تطوير نظام آلي فعال لتحويل الكلام الكويتي المنطوق إلى مكتوب (باللغة الإنجليزية)

إيمان الشرهان
بشاير العتيبي

الملخص

تقدم الدراسة أول نظام آلي فعال يعمل على تحويل الكلام العربي الكويتي المنطوق إلى نص مكتوب باستخدام أحدث تقنيات التعرف الآلي على الكلام، ويأتي هذا النظام المُطوَّر ليحل محل نظام التفريغ اليدوي التقليدي؛ الأمر الذي يجعل عملية التفريغ أكثر دقة وسرعة وتزامناً؛ ما يضمن تقليل نسبة الخطأ ويحقق الاستخدام الأمثل للوقت. وقد تم اختيار اللهجة العربية الكويتية بوصفها مثالاً على اللهجات العربية التي تعاني من نقص الموارد اللازمة لتطوير أنظمة التعرف الآلي على الكلام، ويتمثل أهم تحديين يواجهان تطوير أنظمة التعرف الآلي على الكلام في ما يلي: الأول، نقص البيانات المتاحة من نصوص مكتوبة ومنطوقة مطلوبة للنمذجة الفعالة، وقد تم التصدي لهذا التحدي من خلال استخدام مزيج من بيانات اللغة العربية الفصيحة، بالإضافة إلى بيانات من اللهجة الكويتية عند بناء النماذج الصوتية واللغوية في النظام، والتحدي الثاني، هو غياب القواميس النطقية للهجة المستهدفة نتيجة لعدم وجود نظام إملائي واضح المعالم، وقد استخدمت الدراسة نسخة موسعة من أداة التحليل الصرفي **MADAMIRA** الذي يغطي اللهجة الكويتية لإنشاء قاموس النطق المطلوب تلقائيّاً عند بناء النظام. وقد تم تطوير النظام الآلي المبتكر في هذه الدراسة استناداً إلى بيانات من **GALE** (المرحلة 3)، التي تحتوي على ما يقارب **22** ساعة مسجلة من الكلام الكويتي، متفاوتة بين الأخبار الإذاعية والبرامج الحوارية، بالإضافة إلى **29** ساعة من البرامج التلفزيونية تم الحصول عليها من القنوات التلفزيونية الكويتية، أما بيانات اللغة العربية الفصيحة؛ فقد استخرج ما يقارب **17.1** ساعة من الكلام الفصيح من قاعدة البيانات **GALE** نفسها، وهو كلام صادر عن متحدثين خليجيين. وقد اختبر هذا النظام الآلي المبتكر في ظروف مختلفة حقق في أفضلها أداء ممتازاً، وبنسبة خطأ لا تتجاوز **7.9 %**؛ مما يبشر بإمكانية استخدامه والاستفادة منه في تطوير تطبيقات متنوعة في المستقبل.

الكلمات المفتاحية: نظام التفريغ التلقائي، التعرف الآلي على الكلام، المعالجة الآلية للغة الطبيعية، اللهجات العربية، الصوتيات، نموذج ماركوف المخفي، التعلم العميق.

The Development of an Efficient Transcription System for Kuwaiti Broadcast News and Conversational Speech

Eiman Alsharhan
Bashayer Alotaibi

Assistant Professor, Department of Arabic Language and Linguistics, College of Arts, Kuwait University

Abstract

The research aims to introduce the first efficient speech transcription system for Kuwaiti Arabic (KA) using speech recognition technologies. The system replaces the conventional manual transcription scheme, which improves reliability, achieves the best use of time, and streamlines the process simultaneously. The research also presents two practical solutions for two fundamental challenges facing the development of speech recognition systems for Arabic dialects. The first challenge is the shortage of dialectal data that is required for efficient modeling. This challenge is addressed by using a combination of available Modern Standard Arabic (MSA) data and the dialectal data when building the acoustic and language models. The second challenge is related to the linguistics specifications of the targeted dialect, which can be seen in the absence of a well-defined orthography system and consequently, the lack of pronunciation dictionaries. The research uses an extended version of the MADAMIRA morphological analyzer that covers KA to automatically generate the pronunciation dictionary needed to build the model. It also uses data from the GALE (phase 3), which contains approximately 22 hours of Kuwaiti speech, varying among broadcast news, talk shows, and conversational programs, as well as 29 hours of TV shows obtained from Kuwaiti TV channels. For MSA data, the researchers retrieved approximately 17.1 hours of speech produced by Gulf speakers from the GALE (phase 3) database. The best performing model reported in this research achieves 7.9% of the word error rate (WER), which is anticipated to deliver a good performance when used in varied applications. The paper recommends for future research to pursue similar studies on other dialects.

Keywords: Automatic transcription system; Speech recognition; Speech processing; Dialectal Arabic; Deep neural networks; Hidden Markov model.

ISSN : 1026-9576
DOI : 10.34120/0117-039-155-009

1. Introduction

Automatic Speech Recognition system (ASR) plays an important role in a variety of applications, such as query response, voice-based search, and transcription system. Generally, developing reliable ASR systems is a complex and demanding task, as it needs to control all sorts of variations in speech signals. This task becomes even more challenging when investigating a language with many varieties, such as Arabic.

Arabic can be seen as a family of related varieties and is treated as such. For instance, there is Modern Standard Arabic (MSA), which is taught in schools and is widely used in any kind of formal communication. MSA is considered the official language of every Arabic speaking country. However, native Arabic speakers use their distinctive dialects in daily communication rather than MSA. Arabic dialects vary from country to country, and often more than one dialect exists within a country. Arabic dialects are commonly divided by region into five major groups: (1) The Gulf dialect, which is spoken by people who live around the shores of the Arabian Gulf; (2) the Iraqi dialect, which is used only in Iraq; (3) the Egyptian dialect, which is spoken in Egypt and some areas in Sudan; (4) the Levantine dialect, which is spoken by people in Lebanon, Syria, Iraq, Palestine, and Jordan; (5) and the Maghrebi dialect, which is a vernacular Arabic dialect continuum spoken in the Maghreb region (Morocco, Algeria, Tunisia, Libya, Western Sahara, and Mauritania).

