

Chapter 2

Fundamentals

2.1 Introduction

Working in the area of Arabic speech recognition requires a good knowledge about the characteristics of the Arabic language and the principals of automatic speech recognition (ASR) systems. In this chapter, the fundamentals of ASR are introduced. Furthermore, we highlight the characteristics of the Arabic language with relevance to ASR. This includes, characteristics on the acoustic and the language level of Arabic that have to be taken carefully into consideration.

2.2 Automatic Speech Recognition

2.2.1 *Speech Recognition Architecture*

The goal of an ASR system is to map from an acoustic speech signal (input) to a string of words (output). In speech recognition tasks, it is found that spectral domain features are highly uncorrelated compared to time domain features. The Mel-Frequency Cepstral Coefficient (MFCC) is the most widely used spectral representation for feature extraction.

Actually, speech is not a stationary signal, that is why we cannot calculate the MFCCs for the whole input speech signal O . We divide O into small enough overlapping frames o_i where the spectral is always stationary. The frame size is typically 10–25 milliseconds. The frame shift is the length of time between successive frames. It is typically 5–10 milliseconds.

$$O = o_1, o_2, o_3, \dots, o_t \quad (2.1)$$

Similarly, the sentence W is treated as if it is composed of a string of words w_i :

$$W = w_1, w_2, w_3, \dots, w_n \quad (2.2)$$

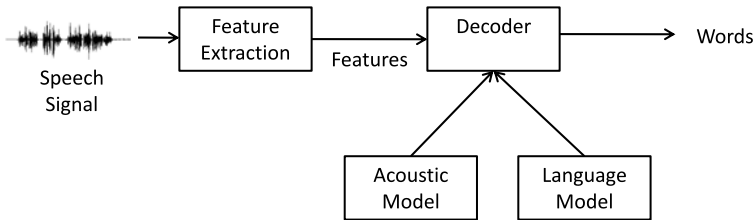


Fig. 2.1 High level block diagram for a state-of-the-art automatic speech recognition system

Now, the task of an ASR system is to estimate the most probable sentence $\hat{W}$ out of all sentences in the language L given the input speech signal O . This can be expressed as:

$$\hat{W} = \operatorname{argmax}_{W \in L} P(W|O) \quad (2.3)$$

Since there is no direct solution for the above equation, Bayes rule is applied to break Eq. (2.3) down as:

$$\hat{W} = \operatorname{argmax}_{W \in L} \frac{P(O|W)P(W)}{P(O)} \quad (2.4)$$

Actually, the probability of observations $P(O)$ is the same for all possible sentences, that is why it can be safely removed since we are only maximizing all over possible sentences:

$$\hat{W} = \operatorname{argmax}_{W \in L} P(O|W)P(W) \quad (2.5)$$

$P(W)$ is the probability of the sentence itself and it is usually computed by the language model. While the conditional probability $P(O|W)$ is the observation likelihood and it is calculated by the acoustic model. See Fig. 2.1 for a high level diagram of an ASR system.

2.2.2 Language Modeling

The probability of the sentence (string of words) $P(W)$ is defined as:

$$P(W) = P(w_1, w_2, w_3, \dots, w_n) \quad (2.6)$$

Usually $P(W)$ is approximated by an N-gram language model as follows:

$$P(w_1^n) \approx \prod_{k=1}^n P(w_k | w_{k-N+1}^{k-1}) \quad (2.7)$$

Where n is the number of words in the sentence and N is the order of the N-gram model. To estimate the probability of the different N-grams in the language, a large text corpus is required to train the language model. Typically in ASR, the language model order N is either two (bi-gram) or three (tri-gram).

Fig. 2.2 Left-to-right (Bakis) hidden Markov Model (HMM)

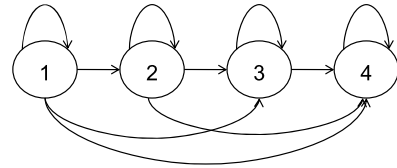
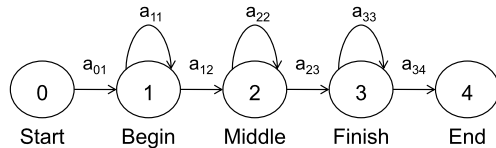


Fig. 2.3 Three states HMM to model one phoneme, where a_{ij} is the transition weight from state i to j



2.2.3 Acoustic Modeling

The conditional probability $P(O|W)$ is computed by the acoustic model. It is typically a statistical model for all the phonemes in the language to be recognized. This type of acoustic models is known as phoneme-based acoustic models. Statistical modeling for the different phonemes is performed using Hidden Markov Model (HMM). Since the speech signal is causal, the left-to-right HMM topology is used for acoustic modeling as shown in Fig. 2.2.

Each phoneme has three sub-phones (begin, middle, and finish), that is why each phoneme is typically modeled using three HMM states besides two non-emitting states (start and end). Skip-state transitions are usually ignored to reduce the total number of parameters in the model as shown in Fig. 2.3.

