Lang Resources & Evaluation (2020) 54:975–998
https://doi.org/10.1007/s10579-020-09505-5

ORIGINAL PAPER

Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition

Eiman Alsharhan[1] · **Allan Ramsay**[2]

Accepted: 11 September 2020 / Published online: 12 October 2020

Abstract Research in Arabic automatic speech recognition (ASR) is constrained by datasets of limited size, and of highly variable content and quality. Arabic-language resources vary in the attributes that affect language resources in other languages (noise, channel, speaker, genre), but also vary significantly in the dialect and level of formality of the spoken Arabic they capture. Many languages suffer similar levels of cross-dialect and cross-register acoustic variability, but these effects have been under-studied. This paper is an experimental analysis of the interaction between classical ASR corpus-compensation methods (feature selection, data selection, gender-dependent acoustic models) and the dialect-dependent/register-dependent variation among Arabic ASR corpora. The first interaction studied in this paper is that between acoustic recording quality and discrete pronunciation variation. Discrete pronunciation variation can be compensated by using grapheme-based instead of phone-based acoustic models, and by filtering out speakers with insufficient training data; the latter technique also helps to compensate for poor recording quality, which is further compensated by eliminating delta-delta acoustic features. All three techniques, together, reduce Word Error Rate (WER) by between 3.24% and 5.35%. The second aspect of dialect and register variation to be considered is variation in the fine-grained acoustic pronunciations of each phoneme in the language. Experimental results prove that gender and dialect are the principal components of variation in speech, therefore, building gender and dialect-specific models leads to substantial decreases in WER. In order to further explore the degree of acoustic differences between phone models required for each of the dialects of Arabic, cross-dialect experiments are conducted to measure how far apart Arabic

✉ Eiman Alsharhan
eiman.alsharhan@ku.edu.kw

[1] Kuwait University, Kuwait City, Kuwait

[2] University of Manchester, Manchester, UK

dialects are acoustically in order to make a better decision about the minimal number of recognition systems needed to cover all dialectal Arabic. Finally, the research addresses an important question: how much training data is needed for building efficient speaker-independent ASR systems? This includes developing some learning curves to find out how large must the training set be to achieve acceptable performance.

1 Introduction and objectives

Arabic is a Semitic language and one of the six official languages of the United Nations (UN). It is spoken by perhaps as many as 422 million speakers (native and non-native) in the Arab region, making it the fifth most spoken language in the world (Lewis and Gary 2015).

Arabic can be viewed and treated as a family of related languages. There is Modern Standard Arabic (MSA), which is widely taught at schools and universities, and is used in the media, formal speeches, courtrooms, and indeed in any kind of formal communication. MSA is considered to be the official language in all Arabic speaking countries. In addition, people generally speak in their own dialects in daily communication. These dialects are neither taught at schools nor even have any organised written form. Arabic dialects differ substantially from MSA in terms of phonology, morphology, vocabulary, and syntax. Dialectal Arabic (DA), also known as Colloquial Arabic, is the natural spoken language in everyday life. It varies from one country to another and sometimes more than one dialect can be found within a country.

DAs are commonly divided by region into five main groups: (1) Gulf, which is spoken by people who live around the shores of the Arabian Gulf. (2) Iraqi, which is used only in Iraq. (3) Egyptian, which is the dialect spoken in Egypt and some areas in Sudan. (4) Levantine, which is spoken by Arabs near the Mediterranean east coast. (5) Maghrebi, which is the dialect spoken in western Arab countries.

Building robust speech recognition systems requires feeding the training model with spoken and written data of the targeted language. As will be discussed in Sect. 3, the use of statistical modelling in building acoustic, pronunciation, and language model motivates the need for large quantities of data. Finding such resources for Arabic, which is known for its complex morphology and high vocabulary growth, is a challenging task. Adding to this the extreme dialectal variation and significant differences between the spoken and the written language caused mainly by the absence of diacritics[1] makes the development of ASR systems for Arabic particularly challenging.

All these issues make it difficult to collect large and balanced amounts of speech and text data for Arabic. Researchers have generally emphasised the need for large sized speech corpora, with associated transcriptions, for building Arabic speech recognisers. They describe the resources available in the literature as being

[1] Diacritics are mainly markers for both vowels and consonants. These markers include short vowels, dagger Alif, sukun, nunation, and gemination marker.

expensive, lacking adaptability, reusability, quality, coverage, and adequate information types (Abushariah et al. 2010).

This paper investigates a range of straightforward approaches that can exploit sparse training data for Arabic in a more efficient way to guarantee the best use of data. By applying the suggested approaches, we found that using an informative and balanced subset of data can produce ASR systems that are comparable to those using substantial amounts of data.

The investigation includes applying some general conditions to the data in order to accommodate variation in speech and thereby improve the ASR system's performance. For instance, modifying the settings of the feature-extracting technique applied to the speech signals, suggesting a surrogate proposal to eliminate poor quality recordings, and suggesting the best level of phonetic transcription needed to model pronunciation. In addition, the paper investigates the feasibility of building gender-specific and dialect-specific ASR systems for each of the main five Arabic dialects.

The research also includes carrying out a set of cross-dialectal experiments to examine how well the ASR system performs when trained on one dialect and tested on another. Finally, the research gives recommendations for the amount of data needed for testing and training an ASR system.

The aim of the approaches introduced above is to consider the effects of a collection of fairly straightforward ways of making use of data which is known to be flawed in a number of ways, and in particular of trying, as far as possible, to experiment with these approaches on a single dataset, with no variations in the way that we used this dataset beyond the experiments that we were interested in. A number of these approaches have been tried by other authors, but because they have been investigated in isolation it is very difficult to see from previous studies which are the most effective, and to see how they perform in combination.

All the experiments are conducted with the aid of the latest version of Hidden Markov Model toolkit (HTK) (version 3.5) (Young et al. 2015; Woodland et al. 2015). The research uses the GALE phase 3 dataset of 200 h broadcast news and conversational speech database (LDC 2015) released by the Linguistic Data Consortium (LDC).

2 Availability of data resources for Arabic: problems and solutions

In the natural language processing community, there is a common belief that "there is no data like more data" (Moore 2003). Following this idea, researchers have worked hard to collect large amounts of data to have sufficient training materials to build ASR systems. In addition, researchers have worked on solving some issues related to the process of collecting the required data. Finding sufficient language resources to build ASR systems is particularly difficult for Arabic. This shortage of speech corpora arises for the following reasons:

- Arabic has multiple variants, and the differences between these varieties is like the difference between divergent languages. Many researchers have argued for

treating different Arabic dialects as different languages in building NLP applications.Zaidan and Callison-Burch (2014a), for instance, consider the variation between Arabic dialects to be enormous, to the extent to the need to treat them as different languages. The researchers justified this assumption by pointing to the similarity between the behaviour of a machine translation system when tested on DA and trained exclusively on MSA, with another machine translation system which was tested on Portuguese and trained on Spanish. However, most of the available data in the literature targets MSA and only a few datasets are available for DA. This shortage causes a serious obstacle for researchers working in the field of Arabic ASR.

