Automatic Speech-to-Text Transcription in Arabic

LORI LAMEL and ABDELKHALEK MESSAOUDI and JEAN-LUC GAUVAIN
LIMSI-CNRS

The Arabic language presents a number of challenges for speech recognition, arising in part from the significant differences in the spoken and written forms, in particular the conventional form of texts being non-vowelized. Being a highly inflected language, the Arabic language has a very large lexical variety and with typically several possible (generally semantically linked) vowelizations for each written form. This paper summarizes research carried out over the last few years on speech-to-text transcription of broadcast data in Arabic. The initial research was oriented towards processing of broadcast news data in Modern Standard Arabic, and has since been extended to address a larger variety of broadcast data, which as a consequence results in the need to also be able to handle dialectal speech. While standard techniques in speech recognition have been shown to apply well to the Arabic language, taking into account language specificities help to significantly improve system performance.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Speech recognition

General Terms: Algorithms, Experimentation, Languages

Additional Key Words and Phrases: Arabic language processing, Automatic speech recognition, Mophological decomposition, Speech processing, Speech-to-text transcription

1. INTRODUCTION

This paper summarizes research aimed at speech-to-text transcription (STT) of Arabic broadcast data. Much of this work has been carried out in the context of the DARPA EARS and GALE programs for which speech recognition and machine translation are key supporting technologies.[1] The Arabic language poses challenges somewhat different from the other languages for which we have developed automatic speech recognition systems (mostly Indo-European Germanic or Romance) [Gauvain and Lamel 2000; Gauvain et al. 2002; Lamel and Gauvain 2008]. Modern Standard Arabic is learned in school, used in most newspapers and is considered to be the official language in most Arabic speaking countries. In contrast many people speak in dialects for which there is only a spoken form and no recognized written form. Arabic is a strongly consonantal language with nominally only three vowels,

[1] www.darpa.mil/ipto/Programs/gale

Authors' address: Lori Lamel, Abdelkhalek Messaoudi, Jean-Luc Gauvain, Spoken Language Processing Group, LIMSI-CNRS, BP 133, 91403 Orsay, France, {lamel,abdel,gauvain}@limsi.fr, http://www.limsi.fr/tlp

each of which has a long and short form. Arabic is a highly inflected language, with many different word forms for a given root, produced by appending articles ("the, and, to, from, with, ...") to the word beginning and possessives ("ours, theirs, ...") on the word end. Written texts are by and large non-vowelized, meaning that the short vowels and gemination marks are not indicated. There are typically several possible (generally semantically linked) vowelizations for a given written word, which are spoken. The word-final vowel varies as a function of the word context, and this final vowel is often not pronounced. Thus one of the challenges faced when explicitly modeling vowels in Arabic is to obtain vowelized resources, or to develop efficient ways to use non-vowelized data. It is often necessary to understand the meaning of the text in order to know how to vowelize or pronounce it correctly. To address this problem the Buckwalter Arabic Morphological Analyzer [Buckwalter 2004] is used to propose possible multiple vowelized word forms, and a speech recognizer is used to automatically select the most appropriate one.

Our initial work on transcription of Arabic was carried out using manually vocalized data [Messaoudi et al. 2004], which enabled explicit modeling of the Arabic short vowels. It was shown that even when producing a non-vocalized transcript, explicitly modeling short vowels improves recognition performance [Afify et al. 2005] over a grapheme-based approach where only characters in the non-vocalized written form are modeled [Billa et al. 2002]. However, since only very limited vocalized resources were available, research was carried out to reduce the reliance on such data. Two main directions were pursued. One direction aimed to reduce the supervision needed for acoustic model training, and another investigated how to efficiently combine vocalized and non-vocalized texts when constructing language models. As summarized in Section 4, it was demonstrated that by building a very large vocalized vocabulary of more than 1.2 million words, and by using a language model including a vocalized component, the word error rate (WER)[2] could be significantly reduced [Messaoudi et al. 2006].

Even though pronunciation modeling is generally considered straightforward in Arabic from vocalized texts, there are frequent variants arising in the pronunciation of the definite article 'Al' ('the') depending on the word context which causes the following consonant to be 'doubled'. The 'tanwin', a grammatical mark specifying that a noun is non-definite, causes word final short vowels to be 'doubled' (phonetically realized by adding an 'n' after the vowel – this also referred to as nunation). Studies that address explicitly representing the gemination and tanwin in an attempt to improve the acoustic and lexical models are reported in Section 6.

Research has also explored using morphological decomposition to address the challenges of dealing with the huge lexical variety. For the Arabic language, the combination of compounding, agglutination and inflection generate a large number of surface forms for a given root form. Morphological decomposition [Kirchhoff and

[2]The "word error" rate is commonly used to measure speech recognition performance. It takes into account three types of errors: *substitutions* (a reference word is replaced by another word), *insertions* (a word is hypothesized that was not in the reference) and *deletions* (a word in the reference is missed). The word error rate is defined as $\frac{\#\text{subs}+\#\text{ins}+\#\text{del}}{\#\text{ reference words}}$, and is typically computed after a dynamic programming alignment of the reference and hypothesized transcriptions. Note that the word error can be over 100%.

et al. 2002; Vergyri et al. 2004; Xiang et al. 2006] has been proposed to address this problem and thereby increasing the lexical coverage, and reducing errors that are due to words that are unknown to the system. Our studies on morphological decomposition are described in Section 8. Prior to presenting work specifically directed at processing the Arabic language, an overview of the speech transcription system is given in the next section.

2. RECOGNITION SYSTEM OVERVIEW

Radio and television broadcast data are challenging to transcribe since they are heterogeneous, containing segments of various acoustic and linguistic natures. The signal may be of studio quality or may have been transmitted over a telephone or other noisy channel (i.e., corrupted by additive noise and nonlinear distortions), or can contain speech over music or pure music segments. The speech is produced by a wide variety of speakers with different speaking styles: news anchors and talk show hosts, reporters in remote locations, interviews with politicians and common people, unknown speakers, new dialects, non-native speakers, etc. Speech from the same speaker may occur in different parts of the broadcast, and with different background noise conditions. In recent years the focus of research has moved from broadcast news data (primarily prepared speech in studio conditions) to the transcription of what is referred to as "broadcast conversational" speech (talk shows, debates, and interactive programs). This type of data requires the explicit modeling of spontaneous speech effects, much more common than in broadcast news, and also the ability to deal with speech from a variety of Arabic dialects. The acoustic and language modeling must accurately account for this varied data.

