is no one-to-one assignment between an original stem and a normalised stem. Consequently, a context dependent mapping procedure is required.

To cope with this, a novel approach was adopted in which the MADA-to-word conversion is viewed as a SMT task. In this framework the MADA-domain is the source 'language', while the word-domain is the target 'language'. Put simply, the morpheme sequences are 'translated' into word sequences. This particular 'translation' problem involves many-to-one mappings from the source to the target. Many-to-one mappings from the target to the source do not occur (see Table 1). Further, the approach does not require any reordering, and it gives a linear alignment between the source domain tokens and the target domain tokens. Translation problems structured like this are well defined, and the N-gram SMT approach provides an efficient solution (Mariño et al., 2006). N-gram-based SMT applies a 'bi-lingual' LM trained on so-called 'tuples'. These tuples (t, s), which form the basic bi-lingual units, assign exactly one target token t to one or more source tokens s. They are obtained in training by the unambiguous alignment of the source and target data.

The translation model probability $p(T, S)$ from a source sentence S to a target sentence T is given by an N-gram of tuples. For a sentence pair with

K target tokens, $p(T,S)$ is approximated by:

$$p(T,S) = \prod_{k=1}^{K} p((t,s)_k | (t,s)_{k-1}, ..., (t,s)_{k-N+1}) \tag{1}$$

Decoding is done by likelihood maximisation of $p(T|S)$ with respect to $T = (t_1, t_2, ..., t_K)$ (Crego et al., 2005).

Due to unseen tuples in the training data of the bi-lingual LM, as well as so-called 'stranded' prefixes generated by the ASR system, the N-gram SMT approach fails to translate such morpheme sequences back to the word domain. In these 'Out-of-Tuples' (OOT) cases, possible prefixes are joined to the stem and the resulting MADA domain word is directly transferred to the word domain. When a normalised stem or prefix is involved, but also 'stranded' prefixes, a new word which is not in the vocabulary may be produced. However, since the 'stranded' prefixes problem is very rare and only about 20% of the tokens are affected by stem normalisation, simply recomposing the word by joining the prefixes to the stem provides a reasonable back-mapping procedure.

As a side-effect of the stem back-mapping, the N-gram SMT approach implicitly solves the problem of rejoining the prefixes to their stems. This is effectively caused by the many-to-one assignments from MADA-domain tokens to word-domain tokens to form the tuples. With respect to the example given in Table 1, one of these tuples would be: (*wktb*, *w+ ktb*), effectively rejoining the split-off prefix to its stem.

4. Experimental Setup

In this section, graphemic and phonetic MADA-based systems are discussed together with their corresponding word-based systems which do not make use of any morphological decomposition strategy. The latter provide a baseline reference for system performance. After describing the experimental setup, the MADA-based LM is compared to its word-based counterpart and the process of dictionary generation is discussed.

4.1. Acoustic Models

For both the graphemic and phonetic systems, dedicated word-based and MADA-based acoustic models were required. It is not possible to apply the existing word-based acoustic models to the MADA-domain since the phoneme sequences are modified by MADA stem normalisation, and the morphological decomposition has an impact on the structure of the phone-lattices needed for Minimum Phone Error (MPE) training (Povey and Woodland, 2002). In other words, the use of MADA always involved the complete rebuild of the acoustic model in the MADA-domain. However, with the exception of the MADA processing, the corresponding system builds are identical. Specifically, both graphemic systems model 36 Arabic graphemes: the 28 basic Arabic consonants, plus *hamza* and the 7 modified letters *alif madda*, *alif maqsura*, *alif with hamza below*, *alif with hamza above*, *waw hamza*, *yeh hamza* and *ta marbuta* (Gales et al., 2007). For the phonetic systems, this inventory is expanded by the inclusion of the 3 short vowels *fatha*, *damma* and *kasra*. Both systems are trained on approximately 1825 hours of acoustic training data. The data consists of GALE Broadcast Conversation (BC) and

Broadcast News (BN) data (Linguistic Data Consortium, 2010), from GALE phase 1 to 4, as well as small amounts of FBIS and TDT-4 data.[8]

The audio data was parametrised using a 39-dimensional Perceptual Linear Predictive (PLP) front-end which extracts 13 PLP cepstra (including the zeroth cepstral coefficient with first, second and third delta parameters) followed by a Heteroscedastic Linear Discriminant Analysis (HLDA) projection from 52-dimensions down to 39 (Liu et al., 2003). Cepstral mean normalisation was then applied. Cross-word decision-tree state-clustered triphones (with approximately 9K states and an average of 36 Gaussians per state) were then trained on this data using (initially) an ML-criterion. MPE discriminative training was then performed. For the multi-pass combination and adaptation systems, discriminative-MAP adaptation, (Povey et al., 2003), was used to generate Gender Dependent (GD) models. For the phonetic system, any pronunciations which could not be derived by the Buckwalter morphological analyser were obtained using the automatic pronunciation system described in Gales et al. (2007).