The same phoneme can have different acoustic characteristics in different words. The phoneme /t/ is pronounced in different ways in the words: tea, tree, city, beaten, and steep. The process by which neighboring sounds influence one another is called *Co-articulation Effect*. Modeling phonemes without taking into consideration the context (left and right phonemes) is called *Context Independent (CI)* acoustic modeling where each phoneme (or sub-phone) is modeled using only one HMM.

Acoustic models for each phoneme usually model the acoustic features of the beginning, middle, and end of the phoneme. Usually the starting state is affected by the left phoneme and the final state is affected by the right phoneme. In *Context Dependent (CD)* acoustic models (tri-phones), co-articulation effect is taken into consideration. For the above /T/ example, five different tri-phones models are built because the left and right context of /T/ is not the same in the five words. Actually, CD acoustic modeling leads to a huge number of HMM states. For example, a language with a phoneme set of 30 phonemes, we would need 27,000 tri-phones. The large number of tri-phones leads to a training data sparsity problem. In order to reduce the total number of states, clustering of similar states is performed and *State Tying* is applied to merge all states in the same cluster into only one state. The merged states are usually called *Tied States*. The optimal number of tied states is usually set empirically and it function in the amount of training data and how much tri-phones are covered in the training data.

The most common technique for computing acoustic likelihoods is the *Gaussian Mixture Model* (GMM). Where each HMM state is associated with a GMM that models the acoustic features for that state. Each single Gaussian is a multivariate Gaussian where its size depends on the number of MFCCs in the feature vector. Basically, GMM consists of weighted mixture of n multivariate Gaussians. For speaker independent (SI) acoustic modeling, the number of multivariate Gaussians n is typically ~ 8 or higher. The more number of Gaussians, the more accents that can be modeled in the acoustic model and the more data that is required to train the acoustic model.

To summarize, an acoustic model is defined by the following for each HMM state:

- Transition weights.
- Means and variances of multivariate Gaussians in the GMM.
- Mixture weights of multivariate Gaussians in the GMM.

2.2.4 Training

All acoustic model parameters that define the acoustic model are estimated using training data. Training data is a collection of large number of speech audio files along with the corresponding phonetic transcriptions. Phonetic transcriptions are usually not written explicitly for each audio segment. However, orthographic transcriptions are written and phonetic transcriptions are estimated using a lookup table.

The mapping from orthographic text to phonemes is usually performed using a lookup table called pronunciation dictionary or lexicon. A pronunciation dictionary includes all the words in the target language and the corresponding phonemes sequence. The word HMM is simply the concatenation of the phoneme HMMs. Since the number of phonemes is always finite (50 phonemes in American English), it is possible to build the word HMM for any given word. In Fig. 2.4, the mapping from orthographic transcription to phonetic transcription is clarified.

In order to train a speaker independent acoustic model, a large number of speakers is required (~ 200) in order to be able to model the variability across speakers.

2.2.5 Evaluation

Word error rate (WER) is a common evaluation metric of the performance of ASR systems and many other natural language processing tasks. The WER is based on how much the word string returned by the recognizer (called the hypothesized word string) differs from a correct or reference transcription. The general difficulty of measuring performance is due to the fact that the recognized word string may have a different length from the reference word string (the correct one). This problem is tackled by first aligning the recognized word sequence with the reference word

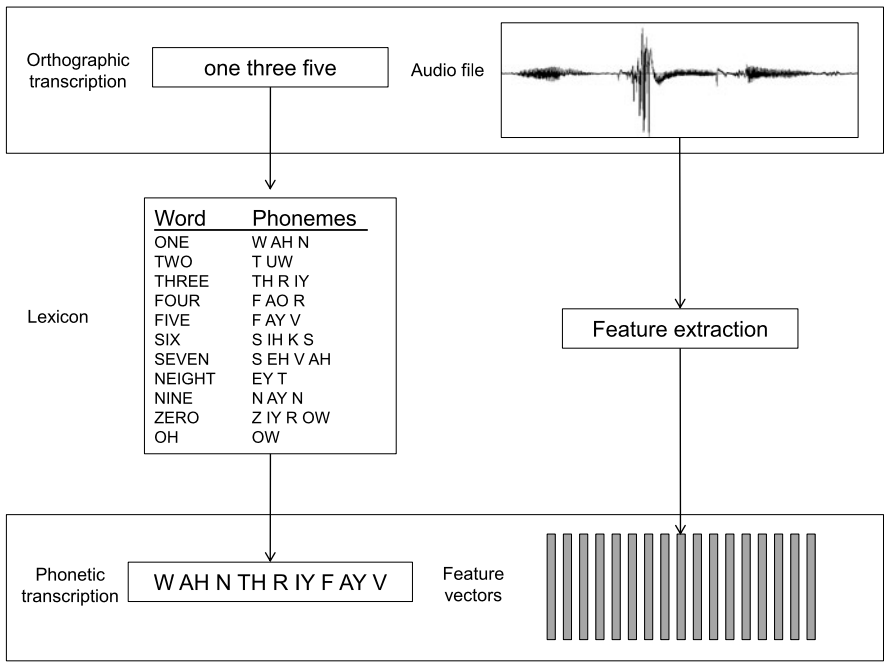


Fig. 2.4 Acoustic model training inputs (audio file of digits and corresponding orthographic transcription). The phonetic transcription is estimated from the lexicon and aligned with the feature vectors