The broadcast news transcription system used in these experiments has two main components, an audio partitioner and a word recognizer. Data partitioning is based on an audio stream mixture model [Gauvain et al. 1998; 2002], and serves to divide the continuous stream of acoustic data into homogeneous segments, associating cluster, gender and labels with each non-overlapping segment. For each speech segment, the word recognizer determines the sequence of words in the segment, associating start and end times and an optional confidence measure with each word. The recognizer makes use of continuous density HMMs for acoustic modeling and n-gram statistics for language modeling [Young and Bloothooft 2000; Chou and Juang 2003; Gauvain and Lamel 2000; Lamel and Gauvain 2003]. Each context-dependent phone model is a tied-state left-to-right CD-HMM with Gaussian mixture observation densities where the tied states are obtained by means of a decision tree.

Word recognition is performed in multiple passes, where each decoding pass generates a word lattice which is expanded with a 4-gram LM. Then the posterior probabilities of the lattice edges are estimated using the forward-backward algorithm and the 4-gram lattice is converted to a confusion network with posterior probabilities by iteratively merging lattice vertices and splitting lattice edges until a linear graph is obtained. This last step gives comparable results to the edge clustering algorithm proposed in [Mangu et al. 1999]. The words with the highest posterior in each confusion set are hypothesized.

The first decoding pass generates initial hypotheses which are then used for

cluster-based acoustic model adaptation. This is done via one pass (less than 1xRT) cross-word trigram decoding with gender-specific sets of position-dependent triphones (typically 5k tied states) and a trigram language model. The trigram lattices are rescored with a 4-gram language model. These hypothesis are used to carry out unsupervised acoustic model adaptation for each segment cluster using the MLLR technique [Leggetter and Woodland 1995] with one regression class. Then a second lattice is generated for each segment using a bigram LM and position-dependent triphones with 11500 tied states (32 Gaussians per state). The word graph generated in this second decoding pass is rescored after carrying out unsupervised MLLR acoustic model adaptation using a variable number of regression classes.

3. PRONUNCIATION LEXICON

Letter to sound conversion is quite straightforward when starting from vowelized texts. A grapheme-to-phoneme conversion tool was developed based on a set of 37 phonemes and three non-linguistic units (silence/noise, hesitation, breath). The phonemes include the 28 Arabic consonants (including the emphatic consonants and the hamza), 3 foreign consonants (/p,v,g/), and 6 vowels (short and long /i/, /a/, /u/). In a fully expressed vowelized pronunciation lexicon, each vowelized orthographic form of a word is treated as a distinct lexical entry. The example entries for the word "kitaAb" are shown in the top part of Figure 1. As reported in [Messaoudi et al. 2004], initial speech-to-text transcription studies were carried out using vocalized word lists. Some example entries of a vocalized lexicon are given in the left part of Figure 1. In 50 hours of manually transcribed vocalized data, there were only 57k distinct lexical forms. The out-of-vocabulary (OOV) rate on an independent set of 12 hours of test data is about 15%, which is very high [Messaoudi et al. 2004].[3]

In order to extend the recognition vocabulary size, an alternative representation is to use the non-vowelized orthographic form as the entry, allowing multiple pronunciations, each being associated with a particular written form. Each entry can be thought of as a word class, containing all observed (or even all possible) vowelized forms of the word. This representation is illustrated in the right side of Figure 1, where the left column contains the non-vocalized orthographic form or word class, and the right column associates each vocalized word with a pronunciation. The pronunciation is on the left of the equal sign and the vowelized written form is on the right. This latter representation was used to create a word lexicon, where a pronunciation graph is associated with each word so as to allow for alternate pronunciations [Messaoudi et al. 2005]. Since multiple vowelized forms are associated with each non-vowelized word entry, the Buckwalter Arabic Morphological Analyzer was used to propose possible forms that were then manually verified[4]. The morphological analyzer was also applied to words in the vowelized training data in order to propose forms that did not occur in the training data. A subset of the words

[3]In speech recognition, there are typically 1.2 to 1.6 errors due to each out-of-vocabulary word.
[4]When the first release of the Buckwalter Arabic Morphological Analyzer (v1.0) was used a series of rules were developed to produce all possible forms [Messaoudi et al. 2006]. These rules were no longer needed with the v2.0 release.

Vowelized lexicon	
Traansliterated	Pronunciation
kitaAb	kitAb
kitaAba	kitAba
kitaAbi	kitAbi
kut˜aAbi	kuttAbi

Non-Vowelized lexicon	
ktAb	kitAb=kitaAb
	kitAba=kitaAba
	kitAbi=kitaAbi
	kuttAbi=kut˜aAbi
sbEyn	sabEIna=saboEiyna
	sabEIn=saboEiyn

Fig. 1. Example lexical entries for the vowelized and non-vowelized pronunciation lexicons. In the vocalized lexicon, the transliterated form is on the left and the pronunciation on the right. In the non-vowelized lexicon, the left column contains the non-vocalized orthographic form, and in the right column vocalized words are associated with their pronunciation, where the pronunciation is on the left of the equal sign and the written form on the right.

(about 1k), mostly proper names and technical terms, were manually vowelized. Using this latter representation, the 57k word vocalized entries are replaced by 33k word classes. The generalization enabled by this representation almost halves the OOV rate of the test data (to about 8%). In Section 6 an expanded phone set is explored, and in Section 7 pronunciation variants are introduced to better represent some Arabic dialects.

4. LARGE VOCALIZED LANGUAGE MODELS

The previous section described a method based on word classes which allows the transcription system to explicitly use information about the short vowels in Arabic, while being able to generalize to other word forms so as to able to make use of non-vocalized audio and textual resources. To address the large lexical variety of Arabic, a much larger recognition vocabulary is needed. In [Messaoudi et al. 2006] the lexicon was extended to 200k word-classes (with over 1 million vocalized words). Both vocalized and non-vocalized audio and textual resources are used for language modeling by constructing separate language models and interpolating them. More precisely, a 1.2 million word vocalized word language model was built by interpolating the non-vocalized LM trained on texts (390M words from LDC Arabic Gigaword corpus [Graff 2007] 204M words collected from Internet news sources) and a vocalized LM trained on 1.1M words of vocalized manual transcriptions of data from several broadcast news sources [Messaoudi et al. 2005]. There are a total of 85k different vocalized forms corresponding 50k distinct non-vocalized forms. As described in the previous section, the vocalized vocabulary has been obtained by semi-automatically generating all possible vocalized forms for the 200k non-vocalized word vocabulary.

