Utrecht University

Artificial Intelligence

Faculty of Humanities

Bachelor Thesis (7,5 ECTS)

Exploring Arabic Automatic Speech Recognition Bias

Author:
Remy van Tussenbroek
(6480373)

Supervisor:
Maria Francisca Pessanha MSc

Second Reader:
dr. Rianne van Lambalgen

July 2, 2021

Abstract

Research is limited regarding the field of Arabic Automatic Speech Recognition (ASR) systems and the bias in these systems. This paper expands on this field by conducting a literature review with the aim to discover how the Arabic language and dialects could introduce difficulties for ASR models. Additionally, it aims to describe the current social situation in Arabic speaking countries and it's perception by the media. To add to that, we explain ASR and discuss the limitations it has for under-resourced languages like Arabic. Lastly, we discuss various datasets, data-distributors and software for Arabic and compare them with the norm for English. We concluded that the dialects and the general complexity of Arabic form a challenge for ASR models. To add to that, the conflicts in the Arab world intensified negative stereotyping of Arabs, that were emphasized by the media. Additionally, since Arabic is an under-resourced language it is hard to find labeled data to train ASR models. Moreover, compared to English, Arabic datasets are less widely available in larger size, less diverse and contain less different types of content related to natural speech. Future work could extensively analyse datasets using data analysis methods to discover more potential reasons for bias.

Contents

1 Introduction

1.1 Motivation

Automatic speech recognition (ASR) has many applications like Text-to-Speech Synthesis, Speech Recognition, Speech Segmentation and Labeling (Transcription), Language Identification, Attitude and Emotion recognition and Spoken Dialog Systems (Karpagavalli & Chandra, 2016).
ASR converts a speech signal into a textual representation, i.e. sequence of said words by means of an algorithm implemented as a software or hardware module (Besacier et al., 2013). Modern automatic speech recognizers are built using various techniques from the field of Artificial Intelligence (AI), such as Hidden Markov Models (Young, 2008) and Support Vector Machines (Solera-Urena et al., 2007).

Current research (like Bolukbasi et al., 2016) in ASR is mostly focused on widespread languages like English. The research on Arabic ASR is limited, even though Arabic is currently one of the most widely spoken languages in the world, with more than 310 million people speaking all varieties of Arabic (Eberhard et al., 2019). This thesis builds on this lesser explored field of Arabic ASR. The languages that are understudied, suffer from the lack of resources, or annotation force (i.e. labeled data), and hence ASR algorithms might work not as good in these languages (Bircan & Ceylan, n.d.).

In general, ASR systems promise to deliver an objective interpretation of human speech. However, they still struggle with the large variation in speech due to e.g., gender, age, speech impairment, race, and accents. Many factors can cause the bias of an ASR system (Feng et al., 2021). In particular, it is known that ASR algorithms reflect religious and orientalist biases in automated transcripts by misattributing images related to conflict, war, and religion to Arabic speakers (Bircan & Ceylan, n.d.). We will discuss parts of the research done in the field of Arabic ASR and attempt to discover potential reasons for bias.

1.2 Objectives

In this thesis, we propose to discuss how the Arabic language and its dialects are structured. The complexity of Arabic could introduce difficulties for ASR models. Additionally, this thesis will explain how ASR models work in general and discuss the limitations ASR has for under-sourced languages like Arabic. To add to that, we attempt to discover potential reasons for bias by comparing properties of English and Arabic datasets. The objectives of this thesis could be summed up as the following:

- Discover how the Arabic language and dialects could introduce difficulties for ASR models.
- Describe the current social situation in Arabic speaking countries and it's perception by the media.
- Explain ASR and discuss the limitations ASR has for under-resourced languages like Arabic.
- Discuss various datasets, data-distributors and commonly used software for Arabic ASR and compare them with the norm for English datasets used in the same context. All with the intent to discover potential reasons for bias and points that could improve Arabic ASR.

1.3 Overview

We will first discuss Arabic as a language and the social context of Arabic speaking countries. Automatic Speech Recognition is then explained and we will go over Bias in Automatic Speech Recognition. Next, the focus is on Arabic Automatic Speech Recognition and Bias in Arabic Automatic Speech Recognition. Lastly, we discuss various datasets, data-distributors and commonly used software for Arabic ASR and comparing them with the norm for English datasets used in the same context. A conclusion is then formed to get a clear image of the findings.

2 Literature review

2.1 Arabic as a language and the social context of Arabic speaking countries

2.1.1 Arabic as a language

The macrolanguage Arabic, an overarching language consisting of Arabic varieties, is spoken by more than 340 million people (Eberhard et al., 2021). Currently there are a total of 25 independent states and territories that have Arabic as their native language (World Population Review, 2021). According to ISO 639-3, Arabic knows 30 separate forms of languages/dialects, like Algerian Saharan Arabic, Sudanese Arabic and Standard Arabic (SA). SA is the 5th most spoken language in the world (Eberhard et al., 2021). SA is divided by Western linguists into Classical Arabic (CA) and Modern Standard Arabic (MSA) (Kamusella, 2017). CA is based on the scriptures in the Quran. MSA follows the grammatical standards of CA and uses a large portion of its vocabulary. It has adapted CA to be more fitting to the current society by discarding some grammatical constructions and vocabulary, and adding new words that correspond to the modern era.

The largest variations (figure 1) in versions of Arabic are caused by regional differences. Arabic variants are divided into five major groups: Peninsular; Mesopotamian; Levantine; Egypto-Sudanic; and Maghrebi (Al-Wer, 2018; Eisele, 1987). An important factor in the differentiation of varieties is the influence from other languages previously spoken or still currently spoken in these regions, such as Coptic in Egypt (Wilfong, 2018) and French in North Africa (Aitsiselmi, 2008).

