المجلة العلمية لجامعة الملك فيصل
The Scientific Journal of King Faisal University

العلوم الإنسانية والإدارية
Humanities and Management Sciences

Developing a Stress Prediction Tool for Arabic Speech Recognition Tasks

Eiman Alsharhan and Salah Alnajem
Department of Arabic Language and Literature, Kuwait University, Kuwait City, Kuwait

تطوير أداة لِتمييز مواضع النبر في الكلمات لأنظمة التعرف الآلي على الكلام العربي

إيمان الشرهان وصلاح الناجم
قسم اللغة العربية وآدابها، جامعة الكويت، المدينة الجامعية، الكويت

LINK الرابط	RECEIVED الاستقبال	ACCEPTED القبول	PUBLISHED ONLINE النشر الإلكتروني	ASSIGNED TO AN ISSUE الإحالة لعدد
https://doi.org/10.37575/h/lng/2323	24/04/2020	03/12/2020	03/12/2020	01/09/2021
NO. OF WORDS عدد الكلمات	**NO. OF PAGES عدد الصفحات**	**YEAR سنة العدد**	**VOLUME رقم المجلد**	**ISSUE رقم العدد**
7200	7	2021	22	2

KEYWORDS
الكلمات المفتاحية
Suprasegmental features, stress, phonetic transcription, automatic speech recognition
الخصائص فوق القطعيّة، النبر، الكتابة الصوتية، نظام التعرّف الآليّ على الكلام

ABSTRACT

Developing natural language processing applications for Arabic must consider the different linguistic characteristics found in speech and translate those characteristics to script in order to reduce computational complexity and therefore reduce the word error rate (WER). Suprasegmental features are fundamental properties of speech that can enhance the performance of many natural speech processing applications. The present study considered stress as a prosodic feature comprising the prominence of syllables in speech by developing a tool that generated phonetic transcriptions and predicted the stress position. The generated transcription was used to create the phonetic dictionary necessary for developing an automatic speech recognition (ASR) system. This tool had to be accurate, linguistically motivated, and applicationally useful; therefore, the effectiveness of the generated stress-marked phonetic dictionary was tested by comparing the performance of a standard fixed dictionary-based system with that of one using the automatically generated dictionary. The research reported a 5.6% reduction in WER when using a dictionary with stress markers attached to each phone in the stressed syllable and a 3.5% reduction in WER when using a dictionary with stress markers assigned only to stressed vowels. These results encourage future studies to employ prosodic features of speech when developing different speech processing applications.

الملخص

يتطلّب تطوير تطبيقات المعالجة الآليّة للغة العربيّة مراعاة الخصائص اللغويّة المختلفة الموجودة في الكلام، وترجمة تلك الخصائص الكلاميّة كتابيًّا، من أجل تقليل التّعقيد الحسابيّ، وبالتالي تقليل معدّل الخطأ في التعرّف على الكلمات. تُعتبر الخواصّ الكلاميّة فوق القطعيّة -كالنبر- من الخصائص الأساسيّة للكلام، والتي يمكنها بدورها أن تعزّز أداء العديد من تطبيقات المعالجة الآليّة للكلام. تستهدف الدّراسة الحاليّة تطوير أداة تَقوم بتقسيم الكلمة إلى المقاطع التي تتكوّن منها، وتمييز المقاطع المنبورة بشكل آليّ، وسيتمّ استخدام مخرجات هذه الآلة في بناء القاموس الصّوتيّ اللازم لتطوير نظام التعرّف التلقائيّ على الكلام العربيّ. يجب أن تكون هذه الآلة دقيقة ومبنيّة على أسس لغويّة صحيحة حتّى تكون مفيدة تطبيقيًّا، ولاختبار فعاليّة ودقّة الآلة المطوَّرة قامت الدّراسة بتطوير أنموذجين مختلفين لبرنامج التعرّف الآليّ على الكلام العربيّ بهدف مقارنة أدائهما. يَستخدِمُ الأنموذجُ الأوّل في بناء النّموذجِ الصّوتيّ للنّظام القاموسَ الصّوتيّ الأساسيّ (بدون نبر)، بينما يعتمد الأنموذجَ الثاني على القاموس المولَّد آليًّا باستخدام الأداة التي قمنا بتطويرها. تشير النتائج إلى تفوّق استخدام القاموس المولَّد آليًّا على الاستخدام التقليديّ للقاموس الصوتيّ، وذلك بتخفيض معدّل الخطأ في التعرّف على الكلمات بنسبة (5.6٪) عند تمييز كلّ محتويات المقطع المنبور، وتحسّن بنسبة (3.5٪) عند استخدام القاموس الذي يميّز الحركات المنبورة فقط. تُعتبَر هذه النتائج مشجّعة للقيام بدراسات مستقبليّة تقوم بتوظيف المميّزات العروضيّة للكلام عند تطوير تطبيقات معالجة الكلام الآلية المختلفة.

1. Introduction and Objectives

Arabic is the fifth most widely spoken language in the world; in 2017, estimates suggested that Arabic is spoken by over 350 million people (Lewis, 2015). Arabic is the primary language used in the 22 countries that comprise the Arab league, and the Arabic language has multiple variants, including five main dialects (Gulf, Iraqi, Egyptian, Levantine, and North African). However, modern standard Arabic (MSA) is the standardised dialect used in media, formal correspondence, and education across the Arab world.

The Arabic language, however, poses several challenges for developing automatic speech recognition (ASR) systems. The current study is concerned with the problem of mismatch between the provided phonetic transcription and the speech signal, which leads to the deterioration of the performance of the system. This problem was tackled by employing the stress feature as a suprasegmental characteristic of Arabic speech to develop a set of precise word stress rules. The motivation behind this work is to minimise the gap between the textual form and the speech during the acoustic modelling to enhance the quality of the recogniser. The study also aims at determining the best way to exploit the stress features during the creation of the phonetic dictionary. This is done by comparing the performance of the ASR system when stress markers are added to each phone within the stressed syllables to adding stress marks only to stressed vowels.