The vocalized n-gram probabilities $P(v_i|v_{i-1},...)$ are estimated in the following way (v_i and w_i are respectively the vocalized and non-vocalized forms of ith word):

$$\begin{aligned} P(v_i|v_{i-1},...) \;=\; & \alpha P_a(v_i|v_{i-1},...) + \\ & (1-\alpha)P_v(v_i|w_i)P_t(w_i|w_{i-1},...) \end{aligned}$$

where P_a is the vocalized LM trained only on the vocalized part of the acoustic data, P_v is trained on all the acoustic data after Viterbi alignment, and P_t is the standard non vocalized LM trained on all of the data described above. Adding the

automatically vowelized transcripts to the data used to estimate the vocalized LM did not improve performance. Independent of whether a vocalized or non vocalized language model is used, the decoder outputs a non vocalized transcription. When using the vocalized LM, the posterior probabilities of the vocalized forms corresponding to the same non-vocalized word are summed to compute the word posterior probabilities. This is the same as what is done for consensus decoding [Mangu et al. 1999] with alternate pronunciations. Using pronunciation probabilities is quite important given the large number of possible pronunciation per word class. On average there about 8 forms per lexical entry, and using probabilities can give a relative the word error rate reduction of almost 10%.

5. TRAINING MODELS WITH GENERIC VOWELS

Generally speaking, extending the pronunciation dictionary to include entries for additional training data entails some manual intervention or verification. For Arabic, the difficulty lies in determining the vocalized forms, after which grapheme-to-phoneme conversion is (relatively) straightforward. In the case of a large quantity of training data with non-vocalized transcripts there can be too many words without vocalizations to add these manually or even semi-automatically. One possibility that we considered was to generate all possible vocalized forms, allowing all 3 short vowels or no vowel after every consonant. This idea was quickly rejected since there are too many possible vocalized forms. For example, with words with 4 consonants generate 512 possible pronunciations, and words with 8 consonants have 8192 possible pronunciations.

In order to simplify the problem, we investigated the use of a generic vowel to replace the three short vowels. This does not pose any problem since even though short vowels are represented internally in the system, the Arabic recognizer outputs the non-vocalized word form. Using a generic vowel offers two main advantages. First, the manual work in dealing with words that are not handled by the Buckwalter morphological analyzer (typically proper names, technical words, words in Arabic dialects) is reduced. With this approach these can be automatically processed. Second, the number of vocalizations, and hence pronunciations, per word is greatly reduced (1 vowel instead of 3).

A set of detailed rules were used to generate pronunciations with a generic vowel from the non-vocalized word form. Some rules concern the word initial Alif (support of the Hamza), which can be stable or unstable. For the former case a pronunciation is generated with a glottal attack (denoted /'/) followed by a generic vowel (denoted /@/). These rules also cover word initial letter sequences [wAl, wbAl, wkAl, fAl, fbAl, fkAl] which often correspond to a composed prefix ending in "Al". Different pronunciations are generated to represent both situations. For example, the possible pronunciations for wAl are: w@l w'@l wAl. In word final position, short vowels can be followed by an "n" (tanwin), so two forms are proposed, the generic vowel alone and the generic vowel followed by an "n". Similarly rules handle the pronunciation of words ending in "wA" and a final letter "p" (which symbolizes the ta marbouta). Within a word, a generic vowel is added after each consonant with the exception of the semivowels "w" and "y" which can be realized as respective semivowels or can serve as a support for the long vowels /U/ and /I/. A word internal Alif can

represent the long vowel or a glottal attack.

After applying these rules, each word has multiple pronunciations represented with consonants, long vowels, and the generic vowel. Since vowels may also be absent (written with a Sukoun), additional pronunciations are added by removing one generic vowel at a time. For example, the rules generate the following two generic vowel forms for the word "ktb":

ktb k@t@b@ k@t@b@n

which after allowing each generic vowel to be deleted produces:

ktb k@t@b@ k@t@b@n
kt@b@ kt@b@n
k@tb@ k@tb@n k@t@b

It should be noted that the diacritic for gemination has not been taken into account when generating pronunciations with generic vowels. These decision was taken to limit the number of pronunciations even though the gemination is explicitly represented for most words in the lexicon. In the current system, words with generic vowels are not included in the recognition word list, and are only used during training.

An experiment was carried out to assess the quality of acoustic models with generic vowels by mapping all short vowels in the vocalized lexicon to a generic vowel. Acoustic models were retrained by first mapping all short vowels to a single generic vowel (@), and training context dependent models with the standard consonant set and the single generic vowel. A pronunciation lexicon was then created that used the standard pronunciations with short vowels for the vocalized words and automatically generated pronunciations with the generic vowel for the non-vocalized words. We then segmented all of the audio data using this lexicon with a combined set of acoustic models formed by merging the CD models with short vowels and those with a generic vowel. Note that the basic idea was to use the generic vowel only in training, but not during recognition so a number of CD models are never used. In the future we may consider also extending the recognition lexicon in an analogous manner. In order to assess the feasibility of this, several model sets were built and tested in decoding using only a generic vowel.

