Chapter 12

Cross-Word Arabic Pronunciation Variation Modeling Using Part of Speech Tagging

Dia AbuZeina, Husni Al-Muhtaseb and Moustafa Elshafei

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/48645

1. Introduction

Speech recognition is often used as the front-end for many natural language processing (NLP) applications. Some of these applications include machine translation, information retrieval and extraction, voice dialing, call routing, speech synthesis/recognition, data entry, dictation, control, etc. Thus, much research work has been done to improve the speech recognition and the related NLP applications. However, speech recognition has some obstacles that should be considered. Pronunciation variations and small words misrecognition are two major problems that lead to performance reduction. Pronunciation variations problem can be divided into two parts: within-word variations and cross-word variations. These two types of pronunciation variations have been tackled by many researchers using different approaches. For example, cross-word problem can be solved using phonological rules and/or small-word merging. (AbuZeina et al., 2011a) used the phonological rules to model cross-word variations for Arabic. For English, (Saon & Padmanabhan, 2001) demonstrated that short words are more frequently misrecognized, they also had achieved a statistically significant enhancement using small-word merging approach.

An automatic speech recognition (ASR) system uses a decoder to perform the actual recognition task. The decoder finds the most likely words sequence for the given utterance using Viterbi algorithm. The ASR decoder task might be seen as an alignment process between the observed phonemes and the reference phonemes (dictionary phonemic transcription). Intuitively, to have a better accuracy in any alignment process, long sequences are highly favorable instead of short ones. As such, we expect enhancement if we merge words (short or long). Hence fore, a thorough investigation was performed on Arabic speech to discover a suitable merging cases. We found that Arabic speakers usually augment two consecutive words; a noun that is followed by an adjective and a preposition that is followed by a word. Even though we believe that other cases are found in Arabic speech, we chose two cases to validate our proposed method. Among the ASR components, the pronunciation dictionary and the language model were used to model our above mentioned objective. This means that the acoustic models for the baseline and the enhanced method are the same.

This research work is conducted for Modern Standard Arabic (MSA). So, the work will necessarily contain many examples in Arabic. Therefore, it would be appropriate for the reader if we start first by providing a Romanization (Ryding, 2005) of the Arabic letters and diacritical marks. Table 1 shows the Arabic–Roman letters mapping table. The diacritics Fatha, Damma, and Kasra are represented using a, u, and i, respectively.

Arabic	Roman	Arabic	Roman	Arabic	Roman	Arabic	Roman
ء (hamza)	’	د (daal)	d	ض (Daad)	D	ك (kaaf)	k
ب (baa’)	b	ذ (dhaal)	dh	ط (Taa’)	T	ل (laam)	l
ت (taa’)	t	ر (raa’)	r	ظ (Zaa’)	Z	م (miim)	m
ث (thaa’)	th	ز (zaay)	z	ع (‘ayn)	‘	ن (nuun)	n
ج (jiim)	j	س (siin)	s	غ (ghayn)	gh	ه (haa’)	h
ح (Haa’)	H	ش (shiin)	sh	ف (faa’)	f	و (waaw)	w or u
خ (khaa’)	kh	ص (Saad)	S	ق (qaaf)	q	ي (yaa’)	y or ii

Table 1. Arabic–Roman letters mapping table

To validate the proposed method, we used Carnegie Mellon University (CMU) Sphinx speech recognition engine. Our baseline system contains a pronunciation dictionary of 14,234 words from a 5.4 hours pronunciation corpus of MSA broadcast news. For tagging, we used the Arabic module of Stanford tagger. Our results show that part of speech (PoS) tagging is considered a promising track to enhance Arabic speech recognition systems.

The rest of this chapter is organized as follows. Section 2 presents the problem statement. Section 3 demonstrates the speech recognition components. In Section 4, we differentiate between within-word and cross-word pronunciation variations followed by the Arabic speech recognition in Section 5. The proposed method is presented in Section 6 and the results in Section 7. The discussion is provided in Section 8. In Section 9, we highlight some of the future directions. We conclude the work in Section 10.

2. Problem statement

Continuous speech is characterized by augmenting adjacent words, which do not occur in isolated speech. Therefore, handling this phenomenon is a major requirement in continuous speech recognition systems. Even though Hidden Markov Models (HMMs) based ASR decoder uses triphones to alleviate the negative effects of cross-word phenomenon, more effort is still needed to model some cross-word cases that could not be avoided using triphones. In continuous ASR systems, the dictionary is usually initiated using corpus transcription words, i.e. each word is considered as an independent entity. In this case, speech cross-word merging will reduce the performance. Two main methods are usually used to model the cross-word problem, phonological rules and small-word merging. Even though the phonological rules and small-word merging methods enhance the performance, we believe that generating compound words is also possible using PoS tagging.

Initially, there are two reasons why cross-word modeling is an effective method in speech recognition system: First, the speech recognition problem appears as an alignment process, hence for, having long sequences is better than short ones as demonstrated by (Saon and Padmanabhan, 2001). To illustrate the effect of co-articulation phenomenon (merging of words in continuous speech), let us examine Figure 1 and Figure 2. Figure 1 shows the words to be considered with no compound words, while Figure 2 shows the words with compound words. In both figures we represented the hypotheses words using bold black lines. During decoding, the ASR decoder will investigate many words and hypotheses. Intuitively, the ASR decoder will choose the long words instead of two short words. The difference between the two figures is the total number of words that will be considered during the decoding process. Figure 2 shows that the total number of words for the hypotheses is less than the total words in Figure 1 (Figure 1 contains 34 words while Figure 2 contains 18 words). Having less number of total words during decoding process means having less decoding options (i.e. less ambiguity), which is expected to enhance the performance.

Second, compounding words will lead to more robust language model. the compound words which are represented in the language model will provide better representations of words relations. Therefore, enhancement is expected as correct choice of a word will increase the probability of choosing a correct neighbor words. The effect of compounding words was investigated by (Saon & Padmanabhan, 2001). They mathematically demonstrated that compound words enhance the language model performance, therefore, enhancing the overall recognition output. They showed that the compound words have the effect of incorporating a trigram dependency in a bigram language model. In general, the compound words are most likely to be correctly recognized more than two separated words. Consequently, correct recognition of a word might lead to another correct word through the enhanced N-grams language model. In contrast, misrecognition of a word may lead to another misrecognition in the adjacent words and so on.

For more clarification, we present some cases to show the short word misrecognition, and how is the long word is much likely to be recognized correctly. Table 2 shows three speech files that were tested in the baseline and the enhanced system. Of course, it is early to show some results, but we see that it is worthy to support our motivation claim. In Table 2, it is clear that the misrecognitions were mainly occurred in the short words (the highlighted short words were misrecognized in the baseline system).

In this chapter, the most noticeable Arabic ASRs performance reduction factor, the cross-word pronunciation variations, is investigated. To enhance speech recognition accuracy, a knowledge-based technique was utilized to model the cross-word pronunciation variation at two ASR components: the pronunciation dictionary and the language model. The proposed knowledge-based approach method utilizes the PoS tagging to compound consecutive words according to their tags. We investigated two pronunciation cases, a noun that is followed by an adjective, and a preposition that is followed by a word. the proposed method showed a significant enhancement.

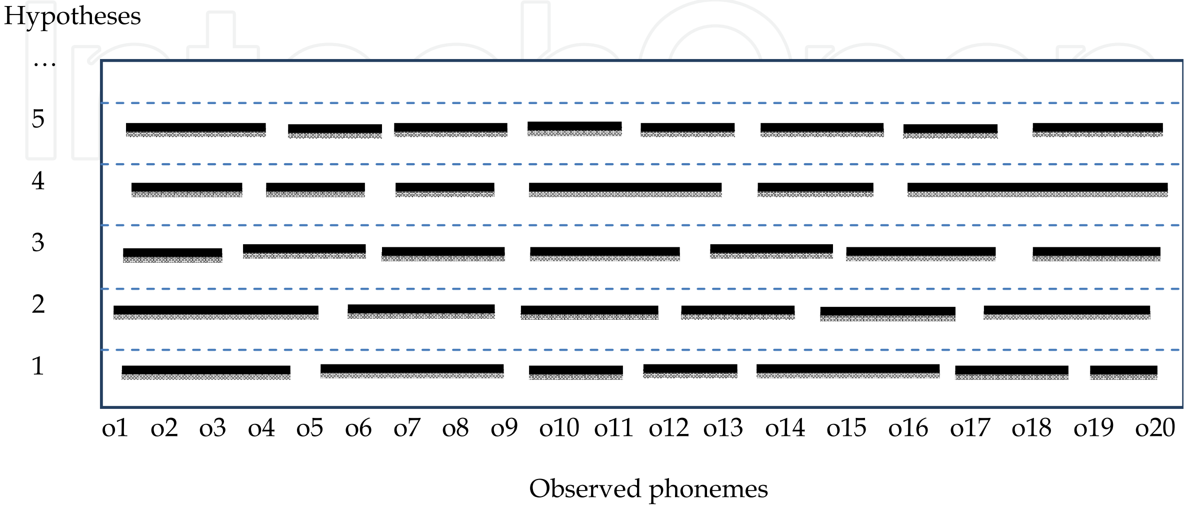

Hypotheses
…
5
4
3
2
1
o1 o2 o3 o4 o5 o6 o7 o8 o9 o10 o11 o12 o13 o14 o15 o16 o17 o18 o19 o20
Observed phonemes