Arabic dialects differ substantially from MSA and even among one another in terms of phonology, morphology, vocabulary, and syntax. For instance, Zaidan emphasizes the importance of treating Arabic dialects as different languages, after showing how a machine translation system trained exclusively on MSA was not able to perform as well when tested on different Arabic dialects (199). The rate of performance parallels the difference detected when testing a machine translation system trained on one language and tested on another from the same linguistic family, such as Spanish and Portuguese. Another research found a huge gap between the dialects that not only indicates the need for considering the five major dialectal groups in building ASR systems but also suggests that it is worth dividing the major dialectal groups into smaller subsets when building ASR systems (Alsharhan and Ramsay, “Investigating the Effects” 995) However, most of the available data in the literature target MSA, while the available resources for dialectal Arabic are noticeably finite. This shortage causes a serious obstacle for researchers working in the field of Arabic ASR. The current research, hence, aims to create language resources necessary to build speech recognition models for Kuwaiti Arabic (KA). Afterwards, these resources are used in developing an efficient transcription system.

KA is the local colloquial version of MSA spoken in Kuwait. It is a sub-variety of

the Gulf dialect. It has many linguistic features that are different from MSA. It is an under-resourced dialect for which there is a significant lack of all kinds of language resources, such as speech data, written text, pronunciation dictionaries, and morphological analyzers. Moreover, KA has a distinct morphological and phonological system that has not been studied thoroughly. All these challenges lead to poor performance of the ASR system with a high out of vocabulary rate (OOV).

This research aims to develop a speech transcription system for KA and create all the necessary resources. The paper proposes a cross-lingual data sharing approach to benefit from existing MSA language resources in building acoustic and language models for KA. This dependency on MSA data is motivated by several observations. First, Kuwaiti speakers, especially those in media, code-switch between MSA and KA, since MSA is the formal variety of Arabic used in educated speech. Therefore, it is expected that their speech would contain many instances of MSA. Second, the Gulf dialects, including KA, share many linguistic features with MSA, probably more than other dialects. According to Versteegh, the Gulf dialects preserve more of MSA's verb conjugation than other Arabic dialects (188). Therefore, the researchers hypothesize that an acoustic model trained with enough MSA and Kuwaiti speech combined will achieve better performance than one using fewer data collected from Kuwaiti speech solely. However, the researchers of this paper restricted the use of MSA speech to MSA spoken by Gulf speakers because speakers from various Arab regions exhibit distinct differences in their use of spoken MSA. Limiting MSA data to that spoken by Gulf speakers would minimize the differences between MSA data and the dialectal data. This decision is also motivated by the findings of a previous research which confirms that using a cross-lingual approach is advantageous only when languages used in modeling share a lot of similarities (Alsharhan and Ramsay, "Investigating the Effects" 991).

Another benefit of the cross-lingual approach is using the available MSA resources, such as the Morphological Analysis and Disambiguation of Arabic (MADAMIRA) tool. MADAMIRA is a morphological analyzer tool designed for MSA. The researchers developed an extended version of this tool named MADAMIRA-KA, which adds Kuwaiti prefixes, suffixes, and stems in its dictionary. The extended version of the morphological analyzer is useful in retrieving the missing short vowels and generating the pronunciation dictionary that is needed for building the transcription system for KA. It is also able to cover MSA data.

This work resulted in designing the first KA transcription tool that can be employed in various fields. The developed tool was tested in different conditions and achieved good performance with 7.9% WER. With this level of accuracy, the developed system is a suitable alternative to a professional human transcriptionist, for it can assist

in captioning meetings, conferences, parliament sessions, and other events in real-time. Compared to audio content, machine-driven transcription is searchable and requires less computer memory, which helps delivering information retrieval systems and automatic translation systems. All experiments were conducted with the help of the latest version of the Hidden Markov Model Toolkit (HTK), which has recently integrated Deep Neural Network (DNN) modules to be used for acoustic modeling and feature extraction (Zhang 3582).

The remaining part of the paper proceeds as follows: the second section provides a brief overview of the recent efforts toward recognizing Arabic speech in recent years. The third section describes the phonetics and phonological specifications of the Kuwaiti Arabic dialect. The fourth section is concerned with the data used in developing the proposed system. The fifth section gives details about the architecture of the automatic transcription system. The sixth section presents the findings of the research and reports the performance of the KA-specific model and the cross-lingual model.

2. Review of Literature

Plenty of work has been done on recognizing Arabic speech in the past few years. A considerable portion of this work was directed towards developing ASR systems for MSA speech, such as Ali et al.; Cardinal; and Menacer et al. However, Arabic dialects have hardly been studied mainly because of the limited speech data for dialects.

The Egyptian Arabic dialect has received considerable attention in comparison to other Arabic dialects (different approaches were introduced to improve the recognition of Egyptian speech). Mousa et al. proposed a novel approach to improve the language modeling by using a mixture of words and morphemes along with their features as an input of the DNN language model. In their study, "Advances in Dialectal Arabic," Ali et al. explored the benefits of using dialectal tweets in improving the vocabulary coverage and reducing the WER. More recently, Li et al. proposed two adaptation models for Recurrent Neural Network Language Models (RNNLMs) for conversational speech recognition, which yields noticeable improvements when applied using Egyptian corpus.

Moreover, several works were dedicated to the Levantine Arabic dialect, such as Soltau et al.'s "From Modern Standard Arabic to Levantine ASR" and Elmahdy et al.'s "A Baseline Speech Recognition System for Levantine Colloquial Arabic". Both works present a cross-lingual approach for acoustic and language modeling using MSA and Levantine data and achieved a noticeable reduction in WER.