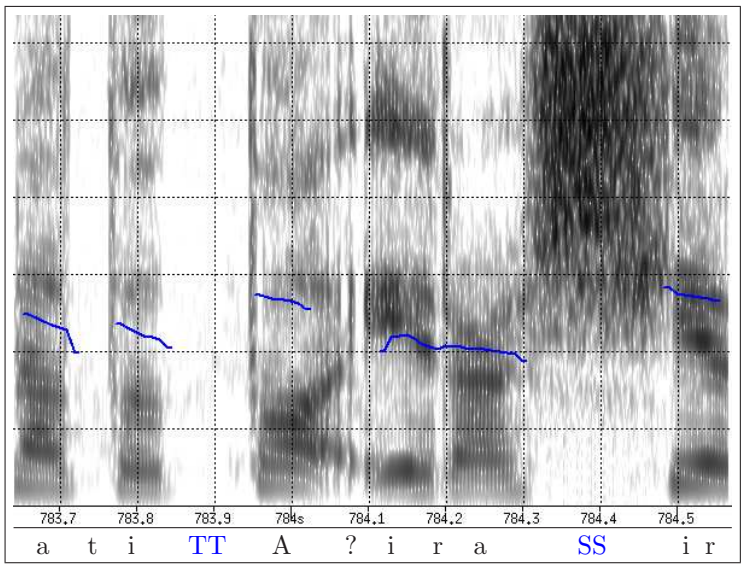
Recognition word error rates with a single pass system (corresponding to the first pass of the evaluation system described below) are given in Table I with the standard phone set including 3 short vowels, and with models trained with only one generic short vowel. Both model sets have 5k tied states (64 Gaussians per state) and covering 5k phone contexts. It can be seen that there is only a slight degradation in performance for both the broadcast news (bnat06) and broadcast conversation (bcat06) data types, when using a generic vowel. Therefore it was decided that the generic vowels provide an effective means to facilitate training on non-vocalized data.

6. MODELING GEMINATES AND TANWIN

The original phone set for Arabic described in Section 3 contained 37 symbols. When pronunciations were produced with this phone set, all consonants with a gemination mark were simply doubled. While this may be a reasonable approximation for some sounds, such as fricatives, if is clearly not well adapted to plosives

Table I. Word error rates on GALE broadcast news (bnat06) and broadcast conversation (bcat06) development data with scoring with small acoustic models, representing 3 short vowels or 1 generic vowel.

	bnat06	bcat06
Standard model	24.4%	35.2%
Generic model	25.7%	35.5%



Solar	
y	50.5%
$	26.9%
S	20.1%
p	19.1%
v	19.0%
Z	18.4%
d	15.3%
s	14.1%
t	9.2%
n	8.9%
T	8.7%
r	8.3%
z	7.4%
l	6.4%
D	4.9%
g	2.6%

Lunar	
k	6.6%
w	4.7%
m	4.2%
q	3.1%
j	2.7%
G	2.5%
b	2.1%
x	1.8%
H	1.2%
c	0.2%
J	1.1%
f	1.7%
h	0.5%
’	0.0%
V	0.0%

Fig. 2. **Left:** Spectrogram illustrating gemination (segments labeled ’TT’ and ’SS’). **Right:** Percentage occurrences of geminates for solar and lunar consonants. The grid is 100ms by 1 kHz. The blue lines represent the estimated pitch (displayed at 10 times its value).

where gemination does not result in multiple bursts.

Figure 2 illustrates a portion of the phrase “(kaAn)ati AlT∼aA}irap Al$∼ir(aAEiyap)”. An aligned approximate phone transcription is shown on the bottom. There are two geminates in this example. The first is the ’T’ (emphatic ’t’) around timee 783.85 and the other a geminate ’S’ (’sh’) is centered at time 784.4. These segments have a duration that is about 50% longer than their non-geminate counterparts.

An additional 30 phone symbols were added to represent the geminate phones. The frequencies of the consonants in single and geminate form were counted in a 100 hour corpus of manually transcribed and vocalized Arabic broadcast news data [Messaoudi et al. 2005]. The right part of Figure 2 lists the solar and lunar consonants, along with the percentage of occurrences as geminates. It can be observed that the solar consonants generally have a higher proportion of geminates than the lunar ones. Figure 3 shows how the geminates are represented in the original pronunciation dictionary (top) and the new dictionary with specific geminate symbols.

ktAb	kitAb=**kitaAb** kitAba=**kitaAba**
	kitAbi=**kitaAbi** kitAbin=**kitaAbK**
	kitAbu=**kitaAbu** kitAbun=**kitaAbN**
	kuttAb=**kut˜aAb** kuttAba=**kut˜aAba**
	kuttAbi=**kut˜aAbi** kuttAbin=**kut˜aAbK**
	kuttAbu=**kut˜aAbu** kuttAbun=**kut˜aAbN**
ktAb	kitAb=**kitaAb** kitAba=**kitaAba**
	kitAbi=**kitaAbi** kitAbin=**kitaAbK**
	kitAbu=**kitaAbu** kitAbun=**kitaAbN**
	ku+Ab=**kut˜aAb** ku+Aba=**kut˜aAba**
	ku+Abi=**kut˜aAbi** ku+Abin=**kut˜aAbK**
	ku+Abu=**kut˜aAbu** ku+Abun=**kut˜aAbN**

Fig. 3. Sample pronunciations for ktb in the original dictionary (top) and with geminate symbols (bottom). Each lexical entry is the non-vocalized word class encompassing all possible vocalized forms. The ˜ signifies gemination in the transliterated form (on the right of the = sign), and the + is the phone symbol for the geminate t (one the left side of the = sign).

Table II. Word error rates without and with explicit modeling of geminates on the GALE 2006 development data sets. bnat06: broadcast news, bcat06: broadcast conversations. Acoustic models were trained on about 1000 hours of Arabic broadcast data.

	bnat06	bcat06
Standard model	22.0%	32.6%
Geminate model	21.7%	32.3%
Combination	21.5%	31.9%

Recognition results are given in Table II on two 3-hour sets of development data used in the GALE community, comparing models trained using the original phone set and the extended one which includes geminates. It can be seen that modeling geminates improves performance for both the broadcast news (bnat06) and broadcast conversation (bcat06) data types, and that a further gain is obtained by combining the two models [Fiscus 1997]. Increasing the phone set also has the added advantage of increasing the number of context-dependent phones that are modeled.

As mentioned above, final short vowels are followed by /n/ for indefinite word forms. These can be realized as a vowel-n sequence or a nasalized vowel. In order to better capture this variability three additional phones were added to the phone set to represent the three tanwin phones (in, an, un) with a single unit. Acoustic models were built using this new phone set, and tested on the development data sets. These models obtained word error rates comparable to that of the non-tanwin models, and when used in system combination [Fiscus 1997] gave a gain of 0.4% absolute over either model set alone. Given the large variability in the realization of tanwin, these results are not surprising.