- As will be described in Sect. 3, the construction of any ASR system requires a corpus of speech data with the associated textual/phonetic transcriptions. DA is mainly spoken and there is no standardised writing system for dialects. This leads to having dialectal speech transcribed following MSA rules in some cases, or transcribed with a great deal of inconsistency in others. The unavailability of adequate written materials is a serious problem faced in the development of language resources.
- The majority of Arabic texts lack diacritics. These are items that carry important information about pronunciation and meaning, and their absence makes it impossible to determine the phonetic structure of the words in a text. For instance, they indicate the presence or absence of one of the Arabic three short vowels, distinguish long vowels from glides or diphthongs, indicate geminated consonants, and also demonstrate the role of the word in the sentence (e.g. whether the word is subject or object of a verb). Retrieving the absent diacritics accurately is one of the main challenges in developing language corpora.
- Many available data resources are poor quality and do not meet the requirements for building robust ASR systems. Handling poor quality speech recordings requires careful inspection before using them in building an ASR system.
- Acquiring a large, high-quality corpus is expensive, therefore, in order to be made available at low cost to researchers, a corpus must be acquired using sources of funding other than user fees. Large, high-quality free corpora exist in, e.g., English (e.g., (Panayotov et al. 2015)), Chinese, (e.g. (Magic Data Technology Co., Ltd. 2019)), and Russian, (e.g. (Andrusenko et al. 2019)), but not in Arabic. Perhaps because of dialect variation corpora in Arabic tend to be smaller (e.g.,Open Speech and Language Resources (2003) contains 11.2 h of Tunisian Arabic) and/or expensive (e.g., GALE Phase 2 Arabic Broadcast Conversation Speech Part 1 (Walker et al. 2013).

In this section we will highlight researchers' efforts directed at compensating for the lack of language resources required for building Arabic ASR systems. Many researchers have found it necessary to face this lack either by starting from zero and constructing their own corpora, and then possibly making it public for interested researchers, or by finding some techniques to better use the available resources. This step is especially important for dialects which have received less attention and pose more challenges compared to MSA. Considerable interest in dialects was found in the literature in recent years. This is mainly due to their wide use in everyday life

and social media (Sadat et al. 2014; Shoufan and Alameri 2015). Several categories of work were conducted in the literature for coping with these dialects to build speech processing systems.

Some work has been conducted on constructing linguistic resources (lexicon and corpus) to deal with this lack in building dialectal Arabic corpora for ASR applications. Masmoudi et al. (2014) and Selouani and Boudraa (2010) have introduced corpora for Tunisian Arabic and Algerian Arabic, respectively. The Tunisian corpus contains audio recordings and transcriptions extracted from dialogues in the Tunisian Railway Transport Network. The Algerian corpus is composed of MSA speech pronounced by 300 Algerian native speakers from different regions. The developed corpora can be used to build dialect-specific Arabic ASR systems.Almeman et al. (2013) developed Arabic parallel texts and speech corpora that cover three major Arabic dialects: Gulf, Egypt and Levantine as well as MSA. The 32 h of recordings were collected with the aid of 52 participants. The researchers chose a specific linguistic domain to work with, namely travel and tourism.

Aiming to build multi-dialectal speech recognition systems to support voice search, dictation, and voice control for the general Arabic speaking public, Biadsy et al. (2012) constructed the largest multi dialectal Arabic speech corpus. This corpus was collected with the aid of more than 125 million people in Egypt, Jordan, Lebanon, Saudi Arabia, and the United Arab Emirates. The main limitation with this corpus is that it contains read speech. The speech was recorded by using an application that displays prompts to the user and asks them to say it in their own dialect. Therefore it is not natural speech and it is likely not be useful for building spontaneous speech recognition systems.

On the other hand, many researchers have dealt with the sparse available resources by using a data sharing approach. Kirchhoff and Vergyri (2005) carried out a thorough investigation into the feasibility of using data from MSA to build an Egyptian ASR system. Researchers found this cross-lingual data sharing approach to lead to significant reduction in WER. Similarly, Elmahdy et al. (2010, 2012) and Elmahdy et al. (2014) proposed a cross-lingual acoustic approach and employed some adaptation techniques to benefit from the existing MSA resources in building Egyptian, Levantine, and Qatari Arabic ASR system, respectively. This cross-lingual technique utilises the available MSA data in combination with dialectal data to overcome the problem of limitation of dialectal speech resources. Using this technique in developing different ASR systems was found to achieve a noticeable reduction in WER.

Huang et al. (2012) proposed a crossed-dialect Gaussian Mixture Model (GMM) training method to learn the maximum likelihood of cross-dialectal data. The researchers used West Point MSA Speech corpus along with Babylon Levantine Arabic speech corpus to build a Levantine ASR system. Researchers demonstrate that the proposed cross-dialectal training improves the system's performance significantly, especially when a small amount of MSA data is transferred. Menacer et al. (2017) presented an ASR system built using combined acoustic models: one for MSA and one for French to compensate for the absence of transcribed speech

data for Algerian dialect. This combination leads to a substantial absolute reduction of the word error of 24%.

3 Architecture of the ASR system

ASR systems work by converting a speech signal into a textual representation with the aid of models built using various techniques. The general architecture of the ASR system is presented in Fig. 1. It can be seen that the system integrates mainly three components: acoustic model, pronunciation model (lexicon), and language model. Modelling these components must be preceded by a front-end process to extract features from the audio signal that are good for modelling the speech. The following is a summary of the Speech features extraction process and the three kinds of knowledge needed for training and decoding the ASR system (acoustic model, pronunciation model, and language model). In this section, we point to the areas where we made changes to the standard settings in our current investigation.

3.1 Speech features extraction

The speech modelling tools cannot directly process waveforms. These waveforms have to be represented in a more compact and efficient way by converting them into a series of acoustical feature vectors. This front-end step is crucial to identify the components of the audio signal that are useful for recognising the linguistic content. The literature shows that there are a variety of feature extraction techniques Wang et al. (2016), for instance, report on experiments using alternative representations of the speech signal such as filterbank features and perceptual linear predictive coding, but the differences between the various representations seem to be marginal),

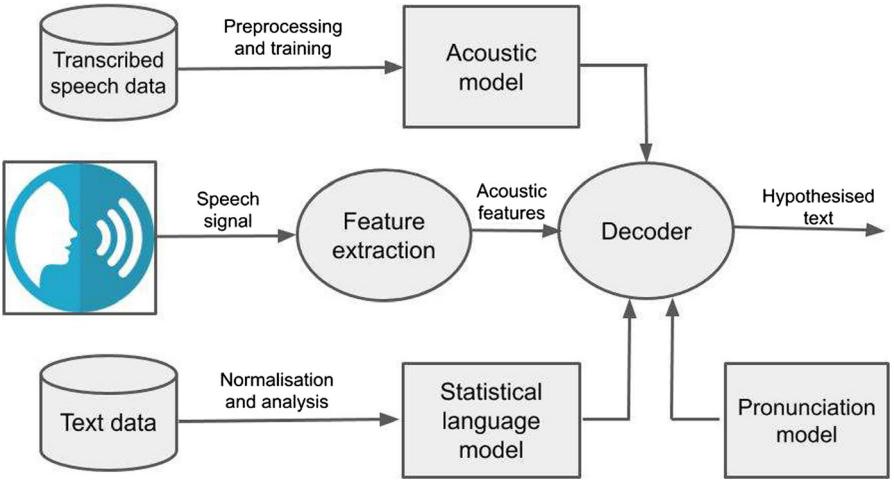

Fig. 1 Architecture and components of the ASR system

however, the use of Mel Frequency Cepstral Coefficient (MFCC) is the predominant one (Sharma and Atkins 2014). Obtaining MFCCs requires a sequence of steps to be applied to an input speech signal. These computational steps of MFCC include Framing, Windowing, Digital Fourier Transform Holography (DFTH), Mel filter bank algorithm, and computing Inverse Discrete Fourier transform (DFT). The speech signal is then converted into a discrete sequence of feature vectors.