4.2. Language Models

The MADA-based LMs were built using 1G words of training data. The 22 available word-based sources consisted of webdata and newswire data, as well as the transcriptions of the acoustic data. All these sources were processed using MADA, and source-specific uni-gram LMs were built, interpolated, and merged. The interpolation test set combined the dev09sub,

[8] All the data (except the 45 hour FBIS data) was prepared by the Linguistic Data Consortium.

dev08, dev07, and eval06 sets which are used by the various sites participating in the DARPA GALE program.[9] The merged uni-gram provided an ordering of the words in the training data, and the top 65k, 130k, 195k and 350k words were chosen as wordlists (or, more accurately, morpheme lists). The word-based LM build followed the same steps, though the 350k wordlist was built directly using the 22 sources without MADA processing.

To facilitate the comparison of the MADA-based and word-based OOV rates, the former were normalised. This normalisation is given by Equation 2. It takes into account the increased number of tokens produced by the MADA decomposition. The normalisation factor was typically around 1.12.

$$\%\mathrm{OOV}_{\mathrm{norm}} = \%\mathrm{OOV} \times \frac{\#\text{ of morphemes}}{\#\text{ of words}}. \tag{2}$$

Table 5 compares the normalised OOV rates for the MADA-based wordlists with the OOV rates for the word-based wordlist. It shows that the morphological decomposition scheme reduces the OOV rate for 350k wordlists by roughly a factor of two. It also shows that the 130k and 350k MADA-based systems have similar word-level OOV rates. This corresponds to a factor 2.7 in the reduction of the vocabulary size.

Using the 350k wordlists for both the MADA-based and word-based systems, 22 source-specific LMs were built, interpolated, and merged to create MADA-based and word-based LMs. Once again the interpolation test set consisted of the combined dev09sub, dev08, dev07, and eval06 sets. The

[9] The dev09sub test set is in fact the so-called dev09sub3 test set defined in the DARPA GALE program. It is the third subset of the dev09 test set and it excludes several snippets which overlap with the LM training data.

Wordlist	**eval07**	**eval08**	**dev09sub**
350k word	1.26	0.63	1.18
65k MADA	2.18	1.27	2.19
130k MADA	1.16	0.58	1.23
195k MADA	0.80	0.38	0.92
350k MADA	0.52	0.27	0.62

Table 5: OOV rates for the 350k word-based wordlist and various MADA-based wordlists.

sources with the highest interpolation weights were the two acoustic sources, gale_bn (weight$_{word}$ = 0.22, weight$_{MADA}$ = 0.21) and gale_bc (weight$_{word}$ = 0.22, weight$_{MADA}$ = 0.21). Since they use different wordlists, the raw Perplexities (PP) for the two LMs are not directly comparable, so *word-level* PPs for both LMs are given in Table 6 since these can be compared. The MADA-based PPs have been normalised as follows: $PP^{y/x}$, where y = # tokens in the MADA-based test sets and x = # tokens in word-based test sets.[10]

LM	**eval07**	**eval08**	**dev09sub**
word-based	599	348	824
MADA-based	502	354	827

Table 6: Perplexities for the unpruned word-based and MADA-based 4-gram LMs.

Any morphological decomposition of a given text increases the number

[10] For completeness, the unnormalised MADA-based PPs for the three test sets were as follows: eval07 = 268, eval08 = 200, dev09sub = 450.

of tokens in the text. This has the negative side-effect that, when keeping the context length of the LM in the word-based and MADA-based domain constant, the LM in the latter domain has reduced coverage. For the experimental setup described in this article, it was found that MADA increases the number of tokens by a factor around 1.12. Consequently, larger LM contexts were explored, and 5-gram word-based and MADA-based LMs were built. These LMs replaced the 4-gram LMs used for lattice rescoring during P3 decoding (see section 4.4). However, in both cases, no consistent improvements in system performance were observed, so these models were not used. Instead all the tests used 4-gram LMs for lattice rescoring. These results are not surprising since limited improvements were obtained when word-based tri-gram LMs were replaced by word-based 4-gram LMs.