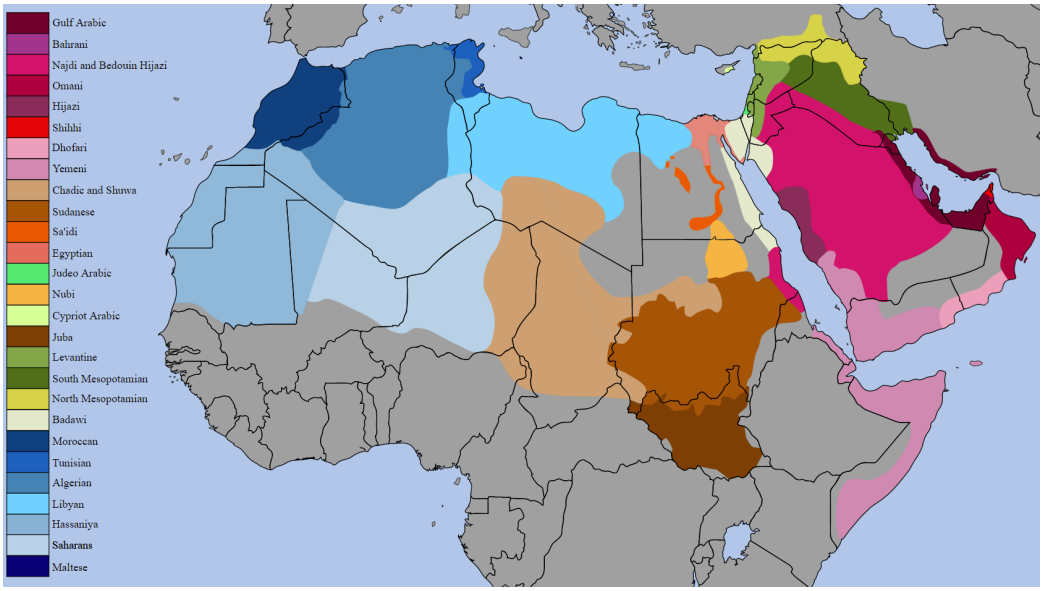

Figure 1: Different Arabic varieties in the Arab world. Reprinted from Varieties of Arabic, In Wikipedia, n.d., Retrieved June 26, 2021, from https://en.wikipedia.org/wiki/Varieties_of_Arabic.

2.1.2 The social context of Arabic speaking countries

The Arab world has known various conflicts, some have been present for several years already. The conflicts in Syria and Yemen, which started in 2011 (Deutsche Welle, 2016) and the territorial disputes between Palestinians and Israelis, which still seem to go on without end (BBC News, 2021). To add to that, there is the impact that external forces had in the political and economic formation of the countries that make up the Middle East, from the post-colonial period to the present day.
The terrorist attacks of 11 September 2001 (The New York Times, 2001) in the USA provoked a new war in Iraq in 2003 (BBC News, 2016). The effects of the Iraq War and the Arab Spring (Safi et al., 2021) in 2011 led to the Syrian Civil War (CNN, 2016). This civil war contributed to the rise of the Islamic State in 2014 (BBC News, 2014), which destabilized several countries in the region. (Farias, 2020)

Many researchers have emphasized the negative stereotyping of Arabs: violent, ruthless, irrational, destructive (Alexander & Brewer, 2005), hostile,

aggressive, victimised (Elayan, 2005), uncivilised, semi- barbaric, cunning, untrustworthy, dogmatic, radical Muslims, strongly supportive of terrorism (Nassar, 2008, as cited in Najm, 2019). The media has played an active role in forming and spreading negative stereotypes about Arabs for over a century. There are more than 900 Hollywood films depicting Arabs as heartless, brutal, uncivilized, religious fanatics, demonstrating a love for wealth and power (Shaheen, 2008, as cited in Najm, 2019). Due to the Western fear, hostility towards Arab and Islamic countries can increase hate crimes and violence against Arabs and Muslims. Arab- American Muslims are perceived as "terrorist neighbors" by American citizens (Shaheen, 1988, as cited in Najm, 2019). Following September 11, 2001, the Federal Bureau of Investigation (FBI) reported a 1,7 percent increase of hate crimes against Muslim Americans between 2000 to 2001 (Anderson, 2002, as cited in Khan and Ecklund, 2012). These crimes against Arabs are often due to misunderstanding, misperception, and stereotypes about Arab culture and heritage. Not all Arabs are Muslims (as American popular culture perceives), and in fact the majority in the United States are Christians (LESS and CSMU, 2015, as cited in Najm, 2019). The terms "Islamophobia" and "Arabophobia" are commonly used, more so, after the September 11 terrorist attacks (Zimbardo, 2014).

The political and military situation in Arab speaking countries and the resulting stereotypes emphasized by the media are important to consider as they could be influential for biases in Arabic ASR. The biases people clearly have regarding people of the Arab world can unintentionally get intertwined with datasets used for ASR models. Moreover, it is known that language corpora in ASR systems actually contain human-like biases (Caliskan et al., 2017).

2.2 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is the process of converting a speech signal to a sequence of words (i.e., spoken words to text) by means of an algorithm implemented as a computer program (Karpagavalli, 2016).
Several dimensions of variation of ASR systems are identified (Jurafsky & Martin, 2009): vocabulary size (speech recognition is easier if the number of distinct words needed to recognize is smaller), channel and noise (low quality audio-equipment and introducing noise makes recognition harder), accent or speaker-class characteristics (dialect that matches training data is easier to

recognize) and lastly the type of natural speech.
Different types of natural speech include spelled speech (with pauses between letters or phonemes), isolated speech (with pauses between words), continuous speech (when a speaker does not make any pauses between words), spontaneous speech (e.g. in a human-to-human dialog) and highly conversational speech (e.g. meetings and discussions of several people).