The feature vector consists of a collection of MFCC coefficients and energy measures. Most researchers use the standard 39 MFCC vectors (12 cepstral features plus a measure of the energy, together with the rates of change and accelerations of these 13 features). In this research we investigate the use of only 25 MFCC vectors (i.e. the 12 cepstral features and their rates of change plus the rate of change of the energy measure) and compare it with the standard use of 39 MFCCs to see if that has any effect on the recognition performance.

3.2 Acoustic model

The development of the acoustic model is based on the HTK 3.5 (Young et al. 2015), which is a portable toolkit for building Hidden Markov Models (HMMs). This version integrates deep neural network (DNN) modules to be used for acoustic modelling and feature extraction. The DNN tools in HTK 3.5 enable 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 converted to a probability distribution, typically by using softmax though other options are available, 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). Wang et al. (2016) report that using DNNs in this way within the HTK produces results that are comparable with, and in some cases better than, other DNN systems tested on the same data.

The actual acoustic modelling takes place in multiple stages, starting from the creation of an initial set of identical monophone HMMs. This pre-defined prototype is used by the HTK, along with the acoustic feature vectors, for initialisation. This is followed by creating short-pause models and extending the silence model to make the system more robust. Since a word may have multiple pronunciations in the dictionary, the created phone models are used to realign the training data and create new transcriptions by selecting from among the alternatives listed in the dictionary. This forced alignment can help to improve the phone-level accuracy as it determines the pronunciation that best matches the acoustic data and then uses this for the phonetic transcription in subsequent rounds of training.

After the creation of a set of monophone HMMs, context-dependent triphone HMMs are created. This is done by (A) converting monophone transcriptions to triphone transcriptions, generating a list of all triphones observed in the training data. The HMM model is now built for each triphone, where there is a separate model for each left and right context for each phoneme and phone. And (B), tying similar acoustic states of these triphones to make robust parameter estimates. These are standard steps when training speech recognisers (Young et al. 2015). Using

triphones can lead to overtraining if the size of the training data is small, but with the GALE data this step makes a significant contribution.

After each step, re-estimation of the parameters must be performed using several rounds of Baum-Welch training. This is done to estimate the optimal values for the HMM parameters (transition probability, mean and variance vectors for each observation function). This iterative step is repeated several times for each HMM to train. The HTK book suggests applying 2 rounds of re-estimation, however, the research reported here found that applying 3 rounds of re-estimation after each HMM training can make the system more effective.

In building the acoustic model, standard DNN-HMM systems are used. Similarly to the monophone system creation explained previously, a prototype model first needs to be defined. Then other tools connect the DNN output units with the monophone HMM states. In addition, the DNNs created in this step are paired up with the cross word triphone HMMs and eventually evaluated.

In this work 6 hidden layers are used, where each hidden layer has 2048 nodes. The input layer has 225 nodes (the 25 MFCC and energy features observed in a window of 9 frames around the current frame) and the output layer has either 96 or 105 nodes. These correspond to the three internal states of each of the HMMs for the phoneme set being used for a given experiment, i.e. 32 phonemes for the grapheme-based experiments and 35 for the dictionary-based ones. Although we do use triphones during the initial GMM training round (used for getting initial estimates of the various HMMs and for aligning the data prior to DNN training), these all get tied to the basic phoneme set because the data does not contain sufficient instances of the individual triphones for them to be assigned distinct models. Stochastic gradient descent (SGD) is applied for pre-training with mini-batch size set to 256 with 0.001 as initial learning rate for all the experiments, using a sigmoid as the activation function for the hidden layers and softmax for the output layer. In order to fine-tune the pre-trained nets we used 18 epochs with a fixed learning rate of 0.001 and the same mini-batch as that used for pre-training. We investigated numerous other settings for these parameters, but the current paper is not primarily aimed at analysing the effects of changing the DNN architecture, and all the results below are obtained using these settings. It is worth noting at this point that the initial GMM models which are used for producing the alignment used when training the DNN models are almost as accurate as the DNN models themselves. DNNs are typically good at extracting detailed patterns from very large datasets. When only a moderate amount of data is available, as in the current case, they do not outperform other models as effectively.

Having sufficient recordings from different speakers is crucial for creating speaker-independent ASR systems. In this research, we investigate building different acoustic models using different settings and different speaker populations. This is carried out by testing the effect of excluding speakers with few utterances, to test the hypothesis that those speakers may have a bad impact on modelling the acoustic. Another way to improve the acoustic modelling is by dividing speakers into sub-populations and training the acoustic model separately on each population.

3.3 Pronunciation model

Pronunciation model provides the link between the language model and the acoustic model. The pronunciation lexicon includes a list of the words with single or multiple phonetic transcriptions. Two levels of textual representation in dictionary can be used; grapheme-based (with no short vowels) and phoneme-based (including short vowels).

The use of grapheme-based transcription was found by Alghamdi et al. (2007); Elmahdy et al. (2010); Abushariah et al. (2012); Alsharhan et al. (2020) to be advantageous in reducing WER. In contrast, other researchers found these non-diacriticised scripts to have considerable ambiguity for the development of NLP tools, such as Vergyri and Kirchhoff (2004). Due to this uncertainty regarding the best level of textual representation, the research investigates the use of both grapheme-based pronunciation modelling and phoneme-based pronunciation modelling.

In the case of building a grapheme-based model, pronunciation modelling is a straightforward process. The phonetic transcription is obtained from the word's graphemes rather than the exact phoneme sequence, and the dictionary is generated from the text without retrieving the diacritics. For any given word, pronunciation modelling is done by splitting the word into letters. In this case, each word is associated with only one graphemic pronunciation variant. Note that in this representation, the properties of the missing short vowels are assumed to be implicitly modeled with the surrounding consonants during the acoustic modelling.

In order to create the phoneme-based model, the research uses the QCRI predefined dictionary, which was developed by Qatar Computing Research Institute[2]. The QCRI dictionary was developed using a collected news archive from many news websites and then processed by MADA. The lexicon has 526k unique grapheme words, with 2M pronunciations, with an average of 3.84 pronunciations for each grapheme word (Ali et al. 2014).

3.4 Language model

Good performance speech recognisers cannot be achieved through acoustic modelling solely; some form of word sequence probability estimate is required to capture the properties of the language. Language modelling is crucial to constrain search by limiting the set of possible HMMs. The role of the language model is to predict the probability of a word occurring in the context during recognition, which can help to improve the recognition accuracy.

This can be done either by using an associated grammar, or probabilistically by computing the likelihood for each possible successor word using n-gram models. The research reported here uses the classical bi-gram language modelling in building all the proposed systems. 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. These tools are applied to the test sets: given that we use

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

the same test sets for each experiment, this provides us with the same set of language models for every experiment.