4.3. Graphemic and Phonetic Dictionaries

Dictionary generation for the phonetic systems is a complex process. As described in section 2.1, the process for the MADA-based system involves grapheme sequences being derived from Buckwalter (albeit modified by MADA), while in the word-based case, Buckwalter is applied directly. However, Buckwalter's coverage is limited: the analyses it produces are derived from morpheme tables which list possible Arabic root forms and affixes, and a set of rules is applied to combine these morphemes. Words which do not follow the patterns predetermined by the tables and rules cannot be analysed by Buckwalter. These include certain words which are of non-MSA origin (e.g., certain proper nouns, foreign words, dialect terms). These lexical items do not conform to the orthographic patterns expected by the Buckwalter analyser. Consequently, they are not analysed, and no pronunciations are provided for them. This accounts for 27% of the entries of the word-based wordlist and for 46% of the entries of the MADA-based wordlist. Entries were generated for these missing pronunciations by applying the rule-based vowelisation method described in Gales et al. (2007). This approach takes advantage of the typical consonant-vowel patterns of Arabic words. It divides the words into their constituent graphemes, and it treats them as individual words with possible pronunciation derived from the analysable words. Pronunciation probabilities are derived by an alignment of the acoustic training data, and word-level pronunciations are obtained by concatenating the most likely pronunciations of the individual graphemes forming a given word.

Though this rule-based vowelisation method generates pronunciations for all missing words in the dictionary, the process is rather complex and the phonetic transcriptions produced are of questionable quality. Graphemic systems provide a way of circumventing these problems (Gales et al., 2007). Instead of modelling the short vowels explicitly in the dictionary (and therefore by means of dedicated acoustic models), a graphemic system models the vowels implicitly. Dictionary generation consists of the direct use of the graphemic word-forms as the corresponding phonetic transcription. Although this procedure reduces the resolution of the acoustic modelling, it means that the dictionary can be produced in an easy and consistent manner. In addition, since there are no pronunciation alternatives, it keeps the dictionary size compact, and this results in faster training and decoding times.

For the two phonetic systems, Buckwalter generated up to 58 vowelised variants per word. Since this is undesirable, a two stage procedure was applied to reduce the number of pronunciations. The acoustic training data was aligned using both the Buckwalter derived dictionaries and the additional rule-based vowelised word-forms. The resulting count statistics enabled the Buckwalter pronunciations to be reduced to the N most likely ones. After preliminary tests, N was set to 15, since no performance improvement was found when a larger number of pronunciation variants per word was used. However, since rule-based pronunciations were generally less accurate and thus acoustically broader, the number of pronunciation alternatives in this case was set to 5.

The alignment statistics also enabled pronunciation probabilities to be estimated. These were simply obtained by normalising the count values from the alignment.[11] For words not seen in the alignment data, a flat count distribution was assumed. In the context of the word-based system, this method constitutes the standard way to obtain pronunciation probabilities since the previously described method using pseudo counts derived from MADA is not available. For the MADA-based systems, this alignment method is a natural alternative to the pseudo counts method.

4.4. System Architecture

All the systems discussed in this paper used a multi-pass combination framework (see Figure 1). This is similar to the framework described at length in (Evermann and Woodland, 2003). P1 is a fast decoding pass which generates hypotheses for least-squares linear regression (LSLR) and variance adaptation for the second stage (P2). The P2 stage generates lattices using a tri-gram based LM. This is then expanded using a 4-gram LM to pro-

[11] The counts were smoothed by adding an initial count of seven.

duce the lattices that are passed to the P3 stage. At this point, 1-best CMLLR (Gales, 1998) and lattice-based MLLR adaptation is applied (Uebel and Woodland, 2001). The lattices are then rescored and confusion networks generated (Mangu et al., 2000). Manual segmentations were used for all the data.[12] The P1 and P2 stages featured only the graphemic models. By contrast, two separate branches were run for the P3 stage. These applied the graphemic models and the phonetic models respectively. The individual outputs of the two branches were either tested directly or were combined using Confusion Network Combination (CNC) (Evermann and Woodland, 2000). System performance was assessed using the dev09sub development set, and the non-sequestered evaluation sets eval07ns and eval08ns. Each of these test sets consists of approximately 20k Arabic words (corresponding to 3 hours of audio) selected from both BN and BC sources. These test sets are used by the various sites participating in the DARPA GALE program.

5. Experiments and Results

In this section, experiments and results are presented. First, the order of the 'bi-lingual' morpheme-to-word mapping LM is investigated. This is followed by a discussion of the stem-normalisation property of MADA, and an

[12] The system architecture does not include the use of neural network LMs for lattice rescoring, multilayer perceptron acoustic features, or boosted maximum mutual information acoustic model training – all techniques discussed as in previous work by the authors. Incorporating these features into the systems described here would have complicated the archiecture considerably without contributing in any way to the discussion of MADA-based techniques for STT.