North African Arabic dialects, which are highly influenced by the French language and believed to be very different from MSA and other Arabic dialects, have received some attention from researchers in the field. A study by Masmoudi et al. presented a tool for automatically generating a pronunciation dictionary for Tunisian Arabic. This rule-based tool includes grapheme to phoneme mapping, a lexicon of exceptions, and a set of phonetic rules. The resulting software was tested on a word list in Tunisian Arabic using two independent test sets and reached an error rate of 9%. Menacer et al. developed an ASR system for MSA and its extension to the Algerian dialect. The research achieved a substantial reduction in WER by combining two acoustic models, one trained on MSA and the other on French. A more recent study carried out by Mouaz introduced the development of an ASR system for Moroccan Arabic. This study's purpose is to verify the ability of the ASR system in distinguishing the vocal prints of speakers to develop a speaker identification system.

For the Gulf dialects, only one study was found, which examines Qatari Arabic (Elmahdy et al., "Development of a TV Broadcasts Speech Recognition System for Qatari Arabic"). The study presented a cross-lingual approach to compensate for the lack of Qatari speech and text data. The suggested approach was found to achieve a 28% relative reduction in WER.

Larger-scale studies were carried out by Biadsy et al. and Khurana and Ali. Biadsy et al.'s study supports multiple dialects from different Arabic regions, namely Egypt, Jordan, Lebanon, Saudi Arabia, and the United Arab Emirates. The study was directed towards building an ASR system to support voice search, dictation, and voice control to be used by the general public employing their dialects. Similarly, Khurana and Ali's study introduces an advanced transcription system for Arabic multi-dialects broadcast media. The data used in this study comprises roughly 70% MSA speech data, and the rest is dialectal data, namely Egyptian, Gulf, Levantine, and North African.

3. Phonetics and Phonological Specifications of the Kuwaiti Arabic Dialect

Usually, descriptions of any Arabic dialect are not presented in isolation when compared with MSA or neighboring dialects due to the range of shared features between them. However, the paper does not aim to provide such comparisons, as this would naturally consume a lot of space in terms of literature. The main purpose is to provide a closer look at the features that make KA different from other Arabic dialects. Additionally, these distinctive features help to build the ASR system dedicated to KA, which is the goal of this research.

Hence, the first part of the research explores many significant linguistic features that are characteristic of the Kuwaiti dialect. Those characteristics are further examined to enhance the ASR system. Since the field of this research focuses on the medium of speech, the main features investigated are phonetic, morphophonological, and morphological features. Moreover, the description covers several functional words that are characteristic of KA, which belong to a closed words class, such as adverbs, prepositions, or participles.

As previously mentioned, literature on KA is relatively scarce in comparison to works on MSA or other Arabic dialects, such as Egyptian, Levantine, or Iraqi Arabic. The current study is informed mainly through the studies of Al-Bahri; Al-Qenaie; and Holes. Furthermore, several linguistic characteristics are reported by analyzing spoken KA found in this research dataset.

MSA has thirty-six phonemes, of which six are vowels, two diphthongs, and twenty-eight are consonants. KA has three more consonants, which are /g/, /tʃ/, and /p/, as well as four more vowels, namely /ɑ/, /ɐ/, /ē/, and /ō/ (Al- Qenaie 84). The differences between MSA and KA phonemes lie in the way a few phonemes are pronounced. The following paragraphs demonstrate these differences.

There are differences between KA and MSA in the pronunciation of some consonants. For instance, the uvular stop ق /q/ is commonly voiced as /g/; the word قال /qa:l/ 'said' is pronounced as /ga:l/ by Kuwaiti speakers. Furthermore, the sound /q/ tends to be palatalized in some words, such as in the name of the city of Sharjah الشارقة. Another difference is the velar stop consonant ك /k/, which is frequently palatalized as /tʃ/ in some words in KA and with some functional morphemes. For example, the second person feminine pronoun in a word like كِتابِك is pronounced /kita:bitʃ/ 'your book'. Additionally, the consonant ج /dʒ/ is palatalized as /j/ in many words, such as جِمعَة /jimʕa / 'Friday' and يِبب /ji:b / 'bring'. Also, the emphatic consonants ض /dˤ/ and ظ /ðˤ/ are merged in most Arabic varieties, including KA. Many speakers do not distinguish between the two consonants, and all words containing ض /dˤ/ are pronounced as ظ /ðˤ/. For example, the word ضَابِط /dˤabitˤ/ 'officer' is pronounced as /ðˤabitˤ/ instead. Finally, the phoneme /p/ is only attested in words that are borrowed from other languages such as بِيب /pi:p/ 'pipe,' borrowed from English (Holes 616).

Moreover, in MSA there are six vowel phonemes forming three pairs of corresponding short and long vowels, namely /a/, /i/, /u/, /a:/, /i:/, and /u:/. KA has an additional inventory of short and long vowels. For example, /ɑ/ and its long version /ɑ:/ are attested in KA in words like قَبِل /gɑbəl / 'before' and خال /xɑ:l:/ 'uncle,' respectively. Also, long vowels /o:/ and /e:/ are attested in words such as موت /mo:t/ 'death' and زيت /ze:t / 'oil.' These words are pronounced as diphthongs in MSA.

Furthermore, compared to MSA, KA shows significant vowel-shifting cases. One such case is the pronouns of the second singular person which are usually pronounced as إنتَ /ʔinta/ 'you-masculine' and إنتِ /ʔinti/ 'you-feminine' instead of the standard pronunciation /ʔanta/ and /ʔanti/. Another case is the shifting of the first vowel in the perfective verb form; verbs that have the template 'faʕala'. This is clear in verbs like كَتَبَ /kataba/ 'wrote,' سَمِعَ /samiʕa/ 'heard,' and سَبَحَ /sabaħa/ 'swam' in MSA, which are pronounced in KA as كِتَب /kitab/, سِمَع /simaʕ/, and سِبَح /sibaħ/, respectively. Yet a further case of vowel shifting is the first syllable of the imperfective verb. In several imperfective verbs, the first vowel /a/ is shifted to /i/, such as in يَلعَب /jalʕab/ 'plays,' يَسبَح /jasbaħ / 'swims,' and يَشرَح /jaʃraħ/ 'explains', which are pronounced as /jilʕab/, /jisbaħ/, and /jiʃraħ/ in KA with vowel shifting.