4 GALE corpus specifications

The research uses the GALE (phase 3) Arabic broadcast news and broadcast conversational speech dataset. This dataset consists of two major parts: the first one contains approximately 132 h of Arabic broadcast news speech (BN) collected from 13 Arabic channels. The second part contains approximately 129 h of Arabic broadcast conversation speech (BC) collected from 17 channels.

This data consists mainly of MSA speech, but with a substantial amount of DA speech, especially in the conversational part. After applying orthographic normalisation, the transcripts were segmented into manageable and well-defined segments according to the given time stamps, along with the associated recordings. Each segment is given a specific label. All unnecessary information, such as non-speech segments and non-Arabic text was removed. However, interjection, hesitation and broken words were kept as long as they have orthographic transcriptions.

In the GALE data, a quick rich transcription (QRTR) is used. This kind of transcription can be carried out more quickly, but with fewer quality checks performed on the finished product, compared to careful transcription. Therefore, some limitations were faced when using this data. For example:

- Time stamps are sometimes placed in the middle of the sentence.
- Some speakers are labeled with the wrong gender or sometimes with no gender identification.
- Transcripts have some irregularities such as the use of non-Arabic punctuation marks and non-Arabic graphemes.
- Some recordings were found to have very noisy background.
- Many cases were found with discrepancies and spelling mistakes.

We have compensated for these errors as far as we can—most of the orthographic errors were treated by using regular expression rules, and we found that many of the noisier segments were produced by people for whom only a very small amount of data was recorded, so that we could use this as a simple way of removing noisy recordings. Problems like these are, however, to be found in almost any large corpus. The experiments reported here are in large part an attempt to find out the best way to cope in the face of data that contains these problems.

The GALE data was used for training and testing the ASR systems through the investigation carried out in this research. Three datasets were used: the full BN dataset with approximately 55.6 h. of speech data (after preprocessing) and 401k of vocabulary size; the full BC dataset after preprocessing it, which consists of approximately 91 h. of speech data and 679k of vocabulary size; and the combination of both BN and BC datasets.

In conducting dialect-specific experiments, information about speakers' dialects was extracted from an annotated version of the GALE (phase 3) corpus provided by

Table 1 Details of amount of data available for each dialect

Dialect	Amount of speech (mins)		
	Total	By gender	
Gulf	578	Male	493
		Female	85
Iraqi	448	Male	427
		Female	21
Egyptian	303	Male	255
		Female	48
Levantine	1265	Male	741
		Female	524
Maghrebi	70	Male	65
		Female	5

Alsharhan and Ramsay (2020). Arabic dialects can be classified based on different aspects, in terms of geography and social class. The dialectal labels provided by Alsharhan and Ramsay (2020) are based on dividing regional dialects into five main groups. This classification of Arabic dialects is commonly used by researchers in the area of Arabic language processing. However, Zaidan and Callison-Burch (2014b) stated that this breakdown is one possible classification but it is relatively coarse and can be further divided into more dialect groups, especially in large regions such as the Maghreb. The annotations are fully available online for searching and downloading[3]. The proposed annotation process resulted in assigning a dialect label of about 2,900 speakers. Each speaker was assigned to an accent group by three annotators. Annotators did not always agree, and hence the final database records all three judgements. In the current research, we rigorously used dialect labels provided by unanimous annotators to achieve more reliable results. This leads to having 88.6 hrs of annotated speech from the GALE data. Table 1 provides information about the amount of data we have for each dialectal group for both genders.

5 Dividing data into training and testing

In order to carry out a thorough investigation and achieve a fair comparison between the varied systems, the research uses a 5-fold cross-validation approach as a way of assessing the proposed systems. This includes shuffling the dataset and extracting five random sets for training and testing. In each fold, a random set of test data is selected that is disjoint from the training data. The results from the five folds can then be averaged to compute a single estimation. This is particularly important when carrying out experiments with limited data sources, such as dialect-specific and gender-specific systems.

[3] https://github.com/AllanRamsay/ACCENTS

However, we cannot follow standard practice by selecting the test set by taking 20% of the data for testing and 80% for training in the main experiments because different experiments involve different amounts of training data: if one experiment involves taking the full 55.6 h of the BN data then 20% would be 678 minutes of test data, if a different experiment involved taking 5 h of the BN data then 20% would be 60 minutes of test data. We therefore wanted to find out what would be a sensible amount of test data to reserve for every experiment in order to have reliable results without making testing take an excessively long time and whilst retaining consistency over the various experiments.

The question is how much data do we need to reserve and use in order to achieve sufficient testing? To answer this question we carried out a series of experiments with varied amount of testing data to develop learning curves and find how much data we need to be able to fairly estimate the error rate. The experiments involve training four different models using different datasets. Each model is then tested with varied amounts of testing data ranging from 3 to 60 minutes. The BN data, which consists of about 55.6 hrs of speech recordings and associated transcriptions, is used in these experiments. In preparing the data for training and testing the main four systems, the whole dataset is shuffled and a different 60 minutes of speech is reserved every time to carry out the gradual testing. Results are shown in Fig. 2 where the Y axis indicates the Word Error Rate (WER), the color of the line distinguishes between four different models (trained using different subsets of the training data), and the X axis indicates the amount of data used in testing each system. The experiments confirm that the WER starts to be stable between 15 and 20 min.

It is important to be clear that the point of this experiment was to ascertain how much data we should reserve for testing when carrying out the main experiments in Sect. 6. For each model, we randomly extracted 60 minutes of potential test data from the full 55.6 h in the GALE BN set and used the remaining 54.6 h for training the model. We then tested the model on increasingly large chunks of the reserved data in order to see the point at which the results became stable. We are not at this point concerned with the performance of the models themselves. The issue is to find the point at which the test results level out, since it is reasonable to assume that if the results from the four experiments all become level at a certain point then that indicates that this is enough data to provide reliable results. In each case in Fig. 2 the curve levels out after we use about 20 minutes of testing data. We therefore used this amount of data for testing in all subsequent experiments.

6 Experiments

6.1 Testing general experimental conditions

In the first experimental phase we want to test the effect of applying some general conditions to the data available to us. In the reported experiments, we use three datasets: BN, BC, and the union of the two more specific datasets. 20 minutes of

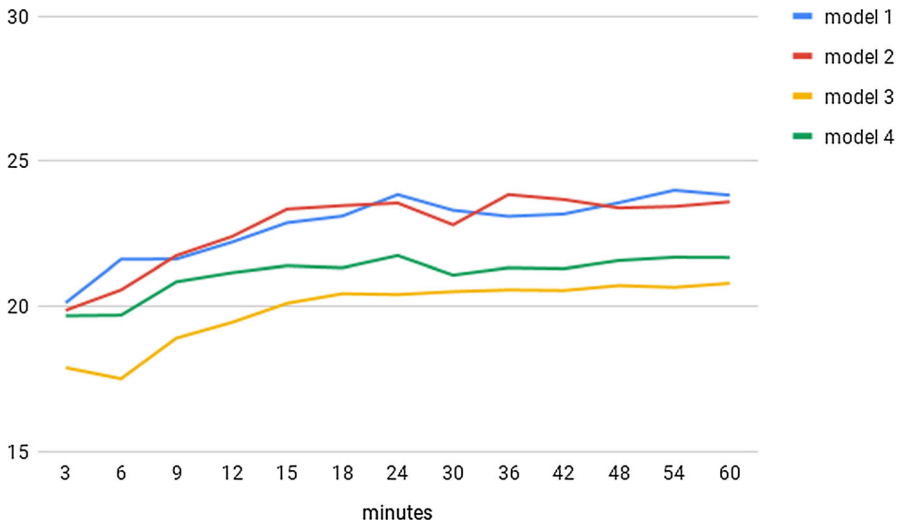

Fig. 2 WER for testing different ASR systems with varied amounts of testing data

data is reserved in each experiment for testing. The testing is carried out using 5-fold cross validation approach as explained in Sect. 5.