sequence to compute the minimum edit distance (minimum error). This alignment is also known as the maximum substring matching problem, which is easily handled by dynamic programming. The result of this computation will be the minimum number of word substitutions, word insertions, and word deletions necessary to map between the correct and the recognized strings. Word error rate can then be computed as:

$$WER = 100 \times \frac{I + S + D}{N} \tag{2.8}$$

Where I , S , and D are the total number of Insertion, Substitution, and Deletion errors respectively, while N is the number of words in the correct (reference) transcriptions.

More details about the theory of speech recognition can be found in (Huang et al. 2001; Jurafsky and Martin 2009; Rabiner 1989).

2.3 The Arabic Language

The Arabic language is the largest still living Semitic language in terms of number of speakers. Arabic speakers exceed 250 million first language speakers and the number of speakers using Arabic as a second language can reach four times that



Fig. 2.5 Map of the Arabic world

number. Arabic is the official language in 22 countries known as the Arab world (see Fig. 2.5) and ranked as the 6th most spoken language based on the number of first language speakers and it is one of the six official languages of the United Nations.

The Arabic alphabet is also used in many other languages such as Arabic, Azerbaijani, Baluchi, Eastern Cham, Comorian, Dogri, Hausa, Kashmiri, Kurdish, Lahnda, Pashto, Persian (Iranian and Dari), Punjabi, Sindhi, Uighur, and Urdu.

Arabic script is written from left to write, and some letters may change shape according to position within the word.

Some work has been done to represent Arabic letters using Roman character sets. This Roman representation is usually called *transliteration* or *Romanization*. One of the popular Arabic transliterations is the *Buckwalter* transliteration (Buckwalter 2002a).

The commonly used binary encoding standards for Arabic characters are: ISO 8599-6, the Windows based CP-1256, and the Unicode standard (see Table A.1).

The Arabic alphabet has 28 basic letters (see Table 2.1), not having upper and lower case like the Roman alphabet. The following is the sentence “The boy went to the market” in Arabic and the corresponding transliteration:¹

¹All transliterations in this chapter are following the ZDMG transcription (DIN 31635), generated by the ArabTeX package unless otherwise mentioned (Lagally 1992).

Table 2.1 The Arabic alphabet

Letter name	Form	Letter name	Form
ألف <i>ʾlif</i>	ا	ضَا ضَا ضَا <i>ḍāḍ</i>	ض
بَاء <i>bāʾ</i>	ب	طَاء <i>ṭāʾ</i>	ط
تَاء <i>tāʾ</i>	ت	ظَاء <i>ẓāʾ</i>	ظ
ثَاء <i>ṭāʾ</i>	ث	عَيْن <i>ʿayn</i>	ع
جِيم <i>ǧiym</i>	ج	غَيْن <i>ǧayn</i>	غ
حَاء <i>ḥāʾ</i>	ح	فَاء <i>fāʾ</i>	ف
خَاء <i>ḫāʾ</i>	خ	قَاف <i>qāf</i>	ق
دَال <i>dāl</i>	د	كَاف <i>kāf</i>	ك
ذَال <i>ḏāl</i>	ذ	لَام <i>lām</i>	ل
رَاء <i>rāʾ</i>	ر	مِيم <i>myim</i>	م
زَاي <i>zāy</i>	ز	نُون <i>nuwn</i>	ن
سَيْن <i>syn</i>	س	هَاء <i>hāʾ</i>	ه
شَيْن <i>šyn</i>	ش	وَاو <i>wāw</i>	و
صَاد <i>šād</i>	ص	يَاء <i>yāʾ</i>	ي

ḍahaba 'l-walado ʾilaā 'l-ssuwqi.

ذَهَبَ الْوَلَدُ إِلَى السُّوقِ.

In Arabic there are three diacritic marks that symbolize short vowels; “َ” (*fatḥat* فَتْحَت), “ِ” (*ḍammāt* ضَمَّت) and “ُ” (*kasrat* كَسَرَتْ). In addition to these three, there is also the “ْ” (*sukuwn* سُكُون) which symbolizes the absence of a vowel, the “ّ” (*šaddat* شَدَّت), which symbolizes the duplication of a consonant, and the three “َ”, “ِ” and “ُ” (*tanwiyn* تَنْوِين) marks which have the effect of adding one of the short vowels followed by “n”, and can only be placed at the end of the word. These marks are shown in Table 2.2. It should also be noted that the shadda marks are always followed by a *ḍammāt*, *fatḥat*, *kasrat* or a *tanwyn* mark. This means that one letter can be followed by more than one mark.