Modern automatic speech recognizers are constructed using various approaches (Besacier et al.,2013), like Hidden Markov Models (HMM) (Young et al., 2008), Dynamic Time Warping or Dynamic Programming (Jing & Min, 2010), Dynamic Bayesian Networks (Stephenson et al., 2002), Support Vector Machines (Solera-Urena et al., 2007) or specific hybrid models (Trentin Gori, 2001; Ganapathiraju et al., 2015). The dominant paradigm for Large-Vocabulary Continuous Speech Recognition is the HMM. HMM-based speech recognition systems make use of a noisy-channel model. This model searches through a huge space of potential "source" sentences and chooses the one which has the highest probability of generating the "noisy" sentence. The noisy-channel introduces several probabilistic components to ASR:

- Prior probability of a source sentence - computed by N-grams
- Probability of words being realized as certain strings of phones - computed by HMM lexicons
- Probability of phones being realized as acoustic or spectral features - computed by Gaussian Mixture Models

The general probabilistic structure of ASR can be formalized as follows:
The acoustic input O consists of a sequence of observations, measured every 10 milliseconds:

$$O = o_1, o_2, o_3, \cdots, o_t \tag{1}$$

The source sentence is composed of a string of words:

$$W = w_1, w_2, w_3, \cdots, w_n \tag{2}$$

The main goal is to find the string of words with the highest probability, given the set of observations:

$$\hat{W} = \text{argmax}_{W \in L} P(W|O) \tag{3}$$

This equation implies that for a given sequence W and acoustic input sequence O, the probability $P(W|O)$ needs to be determined. Bayes' theorem can be applied to this probability. After ignoring $P(O)$ (since it does not change for each sentence), we arrive at the following equation:

$$P(W|O) = P(O|W)P(W) \tag{4}$$

$P(O|W)$ is the observation likelihood, which is computed by the acoustic model. $P(W)$ is defined as the prior probability, which is computed by the language model.

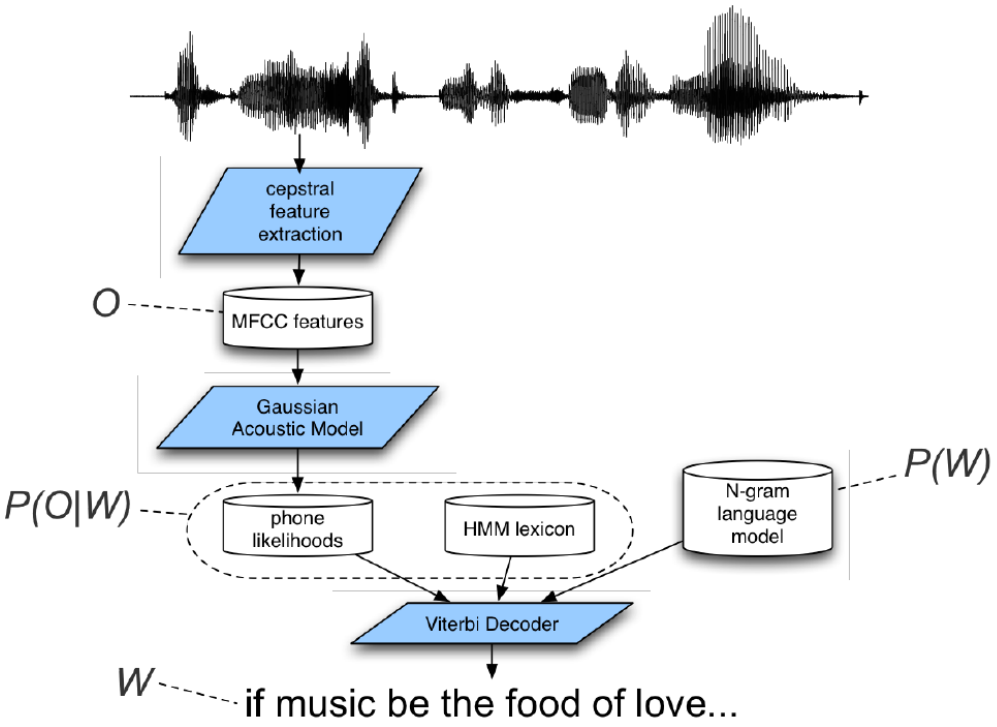

Figure 2: Schematic architecture for a speech recognizer decoding a single sentence. This schematic does not include pruning, fast-match and tree-structured lexicons done by the Viterbi decoder. Reprinted from *Speech and Language Processing* (Second Edition, p. 325), by Jurafsky, D., & Martin, J. H., 2009, New Jersey, United States: Prentice Hall. Copyright 2009 by Pearson Education

In figure 2 the general decoding architecture for speech recognition is displayed. The main components explained by Jurafsky & Martin (2009) will be discussed separately below.

- **Feature extraction:** In the feature extraction stage, the acoustic waveform is sampled into frames that are transformed into spectral features. A portion of speech (window) is chosen and sampled into frames every 10 (usually 10, 15 or 20) milliseconds. Overlapping the windows provides a more reliable measurement as the portions of speech are covered multiple times. Instead of using rectangular windows, that cut off the signal at its boundaries, the Hamming window is used. The Hamming window shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.

- **Acoustic model:** In the acoustic modeling stage, the likelihood of the observed spectral feature vectors given linguistic units (words, phones, subparts of phones) is computed. For example, Gaussian mixture model classifiers are used to compute for each HMM state q, corresponding to a phone or subphone, the likelihood of a given feature vector given this phone $p(o|q)$.

- **Language model:** In the language modeling stage, the prior probability that a given string of words is a sentence in the language is computed. For example, N-gram grammars are used to assign a probability to a sentence by computing the following equation:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k | w_{k-N+1}^{k-1}) \tag{5}$$

- **Decoding:** In the decoding phase, the acoustic model is taken, which consists of a sequence of acoustic likelihoods, plus an HMM dictionary of word pronunciations, combined with the language model, the output is the most likely sequence of words. Most ASR algorithms are used for large vocabulary data sets. This brings up a problem as the search space becomes immense. To limit the search space, a efficient algorithm is needed that will not search through all possible sentences, but only consider ones that have a good chance of matching the input. This is known as the decoding or search problem. Most ASR systems use the Viterbi algorithm for decoding. This efficient algorithm speeds up the

decoding with a variety of sophisticated augmentations such as pruning, fast-match and tree-structured lexicons.

Thus far the general decoding architecture for an ASR system has been discussed. The training aspects of an ASR system will now be considered. Training ASR models is one of most vital parts of an ASR system, as it has great influence on the performance of the recognizer. As noted before the focus is on HMM-based speech recognizers, since the HMM is the dominant paradigm for Large-Vocabulary Continuous Speech Recognition. Therefore, the embedded training procedure, used in training acoustic models for HMM-based speech recognition systems, will be examined below.