In addition, vowels in KA exhibit deletion or insertion in specific contexts in opposition to MSA. KA, like Najdi Arabic, possesses aspects of what is known as the Gahawa-syndrome (De Jong and Versteegh 100). The Gahawa-syndrome refers to both the deletion of /a/ in CaC_2 first and non-final syllables when C_2 is a guttural consonant and the epenthesis of an /a/ after C_2. Examples of this are صَخرَة /sˁax-rah/ 'stone,' which is pronounced as /sˁxa-rah/ in KA. The /a/ from the first syllable is deleted, and another /a/ is inserted after C_2 (or it moved from its position before C_2 to the position after it). This may also trigger epenthesis of another vowel at the beginning of the word, rendering the new pronunciation as follows: /əsˁ-xa-rah/.

However, there is disagreement among researchers about whether the epenthetic vowel is inserted at the beginning of the word or that the word starts directly with a consonant cluster (Al-Qinaie 239; Hole 616). Another example of vowel insertion is found in words that have two consonants (consonant cluster) in their final syllable, usually caused by deletion of a case short vowel in pause context, such as the word بَحْر /baħr/ 'sea,' صَخْر /sˁaxr/ 'stone,' نِسْر /nisr/ 'eagle,' and فَهْد /fahd/ 'jaguar'. In KA, these words are pronounced with an additional vowel inserted to break the consonant cluster as follows: /baħar/, /sˁaxar/, /nisir/, and /fahad/. Short vowels representing a case are usually dropped in Arabic varieties and that causes syllabification in the structure of some words, as explained in the previous examples. Another characteristic of KA is the assimilation of hamza to a glide, such as in بِئْر / biʔr/ 'water-well,' which is pronounced as/bi:r/ and رَأس / raʔs/ 'head,' which is pronounced as /ra:s/. Furthermore, a hamza at the end of a word is usually deleted, such as in سَماء / sama:ʔ/ 'sky' and صَحراء /sˁaħra:ʔ/ 'desert,' which are pronounced as/ sima/ and /sˁaħra/ with shortened final vowels.

These are the main differences between KA and MSA that have a direct effect on the data being examined for the purpose of this research. There are, of course, other lexical and syntactic differences that are not discussed here since they are beyond the scope of this paper. For a detailed description, see Alotaibi and Alsharhan's study,

4. Data Description

The research used the GALE (phase 3) Arabic broadcast news and conversational speech dataset released by Linguistic Data Consortium (LDC) and the conversational programs obtained from Kuwaiti channels. Two separate datasets were used: one contains Kuwaiti speech, while the other contains MSA speech. These datasets were used in training and testing the developed system. Details of the two datasets are given below.

4.1 KA Dataset

This dataset contains approximately twenty-two hours of speech obtained from the GALE dataset and 29.6 hours of speech obtained from KTV, a national broadcast station in Kuwait, and the Kuwait sports channel. This amounts to approximately 51.6 hours of speech data and a vocabulary of 379k items. This data was preprocessed to remove anomalous recordings, noisy recordings, and music. Table 1 provides more details about the names of TV programs that make up this dataset. The wave files in this dataset were produced with the help of 109 male and female speakers.

Table 1

Contents of KA Dataset

	Program name		Duration (in hours)
1	Kuwait news	البرنامج الإخباري	9.7
2	Valid for publication	صالح للنشر	1.2
3	Six by six	ستة على ستة	4.9
4	Weekly diwaniya	ديوانية الأسبوع	6.2
5	Good evening Kuwait	مساء الخير يا كويت	6.1
6	Alsoor street	شارع السور	2.3
7	Kuwait nights	برنامج ليالي الكويت	10.1

	Program name		Duration (in hours)
8	The best question	أحلى سؤال	1
9	Kuwait today	الكويت اليوم	2.4
10	Tonight show	برنامج الليلة	3.6
11	The motor	برنامج الموتر	2.4
12	super	برنامج سوبر	1.7

Transcriptions of the wave files were performed using traditional Arabic orthography. From this, a problem could arise, as many KA phonemes would not be correctly represented in the text, which would eventually lead the acoustic model to model these phonemes incorrectly. To overcome this problem and to achieve the desired level of accuracy in the transcription, all transcriptions were re-annotated manually to give each phoneme a single character. Four more KA letters were added to indicate non-standard Arabic consonants: چ denotes the /tʃ/ consonant, گ denotes /g/, پ denotes /g/, and ڤ denotes /v/.

After applying orthographic normalization, the transcripts were segmented into manageable and well-defined segments with the required timestamps. Each segment was given a specific label that includes the speaker's name and gender with the prompt number. All unnecessary information, such as non-speech segments and non-Arabic texts, were removed.

4.2 MSA Dataset

The MSA speech and textual data contained in this dataset were obtained exclusively from the GALE (phase 3) data. One restriction was applied in the collection of this data: only speakers from the Gulf dialectal backgrounds were chosen. The Gulf speakers were recognized using an annotated corpus of the GALE data reported in Alsharhan and Ramsay's study "An Exploratory Study of the Development of a Speech Corpus Annotated for the Main Arabic Dialects". In this corpus, speakers were labeled with their regional origin. In order to achieve more reliable results, the researchers of this paper rigorously chose speakers who were unanimously recognized to be from the Gulf. This yielded a total of 17.1 hours of speech data (15.3 hours for male speakers and 1.8 hours for female speakers).

In the GALE data, a quick rich transcription (QRTR) was used. This method of transcription is advantageous in comparison to careful transcription since it is accomplished in a shorter timeframe. However, fewer quality checks are performed on the finished product, which leads to certain transcript limitations. For instance, many cases were found to have discrepancies and spelling errors. These orthographic errors were compensated by using regular expression rules when necessary.