The Arabic language has many varieties. These are classified into two main classes: Standard Arabic and Dialectal Arabic as shown in Fig. 2.6. Standard Arabic includes *Classical Arabic* and *Modern Standard Arabic (MSA)* while dialectal Arabic includes all forms of spoken Arabic in everyday life communications. Arabic dialects significantly vary among countries and deviate from standard Arabic to some extent and even within the same country we can find different dialects. While there are many forms of Arabic, there are still many common features on the acoustic level and the language level.

In this chapter, we have chosen classical Arabic and MSA forms to highlight the characteristics of standard Arabic. For dialectal Arabic we have chosen Egyptian

Table 2.2 The Arabic diacritic marks

Diacritic name	Form
فَتْحَت <i>fatḥhat</i>	َ
ضَمَّت <i>ḍammat</i>	ُ
كَسَرَت <i>kasrat</i>	ِ
شَدَّت <i>šaddat</i>	ّ
سُكُون <i>sukuwn</i>	ْ
تَنْوِين فَتْحَت <i>tanwyn fatḥhat</i>	ً
تَنْوِين ضَمَّت <i>tanwyn ḍammat</i>	ٌ
تَنْوِين كَسَرَت <i>tanwyn kasrat</i>	ٍ

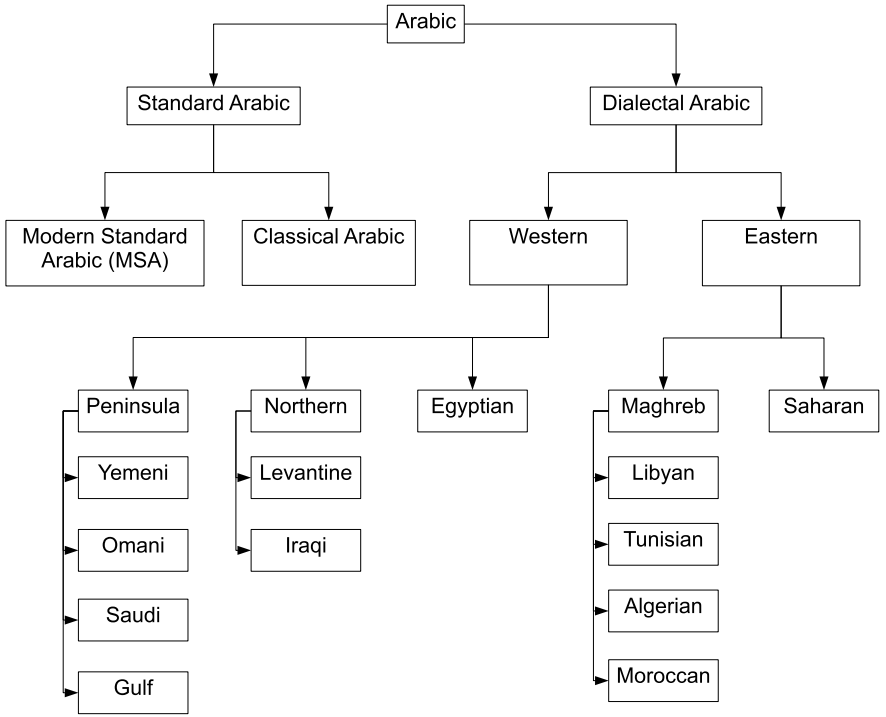


Fig. 2.6 Arabic varieties geographical classification

colloquial Arabic (ECA) (ECA usually stands for the spoken language in Cairo) as a typical dialectal Arabic form. The main ASR problems for Arabic discussed in this chapter are: morphological complexity, dialectal forms, and grapheme-to-phoneme.

2.3.1 Modern Standard Arabic

MSA is also known as Fussha and it is the current standard form of Arabic. Almost all Arabic text resources are written in MSA. It is the formal spoken Arabic language. MSA is used in books, newspapers, news broadcasts, formal speeches, movies subtitling, etc. MSA can be considered as a second language for all Arabic speakers. They can understand spoken MSA as in news broadcasts. MSA is the only commonly accepted Arabic form throughout all native Arabic speakers. That is why many radio and TV broadcasters use MSA in order to target a broad Arabic audience.

The allowed syllables in Arabic are CV, CVC, and CVCC where C is a consonant and V is a long or a short vowel. Therefore, Arabic utterances and words can only start with a consonant and the CVCC pattern is only allowed at the end of a word.

The MSA phonemes inventory consists of 38 phonemes. These include 29 original consonants, three foreign consonants, and six vowels as shown in Tables 2.3 and 2.4. The SAMPA notation is used in our work for the phonetic transcription (Gibbon et al. 1997) in conjunction with the IPA notation. The foreign consonants are /g/, /p/, and /v/. They are rarely found in MSA and appear in loan words. The phoneme /l'/ is also rare as it appears only in the word /ʔal'1'a:h/ (The God) and its derivatives.

Usually the duration of long vowels is approximately twice the duration of short vowels. Long vowels are not usually treated as two successive short vowels because the duration of long vowels is not exactly twice the duration of short ones (Djoudi et al. 1989).