In embedded training, the ASR system trains each phone HMM embedded in an entire sentence, and the segmentation and phone alignment are done automatically as part of the training procedure. Jurafsky & Martin (2009) include a summary of a basic embedded training routine given a phoneset, pronunciation lexicon, and the transcribed wavefiles. The phoneset consists of a set of (untrained) phone HMMs. The pronunciation lexicon is a collection of words together with their pronunciations (one - w ah n). The transcribed wavefiles are the acoustic waves of the speaker transcribed into a sentence or a word. The procedure, given the three components described, involves the following steps:

1. Build a "whole sentence" HMM for each sentence.

2. Initialize A (a subphone transition probability matrix) probabilities to 0.5 (for loop-backs or for the correct next subphone) or to 0 (for all other transitions).

3. Initialize B (a set of observation likelihoods) probabilities by setting the mean and variance of each Gaussian to the global mean and variance for the entire training set.

4. Run multiple iterations of the Baum-Welch algorithm.

The Baum-Welch algorithm computes the probability of being in a certain state at a certain time, by using forward-backward to sum over all possible paths that were in a certain state emitting a certain symbol at a certain time. This allows accumulation of counts for re-estimating the emission probability

from all the paths that pass through another state at a certain time. However, Baum-Welch can be time consuming and as multiple iterations are ran it is interesting to look at an alternative. Viterbi alignment is used, which instead of accumulating counts by a sum over all paths that pass through a state at a certain time, approximates this by choosing only the Viterbi (most probable) path.

Next to training acoustic models, training the language models is evenly critical. Language models are trained by n-grams. The probabilities in n-gram language models are often computed using maximum likelihood estimation. This makes the probability distribution dependent on the available training data. Hence, to ensure statistical significance, large training data are required in language modeling (Besacier et al., 2013).

The entire system of an automatic speech recognizer knows many facets. The problem of bias can therefore be approached from many angles like the structure of the models, the different methods of training and the datasets used for training acoustic and language models. In this paper we will explore the datasets used for training as a reason for bias in ASR.

2.3 Bias in Automatic Speech Recognition

In general, ASR systems promise to deliver an objective interpretation of human speech. However, they still struggle with the large variation in speech due to e.g., gender, age, speech impairment, race, and accents (Feng et al., 2021). Racial disparities have been found in five state-of-the-art ASR systems (developed by Amazon, Apple. Google, IBM and Microsoft) that were used to transcribe structured interviews of white and black speakers. (Koenecke et al.,2020). It is also known that ASR models are often trained on speech from the standard dialect of a language, which results in lower accuracy for other dialect speech (Tatman, 2017). It is well established by now that large language models (LM) exhibit various kinds of bias, including stereotypical associations (Basta et al. 2019, Beltagy et al., 2019, Kurita et al., 2019, Sheng et al., 2019, Zhang et al., 2020, Zhao et al., 2019, as cited in Bender et al., 2021), or negative sentiment towards specific groups (Hutchinson et al., 2020 as cited in Bender et al., 2021). The increase in NLP research has led to the deployment of increasingly larger language models, especially for English. It is argued that the increasing size of LM's comes with costs and

risks such as environmental costs, financial costs, stereotyping, denigration and increases in extremist ideology (Bender et al., 2021). The composition of the training data plays an important role aswell. A speaker with a differing dialect from the training data can lead to a mismatch with the language model (Feng et al., 2021). To add to that, Caliskan et al. (2017) showed that language corpora actually contain human-like biases.

Various bias mitigating strategies have been proposed. These strategies range from specific approaches that aim to reduce a certain type of bias to general approaches that aim to provide a structure to ASR starting from its basis. Savoldi et al. (2021) discuss numerous procedures to mitigate gender bias such as gender tagging, adding context, debiased word embeddings and balanced fine-tuning. To reduce performance differences and ensure that speech recognition technology is inclusive Koenecke et al (2020) propose using more diverse training datasets. An important side that is often neglected in ASR models is the social aspect. Current NLP focusses on information content while ignoring language's social factors. Modeling social factors like social relations, context, social norms, culture and ideology shows substantial improvements in a wide range of ASR applications (Hovy & Yang, 2021). Feng et al. (2021) advices to start thinking about bias at the start of ASR system development. Next to a direct bias mitigation strategy like diversifying the dataset, an indirect bias mitigation strategy should be adopted. This includes framing the problem, developing the developer team composition and the implementation process from a point of anticipating, proactively spotting, and developing mitigation strategies for prejudice. The variety in age, regions, gender, etc. provides additional insights in spotting potential bias in design.

Although it is interesting to look at strategies to reduce bias in ASR, detecting bias and its possible reasons is a vital step to be able to correct it in the future. Correcting bias especially if its related to datasets, means a new dataset has to be created. Considering how time consuming it is to create novel datasets and that current research works with already very established datasets, it is important to be able to detect when biases occur and what the reasons for bias are, so they can be corrected later on.

2.4 Arabic Automatic Speech Recognition

Due to the lack of resources for Arabic ASR, Arabic is seen as an under-resourced language according to its definition by Krauwer (2003) and Berment (2004). ASR for under-resourced languages asks for novel approaches to collecting data like crowdsourcing (Gelas et al., 2011) and multilingual acoustic models (Schultz, 2006; Schultz & Waibel, 2001; Le & Besacier, 2009, as cited in Besacier et al. 2013). It is also known that languages with many dialects and code-switching between these variations(when a speaker switches between languages or dialects) initiate problems for ASR systems (Besacier et al., 2013). According to Hussein et al. (2021) the Arabic language and its dialects introduce three major challenges when developing speech recognition models:

1. Arabic is a consonantal language with most of the available text being non-diacritized. This makes it challenging to determine the location of the vowels, which can convey different meanings.

2. There is limited labeled data available for the different Arabic dialects. Each dialect is a native Arabic language that is spoken, but not written, as it does not have standardized orthographic rules.