5. System Description

The transcription system reported in this project was developed with the help of the latest version of the Hidden Markov Model Toolkit (HTK - version 3.5). This version enables Deep Neural Network (DNN) models to be used beside the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM) for acoustic modeling. The recent application of DNN in the ASR area is a new milestone, as DNN-HMM acoustic modeling has become the-state-of-the-art modeling instead of GMM-HMM modeling. The researchers Du et al. argue that the first several layers of DNN are known to extract highly nonlinear and discriminative features that make DNN-HMM a noise-robust system to a certain extent.

HTK includes many widely used recent speech-processing techniques to cover the entire ASR pipeline. The overall flowchart of the designed automatic transcription system is illustrated in Figure 1. It can be observed that the framework of the system primarily combines three components: the acoustic model, the pronunciation model (dictionary), and the language model. However, modeling these components must be preceded by the front-end feature extraction process. The following sections provide a brief review of the front-end process and the three types of knowledge required for training and decoding the ASR system.

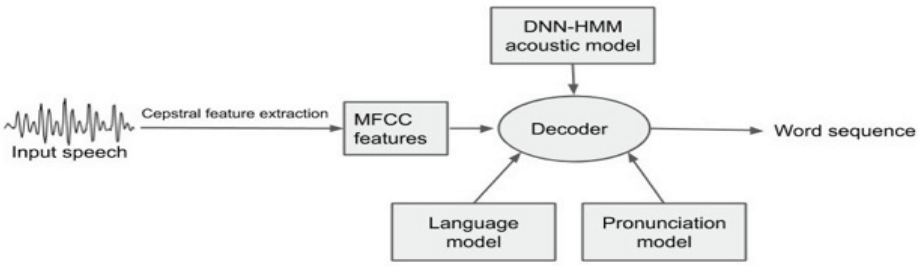

Fig. 1. The overall flowchart of the developed speech transcription system.

5.1 Front-end Feature Extraction

The developed transcription system tool automatically transcribes the contents of wave files with the help of models built using various techniques. These raw speech signals cannot be processed directly. They must be interpreted more compactly and efficiently. This is achieved by converting the raw speech signals into a sequence of acoustical feature vectors. This front-end stage is critical in defining the components of the audio signal that are useful for linguistic content recognition. There is a variety of feature extraction techniques shown in the literature. However, according to Sharma and Atkins, the use of Mel-frequency cepstral coefficients (MFCCs) is the most predominant one (227). Obtaining MFCCs involves a series of steps to be applied to the input speech signal. This includes framing, windowing, DFTH, Mel filter bank algorithm, and computing the inverse of DFT. The speech signal is then transformed into a discrete sequence of feature vectors. The feature vector consists of the standard 39-dimensional MFCC coefficients (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).

5.2 Acoustic Model

The process of acoustic modeling takes place in multiple stages, beginning with the development of an initial set of identical monophonic HMMs, which are used by HTK for initialization. The next step involves building short-pause models and extending the silence model to make the system more robust. Since the word in the dictionary can have several pronunciations, the phone models created are used to realign the training data and generate new transcriptions. This forced alignment can help to improve the phone-level accuracy by outputting the pronunciation that best fits the acoustic data.

The creation of a set of monophonic HMMs is followed by generating context-dependent triphone HMMs. First, the monophonic transcriptions are converted to triphone transcriptions to generate a list of all triphones found in the training data. Second, similar acoustic states of these triphones are tied to allow accurate estimates of the parameters.

Re-estimation of the parameters must be performed after each step by applying multiple rounds of Baum-Welch training. This is performed to calculate the optimal values for the HMM parameters in order to develop the efficiency of the system.

Once the speech signal has been aligned using GMM-HMMs, a DNN-based model can be trained. Similar to the previously described monophonic system

construction, a prototype model needs to be specified first. Those monophonic HMM states are then linked to the DNN output units. Furthermore, the DNNs produced in this phase are paired with the HMMs of the crossword triphone and eventually assessed.

In this project, six hidden layers were used, with 1024 nodes in each layer. The input layer has 351 nodes (the 39 MFCC and energy features observed in a window of 9 frames around the current frame), and the output layer has 114 nodes. Although triphones were used during the initial GMM training round (to obtain initial estimates of the different HMMs and to align the data before DNN training), they were tied to the basic phoneme set, seeing as the data does not contain sufficient instances of the individual triphones that would allow separate models to be allocated to them.

For the pre-training stage, Stochastic Gradient Descent (SGD) was applied with a mini-batch size set to 384 with 0.001 as the initial learning rate for all the experiments. A sigmoid was used as the activation function for the hidden layers and a Softmax for the output layer. To fine-tune the pre-trained nets, 18 epochs with a fixed learning rate of 0.001 and the same mini-batch that was used for pre-training were employed. Several experiments were applied to find the best parameter settings, and since the previously mentioned DNN architecture were found to perform the best, all the results reported in this paper were obtained using these settings.

In this paper, two acoustic models were built. One was trained using the KA dataset and the other one used both the KA dataset and the MSA dataset.

5.3 Pronunciation Model

The pronunciation model, which is introduced in the dictionary, establishes the connection between the acoustic model and the language model. Thus, any mismatch between spoken and written language leads to a weaker connection between the acoustic and language models (El-Desoky et al. 2680). The pronunciation dictionary contains a list of single or multiple phonetic transcriptions of all given words. Due to the morphological complexity of Arabic and the absence of diacritics in the text materials, there is a lack of predefined dictionaries in the literature, and there is no pronunciation dictionary for KA. Therefore, the researchers generated the pronunciation dictionary automatically with the help of an extended version of MADAMIRA. This extended version includes prefixes, suffixes, and stems of KA to allow the morphological analysis of KA text materials. Details about the extended version of the MADAMIRA can be found in Alotaibi and Alsharhan's previously mentioned study.