Figure 1. A list of hypotheses without compounding words

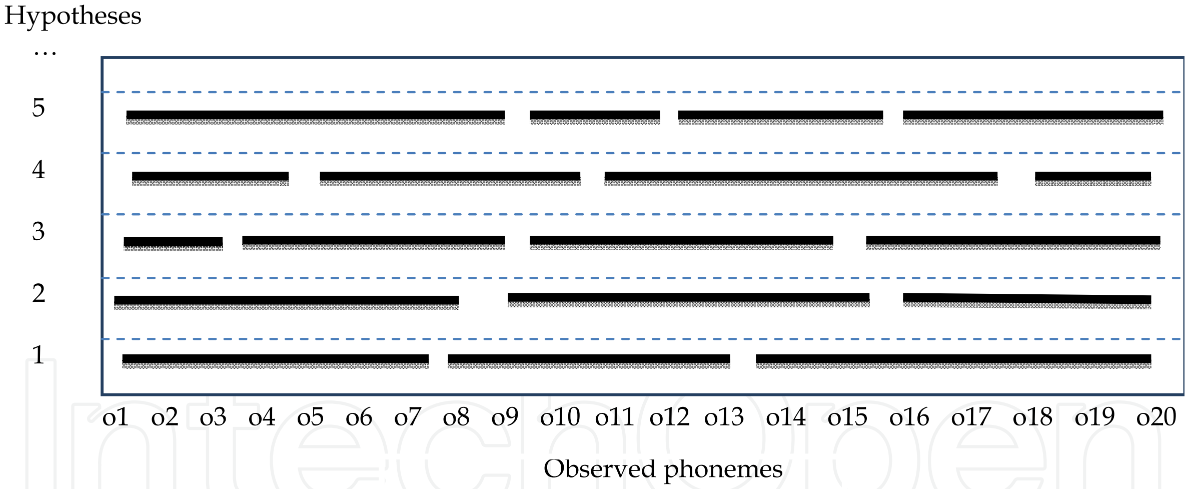

Hypotheses
…
5
4
3
2
1
o1 o2 o3 o4 o5 o6 o7 o8 o9 o10 o11 o12 o13 o14 o15 o16 o17 o18 o19 o20
Observed phonemes

Figure 2. A list of hypotheses with compounding words

3. Speech recognition

Modern large vocabulary, speaker-independent, continuous speech recognition systems have three knowledge sources, also called linguistic databases: acoustic models, language model, and pronunciation dictionary (also called lexicon). Acoustic models are the HMMs of the phonemes and triphones (Hwang, 1993). The language model is the module that provides the statistical representations of the words sequences based on the transcription of the text corpus. The dictionary is the module that serves as an intermediary between the acoustic model and the language model. The dictionary contains the words available in the language and the pronunciation of each word in terms of the phonemes available in the acoustic models.

Figure 3 illustrates the sub-systems that are usually found in a typical ASR system. In addition to the knowledge sources, an ASR system contains a Front-End module which is used to convert the input sound into feature vectors to be usable by the rest of the system. Speech recognition systems usually use feature vectors that are based on Mel Frequency Cepstral Coefficients (MFCCs), (Rabiner and Juang, 2004).

The speech files to be tested	سَيَتَقَابَلانِ وَجهًا لِوَجه فِي المُبَارَاةِ النِّهَائِيَّة sayataqabalani wajhan liwajh fy ’lmubarah ’lniha’iya وَمُمَثِّلِينَ عَن عَدَدٍ مِن الدُّوَلِ الأُورُوبِيَّة wamumathilyna ‘an ‘adadin mina ’lduwali ’l’wrubiya
The baseline system results	سَيَتَقَابَلانِ وَجهًا لِوَجه المُبَارَاةِ النِّهَائِيَّة sayataqabalani wajhan liwajh ’lmubarah ’lniha’iya وَمُمَثِّلِين عَن إِنَّ الدُّوَلِ الأُورُوبِيَّة wamumathilyna ‘an ’inna ’lduwali ’l’wrubiya
The enhance system results	سَيَتَقَابَلانِ وَجهًا لِوَجه فِي المُبَارَاةِ النِّهَائِيَّة sayataqabalani wajhan liwajh fy ’lmubarah ’lniha’iya وَمُمَثِّلِينَ عَن عَدَدٍ مِن الدُّوَلِ الأُورُوبِيَّة wamumathilyna ‘an ‘adadin mina ’lduwali ’l’wrubiya

Table 2. Illustrative cross-word misrecognition results

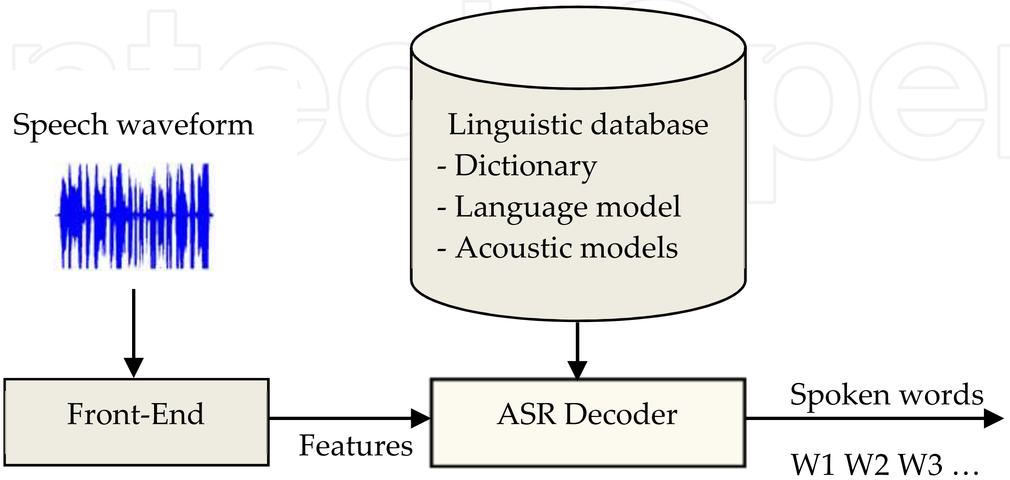


Figure 3. An ASR architecture

The following is a brief introduction to typical ASR system components. The reader can find more elaborate discussion in (Jurafsky and Martin, 2009).

3.1. Front-end

The purpose of this sub-system is to extract speech features which play a crucial role in speech recognition performance. Speech features includes Linear Predictive Cepstral Coefficients (LPCC), MFCCs and Perceptual Linear Predictive (PLP) coefficients. The Sphinx engine used in this work is based on MFCCs.

The feature extraction stage aims to produce the spectral properties (features vectors) of speech signals. The feature vector consists of 39 coefficients. A speech signal is divided into overlapping short segments that are represented using MFCCs. Figure 4 shows the steps to extract the MFCCs of a speech signal (Rabiner & Juang, 2004). These steps are summarized below.

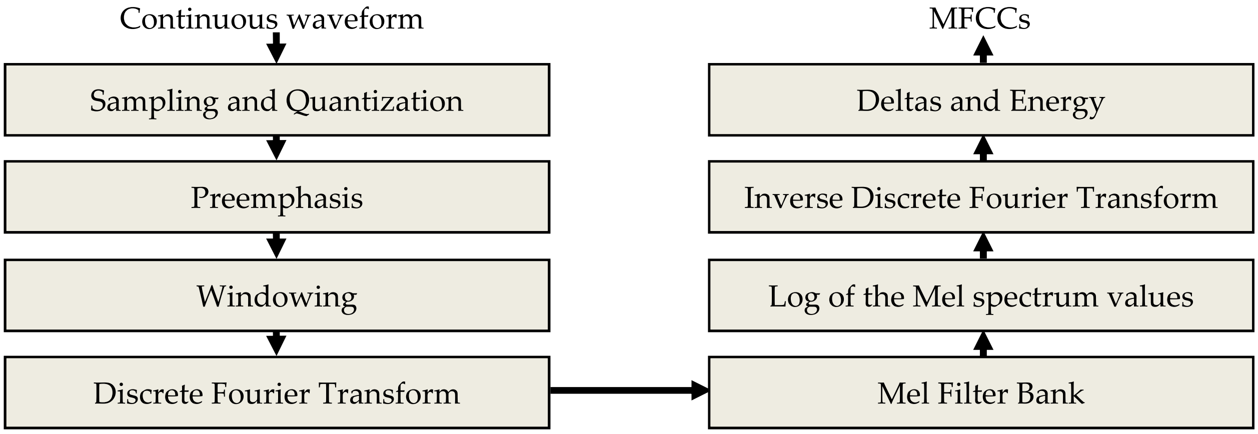

Continuous waveform
Sampling and Quantization
Preemphasis
Windowing
Discrete Fourier Transform
Mel Filter Bank
Log of the Mel spectrum values
Inverse Discrete Fourier Transform
Deltas and Energy
MFCCs

Figure 4. Feature vectors extraction

Sampling and Quantization: Sampling and quantization are the two steps for analog-to-digital conversion. The sampling rate is the number of samples taken per second, the sampling rate used in this study is 16 k samples per seconds. The quantization is the process of representing real-valued numbers as integers. The analysis window is about 25.6 msec (410 samples), and consecutive frames overlap by 10 msec.

Preemphasis: This stage is to boost the high frequency part that was suppressed during the sound production mechanism, so making the information more available to the acoustic model.

Windowing: Each analysis window is multiplied by a Hamming window.