7. PRONUNCIATION VARIANTS FOR DIALECTAL SPEECH

An analysis of the errors made by the STT system showed that many of the errors involve the insertion or deletion of a prefix or a suffix, such as the confusion of ktAb and wktAb or ktAbh and ktAb. The article 'Al' is found in 37% of the prefix

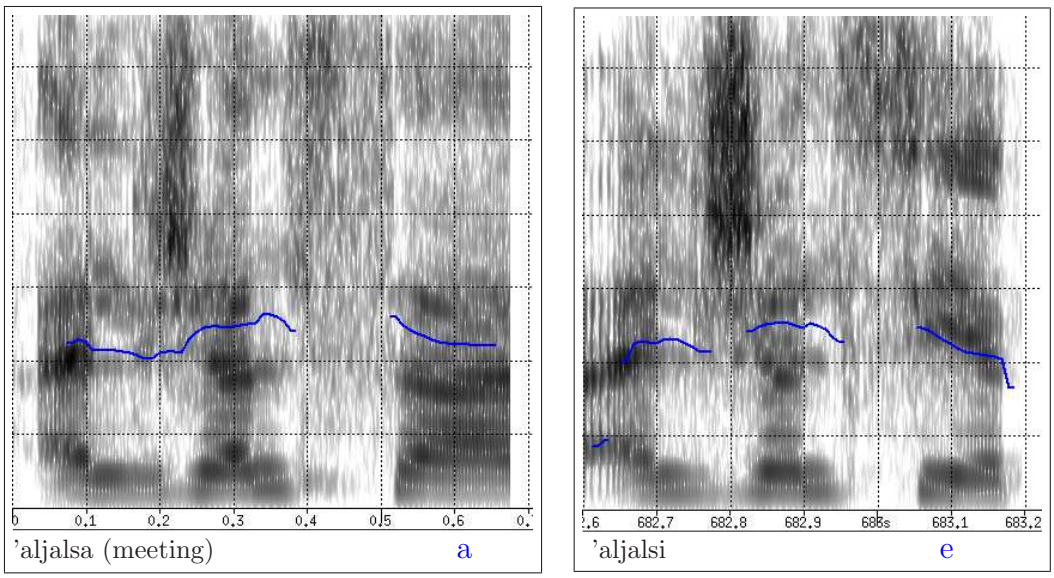


Fig. 4. Two example spectrograms of the word 'aljalsa (meeting), the rightmost illustrating the Lebanese realization of the final vowel. The grid is 100ms by 1 kHz. The blue lines represent the estimated pitch (displayed at 10 times its value).

errors, and contributes an absolute error of 1%. In examining the errors a number of dialectal pronunciation variants were observed, that were not represented in the lexicon. Figure 4 shows two spectrograms of the word 'aljalsa (meeting). The final short vowel in the example on the left is an /a/. The right example is the same word, but the final vowel is not produced in the same manner. Arabic speakers consider this to be an /i/, whereas it appears more like an /e/ in the spectrogram.

Systematically adding pronunciation variants to the lexicon resulted in an absolute WER reduction of 0.3% on broadcast news data and 0.6% on broadcast conversation data which contains more dialectal speech.

8. MORPHOLOGICAL DECOMPOSITION

As for other morphologically-rich languages such as Estonian, Finnish, German, Korean and Turkish [Carki et al. 2000; Whittaker and Woodland 2000; Adda-Decker 2003], one of challenges of Arabic speech recognition is to deal with the huge lexical variety. For Arabic the combination of compounding, agglutination and inflection generate a large number of surface forms for a given root form. Morphological decomposition [Kirchhoff and et al. 2002; Vergyri et al. 2004; Xiang et al. 2006] has been proposed to deal with this characteristic, resulting in increased lexical coverage, thereby reducing errors that are due to words that are unknown to the system.

Generally speaking, a recognition vocabulary is simply a list of words as found in texts of the language. This view is a bit simplistic as it assumes that the texts have already been normalized, which in turn entails a variety of more or less important decisions [Adda et al. 1997; Adda-Decker and Lamel 2000]. For morpho-

logically rich languages there has been growing interest in using sub-word units to reduce the needed vocabulary size for a given lexical coverage. There are two main approaches to morphological decomposition, those based on the use of explicit linguistic knowledge and rules (for example, [Schmid 1994; Vergyri et al. 2004; Xiang et al. 2006]), and unsupervised methods (for example, [Harris 1955; Goldsmith 2001; Adda-Decker 2003; Creutz and Lagus 2005]). Since the Arabic language has a relatively limited number of affixes, and rules can capture the manner in which they are applied, in this work the rules as implemented in the Buckwalter morphological analyzer are used [Buckwalter 2004; Ghaoui et al. 2005].

In the next section the variant methods for morphological decomposition are described, followed by a description of the audio and text training corpora used in the recognition experiments.

8.1 Methodology

Three variant methods for morphological decomposition were investigated. For all three the basis for decomposition is derived from the results of the Buckwalter morphological analysis [Buckwalter 2004]. In Buckwalter, the following affixes are decomposed (the Buckwalter transliteration codes are used here):

- 12 prefixes with 'Al': Al wAl fAl bAl wbAl fbAl ll wll fll kAl wkAl fkAl
- 11 prefixes without 'Al': w f b wb fb l wl fl k wk fk
- 6 negation prefixes: mA wmA fmA lA wlA flA
- 3 prefixes future tense: s ws fs
- suffixes (possessive pronouns): y, ny, nA, h, hm, hmA, hn, k, kmA, km, kn

In total there are 32 prefixes, 6 for negation, 3 for the future formed and 12 formed with the definite article, and 11 others without 'Al'. The suffixes in Arabic are personal pronouns, the objective form serves as a direct object of a verb, and as the possessive form serves as the complement of a noun.

In the first variant, a set of decomposition rules were applied to all words in the training texts that were identifiable by the Buckwalter morphological analyzer. Of the 1137k distinct words in the training texts, 880K can be decomposed with the rules. About half of the remaining words are simple words, and the remainder have several possible decompositions (29%) or have a root that is not in the recognition dictionary (12%). Decomposition of a 200K lexicon results in a lexicon with 79K entries and reduces the out-of-vocabulary rate from 4.4% to 2%. If the decomposition rules are applied to the entire 1.1 M words, it is reduced to 270k forms (stems, affixes and decomposed words). During decomposition, each affix that is split from the word root is marked by adding a "+" (to the end of prefixes and the start of suffices) to signify that it should be recomposed with the following or preceding word in the recognizer hypothesis.