3. Arabic is morphologically complex and has a high level of affixation and derivation. This makes it hard to estimate probabilities for the language model. Also, it increases the out-of-vocabulary rate, which can be defined as the number of unknown words in a new sample of language, usually expressed in percentage (IGI Global,2021).

Developing Arabic ASR systems is no easy task, as it comes with many challenges that are novel and not seen in resourced languages like English. These obstacles are crucial to consider in relation to the introduction of possible bias, as the complexity of the ASR task may unintentionally bring in various types of biases.

2.5 Bias in Arabic Automatic Speech Recognition

Even though Arabic is one of the most widely spoken languages in the world, ASR research on Arabic in general is limited when compared to other languages. Concurrently, exploration of bias in Arabic ASR is minimal.

Bircan & Ceylan (n.d.) expand on the limited field of bias in Arabic ASR, bringing 'religion'-bias into the picture and expressing the need to explore biases in Arabic training corpora. They compared ASR based transcriptions of Arabic refugee interviews with human transcription through ideological textual analysis. Their findings indicated that the problem in ASR for Arabic is beyond a grammatical or accuracy problem but is based on bias at the contextual level. Four different ASR softwares (HappyScribe, Sonix, Vocalmatic, Amberscript) had the same semantic shift due to substituted sentences. Moreover, insertion and substitution errors were found. To add to that, the Arabic language and being an Arab were explicitly linked with Islam and being a Muslim, while respondents' migratory experiences were aligned with conflict and violence related topics despite the absence of any contextual or conceptual references. According to Bircan & Ceylan (n.d.) this asks for the need to investigate biases in training corpora. Also, the orientalist stereotypes and Islamophobia that were found in the automated transcriptions can be seen as examples of 'religion'-bias in Arabic ASR.

Specific biases (e.g. gender, dialect) that occurred in non-Arabic ASR have also been found in Arabic ASR. Sawalha & Shariah (2013) noted that female Arabic speech is recognised better than male Arabic speech. Furthermore, they examined the effects of speakers' country and region. Speakers from the Levant region were recognized better than speakers living in Gulf and Africa region, although all of them were asked to record in MSA. Thus, the speaker's region can affect the performance of ASR systems. Alsharhan & Ramsay (2020) found that building gender and dialect-specific Arabic ASR models leads to substantial decreases in Word Error Rate.

By now it is clear that research in the field of bias in Arabic ASR is fairly sparse. This provides an opportunity for researchers to develop this branch of ASR more. Luckily we have seen recent papers like Bircan & Ceylan (n.d.) that try to do so.

3 Discovering potential reasons for bias

In this section we will discuss various datasets, the distributors and commonly used software for Arabic ASR and compare them with the norm for English datasets used in the same context. Based on the comparison, we discuss certain points that could improve Arabic ASR.

Multiple datasets are used in Arabic ASR model training. They are distributed by different databases. To train their acoustic model, Menacer et al. (2017) used two datasets of several hours of Standard Arabic new broadcasts called Nemlar and NetDC that were distributed by the European Language Resources Association (ELRA). Nemlar has 40 hours of data with news broadcasts from radio as its source. Broadcasts were recorded from four different Arabic speaking radio stations: Medi1, Radio Orient, Radio Monte Carlo (RMC), Radio Television Maroc (RTM). Each broadcast contains between 25 and 30 minutes of news and interviews (259 distinct speakers identified). The data was recorded between 2002 and 2005 (ELRA, 2021). NetDC consists of 22.5 hours of broadcast news speech recorded from Radio Orient (France). The project was developed in the framework of the European-funded project Network of Data Centres, NetDC. The data was recorded between 2001 and 2002 (ELRA, 2021).
The language model of Menacer et al. (2017) was trained by the GigaWord Arabic corpus distributed by the Linguistic Data Consortium (LDC), which consists of 1.000 million word occurrences collected from nine distinct sources of Arabic newswire.

Using news related sources is a common approach to data gathering in ASR as it is a quick way to gather quite a significant amount of data. However, it also neglects the fact that typical language used in news differs from usual day-to-day conversational language in structure and content. Some datasets are introduced to try to fill this gap by using datasets gathered from (scripted) telephone conversations.
Besacier et al. (2013) names two dataset distributors, AppenButlerHill (renamed to Appen) and SpeechOcean. Appen has nine Arabic ASR datasets available, including Eastern Algerian Arabic, Egyptian Arabic, MSA, Moroccan Arabic, Saudi Arabian Arabic and United Arab Emirates (UAE) based Arabic (Appen, 2021). Three relatively large datasets are accessible. Arabic (Egypt) scripted smartphone (ARE_ASR001_CN) consists of 352 hours of

scripted smartphone conversations and Arabic (UAE) scripted smartphone (ARU_ASR001_CN) has 170 hours of data. However, the content of these conversations is not clarified. Arabic (Saudi Arabia) scripted smartphone (ARS_ASR001_CN) consists of 322 hours of scripted smartphone conversations covering general content like education, sports, entertainment, travel, culture and technology. The other datasets are relatively smaller but contain more variations of source, i.e. non-scripted telephone conversations.
SpeechOcean has only four Arabic speech recognition datasets (King-ASR-109/293/318/L-109) but the total volume, 1000 hours, is quite large compared to other Arabic datasets. However, the sources and type of content for these datasets are not mentioned (SpeechOcean, 2021). It is beneficial to have more conversational datasets next to news related datasets. Having said that, it has a few disadvantages: the average size of these sets is quite small and if there is a dataset with a considerable size, the content of that dataset is often unknown.

Most of the available data in the literature targets MSA and only a few datasets are available for Arabic dialects, according to Alsharan & Ramsay (2020). Due to the lack of dialectal Arabic corpora, many researchers start from zero and construct their own corpora. Masmoudi et al. (2014) and Droua-Hamdani (2010) have made corpora for Tunisian Arabic and Algerian Arabic. The Tunisian corpus consists of audio recordings and transcriptions from dialogues in the Tunisian Railway Transport Network. The Algerian corpus contains MSA speech from 300 Algerian native speakers from different regions. The developed corpora can be used to build dialect-specific Arabic ASR systems. A corpus was also introduced to be able to build multi-dialectal speech recognition systems. Biadsy et al. (2012) constructed the largest multi-dialectal Arabic speech corpus. The data was gathered with the help of more than 125 million people in Egypt, Jordan, Lebanon, Saudi Arabia, and the United Arab Emirates. However, the main problem with this corpus is that it contains read speech. The speech was recorded by using software that displays prompts to the user and asks them to say it in their own dialect. Thus, it is not spontaneous speech and it is probably not useful for speech recognition systems, according to Alsharan & Ramsay (2020). In contrast, Besacier et al. (2013) state that prompted speech could be more useful than spontaneous speech as a starting point for ASR development in an under-resourced language.