Lexical stress is the relative emphasis or compression given to a certain syllable within a word to make this syllable more clearly perceived than other syllables of the same word. However, stress is a non-phonemic feature in Arabic, which means it cannot be used to distinguish meaning; phones are articulated in a different way when stressed (Betti, 2018). Researchers have confirmed that well-defined stress rules can help improve the performance of various Arabic speech processing applications (Alsharif, 2016). Previous research has confirmed that vowels, which construct approximately 40% of speech, are articulated differently if they are part of a stressed syllable (Halpern, 2009; Vetulani, 2011).

Generally, stress is predictable in Arabic and is assigned according to the syllabic structure of the word (Holes, 2004). Therefore, the current study developed a tool that could automatically assign a stress marker to stressed syllables. The significance of the developed stress prediction tool lies in its ability to automatically generate accurate and fine-grained phonetic transcriptions. The generated transcriptions are used to build a prosodic word dictionary which is essential for a wide range of technologies.

The effectiveness of incorporating the stress value during acoustic modelling was tested by comparing the performance of different models that use different sources of linguistic knowledge. Additionally, the effectiveness of the alternative dictionary was tested by running different Hidden Markov Models and deep neural network (HMM–DNN) based experiments using version 3.5 of the Hidden Markov

Corresponding Author: Eiman Alsharhan eiman.alsharhan@ku.edu.kw ,0096599635445 المؤلف المراسل: إيمان الشرهان

Model ToolKit (HTK). [1] The findings of this study suggested that assigning stress positions in speech is crucial not only for speech recognition applications but also for other natural language applications, such as speech syntheses. Although stress is not a phonemic feature in Arabic, mis-assigning stress positions generates an unnatural utterance.

2. Relationship to Previous Work

This study was an extension of previous efforts to integrate linguistics information and enhance ASR system performance (Alsharhan, 2019). Many researchers in this field have recently attempted to integrate different linguistic knowledge sources into state-of-the-art ASR systems to improve their performance. This section reviews some of these works, giving special attention to those that have been specifically dedicated to the Arabic language.

Integrating prosodic information has been found to be useful in different natural language processing applications, such as emotion recognition, accent classification, speech synthesis, and speech recognition. A large and growing body of literature has investigated the influence of integrating prosodic information on the development of emotion and sentiment recognition systems. For instance, Meftah (2016) and Mannepalli (2018) extracted and compared some significant acoustic features from speech to determine the most significant acoustic feature that should be used in developing emotion recognition systems. Their results emphasised the importance of employing certain acoustic features in order to enhance the performance of these systems.

Integrating prosodic features was also found to be useful in developing language identification tools (Chittaragi, 2018). For instance, Reddy (2013) extracted spectral and prosodic features to analyse language-specific information found in speech. The features included in this study were intonation, rhythm, and stress. The evaluation results confirmed that language identification tool performance is improved by 3% and 6% when prosodic features are employed. Similarly, another study combined prosodic and formant information to build a generative language identification system; the study shows that the inclusion of formant information leads to around 50% relative improvement in the performance of the language identification system (Martinez, 2013).

Identifying Arabic dialects with prosodic information has been investigated by a few researchers, including Lounnas (2018) and Ibrahim (2019). The study by Lounnas (2018) targeted the distinctions between MSA and Algerian dialects and reported an improvement of 1.67% in performance when combining prosodic and acoustic features in the designed system. Ibrahim (2019) focused on identifying the accent of Malay speakers when reciting the Quran by using a combination of spectral and prosodic features from the speech in order to determine the variability of accents; they reported 5.5%–7.3% improvement when the prosodic is integrated with Mel-frequency cepstral coefficients (MFCC) compared to MFCC alone.

A few studies have investigated the significance of spectral and prosodic behaviours of speech signals for the purpose of developing a well-performing Arabic ASR system. The studies have confirmed that the employment of prosodic features can help enhance Arabic speech technology applications (Brierley, 2019). Amrous (2011) investigates the contribution of formants and prosodic features, such as pitch and energy, in Arabic speech recognition under real-life conditions using the HTK Toolkit. Khelifa (2017) addresses the integration of complementary features into standard HMMs, aiming to build a robust speech recognition system. A series of experiments with different features combinations was carried out in this research to determine which of these integrated features have the highest effect on the systems performance when using HTK. The experimental results showed a noticeable reduction in WER.

3. Arabic Phonology

Like most Semitic languages, Arabic has a relatively rich consonantal inventory but a limited vocalic one (Watson, 2002; Holes, 2004). For instance, the phonemic inventory of MSA consists of 34 phonemes, six of which are vowels. These sounds are distributed in syllables. Some words contain only one syllable (monosyllabic), some contain two syllables (disyllabic), while others contain more (polysyllabic).

MSA has explicit phonotactic rules for syllable structure. Below is brief review of these rules:

- Syllables cannot start with a vowel.
- Syllables cannot start with a consonant cluster.
- A cluster of three consonants or more is not acceptable anywhere in a syllable.