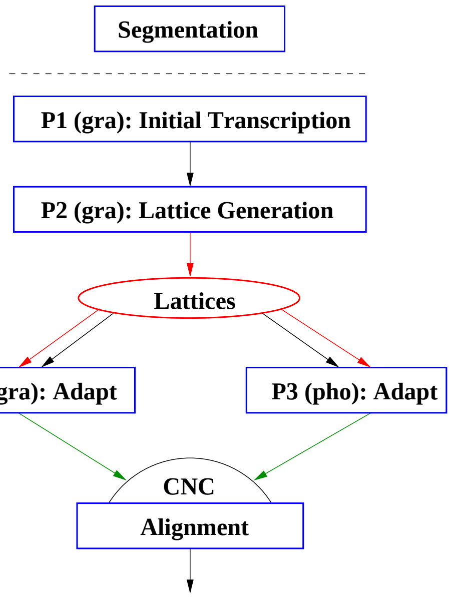

Figure 1: Multi-pass combination framework

in-depth investigation of MADA-based pronunciation probabilities. Finally, performance figures for single branch and combination systems are given.

5.1. Morpheme-to-word Back-mapping LM

In section 3.1, it was pointed out that the word-to-morpheme conversion requires a context-sensitive morpheme-to-word back-mapping. This section describes the 'bi-lingual' mapping LM that is needed for the N-gram SMT back-mapping approach. In addition, the possibilities concerning context length are explored.

The 'bi-lingual' mapping LM was trained using the acoustic training data transcriptions (∼11M words). After MADA processing, the MADA-domain data was aligned with the original word-domain data. The resulting streams of 'source-target' token pairs were used to train the LM which generated the mappings. The OOT rates were 3.63 for eval07, 0.75 for eval08, and 1.44 for dev09sub. Applying the fall-back procedure for unseen OOT tokens described in section 3.1, this was effectively lowered to 2.77 for eval07, 0.40 for eval08, and 0.51 for dev09sub. To investigate the implications of context length, three LMs were built – a uni-gram, a bi-gram, and a tri-gram – and for the actual morpheme-to-word conversion the MARIE N-gram SMT decoder was used.[13] Decoding results for the graphemic MADA-system after morpheme-to-word back-mapping are presented in Table 7.

For all three test sets, Table 7 shows a reduction in WER of 0.2% absolute when increasing the context length from zero to one (i.e., uni-gram to bi-gram). However, no further reductions were obtained for a context length of two (i.e., tri-gram). These results confirm the context dependency of the MADA process and indicate that it may be covered by a bi-gram LM. However, for safety, the tri-gram back-mapping LM was applied for morpheme-to-word conversion of the decoding results in all the tests reported in this article. In additional tests, the amount of data used to train the bi-lingual mapping LMs was increased from the original ∼11M words to ∼154M words. However, although the OOT rates were reduced slightly, no improvements in system performance were observed.

5.2. Stem Normalisation

In section 2, the stem-normalisation property of MADA together with its impact on the morpheme-to-word conversion was discussed. The present section investigates this property further in order to determine its contribution

[13] Available from http://gps-tsc.upc.es/veu/soft/soft/marie.

System	**Test Set**		
n-gram	**eval07ns**	**eval08ns**	**dev09sub**
1	13.6	11.9	18.1
2	13.4	11.6	17.9
3	13.4	11.6	17.9

Table 7: Impact of different context length for the MADA-2-word back-mapping LM (in % WER).

to the system performance. For these tests, graphemic systems were used. Three systems are compared: the word-based baseline system, a MADA-based system, and a word-based stem-normalised ('*WordNorm*') system. The *WordNorm* system was developed from a post-processed version of the MADA training data transcriptions. The post-processing involved rejoining stem-separated proclitics with their corresponding (possibly normalised) stems, giving a word-based stem-normalised transcript. Starting from these new transcriptions, corresponding LMs and acoustic models were built and tested. As the '*WordNorm*' system is based on stem-normalised transcripts this system is formally handled in the same way as a MADA-based system. It includes in particular the MADA-to-word conversion as decribed in chapter 3.1. Table 8 gives the associated test results.