In the first version (v1), the decomposition was applied to a list of 1.1 M words that were recognized by Buckwalter. Of these 880k were decomposed, and 256k remained unchanged. After decomposition, the word list was reduced to 270k forms (stems, affixes and decomposed words). Following what has been done by others, in the second version (v2), the most frequent 65k words were never decomposed. This had the effect of blocking the decomposition of 35k words, which when added to the word list increased its size to 300k words.

In the third version (v3), on top of v2, the prefix 'Al' is not decomposed if the

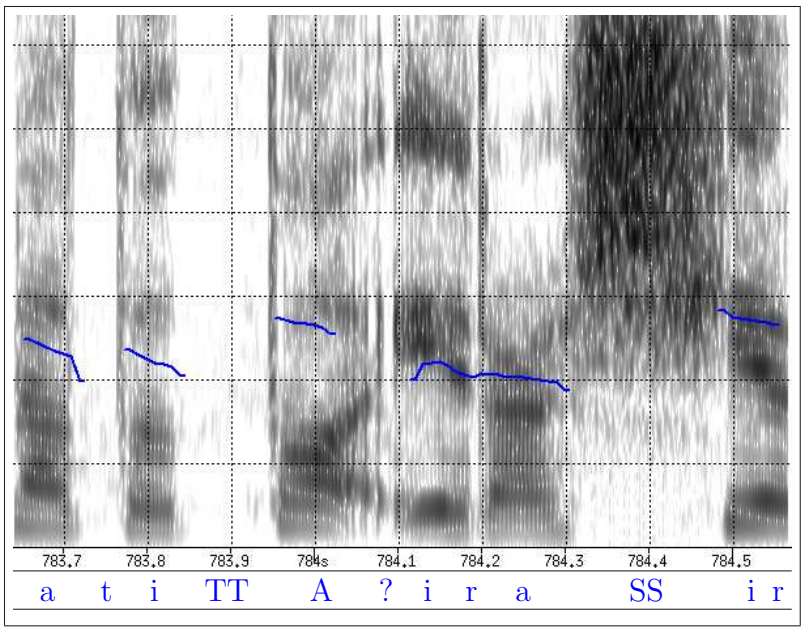


Fig. 5. Example of assimilation of 'Al' preceding a solar consonant. The segment corresponds to the sequence (kaAn)ati AlT∼aA}irap Al$∼ir(aAEiyap) in Buckwalter code and is taken from the segment LBC_NEWS_ARB_20060601_195801-0783.64-784.57. The grid is 100ms by 1 kHz. The blue lines represent the estimated pitch (displayed at 10 times its value).

word begins with a solar consonant (the solar consonants in the Buckwalter code are: t, v, d, g, r, z, s, $, S, D, T, Z, l, n.). The reason to forbid the decomposition of 'Al' preceding words starting with a solar consonant is because the 'l' is assimilated with the following consonant and it is difficult to isolate a portion of the signal that clearly corresponds to the 'Al'. This problem is illustrated by the spectrogram in Figure 5 which is the same excerpt (kaAn)ati AlT∼aA}irap Al$∼ir(aAEiyap) in Buckwalter code shown in Figure 2. The letters in parenthesis at the start and end provide the context. For the portion of interest ati AlT∼aA}irap Al$∼ir the first i was underlyingly a Sukoun (a mark which inhibits the pronunciation of a vowel). However, preceding the Al it is realized as an i (which is reduced to more or less a schwa) and the Al causes the following consonant to be realized as a geminate TT. This example shows a second gemination SS corresponding to the second Al. These type of phenomena are extremely difficult to model when the Al is allowed to be decomposed from the word, and explains why the Al was involved in so many of the errors in the first version. This restriction blocks decomposition of the prefix 'Al' preceding a solar letter if it is a simple prefix. If the prefix 'Al' is preceded by other prefixes, the other prefixes are split off and the 'Al' is kept with the stem.

For example, the original decomposition rules split the word wbAlslAm which has 3 prefixes w+b+Al+slAm into wbAl+ slAm, whereas the version 3 decomposition gives wb+ AlslAm.

Some of the words that were not able to be analyzed with Buckwalter were found to be in dialectal Arabic. Adding seven dialect prefixes to the Buckwalter prefix table allows over 85% of these words to be decomposed.

In addition to processing the training texts and constructing language models, in order to build a complete system using the morphological decomposition, the affixes needed to be added to the pronunciation dictionary, and acoustic models trained with the decomposed lexical units. The pronunciation lexicon was extended to include all possible pronunciations of the affixes. One particular problem is handling the article 'Al' when it is followed by a solar consonant, since in this case the 'l' assimilated with the consonant. This phenomenon is taken into account within words by assigning a gemination mark to the consonant. To represent this in the decomposed prefix 'Al', contextual pronunciations are included for all solar consonants. For the acoustic models, the decomposition rules were applied to the transcripts of the audio training data, and several iterations segmentation and model estimation were carried out.

8.2 Experimental Results

The training and test data are all from the Gale program, and distributed by LDC [5]. The audio training data used in this work are comprised of 1200 hours of manually transcribed broadcast data (1200h train). Roughly 60% of the data are classed as broadcast news (BN), that is typically well-prepared speech from announcers and reporters in speaking Modern Standard Arabic, and 40% is classified as broadcast conversation (BC), which tends to be more casual in style and has a higher proportion of dialectal Arabic. Results are reported on the Gale development and evaluation data sets from 2006 and 2007 (bnat06, bnad06, bcat06, bcad06, eval06, dev07, eval07), each set containing 2 to 3 hours of audio data.

The texts used for language model training are obtained from written sources and transcriptions of audio data. The written texts comprise more that 1.1 billion words from a variety of news sources, predominantly newspapers and news wires in Arabic. The transcriptions of audio data contain over 11 M words: 6.3 M words from BN and 4.8 M words of BC, and an additional 3.8 M words of Web transcripts of Aljazeera BC data.

The baseline recognition lexicon has 200k non-vocalized entries, each of which is associated with multiple vocalized forms, which in turn are associated with one or more phone pronunciations [Messaoudi et al. 2006]. The pronunciations make use of 71 symbols, including 31 simple consonants, 30 geminate consonants, 3 long and 3 short vowels, plus 3 pseudo phones for non-linguistic events (breath, filler, silence). There are on average 8.6 pronunciations/word.