Understanding the syllabic structure is essential to understanding stress rules. That is because stress position depends mainly on the number and weight of containing syllables. Three categories of syllables can be found in Arabic: light, heavy, and superheavy. These categories result in having six structural types as shown in the following table:

Table 1: Syllable structure of spoken Arabic

1	Cv	/ka-ta-ba/ كتب(wrote)	Light syllable
2	CvC	/mak-tab/ مكتب (office)	Heavy closed syllable
3	Cv:	/ka:-tib/ كاتب (writer)	Heavy open syllable
4	CvCC	/bint/ بنت (girl)	Superheavy syllable
5	Cv:C	/ba:b/ باب (door)	Superheavy syllable
6	Cv:CC	/ ħa:dd/ حاد (sharp)	Superheavy syllable

A word can only have one superheavy syllable, which always occurs at the end of the word. An exception to this rule is those cases where a long vowel and a geminated consonant occur next to each other, such as */ħa:d-da/ حادَّ* , where the superheavy syllable appears in the word initially.

Stress is a suprasegmental feature of speech that gives emphasis to a certain syllable in a word. Stress is a non-phonemic feature in Arabic, which means it cannot affect the meaning of the word. However, stress position strongly affects the naturalness of the speech. Placement of stress is dependent on the nature of syllable structure of the word as mentioned earlier. The stress rules presented here are based on considerable research into the literature and on observations of how MSA is actually spoken by the different speakers from our dataset. Before exploring the stress rules in Arabic, two facts need to be mentioned:

- Stress is always measured from the end of an Arabic word. In addition, stress never falls further back than the third syllable from the end of the word (antepenult) (Ryding, 2014).
- Proclitics, which include a definite article, and prepositions attached to the beginning of a word, are ignored when counting the syllables of a word. For instance, the word */wa-lam/* (and did not) ولم is considered to be a monosyllabic word when assigning stress, while the word */qa-lam/* (pen) قلم is considered a disyllabic word. On the other hand, although /ʔ*al-wa-lad/* (the boy) الولدand */wal-wa-lad/* (and the boy) والولد are obviously polysyllabic, they are stressed like disyllabic words because the proclitics */wa/* and /ʔ*al/* are ignored in counting syllables.

Accepted stress patterns for MSA that comply with the following works (Angoujard, 1990; Holes, 2004; Halpern, 2009) are:

1. Stress falls on the ultimate syllable if:
   - The ultimate syllable is superheavy. For example, */ʃa-di:d/* (strong) شديد and /ʔ*a*tˤ*-fa:l/* (kids) *أطفال*.
   - The word is monosyllabic. By saying ‘monosyllabic’, we also include disyllabic words that have proclitics; those proclitics are ignored for stress assigning purposes. For example:

[1] http://htk.eng.cam.ac.uk

  - **/la:/** (no) لا
  - **/min/** (from) من
  - /wa-**lam/** (and did not) ولم
  - /bi-**kam/** (how much) بكم

2. Stress falls on the penultimate syllable if:
   - The word is disyllabic. For example:
     - **/si**-wa:/ (except) سوى
     - **/ba**-lad/ (town) بلد
     - /ʔal-**ba**-lad/ (the town) البلد
     - /wal-**ba**-lad/ (and the town) والبلد
   - The word is polysyllabic and the penultimate syllable is heavy. For example:
     - /sa-**di:**-dun/ (correct) سديد
     - /ba-**na:**-ti:/ (my girls) بناتي
     - /sa:-**ʕa:**-ti:/ (my watches) ساعاتي
     - /ku-waj-**tij**-jun/ (Kuwaiti) كويتي
3. Stress falls on the antepenultimate syllable if the penultimate is light in polysyllabic words. For example:
   - **/da**-ra-sa/ (he studied) درس
   - **/da:**-ri-su/ (a learner) دارس
   - **/da**-ra-sat/ (she studied) درست
   - /wa-**ra**-qa-tun/ (a paper) ورقة
   - /mad-**ra**-sa-tun/ (a school) مدرسة

McCarthy (1990) summarised the basic rules for stress placement in Arabic in the following statement, 'the stress system is obviously weight-sensitive: final syllables are stressed if super heavy CvvC or CvCC; penults are stressed if heavy Cvv or CvC; otherwise the antipenult is stressed.'

4. Architecture of the Developed Stress Prediction Tool

This section seeks to explain the development of the stress prediction tool. Acquiring information about lexical stress requires many initial steps. The details of which are given below:

4.1. Pre-Processing the Text

This is an essential front-end requirement for any system that transcribes text. Basically, it essentially manipulates the information from textual input and prepares the text to be further processed by the system. This includes: restoring the missing diacritics, deleting case markers, dealing with symbols such as *sukoon* and *shadda,* deleting letters which do not match any sound in the pronunciation of the word, unifying the graphemic representation of identical sounds, and labelling the graphemes with the consonant and vowel value.

4.1.1. Diacriticising the text

In the Arabic orthography system, short vowels do not have letters but are instead written as diacritic marks above or under the letters. These diacritics are often omitted from the text and it is the job of the reader to retrieve them by referring to the contextual features and information. This is different from most other alphabetic languages, such as English, where all vowels are indicated in the script.

The absence of diacritics in Arabic texts is a crucial problem for developing our stress prediction tool. The issue here is that the placement of the stress depends mainly on the syllabic structure of the word. Given the fact that the nucleus (which forms the core of the syllable) is a vowel, it is essential to retrieve these vowels in order to syllabify words and assign stress accordingly.

In addition, diacritics do not only include short vowels. They also include other marks such as *shadda* (consonant gemination mark), *tanween* (nunation), and *sukoon* among others. Those diacritics are important in identifying the pronunciation of words with similar forms. For example, the word **kr* ذكر when diacriticised can be*:*akar* (male) ذَكَر, **ikr* (citation) ذِكر, **akkir* (remind) ذكِّر, **ukir* (mentioned) ذُكِر, or **akir* (remembered) ذَكَر.

In this research, the state-of-the-art morphological analysis tool MADAMIRA[2] is used to obtain diacritics for the input text (Pasha, 2014). Running MADAMIRA on GALE data transcripts has shown that it succeeded in recognising 94.3% of the input words.