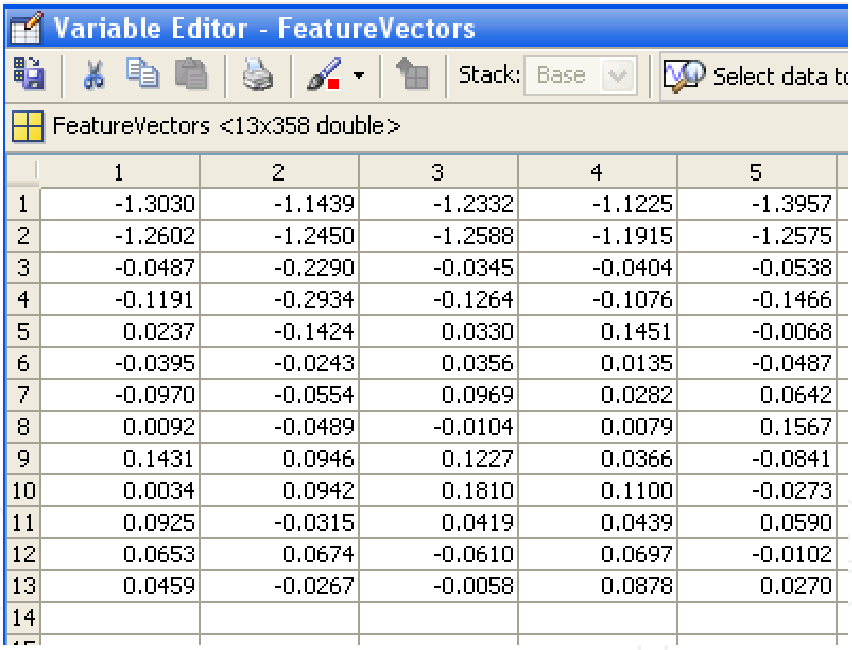
Discrete Fourier Transform: The goal of this step is to obtain the magnitude frequency response of each frame. The output is a complex number representing the magnitude and phase of the frequency component in the original signal.

*Mel Filter Bank***:** A set of triangular filter banks is used to approximate the frequency resolution of the human ear. The Mel frequency scale is linear up to 1000 Hz and logarithmic thereafter. For 16 KHz sampling rate, Sphinx engine uses a set of 40 Mel filters.

Log of the Mel spectrum values: The range of the values generated by the Mel filter bank is reduced by replacing each value by its natural logarithm. This is done to make the statistical distribution of spectrum approximately Gaussian.

*Inverse Discrete Fourier Transform***:** This transform is used to compress the spectral information into a set of low order coefficients which is called the Mel-cepstrum. Thirteen MFCC coefficients are used as a basic feature vector, $x_t(k) \quad 0 \le k \le 12$.

*Deltas and Energy***:** For continuous models, the 13 MFCC parameters along with computed delta and delta-deltas parameters are used as a single stream 39 parameters feature vector. For semi-continuous models, x(0) represents the log Mel spectrum energy, and is used separately to derive other feature parameters, in addition to the delta and double delta parameters. Figure 5 shows part of the feature vector of a speech file after completing the feature extraction process. Each column represents the basic 13 features of a 25.6 milliseconds frame.



FeatureVectors <13x358 double>

	1	2	3	4	5
1	-1.3030	-1.1439	-1.2332	-1.1225	-1.3957
2	-1.2602	-1.2450	-1.2588	-1.1915	-1.2575
3	-0.0487	-0.2290	-0.0345	-0.0404	-0.0538
4	-0.1191	-0.2934	-0.1264	-0.1076	-0.1466
5	0.0237	-0.1424	0.0330	0.1451	-0.0068
6	-0.0395	-0.0243	0.0356	0.0135	-0.0487
7	-0.0970	-0.0554	0.0969	0.0282	0.0642
8	0.0092	-0.0489	-0.0104	0.0079	0.1567
9	0.1431	0.0946	0.1227	0.0366	-0.0841
10	0.0034	0.0942	0.1810	0.1100	-0.0273
11	0.0925	-0.0315	0.0419	0.0439	0.0590
12	0.0653	0.0674	-0.0610	0.0697	-0.0102
13	0.0459	-0.0267	-0.0058	0.0878	0.0270
14					

Figure 5. snapshot of the MFCCs of a speech file

3.2. Linguistic database

This part contains the modifications required for a particular language. It contains three parts: acoustic models, language model, and pronunciation dictionary. Acoustic models contain the HMMs used in recognition process. The language model contains language's words and their combinations, each combination has two or three words. A pronunciation dictionary contains the words and their pronunciation phonemes.

3.2.1. Acoustic models

Acoustic models are statistical representations of the speech phones. Precise acoustic model is a key factor to improve recognition accuracy as it characterizes the HMMs of each phone. Sphinx uses 39 English phonemes (The CMU Pronunciation Dictionary, 2011). The acoustic models use a 3- to 5-state Markov chain to represent the speech phone (Lee, 1988). Figure 6 shows a representation of a 3-state phone's acoustic model. In Figure 6, S1 is the representation of phone at the beginning, while S2 and S3 represent of the phone at the middle and the end states, respectively. Associated with S1, S2, and S3 are state emission probabilities, $b_j(x_t) = P(o = x_t \mid S_t = j)$, representing the probability of observing the feature vector in the state j. The emission probabilities are usually modeled by Gaussian mixture densities.

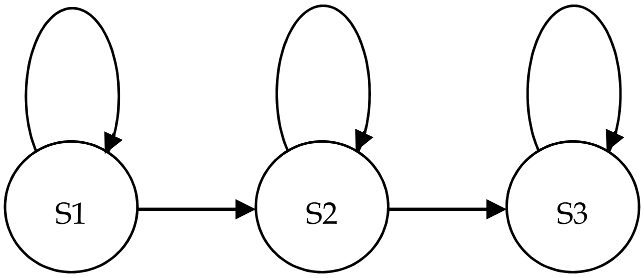


Figure 6. 3-state phone acoustic model

In continuous speech, each phoneme is influenced in different degrees by its neighboring phonemes. Therefore, for better acoustic modeling, Sphinx uses triphones. Triphones are context dependent models of phonemes; each triphone represents a phoneme surrounded by specific left and right phonemes (Hwang, 1993).

3.2.2. Language model

The N-gram language model is trained by counting N-gram occurrences in a large transcription corpus to be then smoothed and normalized. In general, an N-gram language model is used to calculate the probability of a given sequence of words as follows:

$$P(\mathrm{w}_1^{\mathrm{n}}) = \prod_{k=1}^{n} p(w_k \middle| w_1^{k-1})$$

Where n is limited to include the words' history as bigram (two consequent words), trigram (three consequent words), 4-gram (four consequent words), etc. for example, by assigning n=2, the probability of a three word sequence using bigram is calculated as follows: $P(\mathrm{w}_1 w_2 w_3) = p(w_3 \mid w_2) p(w_2 \middle| w_1) p(w_1)$

The CMU statistical language tool is described in (Clarkson & Rosenfeld, 1997). The CMU statistical language tool kit has been used to generate our Arabic statistical language model. Figure 7 shows the steps for creation and testing the language model, the steps are:

- Compute the word unigram counts.
- Convert the word unigram counts into a vocabulary list.
- Generate bigram and trigram tables based on this vocabulary.

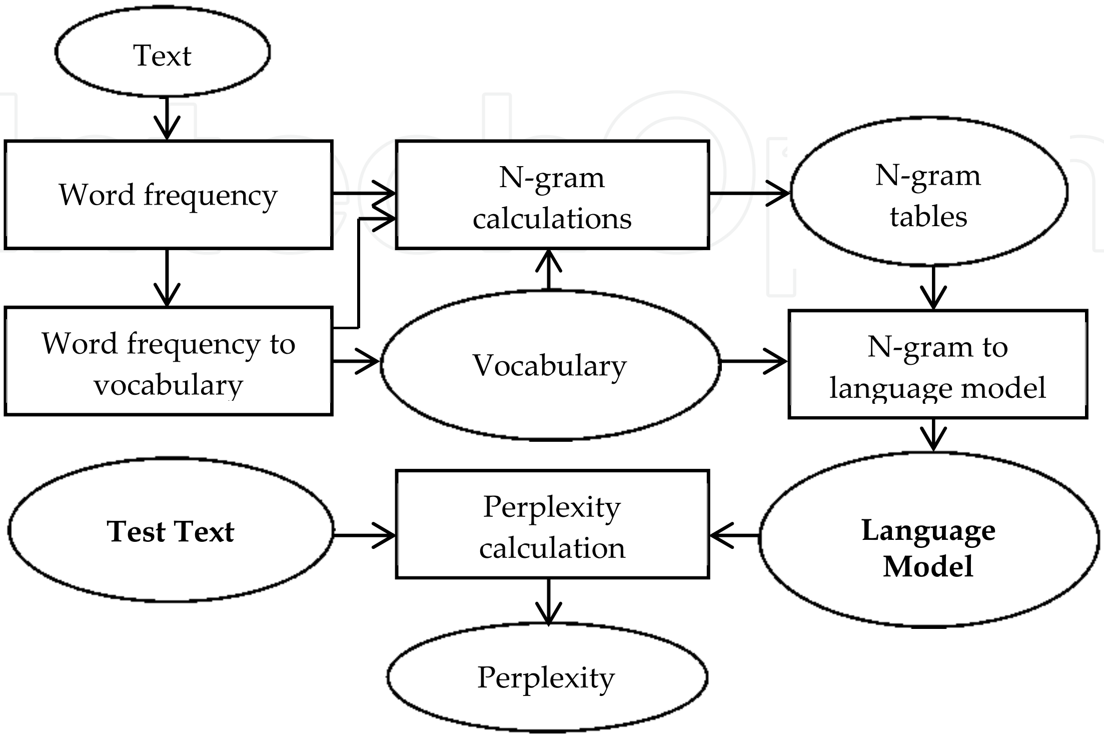


Figure 7. Steps for creating and testing language model

The CMU language modeling tool comes with a tool for evaluating the language model. The evaluation measures the perplexity as indication of the convenient (goodness) of the language model. For more information of the perplexity, please refer to Section 7.

3.2.3. Pronunciation dictionary

Both training and recognition stages require a pronunciation dictionary which is a mapping table that maps words into sequences of phonemes. A pronunciation dictionary is basically designed to be used with a particular set of words. It provides the pronunciation of the vocabulary for the transcription corpus using the defined phoneme set. Like acoustic models and language model, the performance of a speech recognition system depends critically on the dictionary and the phoneme set used to build the dictionary. In decoding stage, the dictionary serves as intermediary between the acoustic model and the language model.