Recognition results of a single decoding pass with unsupervised acoustic model adaptation are given in Table III. The acoustic models were trained on about 1200 hours of manually transcribed speech data distributed by LDC. The three versions of decomposition were applied to the training transcripts, and three sets of word-position dependent acoustic models were estimated, specific to each versions. The WER of the reference word based system with MLE training was 20.9%. With the first decomposition method that simply splits all affixes, the WER is increased

[5] `http://projects.ldc.upenn.edu/gale/index.html`

Table III. WER on the Gale bnat06 data set with the reference 200k word based system, the system with three morphological decomposition versions for a single decoding pass with acoustic model adaptation (1200h train).

Condition	*WER*	*Vocabulary size*
Reference word based	20.9	200k
Decomposition v1	23.7	270k, baseline
Decomposition v2	21.8	300k
Decomposition v3	20.5	320k

by 2% absolute. By forbidding the decomposition of the most frequent 65k words (v2) most of these errors are avoided. Applying the 3rd version of decomposition rules prevents the decomposition of the prefix 'Al' preceding solar 11k solar words. After applying these to the full 1.1 M word list, the recognition vocabulary contains 320k entries (stems, affixes and decomposed words). The WER is reduced by 1.3% compared to the v2 decomposition, but there is only a small gain relative to the word based system.

The above sections have introduced ideas, that were validated during different stages of system development. A complete system was developed based on the prior results, and on complementary work on using multi-layer perceptrons to provide discriminative acoustic feature extraction [Zhu et al. 2005; Stolcke et al. 2006; Fousek 2007; Grézl and Fousek 2008; Fousek et al. 2008a; 2008b] and neural net language models to cope with the data sparseness problem in estimating n-gram probabilities [Schwenk and Gauvain 2005; Schwenk 2007]. Standard techniques used in state-of-the-art speech transcriptions such as speaker adaptive training (SAT) [Anastasakos et al. 1996] and Maximum Mutual Information (MMI) training, Constrained Maximum Likelihood Linear Regression (CMLLR) and MLLR [Leggetter and Woodland 1995] adaptation are all used.

Table IV reports results using MMI trained acoustic models (on the 1200 hours of manually transcribed data), developed for the word-based system, that is the training transcriptions use a word representation. Results are given for all Gale development sets with neural net language models that has been estimated on the texts that have been morphologically decomposed and for the baseline word based NN LM [Schwenk 2007]. Comparing the first two entries, it can be seen that the baseline word-based and morphologically decomposed language models give quite comparable results. The results obtained by combining the two models using Rover [Fiscus 1997] are given in the third row of this table. Compared to the baseline system the average word error reduction across all test sets is about 0.6%. The final entry in the table shows the results of a 4-way Rover obtained using the 290k word based LM and the 290k LM with morphological decomposition each with two acoustic model sets, one using standard cepstral features and the other MLP based features [Fousek et al. 2008a; 2008b]. The word error rate is reduced on all test sets by over 1% compared to the 2-way combination.

9. CONCLUSIONS

This paper has described the incremental improvements to a system for the automatic transcription of broadcast data in Arabic, highlighting techniques developed

Table IV. Word error rates for the 290k word based LM (baseline) and the 290k LM with morphological decomposition for different data sets. All conditions use MMIE trained acoustic models and a NN language model.

Conditions	*bnat06*	*bnad06*	*bcat06*	*bcad06*	*eval06*	*dev07*	*eval07*
Baseline	16.7	15.5	22.8	20.4	19.3	12.4	13.7
Decomposition	16.7	15.3	23.1	20.6	19.4	12.2	13.8
Combination	16.1	14.9	22.3	19.7	18.5	11.8	13.2
4-way Rover	14.5	13.2	20.2	17.9	17.1	10.6	11.9

to deal with specificities of the Arabic language. One of the challenges is training with incomplete information since most Arabic texts are written without diacritics, yet the diacritics provide useful information for pronunciation modeling and higher level processing. After initial studies which focused on Modern Standard Arabic broadcast news data using a completely vocalized representation, different methods were explored to reduce the reliance on vocalized data and to handled more varied data. Many vocalized word forms can be derived using the Buckwalter morphological analyzer and modifications thereof. However it is necessary to also be able to generate pronunciations for words that Buckwalter is not able to process. Rules to generate pronunciations with a generic vowel have been proposed, and this method has been used to significantly facilitate training on non-vocalized data. Concerning pronunciation modeling, explicit rules were developed to handle frequent dialectal variants, as well as systematic variations in the language. The explicit modeling of gemination and the introduction of pronunciation variants led to significant improvements in speech-to-text transcription performance.

ACKNOWLEDGMENTS

This work was in part supported under the GALE program of the Defense Advanced Research Projects Agency, (Contract No. HR0011-06-C-0022) and by OSEO under the Quaero program. The authors gratefully acknowledge the participation of Petr Fousek and Holger Schwenk to some parts of this work.

REFERENCES

Adda, G., Adda-Decker, M., Gauvain, J.-L., and Lamel, L. 1997. Text normalization and speech recognition in French. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*. Vol. 5. ESCA, Rhodes, 2711–2714.

Adda-Decker, M. 2003. A corpus-based decompounding algorithm for German lexical modeling in LVCSR. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*. ESCA, Geneva, 257–260.

Adda-Decker, M. and Lamel, L. 2000. *The use of lexica in automatic speech recognition*. F. Van Eynde and D. Gibbon (eds.), Kluwer Academic Publushers, Holland, 235–266. Based on Course notes for ELSNET's 5TH European Summer School on Language and Speech Communication: Leuven, July 1997.

Afify, M., Nguyen, L., Xiang, B., Abdou, S., and Makhoul, J. 2005. Recent Progress in Arabic Broadcast News Transcription at BBN. In *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*. ESCA, Lisbon, 1637–1640.

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. 1996. A Compact Model for Speaker Adaptation Training. In *International Conference on Speech and Language Processing*. Philadelphia, 1137–1140.

Billa, J., Noamany, N., Srivastava, A., Liu, D., Stone, R., Xu, J., Makhoul, J., and Kubala, F. 2002. Audio Indexing of Arabic Broadcast News. In *Proceedings of ICASSP*. Vol. 1. Orlando, 5–8.