Arabic is characterized by the existence of pharyngeal and emphatic nature of some consonants (El-Halees 1989; Djoudi et al. 1990). Emphatic consonants in Arabic are /XV/, /t'/, /d'/, /D'/, and /s'/. Those types of phonemes exist in Semitic languages (Holes 2004). Emphatic sounds have also significant effect over the whole word containing the emphatic consonant. Some research papers showed a smaller difference between F1 and F2 for all vowels in those words containing emphatic consonant in almost any position (Hassan and Esling 2007). Two diphthongs may also be considered in MSA. These are /ay/ (/a/ followed by /y/) and /aw/ (/a/ followed by /w/) where a vowel is followed by a semi-vowel.

MSA speech corpora are mainly available in the domain of news broadcasts at a relatively low price and those corpora can be used to build speaker independent acoustic models (Yaseen et al. 2006; LDC 2010).

In the MSA transcription, foreign phonemes may not be treated as extra sounds. They can be grouped with the closest phoneme as it is difficult to distinguish between them from the Arabic orthographic transcription. For example /f/ and /v/ may be grouped together and treated as the same phoneme. The same approach may be applied on /b/ and /p/, and also applied on /g/, /Z/, and /dZ/.

The reason for this grouping is mainly because foreign phonemes are rarely used in MSA compared to original sounds. Furthermore, standard Arabic letters do not have any standard letter assigned for foreign sounds. Some efforts were done to differentiate between foreign and original sounds by using non-standard Arabic letters

Table 2.3 Consonants of Modern Standard Arabic in IPA and SAMPA

IPA	SAMPA	Description
b	b	Plosive, voiced bilabial
t	t	Plosive, voiceless dental plain
d	d	Plosive, voiced dental plain
ṭ	tʻ	Plosive, voiceless dental emphatic
ḍ	dʻ	Plosive, voiced dental emphatic
k	k	Plosive, voiceless velar
g	g	Plosive, voiced velar
q	q	Plosive, voiceless uvular
ʔ	ʔ	Plosive, voiceless glottal
p	p	Plosive, voiceless bilabial
f	f	Fricative, voiceless labio-dental
v	v	Fricative, voiced labio-dental
θ	T	Fricative, voiceless interdental plain
ð	D	Fricative, voiced interdental plain
Ṫ	Dʻ	Fricative, voiced interdental emphatic
s	s	Fricative, voiceless alveolar plain
z	z	Fricative, voiced alveolar plain
ṣ	sʻ	Fricative, voiceless alveolar emphatic
ʃ	S	Fricative, voiceless postalveolar
ʒ	Z	Fricative, voiced postalveolar
ʤ	dZ	Affricative, voiced postalveolar
x	x	Fricative, voiceless velar
ɣ	G	Fricative, voiced velar
ħ	X\	Fricative, voiceless pharyngeal
ʕ	ʔ\	Fricative, voiced pharyngeal
h	h	Fricative, voiceless glottal
r	r	Trill, alveolar
l	l	Liquid, dental plain
ḷ	lʻ	Liquid, dental emphatic
w	w	Approximant (semi vowel), bilabial
j	j	Approximant (semi vowel), palatal
m	m	Nasal, bilabial
n	n	Nasal, alveolar

like using the letter پ for the phoneme /p/, the letter ف for the phoneme /v/, and the letter ج for the phoneme /g/. However, these conventions are not standard and even many standard Arabic keyboard layouts do not show these letters and also

Table 2.4 Vowels of Modern Standard Arabic in IPA and SAMPA

IPA	SAMPA	Description
i	i	Short vowel, close front unrounded
a	a	Short vowel, open front unrounded
u	u	Short vowel, close back rounded
i:	i:	Long vowel, close front unrounded
a:	a:	Long vowel, open front unrounded
u:	u:	Long vowel, close back rounded

standard Arabic character sets as in the ISO 8859-6 do not include these additional letters (ISO 1987).

2.3.2 Classical Arabic

Classical Arabic is the most formal and standard form of Arabic and it is the language of the Quran (the holy book for Muslims). Classical Arabic script as used in Quran represents almost completely the phonetic transcription of the word because the script is fully vowelized and includes diacritic marks that are usually omitted in MSA orthography.

The Arabic language is characterized by the presence of the emphatic consonant /dʕ/. It is believed that this phoneme as pronounced in classical Arabic is exclusively appearing in Arabic and not in any other language. That is why Arabic is usually defined as *The language of /dʕ/* (Newman 2002). The Quran script is also characterized by the presence of the *Alif accent* (also called dagger Alif).