Comparing the word-based system with the MADA-based system, Table 8 shows reductions in absolute WER of 0.8%-1.0% in favour of the MADA-based systems, which confirms their superiority. Further, comparing the word-based system with the *WordNorm* system shows that the latter always outperforms the former by 0.1%-0.4% in absolute WER. These results show

System	**Test Set**		
	eval07ns	**eval08ns**	**dev09sub**
Word	14.3	12.4	18.9
WordNorm	14.2	12.0	18.7
MADA	13.4	11.6	17.9

Table 8: Impact of the stem-normalisation property on the system performance (in %WER).

that the WER reductions obtained by the MADA-based system are mainly caused by the morphological decomposition scheme. However, the stem normalisation also contributes to the gains, though to a lesser extent.

5.3. Pronunciation Probabilities

As outlined in section 4.3 and section 2.2, pronunciation probabilities (PProb) for phonetic MADA-based systems can be obtained in two ways: by means of counts based on the alignment of the acoustic training data or by means of pseudo counts based on the rank order information of vowelisation alternatives provided by MADA.

The alignment-counts-based PProb estimation provided pronunciations for 210k words of the word-based dictionary and 136k words of the MADA-based dictionary PProb estimates. For the remaining 40k and 53k words of the two dictionaries for which Buckwalter vowelisations were available, a flat distribution was used. The same flat PProb distribution was also used for the other 90k and 161k words for which rule-based pronunciations had to be applied.

The first four lines of Table 9 compare the word-based and the MADA-based phonetic systems with and without the application of pronunciation probabilities during P3 decoding. PProbs give absolute reductions in WER of 0.3%-0.7% for the word-based system. For the MADA-based system, the WER reductions are generally smaller but reach 0.4% in absolute WER.

System		**Test Set**		
Model	**PProb**	**eval07ns**	**eval08ns**	**dev09sub**
Word	none	13.4	11.8	18.1
Word	align	13.1	11.1	17.5
MADA	none	12.5	10.8	17.6
MADA	align	12.5	10.7	17.2
MADA	MADA-unseen	12.5	10.7	17.2
MADA	MADA-all	12.7	10.9	17.6

Table 9: Impact of alignment count-based PProbs and MADA pseudo count-based PProbs on the word-based and the MADA-based phonetic system (in % WER).

Next the impact of using MADA pseudo count-based PProbs was investigated. As outlined in section 2.2, MADA-based PProbs have the advantage that they are available for all words for which MADA provides Buckwalter vowelisations. Consequently, PProbs were available not only for the 136k words seen in the acoustic alignment data, but also for the remaining 40k words with Buckwalter vowelisations. To test the performance of the MADA-based PProbs, two tests were carried out. First, only the flat PProbs of the 40k 'unseen' words in the acoustic alignment data were replaced by MADA-based PProbs. Next 'all' words for which Buckwalter vowelisations were available were assigned MADA-based PProbs. The corresponding results are labelled 'MADA-unseen' and 'MADA-all' in Table 9. For the 'MADA-unseen' results, no difference to the 'align' results can be observed. However, in the 'MADA-all' case, Table 9 reveals a degradation in system performance of 0.2%-0.4% in absolute WER. For the test sets eval07ns and eval08ns, the system performance is even worse than if the pronunciation probabilities had not been applied at all.

The poor performance of the MADA-based PProbs may be due in part to the fact that the tools implement rules primarily for written MSA Arabic and therefore they do not specifically attempt to cover the full range of pronunciation variants that are present in conversational and dialectal Arabic. To explore this further, the probability distributions derived respectively from the alignment counts and from the MADA pseudo counts were compared by means of a Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951). For the alignment based distribution $p_{jk} = P_{align}(pronunciation\ j|word\ k)$ and the MADA-based distribution $q_{jk} = P_{mada}(pronunciation\ j|word\ k)$ the word-conditional KLD $\mathcal{D}_k(p_{jk}||q_{jk})$ is given by:

$$\mathcal{D}_k(p_{jk}||q_{jk}) = \sum_{j=1}^{J} p_{jk}\, log \frac{p_{jk}}{q_{jk}} \tag{3}$$

The final joint pronunciation-word KLD $\mathcal{D}(p_{jk}||q_{jk})$ is obtained by weighting the individual word-conditional KLDs of Equation (3) by the word priors $p_k = P(word\ k)$:

$$\mathcal{D}(p_{jk}||q_{jk}) = \sum_{k=1}^{K} p_k\, \mathcal{D}_k(p_{jk}||q_{jk}) \tag{4}$$

The KLD $\mathcal{D}(p_{jk}||q_{jk})$ was evaluated using the original MADA-based distribution and a flat distribution for q_{jk}. The flat distribution corresponds to the case of not applying any pronunciation probabilities. Table 10 shows that the MADA derived distribution features a larger KLD than the flat distribution. This indicates that even a flat distribution is closer to the alignment distribution than the MADA-derived one, and this is consistent with the results of Table 9.