The baseline system is based on using 39-dimension MFCCs as our features and a 526k word multiple pronunciation dictionary. All speakers available in the datasets are included in training the baseline system without any restrictions.

The first question in this phase is whether using 25-dimensional MFCC (i.e. using 13 MFCC and their rate of change) is superior to the use of the standard 39-dimensional MFCC (the 13 coefficients, their rate of change and their acceleration). It is widely assumed that using the acceleration as well as the rate of change will produce better performance, but this may not be true for comparatively small training sets, since it may lead to over-training.

The experimental results show that by using 25-dimensional MFCC the WER decreases between 0.7% to 2.6% using different datasets as shown in Table 2. Ignoring the acceleration produces greater benefits for the broadcast news data than for the conversational data. This is slightly surprising, since the conversational data is in general noisier, and hence any information that might be gleaned from the acceleration would at first sight seem to be more unreliable.

The second round of experiments tries to find a surrogate approach to deal with noise and unwanted variation in speech data. We believe that sometimes more data can hurt, with this being particularly true when that data includes unnecessary and disruptive information. This has motivated the research to adopt a data selection strategy which limits the selection of speakers to those who have at least ten recordings. The assumption that speakers with few utterances can negatively affect the performance of the ASR system was driven by general observations made from listening to some recordings in GALE data. Those observations suggest that the majority of speakers with few utterances have noticeably high background noise and

Table 2 WER for the three datasets using different experimental settings

	Phoneme-based			Grapheme-based		
	BN (%)	BC (%)	Combined (%)	BN (%)	BC (%)	Combined (%)
1 baseline system (all speakers, 39 dimension MFCC)	22.7	33.3	28.3	19.9	31.1	26
2 applying 25-dimension MFCC	20.1	32.6	26.4	19	31.2	24.8
2+ eliminating infrequent speakers	19.0	31.5	25.6	17.3	30.0	23.1

poor quality recordings. Experimental evidence shows that by excluding speakers with less than ten utterances the WER decreased between 0.8% to 1.1%. This is not, of course, a direct demonstration that these speakers contribute particularly noisy data. The most that we can say is that informal observation suggests that the recordings for speakers for whom we have only a small amount data tend to be noisy, and that excluding such speakers leads to improved overall performance. The link between the two is tentative, but the concrete results are not in doubt.

While the baseline system uses a predefined multiple pronunciation dictionary, the research investigates the effect of using a simple transcription where each character in the written string is treated as the name of a phoneme and no attempt is made to insert the missing diacritics. We will refer to systems trained using this strategy as 'grapheme-based' systems. Results reported by previous researchers suggested that it was worth carrying out such testing, as some, though not all, researchers reported lower WER when testing on grapheme-based systems.

With the available data and tools, the experimental results confirm that the use of a grapheme-based transcription is superior to the use of the multiple pronunciation predefined dictionary. This superiority is manifested as a lower WER with all datasets and also faster training and testing compared to the use of a multiple pronunciation dictionary. Using the diacriticised multiple pronunciation dictionary leads to increasing the vocabulary size and thereby increasing the perplexity of the language model which explains the longer processing time. When using a grapheme-based dictionary, the properties of the missing short vowels are assumed to be implicitly modeled in the initial and final states of the HMMs for the surrounding consonants during the acoustic modelling. Results show that using a grapheme-based dictionary leads to decreasing WER by between 1.5% and 2.5%.

In summary, the investigation has proved that using 25 MFCCs, eliminating speakers for whom we have little data (which we take to be a reasonable surrogate for eliminating poor quality recordings), and using a grapheme based dictionary leads to reduction in WER by total of between 3.24% for BC data and 5.35% for BN data. A clear general observation in the results obtained is the substantial difference between the performance of systems that use BN data and systems that use BC data. Using the same tools and parameters in training and testing with the BC data resulted in an average of 11.95% increase in WER compared to the use of BN data. This can be explained by the nature of speech materials each dataset includes. The BC dataset is considered to be maximally natural in terms of speaking style and

recording conditions; speakers tend to use their own words, in most cases in their own dialects; and many episodes were recorded in open areas and streets, which increases the negative impact of background noise on the quality of the recordings. On the other hand, in the BN dataset most of the speakers read from source materials and the recording process was carried out in recording studios. However, the effect of applying the suggested conditions on the three datasets was broadly similar. The scores for the combined dataset are close to the averages of the scores for the two individual datasets. Interestingly, the scores for the combined dataset are slightly lower than the average of the individual datasets when using the phoneme-based approach and slightly higher than the average for individual datasets with the grapheme-based system, but the differences are very slight and are unlikely to be significant. Details follow in Table 2.

6.2 Using homogeneous datasets in training and testing the developed systems

A previous study examined the variability among speakers through the application of statistical analysis methods (Huang et al. 2001). This study found that the first two principal components of variation in speech correspond to gender and accent. The current investigation aims at finding the best way to split the data into subgroups when building Arabic ASR systems to limit variation sources and make better use of the data. The key point here is that homogeneous datasets are generally regarded as leading to more accurately trained systems; but dividing the data into small subsets means that there is less data available for training each model, which typically leads to less accurate models. We therefore wanted to investigate the trade-off between using homogeneous datasets vs. the decrease in the size of the training data as we split the dataset into finer and finer subsets. The investigation starts by dividing the three datasets according to speakers' gender groups[4] and applying the experimental conditions suggested in the previous section. Results of building gender-specific ASR systems are reported in Table 3. For comparison reasons, the table also reported the results of testing each gender on a model that is trained on both genders.

The results show that using gender-specific modelling leads to substantial decrease in WER. The average improvement is 2.45% decrease in WER for BN, 2.15% for BC, and 2.95% for the combined datasets. The results also show that despite the dominance of male speakers in the two datasets, models trained and tested on female speakers perform better than ones trained and tested on male speakers. This might arise because there is greater variation of pitch among male speakers than there is among females, or simply because female speakers articulate more clearly than males. We have not investigated this in greater detail.

This suggests that gender features are a great source of acoustic confusability in the construction of ASR systems. Controlling this variability leads to substantial

[4] GALE data files have names like `test-ALAM_NEWSRPT_ARB_20070125_015800-female-*Azza_Zaftaoui-native-166` which include a component that specifies the speaker's gender, which is what we used for the gender-oriented experiments. There are a few cases where this label appears to be wrong, but the labels are generally correct.

Table 3 WER for ASR systems trained on male, female, and both genders using the three datasets

		Training dataset								
		BN			BC			Combined		
		Male	Female	Both	Male	Female	Both	Male	Female	Both
Testing-dataset	Male	18.6%	–	20.3%	28.4%	–	30.2%	19.8%	–	23.1%
	Female	–	10.8%	14.0%	–	24.4%	26.9%	–	14.5%	17.1%

improvement in the overall performance. Despite the fact that the undifferentiated training set is much larger than the gender-specific ones, the latter is more informative and leads to better performance in a shorter time frame. It is, of course, necessary to be able to classify the input speech by gender for this to be helpful. This can be approximated by finding the value for F0—for the Gale Arabic data 84% of speakers with an average F0 of 172Hz or below are labeled male and 84% of speakers with a higher F0 are labeled female. Alternatively, speakers can simply be asked to specify their gender when accessing the tool.