There are two types of dictionaries: closed vocabulary dictionary and open vocabulary dictionary. In closed vocabulary dictionary, all corpus transcription words are listed in the dictionary. In contrast, it is possible to have non-corpus transcription words in the open vocabulary dictionary. Typically, the phoneme set, that is used to represent dictionary words, is manually designed by language experts. However, when human expertise is not available, the phoneme set is possible to be selected using data-driven approach as demonstrated by (Singh et al. 2002). In addition to providing phonemic transcriptions of the words of the target vocabulary, the dictionary is the place where alternative pronunciation variants are added such as in (Ali et al., 2009) for Arabic.

3.3. Decoder (Recognizer)

With help from the linguistic part, the decoder is the module where the recognition process takes place. The decoder uses the speech features presented by the Front-End to search for the most probable words and, then, sentences that correspond to the observed speech features. The recognition process starts by finding the likelihood of a given sequence of speech features based on the phonemes HMMs.

The speech recognition problem is to transcribe the most likely spoken words given the acoustic observations. If $O = o_1, o_2, \dots o_n$ is the acoustic observation, and $W = w_1, w_2, \dots w_n$ is a word sequence, then:

$$\widehat{W} = \underbrace{arg\,max}_{for\ all\ words}\ \mathrm{P(W)P(O|W)}$$

Where $\widehat{W}$ is the most probable word sequence of the spoken words, which is also called maximum posteriori probability. *P(W)* is the prior probability computed in the language model, and *P(O|W)* is the probability of observation computed using the acoustic model.

4. Pronunciation variation

The main goal of ASRs is to enable people to communicate more naturally and effectively. But this ultimate dream faces many obstacles such as different speaking styles which lead to "pronunciation variation" phenomenon. This phenomenon appears in the form of insertions, deletions, or substitutions of phoneme(s) relative to the phonemic transcription in the pronunciation dictionary. (Benzeghiba et al., 2007) presented the speech variability sources: foreign and regional accents, speaker physiology, spontaneous speech, rate of speech, children speech, emotional state, noises, new words, and more. Accordingly, handling these obstacles is a major requirement to have better ASR performance.

There are two types of pronunciation variations: cross-word variations and within-word variations. A within-word variation causes alternative pronunciation(s) of the same word. In contrast, a cross-word variation occurs in continuous speech in which a sequence of words forms a compound word that should be treated as a one entity. The pronunciation variation can be modeled in two approaches: knowledge-based and data-driven. Knowledge-based depends on linguistic studies that lead to the phonological rules which are called to find the possible alternative variants. On the other hand, data-driven methods depend solely on the pronunciation corpus to find the pronunciation variants (direct data-driven) or transformation rules (indirect data-driven). In this chapter, we will use the knowledge-based approach to model the cross-word pronunciation variation problem.

As pros and cons of both approaches, the knowledge-based approach is not exhaustive; not all of the variations that occur in continuous speech have been described. Whereas obtaining reliable information using data-driven is difficult. However, (Amdal & Fossler-Lussier 2003) mentioned that there is a growing interest in data-driven methods over the knowledge-based methods due to lack of domains expertise. Figure 8 displays these two techniques. Figure 8 also distinguishes between the types of variations and the modeling techniques by a dashed line. The pronunciation variation types are above the dashed line whereas the modeling techniques are under the dashed line.

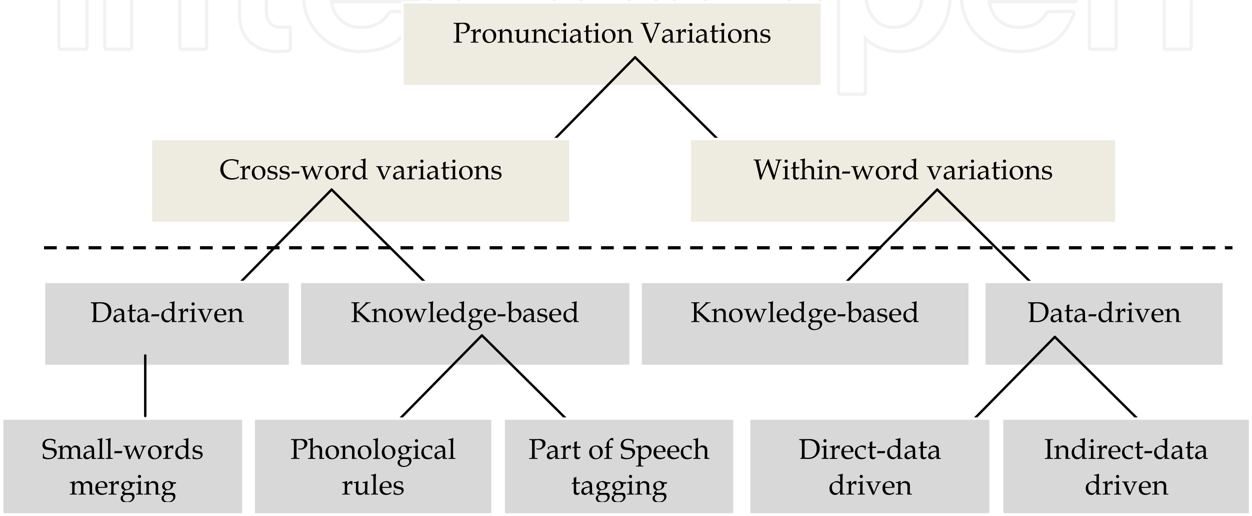


Figure 8. Pronunciation variations and modeling techniques

5. Arabic speech recognition

This work focuses on Arabic speech recognition, which has gained increasing importance in the last few years. Arabic is a Semitic language spoken by more than 330 million people as a native language (Farghaly & Shaalan, 2009). While Arabic language has many spoken dialects, it has a standard written language. As a result, more challenges are introduced to speech recognition systems as the spoken dialects are not officially written. The same country could contain different dialects and a dialect itself can vary from region to another according to different factors such as religion, gender, urban/rural, etc. Speakers with different dialects usually use modern standard Arabic (MSA) to communicate.

5.1. Modern standard Arabic

In this chapter, we consider the modern standard Arabic (MSA) which is currently used in writing and in most formal speech. MSA is also the major medium of communication for public speaking and news broadcasting (Ryding, 2005) and is considered to be the official language in most Arabic-speaking countries (Lamel et al., 2009). Arabic language challenges will be presented in the next section. Followed by the literature review and recent results efforts in Arabic speech recognition. For more information about modern standard Arabic, (Ryding, 2005) is a rich reference.

5.2. Arabic speech recognition challenges

Arabic speech recognition faces many challenges. First, Arabic has many dialects where same words are pronounced differently. In addition, the spoken dialects are not officially written, it is very costly to obtain adequate corpora, which present a training problem for the Arabic ASR researchers (Owen et al., 2006). Second, Arabic has short vowels (diacritics), which are usually ignored in text. The lack of diacritical marks introduces another serious problem to Arabic speech recognition. Consequently, more hypotheses' words will be considered during decoding process which may reduce the accuracy. (Elmahdy et al., 2009) summarized some of the problems raised in Arabic speech recognition. They highlighted the following problems: Arabic phonetics, diacritization problem, grapheme-to-phoneme, and morphological complexity. Although foreign phoneme sounds as /v/ and /p/ are used in Arabic speech in foreign names, the standard Arabic letters do not have standard letter assigned for foreign sounds. Second, the absence of the diacritical marks in modern Arabic text creates ambiguities for pronunciations and meanings. For example, the non-diacritized Arabic word (كتب) could be read as one of several choices, some of which are: (كَتَب,he wrote), (كُتِب, it was written), and (كُتُب, books). Even though, an Arabic reader can interpret and utter the correct choice, it is hard to embed this cognitive process in current speech recognition and speech synthesis systems. The majority of Arabic corpora available for the task of acoustic modeling have non-diacritized transcription. (Elmahdy et al., 2009) also showed that grapheme-to-phoneme relation is only true for diacritized Arabic script. Hence fore, Arabic speech recognition has an obstacle because the lack of diacritized corpora. Arabic morphological complexity is demonstrated by the large number of affixes (prefixes, infixes, and suffixes) that can be added to the three consonant radicals to form patterns. (Farghaly& Shaalan, 2009) provided a comprehensive study of Arabic language challenges and solutions. The mentioned challenges include: the nonconcatenative nature of Arabic morphology, the absence of the orthographic representation of Arabic diacritics from contemporary Arabic text, and the need for an explicit grammar of MSA that defines linguistic constituency in the absence of case marking. (Lamel et al., 2009) presented a number of challenges for Arabic speech recognition such as no diacritics, dialectal variants, and very large lexical variety. (Alotaibi et al., 2008) introduced foreign-accented Arabic speech as a challenging task in speech recognition. (Billa et al., 2002) discussed a number of research issues for Arabic speech recognition, e.g., absence of diacritics in written text and the presence of compound words that are formed by the concatenation of certain conjunctions, prepositions, articles, and pronouns, as prefixes and suffixes to the word stem.