Two types of dictionaries were generated: a single pronunciation dictionary and a multi-pronunciation dictionary. In the single pronunciation dictionary, only one

possible pronunciation of a word was given with no case markers. A previous study confirmed that assigning the right case marker to each word requires a comprehensive understanding of Arabic grammar rules, which many people do not possess (Alsharhan et al., "Evaluating the Effect of Using Different Transcription Schemes" 10). Accordingly, many language users tend to omit the case marker to avoid making errors, even when speaking in MSA. The dropping of case markers is a well-known phenomenon, particularly in the media, and it is called *taskeen*. The second pronunciation dictionary contained a list of all possible pronunciations as provided by the MADAMIRA tool with an average of 3.64 pronunciations for a given word. In case no analysis was produced by the MADAMIRA, the graphemic version of the word was used. In other words, the phonetic transcription was obtained from the word's graphemes by splitting the word into letters with no diacritics.

5.4 Language Model

Robust speech recognition systems cannot be built through acoustic modeling solely; some form of linguistic knowledge is needed to capture the language properties and to restrict the search by limiting the set of possible HMMs. Predicting the probability of a word occurring in the context during recognition increases the chances of accurately recognizing the words.

In this research, the language model was built probabilistically by computing the likelihood for each possible successor word using n-gram models. Bi-gram language modeling was employed in building all the proposed transcription system. This was carried out using several HTK tools, beginning with creating the grammar then producing a word network that listed each word-to-word transition.

6. Results and Discussion

Based on the studies reviewed in Section 2, it is deduced that many researchers believe that cross-lingual modeling is the optimal approach to tackle the limitation of data resources for under-resourced languages (Elmahdy et al., "A Baseline Speech Recognition System for Levantine Colloquial Arabic" 4). Nevertheless, other researchers have shown that building dialect-specific ASR systems using the dialectal data solely is a more effective approach. This argument will be investigated in this section to determine which system performs better, the KA-specific system or the system that uses a combination of KA and MSA.

It is important to mention that the MSA data chosen in this study is the one that is produced by a Gulf accented speaker (from Kuwait, Saudi Arabia, Bahrain, Qatar,

United Arab Emirates, or Oman). This step is crucial to ensure that only dialects with great similarity are trained on. This approach enables the researchers to benefit from the cross-lingual approach and to build a robust system with the least variability.

To achieve optimal use of data, a five-fold cross-validation approach was carried out as a way of evaluating the proposed systems. This was achieved by shuffling the dataset and dividing it into five equal-sized random sets. In each round, four sets were used for training the model, and one set was preserved for testing. Afterwards, the results from the five folds were averaged to calculate a single estimate.

6.1 KA-specific System

This system was trained on 51.6 hours of KA speech and a total of 379k vocabulary size, as mentioned in Section 4. Phoneme-based acoustic modeling was adopted. Two different models were built with different dictionaries: a fixed pronunciation dictionary and a multi-pronunciation dictionary. The results are reported in Table 2.

6.2 Applying Cross-lingual Data Sharing Approach

This system was built using 51.6 hours of KA speech plus 17.1 hours of MSA speech produced by the Gulf accented speakers. Similar to the previous system, two acoustic models were built: one with a fixed-pronunciation dictionary and one with a multi-pronunciation dictionary.

Table 2

WER for the Developed Transcription Systems

		GMM-HMM	DNN-HMM
KA-specific system	Single pronunciation dictionary	19.6%	10.8%
	Multi-pronunciation dictionary	20.9%	11.9%
Cross-lingual system	Single pronunciation dictionary	16.6%	7.9%
	Multi-pronunciation dictionary	16.9%	9.2%

Results reported in Table 2 show the superiority of:

1- DNN-HMM model over GMM-HMM model.

2- The single pronunciation dictionary over the multi-pronunciation dictionary.

3- The model trained with cross-lingual data over the model that is trained with dialect-specific data.

The dominance of the DNN-HMM model over GMM-HMM with an average reduction of 8.55% in the WER is an interesting finding. This finding is consistent with Ali et al.'s study, "A Complete KALDI Recipe for Building Arabic Speech Recognition Systems" and Woodland et al.'s study, "Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge". A previous study attempted to explain how the DNN-HMM model outperforms the conventional GMM-HMM model. The researchers of the study argue that the DNN gain is almost entirely attributed to DNN feature vectors, which are concatenated within a relatively long context window from several consecutive speech frames (Pan et al. 302).

Moreover, the results reported in Table 2 suggest that using a fixed-pronunciation dictionary is better than using a multi-pronunciation dictionary. This result may partly be explained by the fact that the speakers selected for this paper tend to omit case markers. This outcome complies with the linguistic description given for KA in Section 3, which indicates that omitting case markers is a characteristic of KA speech. Furthermore, this result is consistent with several previous studies which argue that augmenting the pronunciation dictionary with all possible pronunciations can cause extra confusion due to rare pronunciations (Jurafsky and Martin 120; Saraclar et al. 151).

Another observed result is the predominance of training with cross-lingual data over training with dialect-specific data. This outcome indicates that:

- Cross-lingual strategy is crucial to overcome the limited availability of data, especially for under-resourced languages like KA.

- There is a common argument in the literature that "there is no data like more data" (Moore). This means that training a model with a substantial amount of data may lead to a well-performed model. However, previous research confirms that consistently feeding the model with arbitrary data can worsen the performance of the system. In other words, researchers must be careful about selecting the data that will be used in training the system. Using carefully selected subsets of data produces recognition systems that are superior to systems that make use of a much larger amount of undifferentiated data. The

strategy followed in this project was to select data from MSA due to the great linguistic similarity between MSA and KA. Restricting speakers to only those from the Gulf region reduces the variability between speakers, for previous research confirms that the speakers' original dialect can be detected even when speaking in MSA (Alsharhan, "Investigating the Effects" 991).

- The superiority of combining KA data with MSA data imply that speakers in media tend to code-switch between MSA and dialectal Arabic to sound more formal.

This finding has important implications for developing speech-processing tools for under-resourced languages. Therefore, researchers may benefit from resources available for other similar languages in case resources for the targeted language are not available.