4.1.2. Removing case markers

Finally, diacritics can also appear in words as short vowels or *tanween* (case markers) to demonstrate the syntactic function of the word in the sentence (e.g. whether the word is the subject or object of a verb). Different case endings are used to indicate the grammatical function of the word. Obviously, correctly assigning case markers is a challenging aspect of Arabic grammar which requires a solid foundation of knowledge.

There are two issues present: firstly, MADAMIRA does not always assign the right case markers to the words. A probable cause for this issue is the nature of the input text, which is a transcript of spoken language. Given that the majority of the transcribed text is for spontaneous speech including all its irregularities, it is necessary to understand the complications faced while doing a syntactic analysis of the text. In addition, the input text has a great deal of dialectal variation with speakers from all around the Arabic regions. It is well-known that Arabic dialects vary greatly in all linguistical aspects.

Secondly, assigning the right case marker to each word in Arabic requires a comprehensive knowledge of Arabic grammatical rules, which many people do not have. As a result, people tend to drop the case marker to avoid making errors. Dropping case markers is a well-known phenomenon in Arabic called *taskeen*. A previous study found that removing case markers from text can yield to an approximately 4% drop in WER (Alsharhan, 2020).

4.1.3. Dealing with *sukoon* and *shadda*

Arabic orthography includes two diacritics that have no phonetic realisation, namely *sukoon*, as in the word بيْت, and *shadda*, as in the word سيِّد. *Sukoon* has no phonetic realisation. It just indicates that the consonant to which it is attached is not followed by a vowel. *Shadda*, on the other hand, is used to duplicate the previous consonant (geminating). For the purpose of this work, all *sukoon* instances are deleted and *shadda* is replaced with a second copy of the consonant to which it is attached. This step is crucial to obtaining the accurate syllabification of the words.

4.1.4. Unifying *hamza* variants

Hamza (glottal stop) appears in Arabic texts in different forms depending on its position in the word and the surrounding vowels. For instance, it can be written in five ways: on its own ء, under an *Alif* إ, or over an *Alif* أ , over *wAw* ؤ, or over *yA* ئ. However, the way that *hamza* is written does not affect the way it is pronounced. Therefore, the phonetic transcription of *hamza* has been unified no matter how it is written in the orthographic form of Arabic.

4.1.5. Disambiguating the semi-vowels

The Arabic orthography system has two graphemes, each of which represents two different sounds. The grapheme (w) و represents both the long vowel /u:/, as in *tqwl* تقول/taqu:l/ (she says), and the consonant /w/ as in *waraq* ورق /waraq/ (paper). In addition, (y) ي is used to represent the long vowel /i:/ as in *kbyr* كبير /kabi:r/ (big) and the consonant /j/ as in *yqwl* يقول /jaqu:l/ (he says). Assigning a different grapheme for each phonetic representation is crucial for the syllabification of the words which depends mainly on information about consonants and vowels. Ignoring this step is likely to cause a

[2] MADAMIRA is a toolkit that provides linguistic information such as tokenisation, lemmatisation, diacritisation, and parts of speech tagging. It contains models for both MSA and Egyptian. What sets MADAMIRA apart from similar tools is that it takes word context into account, which makes the generated analysis more accurate. A non-commercial license is freely available at: http://innovation.columbia.edu/technologies/CU14012

great deal of confusion in syllabification.

For this reason, it is crucial to assign the right phonetic class for each grapheme. For instance, when (w) occurs in a consonantal position, it is transcribed as /w/, otherwise it is transcribed as /u:/. This is not a straightforward task, since deciding whether one of these graphemes is pronounced as a vowel or a consonant depends on various properties of the surrounding context.[3]

4.1.6. Removing silent letters

The Arabic orthography system includes some letters that do not correspond to any sound in the word's pronunciation. An example of this is the silent *Alif*, which follows the *group wAw* indicating masculine plural morpheme in verbs. For example, the verb *daraswA* (they studied) درسواis pronounced /darasu:/. Another example is the silent *Alif* that appears the *fatH* nunation, e.g. *dArisFA* (a learner) دارسا which is pronounced as /da:risan/.

This step also involves the elimination of short vowels that precede a long vowel. Such short vowels typically appear before their long vowel counterparts in automatically diacritised texts. An example of this is *yaktubuwn* (they write) /jaktubu:n/ and *taktubiyn* (you write) /taktubi:n/. Since these short vowels are not phonetic, they need to be omitted from the text.

4.2. Grapheme to Phoneme Mapping

This step is necessary to inspect the graphemes and convert them into phonemes by applying a set of one-to-one rules. This is a straightforward conversion that aims at mapping Arabic graphemes to their aligned phonemes. In the rules provided by the Buckwalter transliteration scheme (Habash, 2007) and the Speech Assessment Methods Phonetic Alphabet (SAMPA) (University College London, 2002), a phonetisation scheme is employed to present the letters and the sounds of the language, respectively. The preference for using the SAMPA notation over other popular notation systems such as IPA in writing the program can be explained by the fact that IPA is not an ASCII compatible scheme.

4.3. Applying Phonological Rules

In writing the phonological rules, the focus was on the changes that affect the syllabic structure of the word. Therefore, only rules that cause the deletion or insertion of sounds were included.

4.3.1. Deletion

The main deletion process in Arabic can be observed in the deletion of *hamzat AlwaSl* همزة الوصل. *hamzat AlwaSl* can appear as part of the definite article *Al*ال*, e.g. Alwalad* (the boy) الولد, it can also appear as a verb or noun initial, e.g. *Aktub* (write) اكتب and *AbtisAmap* (a smile) ابتسامة . *Hamzat AlwaSl* is deleted when it is preceded by a vowel. For example, a phrase like *wa Aktub* (and write) واكتب is pronounced as /waktub/.