5.3. Literature and recent work

A number of researchers have recently addressed development of Arabic speech recognition systems. (Abushariah et al., 2012) proposed a framework for the design and development of a speaker-independent continuous automatic Arabic speech recognition system based on a phonetically rich and balanced speech corpus. Their method reduced the WER to 9.81% for a diacritized transcription corpus, as they have reported. (Hyassat & Abu Zitar, 2008) described an Arabic speech recognition system based on Sphinx 4. Three corpora were developed, namely, the Holy Qura'an corpus of about 18.5 hours, the command and control corpus of about 1.5 hours, and the Arabic digits corpus of less than 1 hour of speech. They also proposed an automatic toolkit for building pronunciation dictionaries for the Holy Qur'an and standard Arabic language. (Al-Otaibi, 2001)] provided a single-speaker speech dataset for MSA. He proposed a technique for labeling Arabic speech. using the Hidden Markov Model Toolkit (HTK), he reported a recognition rate for speaker dependent ASR of 93.78%. (Afify et al. , 2005) compared grapheme-based recognition system with explicitly modeling diacritics (short vowels). They found that a diacritic modeling improves recognition performance. (Satori et al. , 2007) used CMU Sphinx tools for Arabic speech recognition. They demonstrated the use of the tools for recognition of isolated Arabic digits. They achieved a digits recognition accuracy of 86.66% for data recorded from six speakers. (Alghamdi et al., 2009) developed an Arabic broadcast news transcription system. They used a corpus of 7.0 h for training and 0.5 h for testing. The WER they obtained was 14.9%. (Lamel et al., 2009) described the incremental improvements to a system for the automatic transcription of broadcast data in Arabic, highlighting techniques developed to deal with specificities (no diacritics, dialectal variants, and lexical variety) of the Arabic language. (Billa et al., 2002) described the development of audio indexing system for broadcast news in Arabic. Key issues addressed in their work revolve around the three major components of the audio indexing system: automatic speech recognition, speaker identification, and named entity identification. (Soltau et al., 2007) reported advancements in the IBM system for Arabic speech recognition as part of the continuous effort for the Global Autonomous Language Exploitation (GALE) project. The system consisted of multiple stages that incorporate both diacritized and non-diacritized Arabic speech model. The system also incorporated a training corpus of 1,800 hours of unsupervised Arabic speech. (Azmi et al., 2008) investigated using Arabic syllables for speaker-independent speech recognition system for Arabic spoken digits. The pronunciation corpus used for both training and testing consisted of 44 Egyptian speakers. In a clean environment, experiments showed that the recognition rate obtained using syllables outperformed the rate obtained using monophones, triphones, and words by 2.68%, 1.19%, and 1.79%, respectively. Also in noisy telephone channel, syllables outperformed the rate obtained using monophones, triphones, and words by 2.09%, 1.5%, and 0.9%, respectively. (Elmahdy et al., 2009) used acoustic models trained with large MSA news broadcast speech corpus to work as multilingual or multi-accent models to decode colloquial Arabic. (Khasawneh et al., 2004) compared the polynomial classifier that was applied to isolated-word speaker-independent Arabic speech and dynamic time warping (DTW) recognizer. They concluded that the polynomial classifier produced better recognition performance and much faster testing response than the DTW recognizer. (Shoaib et al., 2003) presented an approach to develop a robust Arabic speech recognition system based on a hybrid set of speech features. The hybrid set consisted of intensity contours and formant frequencies. (Alotaibi, 2004) reported achieving high-performance Arabic digits recognition using recurrent networks. (Choi et al., 2008) presented recent improvements to their English/Iraqi Arabic speech-to-speech translation system. The presented system-wide improvements included user interface, dialog manager, ASR, and machine translation components. (Nofal et al., 2004) demonstrated a design and implementation of stochastic-based new acoustic models for use with a command and control system speech recognition system for the Arabic. (Mokhtar & El-Abddin, 1996) represented the techniques and algorithms used to model the acoustic-phonetic structure of Arabic speech recognition using HMMs. (Park et al. , 2009) explored the training and adaptation of multilayer perceptron (MLP) features in Arabic ASRs. They used MLP features to incorporate short-vowel information into the graphemic system. They also used linear input networks (LIN) adaptation as an alternative to the usual HMM-based linear adaptation. (Imai et al.,1995) presented a new method for automatic generation of speaker-dependent phonological rules in order to decrease recognition errors caused by pronunciation variability dependent on speakers. (Muhammad et al., 2011) evaluated conventional ASR system for six different types of voice disorder patients speaking Arabic digits. MFCC and Gaussian mixture models (GMM)/HMM were used as features and classifier, respectively. Recognition result was analyzed for recognition for types of diseases. (Bourouba et al., 2006) presented a HMM/support vectors machine (SVM) (k-nearest neighbor) for recognition of isolated spoken Arabic words. (Sagheer et al., 2005) presented a visual speech features representation system. They used it to comprise a complete lip-reading system. (Taha et al. , 2007) demonstrated an agent-based design for Arabic speech recognition. They defined the Arabic speech recognition as a multi-agent system where each agent had a specific goal and deals with that goal only. (Elmisery et al., 2003) implemented a pattern matching algorithm based on HMM using field programmable gate array (FPGA). The proposed approach was used for isolated Arabic word recognition. (Gales et al., 2007) described the development of a phonetic system for Arabic speech recognition. (Bahi & Sellami, 2001) presented experiments performed to recognize isolated Arabic words. Their recognition system was based on a combination of the vector quantization technique at the acoustic level and markovian modeling. (Essa et al., 2008) proposed a combined classifier architectures based on Neural Networks by varying the initial weights, architecture, type, and training data to recognize Arabic isolated words. (Emami & Mangu, 2007) studied the use of neural network language models (NNLMs) for Arabic broadcast news and broadcast conversations speech recognition. (Messaoudi et al., 2006) demonstrated that by building a very large vocalized vocabulary and by using a language model including a vocalized component, the WER could be significantly reduced. (Vergyri et al., 2004) showed that the use of morphology-based language models at different stages in a large vocabulary continuous speech recognition (LVCSR) system for Arabic leads to WER reductions. To deal with the huge lexical variety, (Xiang et al., 2006) concentrated on the transcription of Arabic broadcast news by utilizing morphological decomposition in both acoustic and language modeling in their system. (Selouani & Alotaibi, 2011) presented genetic algorithms to adapt HMMs for non-native speech in a large vocabulary speech recognition system of MSA. (Saon et al., 2010) described the Arabic broadcast transcription system fielded by IBM in the

GALE project. they reported improved discriminative training, the use of subspace Gaussian mixture models (SGMM), the use of neural network acoustic features, variable frame rate decoding, training data partitioning experiments, unpruned n-gram language models, and neural network based language modeling (NNLMs) . The achieved WER was 8.9% on the evaluation test set. (Kuo et al., 2010) studied various syntactic and morphological context features incorporated in an NNLM for Arabic speech recognition.

6. The proposed method

Since the ASR decoder works better with long words, our method focuses on finding a way to merge transcription words to increase the number of long words. For this purpose, we consider to merge words according to their tags. That is, merge a noun that is followed by an adjective, and merge a preposition that is followed by a word. we utilizes PoS tagging approach to tag the transcription corpus. the tagged transcription is then used to find the new merged words.

A tag is a word property such as noun, pronoun, verb, adjective, adverb, preposition, conjunction, interjection, etc. Each language has its own tags. Tags may be different from language to language. In our method, we used the Arabic module of Stanford tagger (Stanford Log-linear Part-Of-Speech Tagger, 2011). The total number of tags of this tagger is 29 tags, only 13 tags were used in our method as listed in Table 3. As we mentioned, we focused on three kinds of tags: noun, adjectives, and preposition. In Table 3, DT is a shorthand for the determiner article (ال التعريف) that corresponds to "the" in English.

#	Tag	Meaning	Example
1	ADJ_NUM	Adjective, Numeric	الرابعة السابع،
2	DTJJ	DT + Adjective	الجديد النفطية،
3	DTJJR	Adjective, comparative	العليا الكبرى،
4	DTNN	DT + Noun, singular or mass	المنظمة، العاصمة
5	DTNNP	DT + Proper noun, singular	القاهرة العراق،
6	DTNNS	DT + Noun, plural	السيارات، الولايات
7	IN	Preposition subordinating conjunction	حرف جر مثل : في حرف مصدري مثل :أنْ
8	JJ	Adjective	قيادية جديدة،
9	JJR	Adjective, comparative	كبرى أدنى،
10	NN	Noun, singular or mass	إنتاج، نجم
11	NNP	Proper noun, singular	لبنان أوبك،
12	NNS	Noun, plural	طلبات توقعات،
13	NOUN_QUANT	Noun, quantity	الربع، ثلثي

Table 3. A partial list of Stanford Tagger's tag with examples

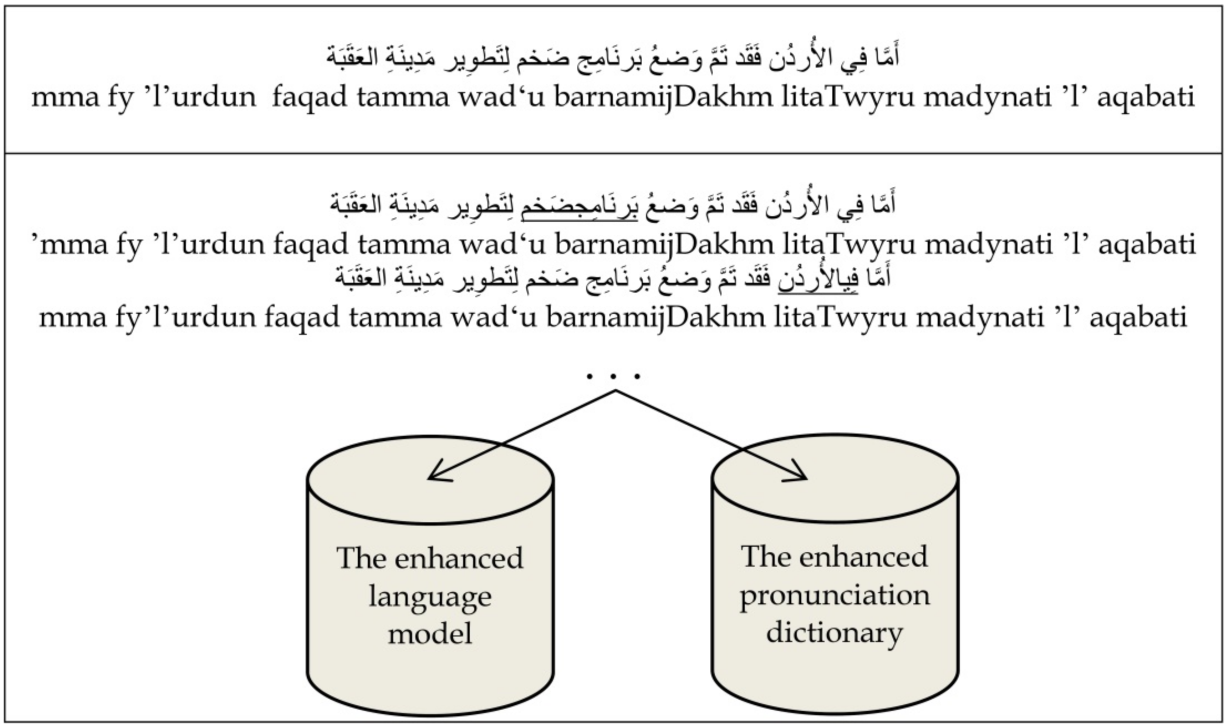
In this work, we used the Noun-Adjective as shorthand for a compound word generated by merging a noun and an adjective. We also used Preposition-Word as shorthand for a compound word generated by merging a preposition with a subsequent word. The prepositions used in our method include:

(منذ ، حتى ، في ، على ، عن ، الى ، من) ➔ (mundhu, Hata, fy, ‘ala, ‘an, ’ila, min), Other prepositions were not included as they are rarely used in MSA. Table 4 shows the tagger output for a simple non-diacritized sentence.

An input sentence to the tagger	وأوضح عضو لجنة المقاولين في غرفة الرياض بشير العظم wa ’wdaHa ‘udwu lajnata ’lmuqawilyna fy ghurfitu ’lriyaD bashyru ’l‘ aZm
Tagger output (read from left to right)	وأوضح/VBD عضو/NN لجنة/NN المقاولين/DTNNS في/IN غرفة/NN الرياض/DTNNP بشير/NNP العظم/DTNN

Table 4. An Arabic sentence and its tags

Thus, the tagger output is used to generate compound words by searching for Noun-Adjective and Preposition-Word sequences. Figure 9 shows two possible compound words: (بَرنَامِجضَخم) and (فِيالأُردُن) for Noun-Adjective case and for Preposition-Word case, respectively. These two compound words are, then, represented in new sentences as illustrated in Figure 9. Therefore, the three sentences (the original and the new ones) will be used, with all other cases, to produce the enhanced language model and the enhanced pronunciation dictionary.



أَمَّا فِي الأُردُن فَقَد تَمَّ وَضعُ بَرنَامِج ضَخم لِتَطوِير مَدِينَةِ العَقَبَة
mma fy ’l’urdun faqad tamma wad‘u barnamijDakhm litaTwyru madynati ’l’ aqabati

أَمَّا فِي الأُردُن فَقَد تَمَّ وَضعُ بَرنَامِجضَخم لِتَطوِير مَدِينَةِ العَقَبَة
’mma fy ’l’urdun faqad tamma wad‘u barnamijDakhm litaTwyru madynati ’l’ aqabati

أَمَّا فِيالأُردُن فَقَد تَمَّ وَضعُ بَرنَامِج ضَخم لِتَطوِير مَدِينَةِ العَقَبَة
mma fy’l’urdun faqad tamma wad‘u barnamijDakhm litaTwyru madynati ’l’ aqabati

. . .

The enhanced language model
The enhanced pronunciation dictionary

Figure 9. The compound words representations

Figure 10 shows the process of generating a compound word. It demonstrates that a noun followed by an adjective will be merged to produce a one compound word. similarly , the preposition followed by a word will be merged to perform a one compound word. It is noteworthy to mention that our method is independent from handling pronunciation variations that may occur at words junctures. That is, our method does not consider the phonological rules that could be implemented between certain words.

The steps for modeling cross-word phenomenon can be described by the algorithm (pseudocode) shown in Figure 11. In the figure, the Offline stage means that the stage is implemented once before decoding, while Online stage means that this stage needs to be repeatedly implemented after each decoding process.

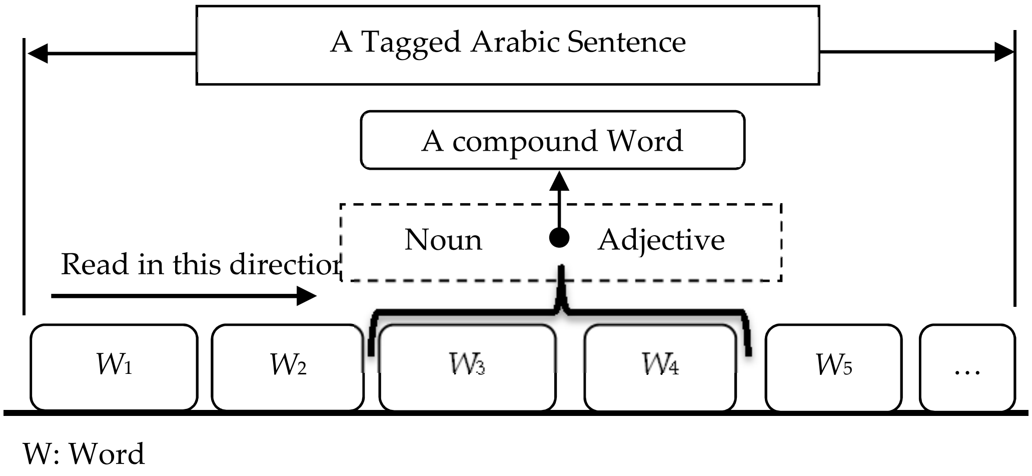


Figure 10. A Noun-Adjective compound word generation

```
                         Offline Stage
Using a PoS tagger, have the transcription corpus tagged
For all tagged sentences in the transcription file
    For each two adjacent tags of each tagged sentence
        If the adjacent tags are adjective/noun or word/preposition
            Generate the compound word
            Represent the compound word in the transcription
        End if
    End for
End for
Based on the new transcription, build the enhanced dictionary
Based on the new transcription, build the enhanced language model
                         Online Stage
Switching the variants back to its original separated words
```

Figure 11. Cross-word modeling algorithm using PoS tagging

7. The results

The proposed method was investigated on a speaker-independent modern standard Arabic speech recognition system using Carnegie Mellon University Sphinx speech recognition engine. Three performance metrics were used to measure the performance enhancement: the word error rate (WER), out of vocabulary (OOV), and perplexity (PP).

WER is a common metric to measure performance of ASRs. WER is computed using the following formula:

$$WER = \frac{S + D + I}{N}$$

Where:

- S is the number of substituted words,
- D is the number of deleted words,
- I is the number of inserted words,
- N is the total number of words in the testing set.

The word accuracy can also be measured using WER as the following formula:

$$\text{Word Accuracy} = 1 - \text{WER}$$

OOV is a metric to measure the performance of ASRs. OOV is known as a source of recognition errors, which in turn could lead to additional errors in the words that follow (Gallwitz et al., 1996). Hence fore, increasing OOVs plays a significant role in increasing WER and deteriorating performance. In this research work, the baseline system is based on a closed vocabulary. The closed vocabulary assumes that all words of the testing set are already included in the dictionary. (Jurafsky & Martin, 2009) explored the differences between open and closed vocabulary. In our method, we calculate OOV as the percentage of recognized words that are not belonging to the testing set, but to the training set. The following formula is used to find OOV:

$$\text{OOV (baseline system)} = \frac{\text{none testing set words}}{\text{total words in the testing set}} * 100$$

The perplexity of the language model is defined in terms of the inverse of the average log likelihood per word (Jelinek, 1999). It is an indication of the average number of words that can follow a given word, a measure of the predictive power of the language model, (Saon & Padmanabhan ,2001). Measuring the perplexity is a common way to evaluate N-gram language model. It is a way to measure the quality of a model independent of any ASR system. Of course, The measurement is performed on the testing set. A lower perplexity system is considered better than one of higher perplexity. The perplexity formula is:

$$\text{PP(W)} = \sqrt[N]{\frac{1}{\text{P}(\text{w}_1,\text{w}_2,\ldots,\text{w}_\text{N})}}$$

Where PP is the perplexity, P is the probability of the word set to be tested W=w_1, w_2, … , w_N, and N is the total number of words in the testing set.

The performance detection method proposed by Plötz in (Plötz,2005) is used to investigate the achieved recognition results. A 95% is used as a level of confidence. The WER of the baseline system (12.21 %) and the total number of words in the testing set (9288 words) are used to find the confidence interval [εl , εh]. The boundaries of the confidence interval are found to be [12.21 – 0.68 , 12.21 + 0.68] ➔ [11.53,12.89]. If the changed classification error rate is outside this interval, this change can be interpreted as statistically significant. Otherwise, It is most likely caused by chance.

Table 5 shows the enhancements for different experiments. Since the enhanced method (in Noun-Adjective case) achieved a WER of (9.82%) which is out of the above mentioned confidence interval [11.53,12.89], it is concluded that the achieved enhancement is statistically significant. The other cases are similar, i.e. (Preposition-Word, and Hybrid cases also achieved a significant improvement).

#	Experiment	Accuracy (%)	WER (%)	Enhancement (%)
	Baseline system	87.79	12.21	----------
1	Noun-Adjective	90.18	9.82	2.39
2	Preposition-Word	90.04	9.96	2.25
3	Hybrid (1 & 2)	90.07	9.93	2.28

Table 5. Accuracy achieved and WERs for different cases

Table 5 shows that the highest accuracy achieved is in Noun-Adjective case. The reduction in accuracy in the hybrid case is due to the ambiguity introduced in the language model. For more clarification, our method depends on adding new sentences to the transcription corpus that is used to build the language model. Therefore, adding many sentences will finally cause the language model to be biased to some n-grams (1-grams, 2-grams, and 3-grams) on the account of others.