	$\mathcal{D}(\mathbf{p_{jk}}\|\|\mathbf{q_{jk}})$
MADA	0.411
flat	0.363

Table 10: KLD between the alignment count-based PProb distribution and the MADA pseudo count-derived PProb distribution and a flat PProb distribution, respectively.

A more detailed analysis was carried out to identify the possible reasons why the MADA-derived pronunciation probabilities are so different to the pronunciation probabilities obtained by the alignment. Consequently, the 10000 words with the highest a-priori probabilities within the alignment data were selected. These words cover 72.2% of the alignment tokens. For each word, the highest ranking pronunciations from both dictionaries were compared. It was found that only in 28.2% of the cases were the pronunciations the same. Crucially, in 40.0% of the cases, a systematic difference in the word-final pronunciations was found. In these cases, the MADA-derived pronunciations indicated word-final nominal cases (i.e., nominative /u/, accusative /a/, and genitive /i/) and nunations (i.e., /an/, /un/, and /in/), whereas these endings were missing in the alignment-derived pronunciations.

This confirms the claim in Biadsy et al. (2009) that the case-ending vowels are either reduced or omitted in conversational and/or dialectal Arabic. A corollary of this observed mismatch in spoken Arabic and MSA is that, contra (El-Desoky et al., 2009), a MADA-derived vowelised LM – that is, a system design which models the short vowels at the graphemic level and not as pronunciation alternatives by the dictionary – does not improve system performance. The use of the pronunciation probabilities corresponds to a vowelised uni-gram LM, and these probabilities already have an adverse impact on the WER.

5.4. Overall System Performance

Table 11 compares the word-based graphemic and phonetic systems with their MADA-based counterparts and presents the combination results. The individual P3 results show that, in the phonetic case, the reduction in absolute WER obtained by introducing MADA ranges between 0.3%-0.6% compared to 0.8%-1.0% for the graphemic system. After combining the graphemic and phonetic branches by CNC, the performance improvement is in the range of 0.6%-0.7% in absolute WER which indicates that the graphemic modelling captures additional information which cannot be modelled by the phonetic system. Crucially, CNC tends to be more effective in the MADA case since, when compared to the corresponding phonetic system, larger reductions in WER are obtained for the MADA setup. Running the ‘Matched Pair Sentence Segment Word Error’ (MAPSSWE) significance test (Gillick et al., 1989; Pallet et al., 1990) on the CNC test results, it was found that the 0.6%-0.7% absolute reductions in WER between the word-based baseline and the MADA-based tests are significant with a 95.0% confidence level. Finally, combining all four branches by ROVER produces further reductions in absolute WER by 0.1%-0.4% compared to the MADA CNC system. With respect to the best word-based system (CNC) the introduction of the MADA branches by ROVER gives relative reductions of 5.4%-8.1% in WER.

System			**Test Set**		
			eval07ns	**eval08ns**	**dev09sub**
P3	**gra**	**Word**	14.3	12.4	18.9
P3	**gra**	**MADA**	13.4	11.6	17.9
P3	**pho**	**Word**	13.1	11.1	17.5
P3	**pho**	**MADA**	12.5	10.7	17.2
CNC	**gra⊕pho**	**Word**	12.3	10.6	16.8
CNC	**gra⊕pho**	**MADA**	11.7	9.9	16.2
ROVER			11.3	9.8	15.9

Table 11: MADA-based systems compared to word-based systems on the P3, CNC and ROVER stage (in % WER).

6. Conclusions

Morphological decomposition for Arabic speech recognition is currently an active area of research in state-of-the-art ASR systems. This article has investigated the use of the MADA tools in the construction of graphemic and phonetic Arabic systems. These tools, which are publicly available for research purposes, make extensive use of morphological and syntactic rules in order to provide a rigorous decomposition of input Arabic word-based data.

A novel n-gram SMT-based morpheme-to-word conversion approach was introduced, simultaneously solving the problems of the back-mapping of stem-normalised morphemes to the word-domain and the concatenation of split-off prefixes to their associated stems. In the course of investigating MADA's stem-normalisation property, it became apparent that written Arabic often lacks orthographic consistency. The MADA stem normalisation homogenises these inconsistencies and this gives better ASR performance. Furthermore, the investigation of MADA-based pronunciation probabilities showed that grammatically correct written Arabic text lacks the variabulity that is present in conversational and/or dialectal spoken Arabic. This precludes the use of a vowelised LM based on the MADA-processed data.