4.3.2. Insertion

Insertion is the process of inserting a phonetic element into a string without providing an orthographic representation, sometimes referred to as 'epenthesis'. Insertion is another phenomenon that can be seen in Arabic which leads to changing the syllabic structure of the word. The general rule in pronouncing two consecutive words in MSA (when the first word ends with a consonant and the second word starts with *hamzat AlwaSl*) is to add a short vowel /i/ after the first word and to remove *hamzat AlwaSl* as stated earlier. This can be seen in phrases such as *man Almutakallim* (who is talking) من المتكلم /manil mutakallim/. The short vowel /i/ is inserted in compliance with MSA's basic phonetic rule that does not allow having two consecutive consonants except at the end of the utterance. The *hamzat AlwaSl* deletion process takes place after the insertion of the short vowel.

4.3.3. Shortening long vowels

As the name implies, this process is concerned with shortening long vowels in some specific contexts. For instance, when there are two words and one of them ends with a long vowel and the other begins with *hamzat AlwaSl*; the long vowel must be alternated with its shorter version. The reason behind this alternation is that there would be super heavy syllables (CV:C) in the middle of the utterance by deleting *hamzat AlwaSl*. This type of syllable is not allowed in the middle of the utterance, and for this reason it is shortened so the resulting syllable is (CVC). An example of this case is *fy Almadrasap* (at the school) في المدرسة , which is pronounced as /fi lmadrasa/.

4.4. Syllabification

Before the process of assigning stress into syllables, a set of rules is applied to syllabify the words. Syllabification is the process of dividing a word into its constituent syllables. All spoken languages have their own rules that control the construction of syllables and their allowed sequences. MSA has explicit structural restrictions on syllables based on the distribution of sounds as discussed in Section 3. In this step, the phonotactic constraints of syllable structure for Arabic are followed for the purpose of developing a set of rules that can be applied to the text generated from the previous steps. Generally speaking, the syllabification of the word depends mainly on the core of the syllable, which is a vowel. In addition to a vowel, a syllable contains consonants that are either prevocalic or postvocalic or both. After the application of this step, words are segmented into the containing syllables and it is possible to look at the stress position.

4.5. Aligning Stress into Syllables

After the application of all processes mentioned previously, a set of rules is introduced to scan the word's syllables and assign the stressed syllable with a stress value. It can be noted that the process of incorporating stress markers comes at the end after a long process of text normalisation, phonological alternations, and syllabification. This can be justified as the stress position primarily depends mainly on the internal structure of the syllables that make up the word, which can significantly alter after the application of the phonological rules. The stress rules applied here are based on MSA's standard pronunciation.

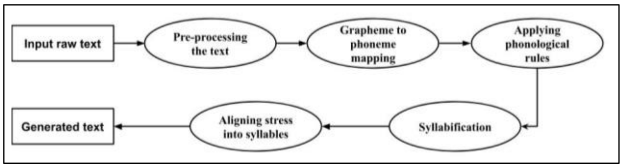
The algorithm for determining stressed syllables in a text transcript is based on a set of rules provided in Section 3. The following conditional rules are applied consequently:

- If the word is monosyllabic then the whole word will be stressed;
- If the word has a word-final super heavy syllable (CVVC or CVCC), then this syllable must be stressed;
- If the word is disyllabic, then stress falls on the penultimate syllable;
- If the word is polysyllabic, and the penultimate syllable is heavy (CVV or CVC), then the stress falls on the penultimate syllable;
- Otherwise, the stress falls on the antepenultimate syllable.

Proclitics, such as *wa-* and *fa-*, are known to pose challenges for automatic stress assignment. To avoid mis-assigning stress, the morphological analysis provided by MADAMIRA is used to identify instances of proclitics. This element will then be disregarded when applying stress assignment rules. Figure 1 summarises the architecture of the developed stress prediction tool.

Figure 1: Architecture of the developed stress prediction tool

[3] **This is particularly challenging when these characters appear adjacent to one another, since it is necessary to know whether the first is a consonant or a vowel before it can be decided which class the second belongs to; but in order to know whether the first is a vowel or a consonant one needs to know which class the second belongs to.**


Input raw text
Pre-processing the text
Grapheme to phoneme mapping
Applying phonological rules
Syllabification
Aligning stress into syllables
Generated text

5. Experimental Design

The development of the ASR system reported in this paper is based on the HTK 3.5 (Young, 2015), which is a portable toolkit for building HMMs. This version integrates DNN modules to be used for acoustic modelling and feature extraction. The Linux operating system Ubuntu has been used for developing the recognition systems reported in this paper. In addition, a Python script was developed to allow the systematic application of HTK with different settings and for the creation of the required files. Details about experimental design are given below.

5.1. Dataset

This study uses the GALE (phase 3) Arabic broadcast news and broadcast conversational speech dataset. This dataset is composed of two major parts: the first part contains about 132 hours of Arabic broadcast news speech (BN) collected from 13 Arabic channels, while the second part contains about 129 hours of Arabic broadcast conversation speech (BC) collected from 17 TV channels.

This data is used for training and testing the developed ASR systems. However, the transcripts have been revised and orthographic normalisation was applied to compensate for some of the orthographic errors found in the data. In addition, the transcripts were segmented into manageable, well-defined segments according to the time stamps provided. Each segment was given a specific label that matches the speech file label. All redundant information, such as non-speech segments and non-Arabic texts, was removed.