The second major source of variability in speech arises from the use of different dialects. Although dialectal groups in Arabic share many similarities, the various dialects show differences at all linguistic levels. Looking at the intelligibility level among Arabic speakers, it can be noticed that Levantine speakers are unintelligible to Moroccan speakers and vice versa, for instance. These divergences suggest the importance of building dialect-specific ASR systems for Arabic as computational tools trained on one dialect will underperform when tested on another dialect. Similarly, a system simultaneously trained with many dialects is not expected to achieve well when tested on any specific dialect.

The experimental work reported here mainly involves building two kinds of recognition systems. In the first one, the selection of the training data is based solely on speakers' gender so it contains two multi-dialect speech recognisers (one for each gender). The second one is based on speakers' gender and dialect, so it contains nine recognition systems for the five main Arabic dialects and the two genders[5]. Table 4 shows the results of testing each dialect and gender-specific subset on gender-specific, multi-dialects based-system and on an appropriate gender-specific plus dialect-specific based system.

To avoid unfair comparisons resulting from the preponderance of some dialects in the dataset, especially Levantine, the training set in these experiments is balanced so that it includes a fixed amount of speech from each dialect and gender group. As with the previous experiments, 5-fold cross-validation with a 20 minute testset is used in carrying out the experiments.

The results clearly show that all gender plus dialect-specific speech recognisers show better performance than the multi-dialects based speech recognisers. That is, for instance, the Gulf (male) recogniser beats the multi-dialect male Arabic

[5] Female Maghrabi speakers were excluded from the training and testing sets because the GALE data contains too few instances of this class for it to be possible to train a model.

recogniser when operating over Gulf male speakers test set and likewise for all the other gender/dialect specific models.

The level of improvement varies depending on the dialect. For instance, the highest improvement was observed in Maghrebi dialect with 5.26 % reduction in WER, while the lowest improvement was observed with Levantine dialect with 0.83% reduction in WER for the male subset and 0.54% WER for the female subset.

The average of the WER for testing different dialect plus gender-specific subsets on multi-dialect, gender-specific based systems is 28.65% for male and 22% for female speakers, while testing on gender and dialect specific based systems results in 25.84% and 20.12% WER for male and female speakers, respectively. The reduction in WER achieved by this approach is good but not overwhelming. This might be explained by the nature of data we are using in these experiments. Many speakers in the GALE (phase 3) dataset are news presenters who are trained to use the standard form of the language and to be linguistically neutral. Those professional speakers adopt the use of MSA speaking rules, hiding their regional dialect as far as possible. We speculate that with fully spontaneous, conversation style dialectal dataset in building dialect-specific ASR systems, the difference between the systems will dramatically increase.

The results also show that models trained and tested on Levantine speakers have the best performance compared to models trained and tested on speakers from other dialects with 16.48% average WER. This might be explained by the high level of proficiency that Levantine speakers show. Most of the Levantine speakers in the dataset are well-trained broadcasters recording in noise-free studios. In addition, it is well-known in the linguistics domain that the Levantine dialect shares a great deal of lexical similarity with MSA and other Arabic varieties. This result supports previous research findings which found that Jordanian Arabic—which is one of the Levantine varieties—achieved relatively lower WER compared to other dialects (Biadsy et al. 2012). On the other hand, the worst performance was observed with Maghrebi speech with 42.3% to 36.77% WER. This is partly due to the fact that Maghrebi dialect is under-represented in the multi-dialect based system. This outcome supports a previous research result which confirms that Maghrebi dialect is the hardest dialect to be recognised by native speakers from other dialectal backgrounds, in addition to the linguistics fact that Maghrebi dialect does not share lots of its vocabulary with other Arabic varieties which makes it not mutually intelligible with other dialectal varieties (Ibrahim 2009). However, the Maghrebi subset was found to underperform even when tested on Maghrebi-specific recognisers. Maghrebi dialect exhibits many borrowed words from French and Spanish, even words from MSA origins have undergone major changes in its structure. Hence, building ASR systems for Maghrebi dialect might require using distinctive linguistic resources. In addition, it can be noticed that Maghrebi speech found in the dataset covers a huge range of speakers who do not actually share a common accent, such as speakers from Tunisia, Libya, Algeria, and Morocco.

Table 4 Average of WER for the multi-dialect based system and gender plus dialect-specific system, when evaluated on its heldout test data (numbers in bold indicates lower WER)

Testing subset	Training subset	WER
Gulf (M)	All dialects (M)	27.22%
	Gulf (M)	**25.30%**
Gulf (F)	All dialects (F)	22.1%
	Gulf (F)	**20.04%**
Iraqi (M)	All dialects (M)	25.47%
	Iraqi (M)	**23.15%**
Iraqi (F)	All dialects (F)	29.93%
	Iraqi (F)	**27.88%**
Egyptian (M)	All dialects (M)	25.77%
	Egyptian (M)	**22.05%**
Egyptian (F)	All dialects (F)	24.41%
	Egyptian (F)	**21.53%**
Levantine (M)	All dialects (M)	22.76%
	Levantine (M)	**21.93%**
Levantine (F)	All dialects (F)	11.57%
	Levantine (F)	**11.03%**
Maghrebi (M)	All dialects (M)	42.03%
	Maghrebi (M)	**36.77%**

6.3 Cross-dialect experiments

The results reported in the previous section prove that gender plus dialect-specific based recognition systems always beat recognition systems based on multi-dialects trained on the same amount of data. The experimental work reported in this section aims to investigate how well an Arabic ASR system trained on a particular dialect performs when tested on other dialects. In other words, we want to measure how far apart Arabic dialects are acoustically. This is particularly important in the research to regroup Arabic dialects based on their acoustic features rather than their geographical region. Eventually, we can come up with a better decision about the minimal number of recognition systems needed to cover all dialectal Arabic.

In order to carry out the cross-dialect evaluation for the five main Arabic dialects, nine ASR systems were built for each dialect and gender group (except female Maghrebi speakers for whom we do not have sufficient data). Those systems were tested with different dialectal subsets from the same gender. Results are reported in Table 5.

By analysing results reported in Tables 5 and 6, we can conclude to two main facts:

[5] Female Maghrabi speakers were excluded from the training and testing sets because the GALE data contains too few instances of this class for it to be possible to train a model.

- The average WER of cross-dialect systems is 30.9%, whilst the average WER for dialect-specific systems is 23.4%. This large difference supports findings reported in the previous section which confirm that the dialectal groups are acoustically different and that dialectal subsets work better with dialect-specific based ASR systems. For instance, all dialectal groups were found to achieve their best performance when tested over their corresponding system.
- The gap between testing on the same dialectal group and testing on other dialectal groups is huge. This gap not only indicates the importance of considering the five major dialectal groups in building ASR systems, but also suggests that it might be worth dividing some of these five dialectal groups into subgroups. For instance, instead of building a model trained on Levantine speech, we may think of building separate models for each sub dialect (Syrian, Palestinian, Lebanese, Jordanian, etc). This kind of investigation is crucial to make a better decision of the number of systems we need to have for Arabic instead of relying on the geographical information. Current state of the art speech recognition systems have multiple models for each language to include the available dialects. For instance, according to Elfeky et al. (2018), Google speech recognition system includes four models for Arabic, five for Spanish, and eight for English.