7. Conclusion

The project attempted to design the first efficient automatic transcription system for Kuwaiti Arabic and evaluate the application of a cross-lingual approach in building the system. The best performance model reported in this paper achieved 7.9% WER, which is considered low in comparison to other systems developed for different Arabic dialects. The proposed system has many applications, such as broadcast news transcription system, automatically archiving emergency calls and parliament sessions, dictation, and information extraction, to mention just a few.

Two approaches were followed and compared to build the system: one employed dialect-specific data, whilst the other employed a combination of the KA dataset and the MSA dataset in training the system. The MSA data chosen for this work consists of speech files produced rigorously by Gulf speakers. The results reported in Section 6 confirm the superiority of combining the MSA dataset with the KA dataset in training the acoustic and language models, which leads to an average reduction of 3.15% in WER in comparison to the dialect-specific system. The experimental evidence suggests that using cross-lingual data is a good strategy to overcome the shortage of dialectal resources. However, it is important to ensure that this data is drawn from a language with a high degree of similarity to minimize the level of variation and achieve better performance.

DNN models are found to outshine the performance of GMM-HMM models, which complement the findings of the previous studies. For future work, it is possible

to extend the current framework to cover other dialects. Another plan is to embed this system into a sentiment analysis system to allow businesses to identify customers' sentiments regarding the provided products or services using transcribed telephone conversations.

Bibliography

– Al-Bahri, Khaled W. *A Grammar of Hadari Arabic: A Contrastive-Typological Perspective*. 2014. University of Sussex, dissertation.
– Al-Qenaie, Shamlan. *Kuwaiti Arabic: A Socio-Phonological Perspective*. 2011. Durham University, dissertation.
– Ali, Ahmed, et al. "Advances in Dialectal Arabic Speech Recognition: A Study Using Twitter to Improve Egyptian ASR." *International Workshop on Spoken Language Translation (IWSLT 2014)*, 2014.
– Ali, Ahmed, et al. "A Complete KALDI Recipe for Building Arabic Speech Recognition Systems." *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 525-529.
– Alotaibi, Bashayer, and Eiman Alsharhan. *The Characteristics of Written Kuwaiti Arabic and Its Use in Creating Resources for Morphological Analysis.* (under publication).
– Alsharhan, Eiman, and Allan Ramsay. "An Exploratory Study of the Development of a Speech Corpus Annotated for the Main Arabic Dialects." *Arab Journal for the Humanities*, vol. 38, no. 150, 2020, pp. 365-386.
– "Investigating the Effects of Gender, Dialect, and Training Size on the Performance of Arabic Speech Recognition." *Language Resources and Evaluation*, vol. 54, no. 4, 2020, pp. 975-998.
– Alsharhan, Eiman, et al. "Evaluating the Effect of Using Different Transcription Schemes in Building a Speech Recognition System for Arabic." *International Journal of Speech Technology,* 2020, pp. 1-14.
– Bezoui, Mouaz, et al. "Speech Recognition of Moroccan Dialect Using Hidden Markov Models." *IAES International Journal of Artificial Intelligence*, vol. 8, no. 1, 2019, pp. 985-991.
– Biadsy, Fadi, et al. "Google's Cross-Dialect Arabic Voice Search." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.
– Cardinal, Patrick, et al. "Recent Advances in ASR Applied to an Arabic Transcription System for Al-Jazeera." *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
– De Jong, Rudolf E., and Kees Versteegh. *Encyclopedia of Arabic Language and Linguistics*. Brill, 2011.
– Du, Jun, et al. "Robust Speech Recognition With Speech Enhanced Deep

Neural Networks." *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

– El-Desoky, Amr, et al. "Investigating the Use of Morphological Decomposition and Diacritization for Improving Arabic LVCSR." *Tenth Annual Conference of the International Speech Communication Association*, 2009.
– Elmahdy, Mohamed, et al. "A Baseline Speech Recognition System for Levantine Colloquial Arabic." *Proceedings of ESOLEC*, 2012.
– "Development of a TV Broadcasts Speech Recognition System for Qatari Arabic." *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014.
– Holes, Clive. "Kuwaiti Arabic." *Encyclopedia of Arabic Language and Linguistics*, 2, 2007, pp. 608-620.
– Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson, 2014.
– Khurana, Sameer, and Ahmed Ali. "QCRI Advanced Transcription System (QATS) for the Arabic Multi-Dialect Broadcast Media Recognition: MGB-2 Challenge." *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016.
– Li, Ke, et al. "Recurrent Neural Network Language Model Adaptation for Conversational Speech Recognition." *Interspeech*, 2018.
– Masmoudi, Abir, et al. "A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition." *LREC*, 2014.
– Menacer, Mohamed, et al. "An Enhanced Automatic Speech Recognition System for Arabic." *The Third Arabic Natural Language Processing Workshop-EACL*, 2017.
– Moore, Roger K. "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners." *Eighth European Conference on Speech Communication and Technology*, 2003.
– Mousa, Amr El-Desoky, et al. "Morpheme-Based Feature-Rich Language Models Using Deep Neural Networks for LVCSR of Egyptian Arabic." *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
– Pan, Jia, et al. "Investigation of Deep Neural Networks (DNN) for Large Vocabulary Continuous Speech Recognition: Why DNN Surpasses GMMs in Acoustic Modeling." *2012 8th International Symposium on Chinese Spoken Language Processing*, 2012.
– Saraclar, Murat, et al. "Pronunciation Modeling by Sharing Gaussian Densities Across Phonetic Models." *Computer Speech & Language*, vol. 14, no. 2, 2000, pp. 137-160.
– Sharma, Davinder Pal, and Jamin Atkins. "Automatic Speech Recognition Systems:

Challenges and Recent Implementation Trends." *International Journal of Signal and Imaging Systems Engineering*, vol. 7, no. 4, 2014, pp. 220-234.