5.2. Feature Extraction

Audio files need to be represented in a more compact and efficient way, as speech modelling tools cannot process raw waveforms. This is done by converting them into a series of acoustical feature vectors. This front-end step is crucial to identifying the useful components of the audio signals for recognising the linguistic content. The literature shows that there is a variety of feature extraction techniques that can be used for this exact purpose. The research uses Mel Frequency Cepstral Coefficients (MFCC) of the speech samples to extract the speech feature vectors. Alternative techniques such as filterbank features and perceptual linear predictive coding are reported (Wang, 2016), but the differences between the various representations seem to be marginal. However, the use of MFCCs is predominant (Sharma, 2014). Obtaining MFCCs requires a sequence of steps to be applied to an input speech signal. These computational steps of MFCC include Framing, Windowing, DFTH, a Mel filter bank algorithm, and computing the inverse of DFT. The speech signal was then converted into a discrete sequence of feature vectors.

The feature vector consists of a collection of MFCC coefficients. Most researchers use the standard 39 MFCC vectors (12 cepstral features, plus an energy feature, 12 delta-cepstral coefficient features plus delta energy coefficient features, and 12 double-delta-cepstral coefficient features plus double-delta energy coefficient features). A previous study found that the use of 25 MFCC vectors (without acceleration) was superior to the use of the standard 39 MFCCs (Alsharhan, 2019). Given that information, this research uses 25 MFCC vectors when extracting the features.

5.3. Acoustic Model

The DNN tools in HTK 3.5 allow for the use of DNNs for constructing acoustic models, i.e. for identifying the phoneme corresponding to a given set of acoustic features. The output of the DNNs is eventually converted to a probability distribution, typically by using softmax, and used as the 'emission probability' in an HMM. This is similar to the way that DNNs are used in other state-of-the-art toolkits, such as KALDI (Ali et al.,, 2014). A previous study reported that using DNNs in this way within the HTK gives results that are comparable with, and in some cases better than, other DNN-based systems tested using the same data (Wang, 2016).

The actual acoustic modelling takes place in multiple stages, starting with the creation of an initial set of identical monophone HMMs. This intialisation step is followed by creating short-pause models and extending the silence model to make the developed system more robust. Phone models are created and aligned with the acoustic data in subsequent rounds of training.

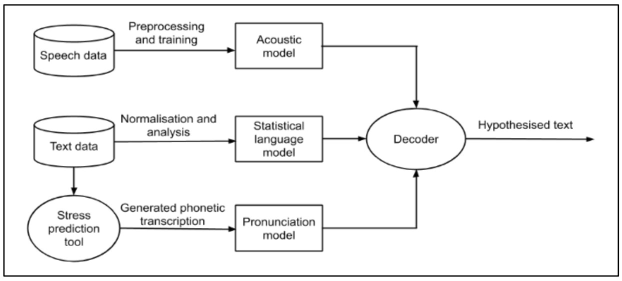
Following the creation of a set of monophone HMMs, context-dependent triphone HMMs are created. This is basically done by generating a list of all triphones observed in the training data and tying similar acoustic states of these triphones to make robust parameter estimates.

In developing the DNN–HMM systems, similar to the monophone system creation explained previously, a prototype model must first be defined. Next, other tools from the HTK work by connecting the DNN output units with the monophone HMM states. In addition, the DNNs created in this step are paired with the cross-word triphone HMMs and eventually evaluated.

5.4. Pronunciation Model

The pronunciation model provides the link between the language model and the acoustic model, as Figure 2 shows. Well-designed pronunciation models contribute greatly to the robustness of the recognition system. Such models can help boost the performance by shrinking the mismatch between the speech and text used in building the acoustic model. The pronunciation model is created with the aid of the transcription tool explained in Section 4. The outcome of this tool is listed in the pronunciation dictionary which includes a list of the words that are present in the language model with their phonetic representation. In this paper, the standard use of the predefined dictionary (namely, the Qatar Computing Research Institute (QCRI) dictionary with multiple phonetic transcriptions) is compared with the use of the generated dictionary. Two versions of the dictionary are created and tested: a phoneme-based dictionary where only stressed vowels are marked, and the syllable-based dictionary where the whole stressed syllable is marked.

Figure 2: The architecture of the developed ASR system when using the stress prediction tool


Speech data
Preprocessing and training
Acoustic model
Text data
Normalisation and analysis
Statistical language model
Decoder
Hypothesised text
Stress prediction tool
Generated phonetic transcription
Pronunciation model

5.5. Language Model

The language model is an essential component of the ASR system. It works on capturing the properties of the language in order to constrain search by limiting the set of possible HMMs. A language model supports the work of the acoustic model by predicting the probability of a word occurring in the context during recognition, which can help to improve the overall performance of the system.

The research reported here uses bi-gram language modelling, which works probabilistically by computing the likelihood for each possible successor word. This is carried out using a set of HTK tools, firstly to create the grammar and then to produce a word network that lists each word-to-word transition.

6. Results and Discussion

The experimental work reported in this study aims at answering the following questions:

- Does the inclusion of a stress feature in the transcription improve the overall performance of the ASR system?
- What is the best way to add stress markers in the dictionary? Should the whole stressed syllable be marked or just the stressed vowels?

The first set of experiments is aimed at answering the first question. The whole corpus is used without stress annotations in one round and with stress annotations included in the next round. Other experiments are carried out to compare the performance of the ASR system when a stress marker is added to the whole stressed syllable and when stress is added only to stressed vowels.

The study uses a five-fold cross validation approach as a way to evaluate the proposed schemes for optimal use of data and to prevent biased testing. This includes partitioning the data randomly into five equal size subsets to execute the training on four subsets and validate the system using the other subset. This process is replicated five times, with each of the five subsets used only once as the validation data. The results from the five folds can then be combined to determine a single calculation. The benefit of this method over the standard evaluation approach is that all observations are used for both training and testing, providing a more rigorous testing.