The experiments reported in this article show that, in terms of relative WER, the MADA-based systems outperformed the word-based systems by 5.3%-6.3% in the graphemic framework, and by 1.7%-4.6% in the phonetic framework. CNC gave a relative WER reduction of 3.6%-6.6% compared to the word-based baseline. Finally, when combining the word-based and the MADA-based branches using ROVER, a relative WER reduction of 5.4%-8.1% compared to the best word-based system was obtained.

The results presented in this article open up many avenues for future research which focuses on the use of the MADA tools in state-of-the-art ASR systems. In particular, the basic approach described in this article could be refined and modified in several different ways. For instance, as mentioned earlier, the MADA tools enable pronominal suffixes to be identified and split off from their lexical roots, and this is a modelling option would provide a revealing contrast with the kind of decomposition presented in this article.

Similarly, decomposition schemes involving the definite article, *Al*, merit closer attention. Specifically, it is well-known that the acoustic realisation of *Al* varies greatly depending upon the phonological contexts in which it appears, and it would be possible to try to model this rule-governed variation more explicitly than we do at present. In addition, it would be interesting to assess the impact of using pronunciation probabilities that were generated by a Grapheme-2-Phoneme (G2P) conversion tool. These are all future research topics that await detailed consideration.

7. Acknowledgements

This work was supported by DARPA under the GALE programme via a subcontract to BN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. We would like to thank the anonymous *CSL* reviewers for their helpful comments and suggestions.

References

Afify, M., Sarikaya, R., Kuo, H.-K., Besacier, L., Gao, Y., 2006. On the use of Morphological Analysis for Dialectal Arabic Speech Recognition. In: Proc. of ICSLP, 2006.

Bender, O., Matusov, E., Hahn, S., Hasan, S., Khadivi, S., Ney, H., 2007. The RWTH Arabic-to-English Spoken Language Translation System. In: Proc. of ASRU, 2007.

Biadsy, F., Habash, N., Hirschberg, J., 2009. Improving the Arabic Pronunciation Dictionary for Phone and Word Recognition with Linguistically-based Pronunciation Rules. In: Proc. of HLT-NAACL, 2009.

Buckwalter, T., 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. In: Linguistic Data Consortium, University of Pennsylvania.

Choueiter, G., Povey, D., Chen, S., Zweig, G., 2006. Morpheme-Based Language Modeling for Arabic LVCSR. In: Proc. of ICASSP, 2006.

Crego, J., Mariño, J., de Gispert, A., 2005. An N-gram-based Statistical Machine Translation Decoder. In: Proc. of Interspeech, 2005.

de Gispert, A., Blackwood, G., Brunning, J., Byrne, W., 2008. The CUED NIST 2008 Arabic-English SMT System. In: Proc. of NIST Open Machine Translation 2008 Evaluation Workshop, 2008.

Diehl, F., Gales, M., Tomalin, M., Woodland, P., 2009. Morphological analysis and decomposition for arabic speech-to-text systems. In: Proc. of InterSpeech.

El-Desoky, A., Gollan, C., Rybach, D. Schlüter, R., H., N., 2009. Investigating the Use of Morphological Decomposition and Diacritization for Improving Arabic LVCSR. In: Proc. of Interspeech, 2009.

Evermann, G., Woodland, P. C., 2000. Posterior Probability Decoding, Confidence Estimation and System Combination. In: Proc. of Speech Transcription Workshop, 2000.

Evermann, G., Woodland, P. C., 2003. Design of Fast LVCSR Systems. In: Proc. ASRU, 2003.

Gales, M., 1998. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Computer Speech & Language 12(2), 75–98.

Gales, M., Diehl, F., Raut, C., Tomalin, M., Woodland, P., Yu, K., 2007. Development of a Phonetic System for Large Vocabulary Arabic Speech Recognition. In: Proc. of ASRU, 2007.

Gillick, L., Cox, S. J., Inc, D. S., A., N. M., 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In: Proc. ICASSP, 1989.

Habash, N., 2010. Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Habash, N., Rambow, O., 2005. Arabic Tokenisation, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In: Proc. of ACL, 2005.

Habash, N., Rambow, O., 2007. Arabic Diacritization through Full Morphological Tagging. In: Proc. of HLT-NAACL, 2007.

Habash, N., Sadat, F., 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In: Proc. of HLT-NAACL, 2006.

Kirchhoff, K., Bilmes, J., Das, J., Duta, S., Egan, M., Ji, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., Vergyri, D.,

2003. Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns Hopkins Summer Workshop. In: Proc. of ICASSP, 2003.

Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K., Stolcke, A., 2006. Morphology-based Language Modeling for Conversational Arabic Speech Recognition. Computer Speech & Language, 2006 20(4), 589–608.

Kullback, S., Leibler, A., 1951. On Information and Sufficiency. Ann. Math. Stat. 22, 79–86.

Kuo, H.-K., Mangu, L., Emami, A., Zitouni, I., 2010. Morphological and Syntactic Features for Arabic Speech Recognition. In: Proc. ICASSP, 2010.

Kuo, H.-K., Mangu, L., Emami, A., Zitouni, I., Lee, Y.-S., 2009. Syntactic Features for Arabic Speech Recognition. In: Proc. ASRU, 2009.

Lamel, L., Messaoudi, A., Gauvain, J.-L., 2008. Investigating Morphological Decomposition for Transcription of Arabic Broadcast News and Broadcast Conversation Data. In: Proc. of InterSpeech, 2008.

Linguistic Data Consortium, L., 2010. Global Autonomous Language Exploitation (GALE). http://projects.ldc.upenn.edu/gale/.

Liu, X., Gales, M., Woodland, P., 2003. Automatic Complexity Control for HLDA Systems. In: Proc ICASSP, 2003.

Mangu, L., Brill, E., Stolke, A., 2000. Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Networks. Computer Speech & Language 14(4), pp.373–400.

Mariño, J., Banchs, R., Crego, J., de Gispert, A., Lambert, P., Fonollosa, J., Costa-Jussa, M., 2006. N-gram-based Machine Translation. Computational Linguistics 32(4), 527–549.

Messaoudi, A., Gauvain, J.-L., Lamel, L., 2006. Arabic Transcription Using a One Million Word Vocalized Vocabulary. In: Proc. of ICASSP, 2006.

Ng, T., Nguyen, K., Zbib, R., 2009. Improved Morphological Decomposition for Arabic Broadcast News Transcription. In: Proc. ICASSP, 2009.

Nguyen, L., Ng, T., Nguyen, K., Zbib, R., Makhoul, J., 2009. Lexical and Phonetic Modeling for Arabic Automatic Speech Recognition. In: Proc. of Interspeech, 2009.

Pallet, D. S., Fisher, W. M., Fiscus, J. G., 1990. Tools for the Analysis of Benchmark Speech Recognition Tests. In: Proc. ICASSP, 1990.

Povey, D., Woodland, P. C., 2002. Minimum Phone Error and I-smoothing for Improved Discriminative Training. In: Proc. ICASSP, 2002.

Povey, D., Woodland, P. C., Gales, M. J. F., 2003. Discriminative MAP for Acoustic Model Adaptation. In: Proc. ICASSP, 2003.

Rybach, D., Hahn, S., Gollan, C., Schlüter, R., Ney, H., 2007. Advances in Arabic broadcast News Transcription at RWTH. In: Proc. ASRU, 2007.

Sadat, F., Habash, N., 2006. Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. In: Proc. of COLING/ACL, 2006.

Saon, G., Soltau, H., Chaudhari, U., Chu, S., Kingsbury, B., Kuo, H.-K., Mangu, L., Povey, D., 2010. The IBM 2008 GALE Arabic Speech Transcription System. In: Proc. ICASSP, 2010.

Soltau, H., Saon, G., Kingsbury, B., Kuo, H.-K., Mangu, L., Povey, D., Emami, A., 2009. Advances in Arabic Speech Transcription at IBM Under the DARPA GALE Program. IEEE Trans. ASLP, 2009 17.

Tomalin, M., Diehl, F., Gales, M., Park, J., Woodland, P., 2010. Recent Improvements to the Cambridge Arabic Speech-to-Text Systems. In: Proc. ICASSP, 2010.

Uebel, L. F., Woodland, P. C., 2001. Speaker Adaptation using Lattice-based MLLR. In: Proc. ITRW on Adaptation Methods for Speech Recognition, 2001.

Vergyri, D., Kirchhoff, K., Duh, K., Stolcke, A., 2004. Morphology-based Language Modeling for Arabic Speech Recognition. In: Proc. of ICSLP, 2004.

Vergyri, D., Mandal, A., Wang, W., Stolcke, A., Zheng, J., Graciarena, M., Rybach, D., Gollan, C., R., S., Kirchhoff, K., Fari, A., Morgan, N., 2008. Development of the SRI/Nightingale Arabic ASR System. In: Proc. of Interspeech, 2008.

Xiang, B., Nguyen, K., Nguyen, L., Schwartz, R., Makhoul, J., 2006. Morphological Decomposition for Arabic Broadcast News Transcription. In: Proc. of ICASSP, 2006.