– Soltau, Hagen, et al. "From Modern Standard Arabic to Levantine ASR: Leveraging Gale for Dialects." *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2011.

– Teller, Virginia. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition." *Computational Linguistics*, vol. 26, no. 4, 2000, pp. 638-641.

– Versteegh, Kees. *Arabic Language*. Edinburgh University Press, 2014.

– Woodland, Philip C., et al. "Cambridge University Transcription Systems for the Multi-Genre Broadcast Challenge." *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

– Zaidan, Omar F., and Chris Callison-Burch. "Arabic Dialect Identification." *Computational Linguistics*, vol. 40, no. 1, 2014, pp. 171-202.

– Zhang, Chao, and Philip C. Woodland. "A General Artificial Neural Network Extension for HTK." *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

مجلة الحقوق

مجلة فصلية أكاديمية
محكّمة تعنى بنشر البحوث
والدراسات القانونية والشرعية

تصدر عن مجلس النشر العلمي - جامعة الكويت

صدر العدد الأول في
يناير ١٩٧٧

الاشتراكات

	في الكويت	في الدول العربية	في الدول الأجنبية
الأفراد	٣ دنانير	٤ دنانير	١٥ دولاراً
المؤسسات	١٥ ديناراً	١٥ ديناراً	٦٠ دولاراً

المراسلات

توجه جميع المراسلات إلى رئيس التحرير على العنوان الآتي:
مجلة الحقوق - جامعة الكويت ص.ب: ٦٤٩٨٥ الشويخ - ب 70460 الكويت
تلفون: ٢٤٨٣٥٧٨٩ - ٢٤٨٤٧٨١٤ فاكس: ٢٤٨٣١١٤٣
E.mail: jol@ku.edu.kw

عنوان المجلة في شبكة الإنترنت http://www.pubcouncil.kuniv.edu.kw/jol

ISSN 1029 - 6069

مجلة الشريعة والدراسات الإسلامية

فصلية علمية محكمة تصدر عن مجلس النشر العلمي بجامعة الكويت
تعنى بالبحوث والدراسات الإسلامية

رئيس التحرير الأستاذ الدكتور: **وليد خالد الربيع**

صدر العدد الأول في رجب ١٤٠٤هـ - أبريل ١٩٨٤م

* تهدف إلى معالجة المشكلات المعاصرة والقضايا المستجدة من وجهة نظر الشريعة الإسلامية.
* تشمل موضوعاتها معظم علوم الشريعة الإسلامية: من تفسير، وحديث، وفقه، واقتصاد وتربية إسلامية، إلى غير ذلك من تقارير عن المؤتمرات، ومراجعة كتب شرعية معاصرة، وفتاوي شرعية، وتعليقات على قضايا علمية.
* تنوع الباحثون فيها، فكانوا من أعضاء هيئة التدريس في مختلف الجامعات والكليات الإسلامية على رقعة العالمين: العربي والإسلامي.
* تخضع البحوث المقدمة للمجلة إلى عملية فحص وتحكيم حسب الضوابط التي التزمت بها المجلة، ويقوم بها كبار العلماء والمختصين في الشريعة الإسلامية، بهدف الارتقاء بالبحث العلمي الإسلامي الذي يخدم الأمة، ويعمل على رفعة شأنها، نسأل المولى عز وجل مزيداً من التقدم والازدهار.

جميع المراسلات توجه باسم رئيس التحرير

ص.ب ١٧٤٣٣ - الرمز البريدي: 72455 الخالدية - الكويت هاتف: ٢٤٨١٢٥٠٤ - ٢٤٩٨٤٧٢٣ - ٢٤٩٨٨٠٩٥
فاكس: ٢٤٨١٠٤٣٤

العنوان الإلكتروني: E-mail - jsis@ku.edu.kw

issn: 1029 - 8908

عنوان المجلة على شبكة الإنترنت: http://pubcouncil.kuniv.edu.kw/JSIS

اعتماد المجلة في قاعدة بيانات اليونسكو Social and Human Sciences Documentation Center

في شبكة الإنترنت تحت الموقع www.unesco.org/general/eng/infoserv/db/dare.html

حوليات الآداب والعلوم الاجتماعية
ANNALS OF THE ARTS AND SOCIAL SCIENCES

• مجلة فصلية محكمة.
• تصدر عن مجلس النشر العلمي بجامعة الكويت.
• صدر العدد الأول سنة ١٩٨٠م.
• تنشر الموضوعات التي تدخل في مجالات اهتمام الأقسام العلمية لكليتي الآداب والعلوم الاجتماعية.
• تنشر الأبحاث والدراسات باللغتين العربية والإنجليزية شريطة أن لا يقل حجم البحث عن ٥٠ صفحة وأن لا يزيد عن ٢٠٠ صفحة مطبوعة من ثلاث نسخ.
• لا يقتصر النشر في الحوليات على أعضاء هيئة التدريس لكليتي الآداب والعلوم الاجتماعية فحسب، بل يشمل ما يعادل هذه التخصصات في الجامعات والمعاهد الأخرى داخل الكويت وخارجها.
• تمنح المجلة الباحث خمسين نسخة من بحثه المنشور كإهداء.

1966

رئيس هيئة التحرير
أ.د يعقوب يوسف الكندري

ثمن الرسالة للأفراد
(٥٠٠ فلس)

نوع الاشتراك	الكويت	الدول العربية	الدول الأجنبية
الأفراد	٤ دنانير	٦ دنانير	٢٢ دولاراً
المؤسسات	٢٢ ديناراً	٢٢ ديناراً	٩٠ دولاراً

جميع المراسلات توجه إلى رئيس تحرير حوليات الآداب والعلوم الاجتماعية
ص.ب 17370 الخالدية 72454 الكويت - هاتف 24830256 (965) فاكس 24830256 (965)
ISSN 1560 - 5248 Key title: Hawliyyat Kulliyyat Al-Adab
www.apc. kuniv. edu.kw/aass E-mail: aass@ku.edu.kw