The common way to evaluate the N-gram language model is using perplexity. The perplexity for the baseline is 34.08. For the proposed cases, the language models' perplexities are displayed in Table 6. The measurements were taken based on the testing set, which contains 9288 words. The enhanced cases are clearly better as their perplexities are lower. The reason for the low perplexities is the specific domains that we used in our corpus, i.e. economics and sports.

#	Experiment	Perplexity	OOV (%)
	Baseline System	34.08	328/9288 = 3.53%
1	Noun-Adjective	3.00	287/9288 = 3.09%
2	Preposition-Word	3.22	299/9288 = 3.21%
3	Hybrid (1 & 2)	2.92	316/9288 = 3.40%

Table 6. Perplexities and OOV for different experiments

The OOV was also measured for the performed experiments. Our ASR system is based on a closed vocabulary, so we assume that there are no unknown words. The OOV was calculated as the percentage of recognized words that do not belong to the testing set, but to the training set. Hence,

$$\text{OOV (baseline system)} = \frac{\text{none testing set words}}{\text{total words in the testing set}} * 100$$

which is equal to 328/9288*100= 3.53%. For the enhanced cases, Table 6 shows the resulting OOVs. Clearly, the lower the OOV the better the performance is, which was achieved in all three cases.

Table 7 shows some statistical information collected during experiments. The "Total compound words" is the total number of Noun-Adjective cases found in the corpus transcription. The "unique compound words" indicates the total number of Noun-Adjective cases after removing duplicates. The last column, "compound words replaced" is the total number of compound words that were replaced back to their original two disjoint words after the decoding process and prior to the evaluation stage.

#	Experiment	Total compound words	unique compound words	compound words replaced
1	Noun-Adjective	3328	2672	377
2	Preposition-Word	3883	2297	409
3	Hybrid (1 & 2)	7211	4969	477

Table 7. Statistical information for compound words

Despite the claim that the Stanford Arabic tagger accuracy is more than 96%, a comprehensive manual verification and correction were made on the tagger output. It was reasonable to review the collected compound words as our transcription corpus is small (39217 words). For large corpora, the accuracy of the tagger is crucial for the results. Table 8 shows an error that occurred in the tagger output. The word, for example, "وقال"(waqala) should be VBD instead of NN.

Sentence to be tagged	هذا وقال رئيس لجنة الطاقة بمجلس النواب ورئيس الرابطة الروسية للغاز إن الاحتكارات الأوروبية hadha waqala ra'ysu lajnati 'lTaqa bimajlisi 'lnuwab wa ra'ysu 'lrabiTa 'lrwsiya llghaz 'ina 'l'iHtikarati 'liwrobiya
Stanford Tagger output (read from left to right)	هذا/DT وقال/NN رئيس/NN لجنة/NN الطاقة/DTNN بمجلس/NN النواب/DTNN ورئيس/NN الرابطة/DTNN الروسية/DTJJ للغاز/NNP إن/NNP الاحتكارات/DTNNS الأوروبية/DTJJ

Table 8. Example of Stanford Arabic Tagger Errors

Table 9 shows an illustrative example of the enhancement that was achieved in the enhanced system. It shows that the baseline system missed one word “مِن”(min) while it appears in the enhanced system. Introducing a compound word in this sentence avoided the misrecognition that occurred in the baseline system.

The text of a speech file to be tested	فِي المَرحَلَةِ السَّابِعَةِ وَالثّلاثِين مِن الدَّورِيِّ الإِسبَانِيِّ لِكُرَةِ القَدَم fy ’lmarHalati ’lsabi‘ a wa ’lthalathyn mina ’ldawry ’l’sbany likurati ’lqadam
As recognized by the baseline system	فِي المَرحَلَةِ السَّابِعَةِ وَالثّلاثِين الدَّورِيِّ الإِسبَانِيِّ لِكُرَةِ القَدَم fy ’lmarHalati ’lsabi‘ a wa ’lthalathyn mina ’ldawry ’l’sbany likurati ’lqadam
As recognized by the enhanced system	فِي المَرحَلَةِ السَّابِعَةِ وَالثّلاثِين مِن الدَّورِيِّالإِسبَانِيِّ لِكُرَةِ القَدَم fy ’lmarHalati ’lsabi‘ a wa ’lthalathyn mina ’ldawry ’l’sbany likurati ’lqadam
Final output after decomposing the merging	فِي المَرحَلَةِ السَّابِعَةِ وَالثّلاثِين مِن الدَّورِيِّ الإِسبَانِيِّ لِكُرَةِ القَدَم fy ’lmarHalati ’lsabi‘ a wa ’lthalathyn mina ’ldawry ’l’sbany likurati ’lqadam

Table 9. An example of enhancement in the enhanced system

According to the proposed algorithm, each sentence in the enhanced transcription corpus can have a maximum of one compound word, since sentences are added to the enhanced corpus once a compound word is formed. Finally, After the decoding process, the results are scanned in order to decompose the compound words back to their original form (two separate words). This process is performed using a lookup table such as:

الكُوَيتالدُّوَلِيِّ ➔ الكُوَيت الدُّوَلِيِّ (’lkuwaytldawly ➔ ’lkuwayt ’ldawly)

فِيمَطَارِ ➔ فِي مَطَارِ (fymatari ➔ fy matari)

8. Discussion

Table 10 shows comparison results of the suggested methods for cross-word modeling. It shows that PoS tagging approach outperform the other methods (i.e. the phonological rules and small word merging) which were investigated on the same pronunciation corpus. The use of phonological rules was demonstrated in (AbuZeina et al. 2011a) while merging of small-words method was presented in (AbuZeina et al. 2011b). even though PoS tagging seems to be better than the other methods, more research should be carried out for more confidence. So, the comparison demonstrated in Table 10 is subject to change as more cases need to be investigated for both techniques. That is, cross-word was modeled using only two Arabic phonological rules, while only two compounding schemes were applied in PoS tagging approach.

The recognition time is compared with the baseline system. The comparison includes the testing set which includes 1144 speech files. The specifications of the machine where we conducted the experiments were as follows: a desktop computer which contains a single processing chip of 3.2GHz and 2.0 GB of RAM. We found that the recognition time for the enhanced method is almost the same as the recognition time of the baseline system. This means that the proposed method is almost equal to the baseline system in term of time complexity.

#	System	Accuracy (%)	Execution Time (minutes)
	Baseline system	87.79	34.14
1	phonological rules	90.09	33.49
2	PoS tagging	90.18	33.05
3	small word merging	89.95	34.31
4	Combined system (1,2,and3)	88.48	30.31

Table 10. Comparison between cross-word modeling techniques

9. Further research

As future work, we propose investigating more word-combination cases. In particular, we expect that the construct phrases *Idafa* (الإضافة) make a good candidate. Examples include: (سلسلة جبال, silsilt jibal), (مطار بيروت, maTaru bayrwt) , (مدينة القدس, madynatu ʾlquds). Another suggested candidate is the Arabic "and" connective (واو العطف), such as: (مواد أدبية ولغوية, mawad ʾdabiyah wa lughawiyah), (yataʿallaqu biqaDaya ʾlʿ iraqi wa ʿlsudan يتعلق بقضايا العراق والسودان،). A hybrid system could also be investigated. It is possible to use the different cross-word modeling approaches in a one ASR system. It is also worthy to investigate how to model the compound words in the language model. In our method, we create a new sentence for each compound word. we suggest to investigate representing the compound word exclusively with its neighbors. for example, instead of having two complete sentences to represent the compound words (بَرنَامِجضَخم , barnamijDakhm) and (فِيالأُردُن , fyʾlʾurdun) as what we proposed in our method:

أَمَّا فِي الأُردُن فَقَد تَمَّ وَضعُ بَرنَامِجضَخم لِتَطوِير مَدِينَةِ العَقَبَة

ʾmma fy ʾlʾurdun faqad tamma wadʿu barnamijDakhm litaTwyru madynati ʾlʾ aqabati

أَمَّا فِيالأُردُن فَقَد تَمَّ وَضعُ بَرنَامِج ضَخم لِتَطوِير مَدِينَةِ العَقَبَة

ʾ mma fyʾlʾurdun faqad tamma wadʿu barnamijDakhm litaTwyru madynati ʾlʾ aqabati

We propose to add the compound words only with their adjacent words like:

وَضعُ بَرنَامِجضَخم لِتَطوِير

waDʿu barnamijDakhm litaTwyr

أَمَّا فِيالأُردُن فَقَد

’ mma fy ’l’urdun faqad

A comprehensive research work should be made to find how to effectively represent the compound words in the language model. In addition, we highly recommend further research in PoS tagging for Arabic.

10. Conclusion

The proposed knowledge-based approach to model cross-word pronunciation variations problem achieved a feasible improvement. Mainly, PoS tagging approach was used to form compound words. The experimental results clearly showed that forming compound words using a noun and an adjective achieved a better accuracy than merging of a preposition and its next word. The significant enhancement we achieved has not only come from the cross-word pronunciation modeling in the dictionary, but also indirectly from the recalculated n-grams probabilities in the language model. We also conclude that Viterbi algorithm works better with long words. Speech recognition research should consider this fact when designing dictionaries. We found that merging words based on their types (tags) leads to significant improvement in Arabic ASRs. We also found that the proposed method outperforms the other cross-word methods such as phonological rules and small-words merging.

Author details

Dia AbuZeina, Husni Al-Muhtaseb and Moustafa Elshafei
King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

Acknowledgement

The authors would like to thank King Fahd University of Petroleum and Minerals for providing the facilities to write this chapter. We also thank King Abdulaziz City for Science and Technology (KACST) for partially supporting this research work under Saudi Arabia Government research grant NSTP # (08-INF100-4).