Quran phonetics (according to Hafs’s narration from Assim) includes the sounds of MSA (except foreign phonemes) plus some extra sounds. The following summarizes the main extra sounds that can appear in the Quran recitation:

- Vowel prolongation with duration of four, five, or six short vowels.
- Necessary prolongation of six short vowels.
- Obligatory prolongation of four or five short vowels.
- Permissible prolongation of two, four, or six short vowels.
- Nasalization (ghunnah) with duration of two short vowels.
- Emphatic pronunciation of the consonant /r/ may happen depending on the context of the consonant /r/.
- Echoing sound in unrest letters (qualqala) for the consonants /q/, /tʕ/, /b/, /dʒ/, and /d/ may happen depending on the context of those consonants.

Using Quran in speech recognition is currently limited to recitation learning applications as in (Abdou et al. 2006; Razak et al. 2008). In these applications, the acoustic model is trained with Quran recitations and the user is asked to utter a specific verse and then the application identifies recitation mistakes.

2.3.3 *Dialectal Arabic*

Colloquial (or dialectal) Arabic is the natural spoken Arabic in everyday life communications which is not the case for MSA. Colloquial Arabic is not used as a standard form of Arabic in writing. There are many Arabic dialects and almost every country has its own colloquial form. Even within the same country we can find different dialects. Dialectal Arabic may be classified into two groups: Western and Eastern Arabic. Western Arabic can be subclassified into Moroccan, Tunisian, Algerian, and Libyan dialects, while Eastern Arabic can be subdivided into Egyptian, Gulf, Damascus, and Levantine. The Damascus Arabic is considered the closest dialect to MSA. Arabic speakers with different dialects usually use MSA to communicate.

Since dialectal Arabic is not formally written and does not have any orthographic standard, transcribing adequate speech corpora for dialectal Arabic for the purpose of acoustic modeling is very costly.

For large vocabulary tasks, it is also difficult to obtain a large dialectal text corpus (in the order of millions of words) that is necessary in statistical language modeling.

The available corpora for dialectal Arabic are expensive compared to MSA and they are either low quality telephony conversations corpora as the CALLHOME Egyptian Arabic Speech (Canavan et al. 1997), or read speech corpora but with limited vocabulary as in the Orientel project (Zitouni et al. 2002).

Egyptian Colloquial Arabic

ECA is the first ranked Arabic dialect in terms of number of speakers. Moreover, ECA is the most known dialect among Arabic speakers. The main characteristics of ECA phonetics compared to MSA are:

- /t/ and /s/ are used instead of /T/. E.g. /Tala:Tah/ (three) in MSA is transformed to /tala:tah/ in ECA.
- /g/ is used instead of /Z/ and /dZ/. E.g. /Zami:l/ (beautiful) in MSA is transformed to /gami:l/ in ECA.
- /ʔ/ is used instead of /q/. E.g. /qabl/ (before) in MSA is transformed to /ʔabl/ in ECA.
- The existence of the mid front unrounded long vowel /E:/ and short vowel /E/ (they do not exist in MSA). E.g. /ʔiTnayn/ (two) in MSA is transformed to /ʔitnE:n/ in ECA.
- The existence of the open back unrounded long vowel /A:/ and short vowel /A/ (they do not exist in MSA). E.g. /ʔarbaʔah/ (four) in MSA is transformed to /ʔArbAʔAh/ in ECA.
- The existence of the mid back rounded long vowel /O:/ (it does not exist in MSA). E.g. /jawm/ (day) in MSA is transformed to /jO:m/ in ECA (Hinds and Badawi 2009; Stevens and Salib 2005).

On the lexical level, some words exist only in ECA and not in MSA like: /t'ArAbE:zA/ (table) in ECA while it is /t'awila/ in MSA. The sentence structure in ECA tends to be VSO while in MSA it tends to be SVO.

The transcription of ECA is difficult because people are quite influenced by MSA and always write the MSA word instead, for example the ECA word /tamanjah/ (eight) is usually wrongly transcribed as /tama:njah/ or /Tama:njah/ and keep including the long vowel /a:/ as in MSA while it is replaced in ECA by the short vowel /a/.

2.3.4 Morphological Complexity

Arabic is a morphological very rich language and hence dealing with Arabic as morphemes rather than words will limit the size of vocabulary and significantly decrease the number of Out-Of-Vocabulary (OOV) words. For instance a lexicon of 65,000 words in the domain of news broadcast leads to an OOV rate of 4% in Arabic whilst in English it leads to an OOV rate of less than 1%. Moreover, increasing the size of the lexicon substantially might only decrease the OOV by a small percentage. For example, a lexicon of 128,000 words will only lower the OOV rate to 2.4% (Billa et al. 2002).

Most words in Arabic have a root that consists of three consonants called radicals (rarely two and four). A large number of affixes (prefixes, infixes, and suffixes) can be added to the three consonant radicals to form patterns. Arabic is a highly inflected language with gender, number, tense, person, and case. A single Arabic word can represent a whole English sentence like: **وَإِسْتَطَاعْتَهُمْ** /wabi?stit'a:ʔatihim/ (and with their ability). Table 2.5 explains the morphological complexity in Arabic by showing some of the affixes that can be added to the word **لَاعِب** (player), and how this changed the meaning of the word.