Apart from having a certain amount of datasets available, it is also important to consider how big datasets should be. Alsharan & Ramsay (2020) investigated how much data is needed to achieve the optimal performance of an Arabic ASR system. Their results indicated that it is more important to ensure that the training data is taken from the same group as the target group, than to maximise the amount of training data. They state that it is beneficial to treat accent groups seperately, eventhough a multi-dialectal corpus can often be much bigger as data from multiple dialects can be included. Thus, when conducting ASR research in a certain Arabic dialect, researchers should not primarily focus on creating the biggest dataset possible. It can be helpful to handle dialects separately.

Besides datasets and its distributors it can be meaningful to investigate ASR softwares. Various types of softwares are often used to perform ASR tasks, as these give the opportunity to have ASR algorithms available without too much hassle. Commonly used ASR softwares to recognize Arabic speech include HappyScribe, Sonix, Vocalmatic, Amberscript and NVivo (softwares used by Bircan et al (n.d.)). It is not publicly known what datasets are used to train the models for the softwares. The companies were contacted to clarify what data was used. However, only HappyScribe responded. HappyScribe does not use an in-house ASR model. Speechmatics is used if the specific language is available there, otherwise Google's ASR model is used. Nevertheless, the fact that all other software companies are not clear in discussing what datasets are used could be problematic as it is not clear on what sources and content the models are trained on.

It is interesting to compare the availability, size, sources and content of Arabic ASR datasets with English ASR datasets as this could help identify certain points that could help improve Arabic ASR. English ASR systems are without a doubt the most researched and therefore many resources are available. When investigating the dataset distributor ELRA, it can be seen that there are more than 330 language corpora for English available. Next to ELRA, the immense amount of data available for English is visible in another well known database, LDC. All ten corpora in the Top Ten LDC Corpora contain English (LDC, 2021). Appen has 13 English datasets available for ASR use, with a total size of more than 2000 hours of speech including various English dialects like United Kingdom, United States and Australian. Another popular corpus is the LibriSpeech ASR corpus (SLR12) containing 1000 hours of

read English speech based on public domain audio books (Panayotov et al., 2015). LibriSpeech ASR corpus is used by Google's ASR model for English (Google, 2021). The Speech Accent Archive (Tatman, 2017) is a dataset containing 2140 speech samples, each from a different speaker reading the same reading passage. Speakers come from 177 countries and have 214 different native languages. Each speaker is speaking in English.
In comparison with Arabic ASR datasets, English ASR datasets are widely available in much larger sizes. The English datasets are more diverse as they involve people from all around the world speaking the language and cover a multitude of content. To add to that, English datasets have more resources that are based on natural and conversational speech like the LibriSpeech ASR corpus.
To improve Arabic ASR models, research should focus on creating datasets of considerable size that involve natural speech. News sources can still be used as this is an easy way to gather data, however diversification of the datasets should be considered. The datasets should cover multiple topics, instead of focusing only on one type of content like telephone conversations.

4 Conclusion

The Arabic language and dialects were discussed including potential difficulties for ASR models. Despite the fact that Arabic is spoken by a large group of people, the research in the field of ASR systems is limited. The great amount of dialects and the complexity of the language form challenges for developing ASR systems.

We described the current social situation in Arab speaking countries and it's perception by the media. The Arab world has known various conflicts, some have been present for several years already. These conflicts intensify the amount of negative stereotypes people have of Arabs. The media has played an active role in forming and spreading these negative stereotypes in the form of news articles and films.

Automatic Speech Recognition (ASR) was explained as the process of converting a speech signal to a sequence of words (i.e.,spoken words to text) by means of algorithm implemented as a computer program. Various approaches from the field of AI are used to construct modern automatic speech

recognizers. HMM-based speech recognizers use a noisy-channel which introduces several probabilistic components. These components are computed by the acoustic model and language model. The acoustic model is trained by embedded training and the language model uses n-grams for training. However, since Arabic is an under-resourced language it is hard to find labeled data to train the models.

To discover potential reasons for bias and points that could improve Arabic ASR, various datasets data-distributors and commonly used software were discussed. Arabic datasets are limited in terms of availability, size, sources and content. Mostly news sources are used to gather a relative large amount of data quickly. However, more day-to-day conversational speech datasets are rare. Scripted telephone conversations are used but their content is often unknown when the datasets have a larger size. To add to that, most of the available data in the literature targets MSA and only a few datasets are available for Arabic dialects. The largest multi-dialectal Arabic speech corpus constructed by Biasdy et al. (2012) has a significant amount of data, but the main problem is the fact that it contains read speech. However, according to Besacier et al. (2013) this should not be an issue for under-resourced languages. Software companies are not clear in discussing what datasets are used, this could be problematic as it is not clear on what sources and content the models are trained on.

Arabic datasets were then compared with English datasets to point out several areas for improvement. English datasets are widely available in much larger sizes than Arabic datasets. The data is more diverse and contain various type of content related to natural and conversational speech. Subsequent Arabic ASR datasets should be of considerable size and involve natural speech. Next to news sources, diversification of data should be considered.

As research in Arabic ASR is very limited, a general increase of interest in this field is encouraged. Since Arabic ASR can be quite complex, researchers should focus on trying to improve seperate components of the Arabic ASR architecture. Future research in Arabic ASR datasets could help to discover more possible reasons for problems that are currently holding back the progress of Arabic ASR. Specific research could analyse certain datasets more in-depth with various data analysis methods to come to new discoveries regarding bias.