11. References

Abushariah, M. A.-A. M.; Ainon, R. N.; Zainuddin, R.; Elshafei, M. & Khalifa, O. O. Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. Int. Arab J. Inf. Technol., 2012, 9, 84-93

AbuZeina D., Al-Khatib W., Elshafei M., “Small-Word Pronunciation Modeling for Arabic Speech Recognition: A Data-Driven Approach”, Seventh Asian Information Retrieval Societies Conference, Dubai, 2011b.

AbuZeina D., Al-Khatib W., Elshafei M., Al-Muhtaseb H., "Cross-word Arabic pronunciation variation modeling for speech recognition" , International Journal of Speech Technology , 2011a.

Afify M, Nguyen L, Xiang B, Abdou S, Makhoul J. Recent progress in Arabic broadcast news transcription at BBN. In: Proceedings of INTERSPEECH. 2005, pp 1637–1640

Alghamdi M, Elshafei M, Almuhtasib H (2009) Arabic broadcast news transcription system. Int J Speech Tech 10:183–195

Ali, M., Elshafei, M., Alghamdi M. , Almuhtaseb, H. , and Alnajjar, A., "Arabic Phonetic Dictionaries for Speech Recognition". Journal of Information Technology Research, Volume 2, Issue 4, 2009, pp. 67-80.

Alotaibi YA (2004) Spoken Arabic digits recognizer using recurrent neural networks. In: Proceedings of the fourth IEEE international symposium on signal processing and information technology, pp 195–199

Al-Otaibi F (2001) speaker-dependant continuous Arabic speech recognition. M.Sc. thesis, King Saud University

Amdal I, Fosler-Lussier E (2003) Pronunciation variation modeling in automatic speech recognition. Telektronikk, 2.2003, pp 70–82.

Azmi M, Tolba H,Mahdy S, Fashal M(2008) Syllable-based automatic Arabic speech recognition in noisy-telephone channel. In: WSEAS transactions on signal processing proceedings, World Scientific and Engineering Academy and Society (WSEAS), vol 4, issue 4, pp 211–220

Bahi H, Sellami M (2001) Combination of vector quantization and hidden Markov models for Arabic speech recognition. ACS/IEEE international conference on computer systems and applications, 2001

Benzeghiba M, De Mori R et al (2007) Automatic speech recognition and speech variability: a review. Speech Commun 49(10–11):763–786.

Billa J, Noamany M et al (2002) Audio indexing of Arabic broadcast news. 2002 IEEE international conference on acoustics, speech, and signal processing (ICASSP)

Bourouba H, Djemili R et al (2006) New hybrid system (supervised classifier/HMM) for isolated Arabic speech recognition. 2nd Information and Communication Technologies, 2006. ICTTA'06

Choi F, Tsakalidis S et al (2008) Recent improvements in BBN's English/Iraqi speech-to-speech translation system. IEEE Spoken language technology workshop, 2008. SLT 2008

Clarkson P, Rosenfeld R (1997) Statistical language modeling using the CMU-Cambridge toolkit. In: Proceedings of the 5th European conference on speech communication and technology, Rhodes, Greece.

Elmahdy M, Gruhn R et al (2009) Modern standard Arabic based multilingual approach for dialectal Arabic speech recognition. In: Eighth international symposium on natural language processing, 2009. SNLP'09

Elmisery FA, Khalil AH et al (2003) A FPGA-based HMM for a discrete Arabic speech recognition system. In: Proceedings of the 15th international conference on microelectronics, 2003. ICM 2003

Emami A, Mangu L (2007) Empirical study of neural network language models for Arabic speech recognition. IEEE workshop on automatic speech recognition and understanding, 2007. ASRU

Essa EM, Tolba AS et al (2008) A comparison of combined classifier architectures for Arabic speech recognition. International conference on computer engineering and systems, 2008. ICCES 2008

Farghaly A, Shaalan K (2009) Arabic natural language processing: challenges and solutions. ACM Trans Asian Lang Inform Process 8(4):1–22.

Gales MJF, Diehl F et al (2007) Development of a phonetic system for large vocabulary Arabic speech recognition. IEEE workshop on automatic speech recognition and understanding, 2007. ASRU

Gallwitz F, Noth E, et al (1996) A category based approach for recognition of out-of-vocabulary words. In: Proceedings of fourth international conference on spoken language, 1996. ICSLP 96

Hwang M-H (1993) Subphonetic acoustic modeling for speaker-independent continuous speech recognition, Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

Hyassat H, Abu Zitar R (2008) Arabic speech recognition using Sphinx engine. Int J Speech Tech 9(3–4):133–150

Imai T, Ando A et al (1995) A new method for automatic generation of speaker-dependent phonological rules. 1995 international conference on acoustics, speech, and signal processing, 1995. ICASSP-95

Jelinek F (1999) Statistical methods for speech recognition, Language, speech and communication series. MIT, Cambridge, MA

Jurafsky D, Martin J (2009) Speech and language processing, 2nd edn. Pearson, NJ

Khasawneh M, Assaleh K et al (2004) The application of polynomial discriminant function classifiers to isolated Arabic speech recognition. In: Proceedings of the IEEE international joint conference on neural networks, 2004

Kirchhofl K, Bilmes J, Das S, Duta N, Egan M, Ji G, He F, Henderson J, Liu D, Noamany M, Schoner P, Schwartz R, Vergyri D (2003) Novel approaches to Arabic speech recognition: report from the 2002 John-Hopkins summer workshop, ICASSP 2003, pp I344–I347

Kuo HJ, Mangu L et al (2010) Morphological and syntactic features for Arabic speech recognition. 2010 IEEE international conference on acoustics speech and signal processing (ICASSP)

Lamel L, Messaoudi A et al (2009) Automatic speech-to-text transcription in Arabic. ACM Trans Asian Lang Inform Process 8(4):1–1822 2 Arabic Speech Recognition Systems

Lee KF (1988) Large vocabulary speaker independent continuous speech recognition: the Sphinx system. Doctoral dissertation, Carnegie Mellon University.

Messaoudi A, Gauvain JL et al (2006) Arabic broadcast news transcription using a one million word vocalized vocabulary. 2006 IEEE international conference on acoustics, speech and signal processing, 2006. ICASSP 2006 proceedings

Mokhtar MA, El-Abddin AZ (1996) A model for the acoustic phonetic structure of Arabic language using a single ergodic hidden Markov model. In: Proceedings of the fourth international conference on spoken language, 1996. ICSLP 96

Muhammad G, AlMalki K et al (2011) Automatic Arabic digit speech recognition and formant analysis for voicing disordered people. 2011 IEEE symposium on computers and informatics (ISCI)

Nofal M, Abdel Reheem E et al (2004) The development of acoustic models for command and control Arabic speech recognition system. 2004 international conference on electrical, electronic and computer engineering, 2004. ICEEC'04

Owen Rambow, David Chiang, et al., Parsing Arabic Dialects, Final Report – Version 1, January 18, 2006 http://old-site.clsp.jhu.edu/ws05/groups/arabic/documents/finalreport.pdf

Park J, Diehl F et al (2009) Training and adapting MLP features for Arabic speech recognition.IEEE international conference on acoustics, speech and signal processing, 2009. ICASSP 2009

Plötz T (2005) Advanced stochastic protein sequence analysis, Ph.D. thesis, Bielefeld University

Rabiner, L. R. and Juang, B. H., Statistical Methods for the Recognition and Understanding of Speech, Encyclopedia of Language and Linguistics, 2004.

Ryding KC (2005) A reference grammar of modern standard Arabic (reference grammars). Cambridge University Press, Cambridge.

Sagheer A, Tsuruta N et al (2005) Hyper column model vs. fast DCT for feature extraction in visual Arabic speech recognition. In: Proceedings of the fifth IEEE international symposium on signal processing and information technology, 2005

Saon G, Padmanabhan M (2001) Data-driven approach to designing compound words for continuous speech recognition. IEEE Trans Speech Audio Process 9(4):327–332.

Saon G, Soltau H et al (2010) The IBM 2008 GALE Arabic speech transcription system. 2010 IEEE international conference on acoustics speech and signal processing (ICASSP)

Satori H, Harti M, Chenfour N (2007) Introduction to Arabic speech recognition using CMU Sphinx system. Information and communication technologies international symposium proceeding ICTIS07, 2007

Selouani S-A, Alotaibi YA (2011) Adaptation of foreign accented speakers in native Arabic ASR systems. Appl Comput Informat 9(1):1–10

Shoaib M, Rasheed F, Akhtar J, Awais M, Masud S, Shamail S (2003) A novel approach to increase the robustness of speaker independent Arabic speech recognition. 7th international multi topic conference, 2003. INMIC 2003. 8–9 Dec 2003, pp 371–376

Singh, R., B. Raj, et al. (2002). "Automatic generation of subword units for speech recognition systems." Speech and Audio Processing, IEEE Transactions on 10(2): 89-99.

Soltau H, Saon G et al (2007) The IBM 2006 Gale Arabic ASR system. IEEE international conference on acoustics, speech and signal processing, 2007. ICASSP 2007

Stanford Log-linear Part-Of-Speech Tagger, 2011. http://nlp.stanford.edu/software/tagger.shtml

Taha M, Helmy T et al (2007) Multi-agent based Arabic speech recognition. 2007 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology workshops

The CMU Pronunciation Dictionary (2011), http://www.speech.cs.cmu.edu/cgi-bin/cmudict, Accessed 1 September 2011.

Vergyri D, Kirchhoff K, Duh K, Stolcke A (2004) Morphology-based language modeling for Arabic speech recognition. International conference on speech and language processing. Jeju Island, pp 1252–1255

Xiang B, Nguyen K, Nguyen L, Schwartz R, Makhoul J (2006) Morphological ecomposition for Arabic broadcast news transcription. In: Proceedings of ICASSP, vol I. Toulouse, pp 1089–1092