6.4 Training data size experiments

Machine learning systems generally give satisfactory results when the training dataset and the testing dataset are similar. There is also a point at which more training data does not make significant improvement to the system's performance. Moreover, if the data used in training the system has a high level of variability, the machine learning system will have difficulties in making the right generalisation with this sparse data. The question here is how much data do we need to sufficiently train a speech recognition system?

In order to answer this question, in the experimental work reported in this section we run a collection of experiments with increasing size of data. In the first set of experiments we did not put any restrictions on the selection of the data. In other words, the data used in running the experiments comes from multiple dialects and both genders. In the second set of experiments, we select the training and testing sets strictly from Levantine male speakers. The reason behind choosing data from Levantine male speakers is its high availability compared to other dialects and gender (Table 1). This will allow us to carry out enough experiments without running out of data. In running the experiments, we are taking random samples of increasing size from the data source. Each of the training sets includes everything that was in the previous training set (e.g. the 30 min training set includes the 15 min one used in the previous experiment, the 60 min set includes the 30 min set from the previous round, ...). The testing set was extracted previously from the dataset to insure that there is no overlap between the testset and the training data. Similar to other experiments, fivefold validation is used in carrying out the test.

Table 5 WER for testing five dialectal subsets of male speakers on each dialect-specific recognition system

		Training dataset				
		Gulf (F)	Iraqi (F)	Egyptian (F)	Levantine (F)	Maghrebi (M)
Testing dataset	Gulf (F)	**25.3**	27.91	29.8	32.25	33.37
	Iraqi (F)	27.48	**23.15**	32.2	31.51	30.52
	Egyptian (F)	29.86	31.98	**23.1**	31.11	31.02
	Levantine (F)	27.65	27.6	25.13	**21.92**	27.87
	Maghrebi (M)	45.8	44.2	42.42	41.55	**36.7**

Bold indicates the error rate

Table 6 WER for testing four dialectal subsets of female speakers on each dialect-specific recognition system

		Training dataset			
		Gulf (F)	Iraqi (F)	Egyptian (F)	Levantine (F)
Testing dataset	Gulf (F)	**20.04**	32.92	27.02	24.86
	Iraqi (F)	31.83	**27.88**	33.58	32.69
	Egyptian (F)	30.82	36.9	**21.8**	24.7
	Levantine (F)	19.31	30.3	13.22	**11.03**

Bold indicates the error rate

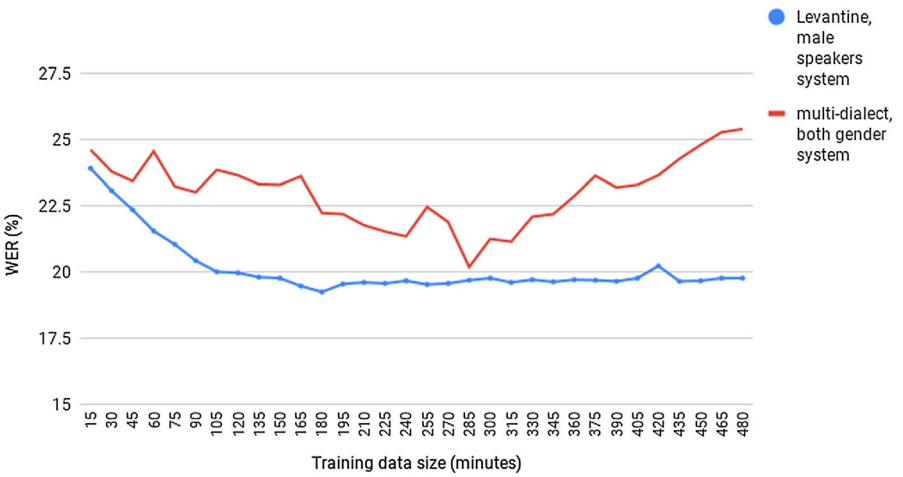

Fig. 3 Results for gradually increasing the amount of data using two resources: undifferentiated data, and data from Levantine, male speakers

Figure 3 shows the results of gradually increasing the amount of training data using two different data resources: undifferentiated dataset, and dialect & gender specific dataset (using samples from Levantine male speakers). The starting point is training the model with 15 minutes of data, extra 15 minutes are then added gradually until we reach 480 minutes. 20 minutes of the data is reserved for testing the developed systems.

The experiments reported in this section address two important questions:

- How much data do we need to achieve the optimal performance of an Arabic ASR system?
- Does applying any selection strategy on the data have any effect on the system's performance?

The experimental evidence suggests that it is more important to ensure that the training data is drawn from the same population as the target population than to maximise the amount of training data. This is, at some level, obvious – you would expect a speech recogniser trained entirely on female recordings to be better at recognising female speech than one that is trained on a mixture of male and female data. The experiments reported above show that this holds even when the populations that the data is obtained from are, at first sight, fairly similar, and where splitting them into sub-populations has a drastic effect on the amount of data that is available for training. Aggregating the training data from the five accent groups gives us five times as much data, but, because of differences among the accents, data aggregation leads to worse performance than we get treating the accent groups separately, even though the differences in accents are not always very clear (our annotators often had considerable difficulty when assigning accents, suggesting that speakers from different accent groups do not always sound very distinct). It is also noteworthy that providing too much detail is also unhelpful when you only have a modest amount of training data. Using 39 MFCC and energy coefficients produces worse performance than using 25 coefficients, and using the full phonetic transcription produces worse performance than using the simple graphemic transcription, which omits the short vowels. It may be that some of these effects would be reversed if more data were available, but annotating transcribed speech data for accent is time-consuming and expensive. At the very least, the work here suggests that if you only have a limited amount of data you should be careful about the tendency for learning algorithms to overtrain if the data contains too much detail.

7 Conclusion and future work

We have presented a thorough investigation of the properties of data that may affect the performance of Arabic ASR systems. The motivation behind this research was to overcome the lack of spoken and transcribed resources in the literature by presenting straightforward approaches for efficiently exploiting the available data. The investigation includes applying some general experimental conditions to the

data which showed that using 25-dimension MFCCs, eliminating poor quality recordings, and using a grapheme based dictionary lead to reduction in WER by total of between 3.24% and 5.35%.

The research also introduced a data selection strategy that presents a multiple modelling approach instead of using a single model to cover all the variability found in the speech. The experimental results showed that building gender- and dialect-specific models leads to substantial decrease in WER. All the gender- and dialect-specific systems consistently outperform the combined system, despite the fact that the latter is trained using about 5 times as much training data. Applying such strategies is crucial if we are to overcome the limited availability of the data, reduce training time, and achieve the best performance. Further research is needed to understand the reasons behind the varied performance of some dialect-specific models and to find the reason behind the difficulty of recognising Maghrebi speech compared to other dialects. Cross-dialect experiments are also carried out in this research to understand how different are Arabic dialects acoustically and to help us to rethink about the minimal number of models needed to be built to cover all Arabic varieties. The experimental results confirmed that all dialect-specific subsets perform better on their corresponding dialect-specific systems. It also confirms that the gap between testing on the same dialectal group and testing on other dialectal groups is huge, which calls for the importance of studying the feasibility of applying further division on the main dialectal groups. Finally, the research carried out a set of experiments to address the question of the amount of training data needed to build good performance ASR systems. The outcome of these experiments confirmed that contrary to the common belief that "there is no data like more data", consistently feeding the model with more arbitrary data can worsen the performance of the system. At the same time, using carefully selected subsets of data produces recognition systems that is superior to a system that makes use of a much larger amount of undifferentiated data.