6.1. Baseline System

The baseline system is developed using a multiple pronunciation dictionary provided by QCRI[4]. This dictionary has over 526K unique grapheme words, with 2 million pronunciations, with an average of 3.84 pronunciations for each grapheme word. The baseline system achieved 13.4% WER as shown in Table 2.

6.2. Including a Stress Marker for Every Stressed Syllable in a Syllable-based Model

In this phase of experimentation, the whole stressed syllable is marked in the dictionary. The dictionary is generated using the developed stress prediction tool. This dictionary is syllable-based, which means that each word is syllabified and only the stressed syllable is marked. An example of this dictionary can be seen in Figure 3.

The motivation behind using syllables as a unit of training is that the pronunciation variation and co-articulation effects can be captured directly and train the acoustic model accordingly. The study has also been motivated by other studies, such as Azmi (2008), which used syllables as units in developing ASR systems and reported promising results. Another source of motivation are some studies that focused on the human perception of speech such as Räsänen (2015). These studies demonstrated the central role that syllables played in human perception and speech generation. Developing a syllable-based system and marking stressed syllables has led to significant improvement with 7.8% WER.

QCRI�multi-pronunciation�dictionary		generated�syllable-based�dictionary	
AlslAm�	A�s�a��A�m	AlslAm��	Qas�sa�laa^�mu
AlslAm�	A�s�a��A�m�a	Elykm�	�Ea�lay^�kum�
AlslAm�	A�s�a��A�m�i	>EzA'nA��	Qa�Eiz�zaa^�Qa�naa
AlslAm�	A�s�a��A�m�u	Alm$Ahdyn��	Qal�mu�$aa�hi�diin^
Elykm�	E�a��a�y�k�u�m		
>EzA'nA�	G�a�E��z�A�G�a�n�A	generated�phoneme-based�dictionary	
>EzA'nA�	G�a�E��z�A�G���n�A	AlslAm�	Q�a�s�s�a�l^�aa^�m�u
>EzA'nA�	G�a�E��z�A�G�u�n�A	Elykm��	E�a�l^�a^�y^�k�u�m
Alm$Ahdyn�	A��m�u�$�A�h�i��d��y�n	>EzA'nA�	Q�a�E��z�z^�aa^�Q�a�n�aa
Alm$Ahdyn�	A��m�u�$�A�h�i��d��y�n�a	Alm$Ahdyn��	a��m�u�$�aa�h��d^��i^�n^

Figure 3: An example of the phonetic transcription for the input: (AlslAm Elykm >EzA'nA Alm$Ahdyn) السلام عليكم أعزاءنا المشاهدين as found in the QCRI multi-pronunciation dictionary and the generated syllable-based and phoneme-based dictionaries.

6.3. Including a Stress Marker for Every Stressed Vowel in a Phoneme-based Model

This method is inspired by the belief that stress affects vowels more than consonants (De Jong, 1999; Biadsy, 2009). Another source of motivation is a study on speech synthesis that confirms the average length of utterances synthesised by systems that include stress features was much closer in duration to the natural utterances. This is a strong indicator that marking stressed vowels can significantly improve the recognition of speech (De Jong, 1999).

Stressed vowels are marked as different phones in the dictionary compared to their unstressed counterparts. Examples are given in Figure 3. This system is developed using a dictionary with stressed vowels marking the reported significant improvement compared to the baseline system with 9.9% WER. However, it did not compare well against the performance of the syllable-based system reported previously.

Table 2: The performance of the ASR system with different levels of stress marking

ASR system	WER
Baseline system (no stress marking)	13.4%
Stress marker attached to stressed syllable	7.8%
Stress marker attached to vowels	9.9%

Generally speaking, including information about stress positions during acoustic modelling was found to be invaluable in enhancing the performance of the ASR system developed for Arabic.

Conclusion

This research presented the development and evaluation of a tool that outputs a phonetically transcribed text including stress marking for Arabic materials. This tool will be useful for generating the phonetic transcriptions that are necessary for developing different speech processing applications, such as an automatic speech recognition system. This work also reported on the steps included in the development of this tool. For evaluation purposes, the first question sought to verify the effectiveness of employing the proposed tool in the development of Arabic ASR systems. In addition, this study set out to investigate the best way of including stress markers in building the dictionary. Two methods were investigated: marking every phoneme in the stressed syllable and marking only stress vowels. The research found that using the developed stress prediction tool to generate the phonetic dictionary led to significant improvement with 5.6% and 3.5% reduction in WER compared to the baseline system. The research also found that marking the whole stressed syllable is more effective than marking only stressed vowels. The research findings have important implications for developing many natural language processing applications such as text-to-speech systems, computer-assisted language learning systems, and emotion recognition systems.

[4] http://alt.qcri.org/resources/speech/dictionary/arar_lexicon_2014-03-17.txt.bz2

Biographies

Eiman Alsharhan

The Department of Arabic Language and Literature, Faculty of Arts, Kuwait University, Kuwait City, Kuwait, eiman.alsharhan@ku.edu.kw, 0096599635445

Dr Alsharhan has a PhD in natural language processing from the School of Computer Science, University of Manchester, UK. She has published many research papers with top ranking journals, such as: *Computer Speech and Language*, *Information Processing and Management*, *International Journal of Speech Technology*, and *Language Resources and Evaluation*. She has participated in many international conferences in the United States, Canada, England, Spain, and UAE. Her research interests include phonetics and phonology, linguistics, computational linguistics, speech recognition, and speaker identification. Orcid iD: https://orcid.org/0000-0001-6351-3805

Salah Alnajem

The Department of Arabic Language and Literature, College of Arts, Kuwait University, Kuwait City, Kuwait, salah.alnajem@ku.edu.kw, 0096590097970

Dr Alnajem is an associate professor and the founder and CEO of Information Age for I.T. Consultations. He received his master's and PhD degrees from the University of Essex in England. His scientific interests are: computational linguistics, natural language processing, text processing, text mining, big data analytics, and information retrieval. He has authored a number of research papers published by Elsevier and other academic publishers. Dr Alnajem won the prize of the World Summit on the Information Society (WSIS) 2017 in the e-learning category. Website: www.alnajem.com

Reference List

Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S. and Glass, J. (2014). A complete KALDI recipe for building Arabic speech recognition systems. In *2014 IEEE Spoken Language Technology Workshop,* South Lake Tahoe, California and Nevada, USA.