Since Arabic is a morphologically rich language, a simple lookup table is not practical for phonetic transcription nor in semantic analysis. The high number of morphemes leads to an exponential increase of the number of words in Arabic. The total number of Arabic words is as large as 6×10^{10} due to the morphological complexity as reported in (Darwish 2002). That is why significant effort has been investigated in the area of morphological analysers for Arabic as in (Buckwalter 2002b; Xiang et al. 2006) and Arabic stemmers to estimate word root as in (Alshalabi 2005).

2.3.5 Diacritization

A strong grapheme-to-phoneme (almost one-to-one mapping) relation only applies for the diacritized Arabic script. Diacritics appear only in sacred books (like Quran) and Arabic language teaching books. MSA is usually written with the absence

Table 2.5 A few affixes of the word لَاعِب *lāʿib* (player)

Word	Transliteration	Meaning
لَاعِب	<i>lāʿib</i>	player
الَلَاعِب	<i>al-lāʿib</i>	the player
لَاعِبَان	<i>lāʿibān</i>	two players
لَاعِبِينَ	<i>lāʿibīyn</i>	players
وَالَلَاعِب	<i>wa-ʿl-lāʿib</i>	and the player
فَالَلَاعِب	<i>fa-ʿl-lāʿib</i>	so the player
كَالَلَاعِب	<i>ka-ʿl-lāʿib</i>	like the player
بِالَلَاعِب	<i>bi-ʿl-lāʿib</i>	with the player
لِلَلَاعِب	<i>li-lāʿib</i>	to the player
لَاعِبِي	<i>lāʿibiy</i>	my player
لَاعِبُكَ	<i>lāʿibuka</i>	your player
لَاعِبُهُ	<i>lāʿibuhu</i>	his player
لَاعِبُهَا	<i>lāʿibuhā</i>	her player

of diacritic marks and the reader infers missing diacritics from the context. Table 2.6 shows a simple lookup table to convert from orthographic Arabic letters (graphemes) to the corresponding phonemes in SAMPA notation. Diacritic marks are shown in Table 2.7 together with the corresponding phonemes in SAMPA.

Non-diacritized Arabic script leads to a high degree of ambiguity in pronunciation and meaning. For instance, the non-diacritized word كَتَب may have different possible diacritizations and for each case a different pronunciation and a different meaning as shown in Table 2.8. Another problem is that the vowel diacritic case marker on the last letter in a word is determined by the word position in the sentence for example whether the word is subject or object. Furthermore, the speaker is free to choose whether to pronounce or to omit the vowel case marker.

Actually the missing diacritics is not only a problem in Arabic speech recognition. Missing diacritics is more crucial in Arabic speech synthetic Text-To-Speech (TTS) systems Gu et al. (2007). Estimation of the correct diacritization will reduce ambiguity in all Arabic speech and language processing tasks. The majority of Arabic speech corpora is available with non-diacritized transcription and hence acoustic modeling for the speech recognition task is not straight forward. Therefore, an automatic diacritizer is needed to estimate those missing diacritic marks.

Previous work in Arabic acoustic modeling as reported in (Vergyri and Kirchhoff 2004) include an important stage for automatic script diacritization. Automatic diacritization is also known as automatic vocalization. Many research studied how to automatically estimate missing diacritics from the context in order to determine the exact phonetic transcription (Google 2010; Vergyri and Kirchhoff 2004; Sarikaya et al. 2006; Gal 2002; Habash and Rambow 2007; Nelken and Shieber

Table 2.6 Arabic grapheme-to-phoneme conversion

Grapheme	Phoneme	Grapheme	Phoneme
أ	ʔ	س	s
إ	ʔ	ش	S
ؤ	ʔ	ص	sʕ
ئ	ʔ	ض	dʕ
ء	ʔ	ط	tʕ
آ	ʔa:	ظ	Dʕ
ى	a:	ع	ʔʌ
ا (wasla)	ʔ	غ	G
ا	a:	ف	f
ب	b	ق	q
ت	t	ك	k
ث	T	ل	l
ج	Z	م	m
ح	Xʌ	ن	n
خ	x	ه	h
د	d	و	w
ذ	D	ي	j
ر	r	ة	t
ز	z		

2005). However, the problem still remains unsolvable and the WER of automatic diacritization systems ranges between 15% and 25%.

Automatic diacritization is usually based on statistical language modeling to estimate the most probable diacritics for a word given the context in which the word appears. Automatic diacritization is a very important area in Arabic speech recognition since it was proven that a fully vowelized Arabic script improves accuracy over non-vowelized script (grapheme-based) (Afify et al. 2005). The majority of research in automatic diacritization is usually performed for the tasks of TTS or word disambiguation as in (Sarıkaya et al. 2006; Gal 2002; Habash and Rambow 2007; Nelken and Shieber 2005). On the other hand, for the task of acoustic modeling, fewer research have been performed as in (Vergyri and Kirchhoff 2004). A high level digram for an automatic diacritizer is shown in Fig. 2.7.