Buckwalter, T. 2004. Arabic Morphology Analysis.

Carki, K., Geutner, P., and Schultz, T. 2000. Turkish LVCSR: towards better speech recognition for agglutinative languages. In *Proceedings of ICASSP*. Istanbul, 3688–3691.

Chou, W. and Juang, F., Eds. 2003. *Pattern Recognition in Speech and Language Processing*. CRC Press.

Creutz, M. and Lagus, K. 2005. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0. Tech. rep., Helsinki University of Technology. March. Computer and Information Science Technical Report, A81.

Fiscus, J. 1997. A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER) . In *Proceeding of IEEE Workshop on Automatic Speech Recognition*. Santa Barbara, 347–352.

Fousek, P. 2007. Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics. Ph.D. thesis, Czech Technical University, Faculty Electrical Engineering, Prague, Czech Republic.

Fousek, P., Lamel, L., and Gauvain, J.-L. 2008a. On the Use of MLP Features for Broadcast News Transcription. In *11th International Conference on Text, Speech and Dialogue (TSD08)*. Number 5246/2008 in Lecture Notes in Computer Science. Springer Verlag, Berlin/Heidelberg, 303–310.

Fousek, P., Lamel, L., and Gauvain, J.-L. September 22-26, 2008b. Transcribing Broadcast Data Using MLP Features. In *Interspeech, Annual Conference of the International Speech Communication Association*. Brisbane, 1433–1436.

Gauvain, J.-L. and Lamel, L. 2000. Large Vocabulary Continuous Speech Recognition: Advances and Applications. *Proceedings of the IEEE 88,* 8 (Aug), 1181–1200.

Gauvain, J.-L., Lamel, L., and Adda, G. 1998. Partitioning and transcription of broadcast news data. In *International Conference on Speech and Language Processing*. Vol. 4. Sydney, 1335–1338.

Gauvain, J.-L., Lamel, L., and Adda, G. 2002. The LIMSI Broadcast News Transcription System. *Speech Communication 37,* 1-2 (May), 89–108.

Ghaoui, A., Yvon, F., Mokbel, C., and Chollet, G. 2005. On the use of morphological constraints in n-gram statistical language model. In *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*. Lisbon, 1281–1284.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics 27,* 2, 153–198.

Graff, D. 2007. Arabic Gigaword Third Edition (LDC LDC2007T40).

Grézl, F. and Fousek, P. 2008. Optimizing bottle-neck features for LVCSR. In *Proceedings of ICASSP*. Las Vegas, 4729–4732.

Harris, Z. 1955. From Phoneme to Morpheme. *Language 31*, 190–222.

Kirchhoff, K. and et al. 2002. Novel approaches to Arabic speech recognition. Tech. rep., John-Hopkins University, Baltimore. Final report from the 2002 JHU summer workshop.

Lamel, L. and Gauvain, J.-L. 2003. Speech recognition. In *OUP Handbook on Computational Linguistics*, R. Mitkov, Ed. Oxford University Press, Chapter 16, 305–322.

Lamel, L. and Gauvain, J.-L. 2008. Speech processing for audio indexing. In *Proceedings of the 6th International Conference on Natural Language Processing, GoTAL 2008 - Advances in Natural Language Processing*. Number 5221/2008 in Lecture Notes in Computer Science. Springer Verlag, Berlin/Heidelberg, 4–15.

Leggetter, C. and Woodland, P. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language 9,* 2, 171–185.

Mangu, L., Brill, E., and Stolcke, A. 1999. Finding Consensus Among Words: Lattice-Based Word Error Minimization. In *Proceedings of the European Conference on Speech Technology, EuroSpeech*. Budapest, 495–498.

Messaoudi, A., Gauvain, J.-L., and Lamel, L. 2006. Arabic Broadcast News Transcription using a One Million Word Vocalized Vocabulary. In *Proceedings of ICASSP*. Vol. I. Toulouse, 1093–1096.

Messaoudi, A., Lamel, L., and Gauvain, J.-L. 2004. Transcription of Arabic Broadcast News. In *International Conference on Speech and Language Processing*. Jeju Island, 521–524.

Messaoudi, A., Lamel, L., and Gauvain, J.-L. 2005. Modeling Vowels for Arabic BN Transcription. In *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*. ESCA, Lisbon, 1633–1636.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Conference on New Methods in Language Processing*. Manchester.

Schwenk, H. 2007. Continuous space language models. *Computer Speech and Language 21*, 492–518.

Schwenk, H. and Gauvain, J.-L. 2005. Training Neural Network Language Models On Very Large Corpora. In *Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Vancouver, 201–208.

Stolcke, A., Chen, B., Franco, H., Gadde, V., Graciarena, M., Hwang, M.-Y., Kirchhoff, K., Mandal, A., Morgan, N., Lei, X., Ng, T., Ostendorf, M., Sonmez, K., Venkataraman, A., D. Vergyri, W. W., Zheng, J., and Zhu, Q. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech and Language Processing 14,* 5 (Sep), 1729–1744.

Vergyri, D., Kirchhoff, K., Duh, K., and Stolcke, A. 2004. Morphology-based language modeling for Arabic speech recognition. In *International Conference on Speech and Language Processing*. Jeju Island, 1252–1255.

Whittaker, E. and Woodland, P. 2000. Particle-based language modelling. In *International Conference on Speech and Language Processing*. Vol. 1. Beijing, China.

Xiang, B., Nguyen, K., Nguyen, L., Schwartz, R., and Makhoul, J. 2006. Morphological Decomposition for Arabic Broadcast News Transcription. In *Proceedings of ICASSP*. Vol. I. Toulouse, 1089–1092.

Young, S. and Bloothooft, G., Eds. 2000. *Corpus-Based Methods in Language and Speech Processing*. Kluwer Academic Publushers, Holland. Based on Course notes for ELSNET's 2nd European Summer School on Corpus-based Methods in Language and Speech Processing: Utrecht, July 1994.

Zhu, Q., A.Stolcke, Chen, B.-Y., and Morgan, N. 2005. Using MLP features in SRI's conversational speech recognition system. In *Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology*. Lisbon, 2141–2144.

Submitted: April 09 ; Accepted: July 09