Acknowledgements The research was supported by the research sector at Kuwait university—(Grant AA01/18). We would like to thank Chao Zhang of the Machine Intelligence Laboratory at the University of Cambridge for his endless patience while we were installing and experimenting with the CUDA version of the HTK.

References

Abushariah, M., Ainon, R., Zainuddin, R., Al-Qatab, B., & Alqudah, A. (2010). Impact of a newly developed modern standard Arabic speech corpus on implementing and evaluating automatic

continuous speech recognition systems. *Spoken Dialogue Systems for Ambient Environments* (pp. 1–12).

Abushariah, M. A.-A. M., Ainon, R., Zainuddin, R., Elshafei, M., & Khalifa, O. O. (2012). Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. *International Arab Journal of Information Technology (IAJIT)*, *9*(1), 84–93.

Alghamdi, M., Elshafei, M., & Al-Muhtaseb, H. (2007). Arabic broadcast news transcription system. *International Journal of Speech Technology*, *10*(4), 183–195.

Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., & Glass, J. (2014). A complete kaldi recipe for building Arabic speech recognition systems. In *2014 IEEE Spoken Language Technology Workshop (SLT)* (pp. 525–529).

Ali, A., Zhang, Y., Cardinal, P., Dahak, N., Vogel, S., & Glass, J. (2014). A complete kaldi recipe for building Arabic speech recognition systems. In *Spoken Language Technology Workshop (SLT), 2014 IEEE* (pp. 525–529). IEEE: New York.

Almeman, K., Lee, M., & Almiman, A. A. (2013). Multi dialect Arabic speech parallel corpora. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (pp. 1–6). IEEE: New York.

Alsharhan, E., & Ramsay, A. (2020). The development of a speech corpus annotated for the main Arabic dialects. *Arab Journal for the Humanities*, 150.

Alsharhan, E., Ramsay, A., & Ahmed, H. (2020). Evaluating the effect of using different transcription schemes in building a speech recognition system for Arabic. *International Journal of Speech Technology*, 1–14.

Andrusenko, A., Laptev, A., & Medennikov (2019). Russian open speech to text (STT/ASR) dataset. https://github.com/snakers4/open_stt. Accessed: 2020-07-7.

Biadsy, F., Moreno, P. J., & Jansche, M. (2012). Google's cross-dialect Arabic voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4441–4444). IEEE: New York.

Elfeky, M. G., Moreno, P., & Soto, V. (2018). Multi-dialectical languages effect on speech recognition: Too much choice can hurt. *Procedia Computer Science*, *128*, 1–8.

Elmahdy, M., Gruhn, R., Minker, W., & Abdennadher, S. (2010). Cross-lingual acoustic modeling for dialectal Arabic speech recognition. In *Eleventh Annual Conference of the International Speech Communication Association*.

Elmahdy, M., Hasegawa-Johnson, M., & Mustafawi, E. (2012). A baseline speech recognition system for levantine colloquial Arabic. In *12th ESOLEC conference on Language Engineering*.

Elmahdy, M., Hasegawa-Johnson, M., & Mustafawi, E. (2014). Development of a tv broadcasts speech recognition system for qatari Arabic. *LREC*, 3057–3061.

Huang, C., Chen, T., Li, S., Chang, E., & Zhou, J. (2001). Analysis of speaker variability. In *Seventh European Conference on Speech Communication and Technology*.

Huang, P.-S. & Hasegawa-Johnson, M. (2012). Cross-dialectal data transferring for gaussian mixture model training in Arabic speech recognition. *constraints*, 1:1.

Ibrahim, Z. (2009). *Beyond lexical variation in modern standard Arabic: Egypt*. Lebanon and Morocco: Cambridge Scholars Publishing.

Kirchhoff, K., & Vergyri, D. (2005). Cross-dialectal data sharing for acoustic modeling in Arabic speech recognition. *Speech Communication*, *46*(1), 37–51.

LDC. (2015). *Gale phase 3 Arabic broadcast and conversation speech*. Philadelphia: Linguistic Data Consortium.

Lewis, M. P. Gary, F. (2015). *Ethnologue: Languages of the world*.

Magic Data Technology Co., Ltd. (2019). MAGICDATA Mandarin Chinese read speech corpus. (http://openslr.org/68/). Accessed: 2020-07-7.

Masmoudi, A., Khmekhem, M. E., Esteve, Y., Belguith, L. H., & Habash, N. (2014). A corpus and phonetic dictionary for tunisian Arabic speech recognition. *LREC*, 306–310.

Menacer, M. A., Mella, O., Fohr, D., Jouvet, D., Langlois, D., & Smaili, K. (2017). Development of the Arabic Loria automatic speech recognition system (ALASR) and its evaluation for Algerian dialect. *Procedia Computer Science*, *117*, 81–88.

Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Eighth European Conference on Speech Communication and Technology*.

Open Speech and Language Resources (2003). The Tunisian MSA corpus. (http://openslr.org/46/). (Accessed: 2020-07-7).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). IEEE: New York.

Sadat, F., Kazemi, F., & Farzindar, A. (2014). Automatic identification of Arabic dialects in social media. In *Proceedings of the first international workshop on Social media retrieval and analysis* (pp. 35–40). ACM: New York.

Selouani, S. A., & Boudraa, M. (2010). Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*, *35*(2C), 158.

Sharma, D. P., & Atkins, J. (2014). Automatic speech recognition systems: challenges and recent implementation trends. *International Journal of Signal and Imaging Systems Engineering*, *7*(4), 220–234.

Shoufan, A., & Alameri, S. (2015). Natural language processing for dialectical Arabic: A survey. *Proceedings of the Second Workshop on Arabic Natural Language Processing* (pp. 36–48).

Vergyri, D., & Kirchhoff, K. (2004). Automatic diacritization of Arabic for acoustic modeling in speech recognition. In *Proceedings of the workshop on computational approaches to Arabic script-based languages* (pp. 66–73). Association for Computational Linguistics.

Walker, K., Caruso, C., Maeda, K., DiPersio, D., & Strasse, S. (2013). *Gale phase 2 Arabic broadcast and conversation speech*. Philadelphia: Linguistic Data Consortium.

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

Woodland, P. C., Liu, X., Qian, Y., Zhang, C., Gales, M. J., Karanasou, P., Lanchantin, P., & Wang, L. (2015). Cambridge university transcription systems for the multi-genre broadcast challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 639–646). IEEE: New York.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., & Povey, D., et al. (2015). Phil woodland, and chao zhang". *The HTK book (for HTK version 3.5)," Cambridge University Engineering Department*.

Zaidan, O. F., & Callison-Burch, C. (2014a). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171–202.

Zaidan, O. F., & Callison-Burch, C. (2014b). Arabic dialect identification. *Computational Linguistics*, *40*(1), 171–202.