The available commercial applications for the task of automatic diacritization as Fassieh (RDI 2007) still need manual review in order to achieve lower WER. Manual reviewing is mandatory in order to achieve an accuracy closer to ~99%. However, it is a time consuming and a costly operation especially when preparing large diacritized corpora. The productivity of a well trained linguist is ~1.5K words per work-day (Atiyya et al. 2005). In other words, for a 40 hours speech corpus that consists of 200K words in average, we will need 133 work-days for reviewing the results of a commercial class automatic diacritization system.

Table 2.7 Arabic diacritics to phoneme conversion, example with the consonant /b/

Diacritic	Phoneme	Diacritic description
بَ	b a	Fatha (short vowel a)
بِ	b i	Kasra (short vowel i)
بُ	b u	Damma (short vowel u)
بْ	b	Sukun (no vowel)
بَبْ	b b a	Shadda (double consonant) & Fatha
بَبِ	b b i	Shadda & Kasra
بَبُ	b b u	Shadda & Damma
بَاَ	b a n	Tanween (Nunation) ² Fatha
بَاِ	b i n	Tanween Kasra
بَاُ	b u n	Tanween Damma
بَابَاَ	b b a n	Shadda, Tanween Fatha
بَابَاِ	b b i n	Shadda, Tanween Kasra
بَابَاُ	b b u n	Shadda, Tanween Damma
بَاَ	b a:	Alif Al-Madd (long vowel a:)
بَاِ	b i:	Yaa Al-Madd (long vowel i:)
بَاُ	b u:	Waaw Al-Madd (long vowel u:)

Table 2.8 Possible diacritizations for the word كَتَبَ with English translation

Word	SAMPA	English meaning
كَتَبَ	/kataba/	He wrote
كُتِبَ	/kutiba/	It was written
كَتَّبَ	/kattaba/	He dictated
كُتِّبَ	/kuttiba/	It was dictated
كُتُبَ	/kutub/	Books

2.3.6 Challenges for Dialectal Arabic Speech Recognition

Actually, training an acoustic model for dialectal Arabic is very challenging. In order to train an acoustic model for a particular Arabic dialect, a large speech data set should be collected for that dialect. Unfortunately, dialectal speech data collection is too difficult compared to MSA and compared to other languages like English. The difficulties are mainly due to the inability to estimate the accurate phonetic transcription for dialectal Arabic. This can be clarified in the following:

- A pronunciation dictionary is required to map words to corresponding phoneme sequence. However, due to the high degree of morphological complexity in Ara-

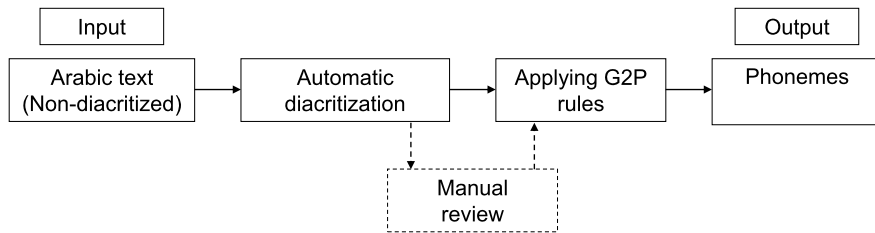


Fig. 2.7 High level block diagram for an Arabic automatic diacritization system

bic, it is impossible to create a pronunciation dictionary that contains all Arabic words.

- Dialectal Arabic is only spoken and not formally written. There is no commonly accepted standard for dialectal Arabic orthography. For instance, the same pronounced letter can be transcribed differently across different transcribers. In other words, from the orthographic transcription, it is not possible to estimate the correct pronunciation.
- Since dialectal Arabic is written without diacritic marks, it is not possible to estimate missing short vowels. Furthermore, automatic diacritization systems do not exist for dialectal Arabic.

2.4 Summary

In this chapter, an overview about the Arabic language was provided from a speech recognition point of view. First, we introduced a brief overview about automatic speech recognition. Second, we provided an overview about the Arabic language. We have seen that there exist a large number of Arabic varieties. These were classified into standard and dialectal. Dialectal Arabic represents all the different Arabic forms as spoken in all Arabic countries. Dialectal Arabic is only spoken in everyday life and there is no standard to represent it in written form. Modern Standard Arabic (MSA) is currently the standard form of Arabic used in all formal situations and it is also a second language for all Arabic speakers. Egyptian Colloquial Arabic (ECA) was chosen as a typical dialectal form and we have shown that there are significant differences between ECA and MSA. Arabic is a morphologically very rich language and a simple lookup table for phonetic transcription is not practical because of the high OOV rate. Grapheme-to-phoneme (G2P) is difficult for Arabic because of missing diacritic marks that are not usually written. Therefore, automatic or manual diacritization is needed in order to estimate the accurate phonetic transcription.

Novel Techniques for Dialectal Arabic Speech
Recognition

Elmahdy, M.; Gruhn, R.; Minker, W.

2012, XXII, 110 p., Hardcover

ISBN: 978-1-4614-1905-1