References

[1] Eiman Alsharhan and Allan Ramsay. "Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition". In: *Language Resources and Evaluation* 54.4 (Dec. 2020), pp. 975–998. ISSN: 15728412. DOI: 10.1007/s10579-020-09505-5.

[2] Appen. *Machine Learning Datasets: 250+ ML Repository Of Speech Datasets*. 2021. URL: https://appen.com/off-the-shelf-datasets/#ProductCatalog.

[3] BBC News. *Israel-Gaza violence: Joe Biden calls for ceasefire*. 2021. URL: https://www.bbc.com/news/world-middle-east-57152723.

[4] BBC News. *The rise of Islamic State*. 2014. URL: https://www.bbc.com/news/world-middle-east-28116033.

[5] BBC News. *Timeline: Iraq War*. 2016. URL: https://www.bbc.com/news/magazine-36702957.

[6] Emily M. Bender et al. "On the dangers of stochastic parrots: Can language models be too big?" In: *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 610–623. DOI: 10.1145/3442188.3445922.

[7] Vincent Berment. *Méthodes pour informatiser les langues et les groupes de langues " peu dotées "*. Tech. rep. 2004. URL: https://tel.archives-ouvertes.fr/tel-00006313.

[8] Laurent Besacier et al. "Automatic speech recognition for under-resourced languages: A survey". In: (2013). DOI: 10.1016/j.specom.2013.07.008. URL: http://dx.doi.org/10.1016/j.specom.2013.07.008.

[9] Fadi Biadsy, Pedro J Moreno, and Martin Jansche. *GOOGLE'S CROSS-DIALECT ARABIC VOICE SEARCH*. 2012. ISBN: 9781467300469. URL: www.google.com.eg..

[10] Tuba Bircan and Duha Ceylan. "Machine Discriminating : Automated Speech Recognition Biases in Refugee Interviews". In: ().

[11] Tolga Bolukbasi et al. "Quantifying and Reducing Stereotypes in Word Embeddings". In: (June 2016), pp. 42–45. URL: http://arxiv.org/abs/1606.06121.

[12] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. *COGNITIVE SCIENCE Semantics derived automatically from language corpora contain human-like biases Downloaded from*. Tech. rep. 2017, pp. 183–186. URL: http://science.sciencemag.org/.

[13] CNN. *Syria crisis*. 2016. URL: https://edition.cnn.com/specials/middleeast/syria.

[14] Edited by Dalila Ayoun. *Studies in French Applied Linguistics*. Tech. rep.

[15] Deutsche Welle. *Syria and Yemen - gaping wounds in the Middle East — Middle East— News and analysis of events in the Arab world*. 2016. URL: https://www.dw.com/en/syria-and-yemen-gaping-wounds-in-the-middle-east/a-36963373.

[16] Ghania Droua-Hamdani, Sid Ahmed Selouani, and Malika Boudraa. "Algerian Arabic Speech Database (ALGASD): Corpus design and automatic speech recognition application". In: *Arabian Journal for Science and Engineering* 35.2 C (2010), pp. 157–166. ISSN: 21914281.

[17] D. M. Eberhard, G. F. Simons, and C.D. Fennig. *Ethnologue: Languages of the World*. 2019. DOI: 10.5860/choice.192005. URL: https://www-ethnologue-com/.

[18] D. M. Eberhard, G. F. Simons, and C.D. Fennig. *Ethnologue: Languages of the World*. 2021. URL: https://www-ethnologue-com.proxy.library.uu.nl/.

[19] John C Eisele. *ARABIC DIALECTOLOGY: A Review Of Recent Literature*. Tech. rep. 1. 1987, pp. 199–269.

[20] Yasmeen Elayan. *Stereotypes of Arab and Arab-Americans Presented in Hollywood Movies Released during 1994 to 2000*. Tech. rep. 2005. URL: https://dc.etsu.edu/etd/1003.

[21] ELRA. *NEMLAR Broadcast News Speech Corpus – ELRA Catalogue*. 2021. URL: http://catalog.elra.info/en-us/repository/browse/ELRA-S0219/.

[22] ELRA. *NetDC Arabic BNSC (Broadcast News Speech Corpus) – ELRA Catalogue*. 2021. URL: http://catalog.elra.info/en-us/repository/browse/ELRA-S0157/.

[23] Igor H. Sabino de Farias. "An Introduction to Middle East Politics". In: *Contexto Internacional* 42.3 (Dec. 2020), pp. 709–712. ISSN: 0102-8529. DOI: 10.1590/s0102-8529.2019420300010. URL: http://doi.org/10.1590/S0102-8529.2019420300010.

[24] Siyuan Feng et al. "Quantifying Bias in Automatic Speech Recognition". In: (Mar. 2021). URL: http://arxiv.org/abs/2103.15122.

[25] Aravind Ganapathiraju and Joseph Picone. "Hybrid SVM / HMM architectures for speech recognition". In: January 2000 (2015), pp. 504–507.

[26] Hadrien Gelas et al. *Quality assessment of crowdsourcing transcriptions for African languages*. Tech. rep. 2011. URL: www.speech.sri.com/projects/srilm/.

[27] Google. *Running the Automated Speech Recognition (ASR) model*. 2021. URL: https://cloud.google.com/tpu/docs/tutorials/automated-speech-recognition.

[28] Dirk Hovy and Diyi Yang. *The Importance of Modeling Social Factors of Language: Theory and Practice*. Tech. rep. 2021. URL: https://public.flourish.studio/visua.

[29] Amir Hussein, Shinji Watanabe, and Ahmed Ali. "Arabic Speech Recognition by End-to-End, Modular Systems and Human". In: (Jan. 2021). URL: http://arxiv.org/abs/2101.08454.

[30] IGI Global. *What is Out-Of-Vocabulary Rate*. 2021. URL: https://www-igi-global-com.proxy.library.uu.nl/dictionary/out-of-vocabulary-rate/21614.

[31] Zhang Jing and Zhang Min. "Speech recognition system based improved DTW algorithm". In: *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, CMCE 2010*. Vol. 5. 2010, pp. 320–323. ISBN: 9781424479566. DOI: 10.1109/CMCE.2010.5609979.