Alsharif, B., Tahboub, R. and Arafeh, L. (2016). Arabic text to speech synthesis using Quran-based natural language processing module. *Journal of Theoretical and Applied Information Technology*, **83**(1), 148–68.

Amrous, A.I., Debyeche, M. and Amrouche, A. (2011). Prosodic features and formant contribution for Arabic speech recognition in noisy environments. In *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011* (465–474). Berlin: Springer.

Angoujard, J.P. (1990). *Metrical structure of Arabic* (Vol. 35). Walter de Gruyter GmbH.

Azmi, M.M. and Tolba, H. (2008). Syllable-based automatic Arabic speech recognition in different conditions of noise. In *2008 9th International Conference on Signal Processing* (601–604). IEEE Beijing, China.

Betti, M.J. and Ulaiwi, W.A. (2018). Stress in English and Arabic: A contrastive study. *English Language and Literature Studies*, **8**(1), 83–102.

Biadsy, F., Hirschberg, J. and Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages* (53–61). Athens, Greece.

Chittaragi, N.B., Prakash, A. and Koolagudi, S.G. (2018). Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, **43**(8), 4289–302.

De Jong, K. and Zawaydeh, B.A. (1999). Stress, duration, and intonation in Arabic word-level prosody. *Journal of Phonetics*, **27**(1), 3–22.

Habash, N., Soudi, A. and Buckwalter, T. (2007). On Arabic transliteration. In A. Soudi, A.V.D. Bosch, N. Günter (eds.) *Arabic Computational Morphology* (15–22). Belin: Springer.

Halpern, J. (2009). Word stress and vowel neutralization in modern standard Arabic. In *2nd International Conference on Arabic Language Resources and Tools* (1–7), Cairo, Eygpt.

Hanna, S., El-Farahaty, H. and Khalifa, A.W. (2019). *The Routledge Handbook of Arabic Translation*. UK: Routledge.

Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Washington, D.C.: Georgetown University Press.

Ibrahim, N.J., Idris, M.Y.I., Yakub, M., Yusoff, Z.M., Rahman, N.N.A. and Dien, M.I. (2019). Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition. *Malaysian Journal of Computer Science*, **31**(n/a), 46–72.

Khelifa, M.O., Elhadj, Y.M., Abdellah, Y. and Belkasmi, M. (2017). Constructing accurate and robust HMM/GMM models for an Arabic speech recognition system. *International Journal of Speech Technology*, **20**(4), 937–49.

Lounnas, K., Demri, L., Falek, L. and Teffahi, H. (2018). Automatic language identification for Berber and Arabic languages using prosodic features. In *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)* (1–4), Algiers, Algeria, 28–31/10/2018.

Mannepalli, K., Sastry, P.N. and Suman, M. (2018). Analysis of emotion recognition system for Telugu using prosodic and formant features. In *Speech and Language Processing for Human-Machine Communications* (137–44). Berlin: Springer.

Martinez, D., Lleida, E., Ortega, A. and Miguel, A. (2013). Prosodic features and formant modeling for an ivector-based language recognition system. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (6847–6851), Vancouver, Canada, 26–31/05/2013.

McCarthy, J.J. and Prince, A.S. (1990). Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language & Linguistic Theory*, **8**(2), 209–83.

Meftah, A., Alotaibi, Y. and Selouani, S.A. (2016). Emotional speech recognition: A multilingual perspective. In *2016 International Conference on Bio-Engineering for Smart Technologies (BioSMART)* (1–4), Paris, France, 04/12/2016.

Pasha, A., Al-Badrashiny, M., Diab, M.T., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O. and Roth, R. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Language Resources and Evaluation Conference (LREC)* (1094–1101), Reykjavik, Iceland, 26-31/05/2014.

Reddy, V.R., Maity, S. and Rao, K.S. (2013). Identification of Indian languages using multi-level spectral and prosodic features. *International Journal of Speech Technology*, **16**(4), 489–511.

Ryding, K.C. (2014). *Arabic: A linguistic introduction*. Cambridge, UK: Cambridge University Press.

Sharma, D.P. and Atkins, J. (2014). Automatic speech recognition systems: challenges and recent implementation trends. *International Journal of Signal and Imaging Systems Engineering*, **7**(4), 220–34.

Vetulani, Z. ed. (2011). Human language technology: Challenges for computer science and linguistics. In *4th Language and Technology Conference (LTC 2009)*, Poznan, Poland, 06/08/11/2009.

Wang, L., Zhang, C., Woodland, P.C., Gales, M.J., Karanasou, P., Lanchantin, P., Liu, X. and Qian, Y. (2016). Improved DNN-based segmentation for multi-genre broadcast audio. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (5700–5704), Shanghai, China, 20–25/03/2016.

Watson, J.C. (2002). *The Phonology and Morphology of Arabic*. Oxford, UK: Oxford University.

Young, S., Gunnar, E., Mark, G., Hain, T. and Kershaw, D. (2015). *The HTK Book Version 3.5 Alpha*. Cambridge, UK: Cambridge University Press.