[32] Daniel Jurafsky and James H. Martin. "Automatic Speech Recognition". In: *Speech and Language Processing*. 2nd ed. Prentice Hall, 2009. Chap. 9, pp. 319–367. ISBN: 9780135041963.

[33] Tomasz Kamusella. "The Arabic Language: A Latin of Modernity?" In: *Memory & Language Politics* 11.2 (2017). DOI: 10.1515/jnmlp-2017-0006.

[34] Mussarat Khan and Kathryn Ecklund. "Attitudes toward Muslim Americans post-9/11". In: *Journal of Muslim Mental Health* 7.1 (Apr. 2012), pp. 1–16. ISSN: 15565009. DOI: 10.3998/jmmh.10381607.0007.101. URL: http://hdl.handle.net/2027/spo.10381607.0007.101.

[35] Allison Koenecke et al. "Racial disparities in automated speech recognition". In: 117.14 (2020), pp. 7684–7689. DOI: 10.1073/pnas.1915768117/-/DCSupplemental.y.

[36] Steven Krauwer. *The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap*. Tech. rep. 2003, pp. 8–15.

[37] LDC. *Arabic Gigaword Fifth Edition - Linguistic Data Consortium*. 2021. URL: https://catalog.ldc.upenn.edu/LDC2011T11.

[38] Abir Masmoudi et al. *A Corpus and Phonetic Dictionary for Tunisian Arabic Speech Recognition*. Tech. rep. 2014.

[39] Mohamed Amine Menacer et al. *An enhanced automatic speech recognition system for Arabic*. Tech. rep. 2017, pp. 157–165. URL: http://catalog.elra.info/product_.

[40] Michael Safi and Antonio Voce and Frank Hulley-Jones and Lydia McMullan. *How the Arab spring engulfed the Middle East – and changed the world — World news*. 2021. URL: https://www.theguardian.com/world/ng-interactive/2021/jan/25/how-the-arab-spring-unfolded-a-visualisation.

[41] Robert W. Livingston Michele G. Alexander Marilynn B. Brewer. "Putting Stereotype Content in Context: Image Theory and Interethnic Stereotypes". In: *INTERETHNIC IMAGES* (2005). DOI: 10.1177/0146167204271550.

[42] Najm A. Najm. "Negative stereotypes of Arabs: The Western case". In: *Indian Journal of Social Work* 80.1 (2019), pp. 87–114. ISSN: 24567809. DOI: 10.32444/IJSW.2018.80.1.87-114.

[43] Vassil Panayotov et al. *LIBRISPEECH: AN ASR CORPUS BASED ON PUBLIC DOMAIN AUDIO BOOKS*. Tech. rep. 2015. URL: http://www.gutenberg.org.

[44] Karpagavalli S and Chandra E. "A Review on Automatic Speech Recognition Architecture and Approaches". In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 9.4 (Apr. 2016), pp. 393–404. ISSN: 20054254. DOI: 10.14257/ijsip.2016.9.4.34.

[45] Beatrice Savoldi et al. "Gender Bias in Machine Translation". In: (Apr. 2021). URL: http://arxiv.org/abs/2104.06001.

[46] M Sawalha and M Abu Shariah. *The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus*. Tech. rep. 2013. URL: http://eprints.whiterose.ac.uk/81859/.

[47] R. Solera-Ureña et al. "Robust ASR using Support Vector Machines". In: *Speech Communication* 49.4 (Apr. 2007), pp. 253–267. ISSN: 01676393. DOI: 10.1016/j.specom.2007.01.013.

[48] SpeechOcean. *SpeechOcean Data center*. 2021. URL: https://en.speechocean.com/datacenter/recognition.html?prosearch=arabic#datacenter_do.

[49] T.A. Stephenson et al. "Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables". In: *Technical Report Idiap-RR-24-2002* (2002), p. 10.

[50] Rachael Tatman. *Speech Accent Archive*. 2017. URL: https://www.kaggle.com/rtatman/speech-accent-archive.

[51] Rachael Tatman and Conner Kasten. "Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions". In: (2017). DOI: 10.21437/Interspeech.2017-1746. URL: http://dx.doi.org/10.21437/Interspeech.2017-1746.

[52] The New York Times. *U.S. Attacked - Hijacked jets destroy Twin Towers and hit Pentagon in day of terror*. 2001. URL: https://archive.nytimes.com/www.nytimes.com/packages/html/nyregion/9-11imagemap.html?pagewanted=all.

[53] Edmondo Trentin and Marco Gori. "A survey of hybrid ANN/HMM models for automatic speech recognition". In: *Neurocomputing* 37.1-4 (Apr. 2001), pp. 91–126. ISSN: 09252312. DOI: 10.1016/S0925-2312(00)00308-8.

[54] E. Al-Wer. "Arabic Languages, Variation in". In: *Concise Encyclopedia of Languages of the World*. Elsevier Science, 2018, pp. 53, 54. ISBN: 978-0080877747. URL: https://books.google.nl/books?id=F2SRqDzB50wC&q=Maghrebi+Egyptian+Mesopotamian+Levantine+Peninsular+Arabic&pg=PA54&redir_esc=y#v=onepage&q&f=false.

[55] T G Wilfong. "Coptic language". In: (2018). DOI: 10.1093/acrefore/9780199381135.013.8219. URL: https://doi.org/10.1093/acrefore/9780199381135.013.8219.

[56] World Population Review. *Arabic Speaking Countries 2021*. 2021. URL: https://worldpopulationreview.com/country-rankings/arabic-speaking-countries.

[57] Steve Young. "HMMs and Related Speech Recognition Technologies". In: *Springer Handbook of Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 539–558. DOI: 10.1007/978-3-540-49127-9{_}27. URL: https://link-springer-com.proxy.library.uu.nl/chapter/10.1007/978-3-540-49127-9_27.

[58] Zara Zimbardo. "Cultural Politics of Humor in (De)Normalizing Islamophobic Stereotypes". In: *Source: Islamophobia Studies Journal* 2.1 (2014), pp. 59–81. DOI: 10.13169/islastudj.2.1.0059.