Int J Speech Technol (2012) 15:487–493
DOI 10.1007/s10772-012-9146-4

Speaker-independent ASR for Modern Standard Arabic: effect of regional accents

Ghania Droua-Hamdani · Sid-Ahmed Selouani · Malika Boudraa

Received: 5 December 2011 / Accepted: 14 May 2012 / Published online: 6 June 2012

Abstract This paper deals with speaker-independent Automatic Speech Recognition (ASR) system for continuous speech. This ASR system has been developed for Modern Standard Arabic (MSA) using recordings of six regions taken from *ALGerian Arabic Speech Database* (ALGASD), and has been designed by using Hidden Markov Models.

The main purpose of this study is to investigate the effect of regional accent on speech recognition rates. First, the experiment assessed the general performance of the model for the data speech of six regions, details of the recognition results are performed to observe the deterioration of the performance of the ASR according to the regional variation included in the speech material. The results have shown that the ASR performance is clearly impacted by the regional accents of the speakers.

Keywords Automatic Speech Recognition systems · Speaker-independent · Continuous speech · Modern Standard Arabic · Algerian Arabic Speech Database (ALGASD) · Regional variability source

G. Droua-Hamdani (✉)
Centre de Recherche Scientifique et Technique pour le Développement de la Langue Arabe (CRSTDLA), Algiers, Algeria
e-mail: gh.droua@post.com

S.-A. Selouani
Information Management, Université de Moncton, Campus de Shippagan, Shippagan, NB, Canada
e-mail: selouani@umcs.ca

M. Boudraa
Houari Boumedienne Science and Technology University, Algiers, Algeria
e-mail: mk_boudraa@usthb.dz

1 Introduction

Automatic Speech Recognition (ASR) research has required many efforts from scientists to progress in development of performing systems. Indeed, it is well-known that intra-linguistics and extra-linguistics features such as: coarticulation, gender, age, regional and social origin, rate of speech, speaking style and spontaneous speech, emotional state, etc. are different sources of speech variability. These speaker's characteristics affecting the speech signal admittedly alter also the performance of the ASR systems. Therefore, improving these systems regarding sources of variability will mostly be a matter of counteracting the effects outlined above (Benzeghiba et al. 2007).

As regards the regional influence, speech recognition under accent variations is a challenging problem whatever the language. The speech for a particular language is rapidly changing depending on the regional accents. The ASR systems suffer from significant performance deterioration when they are operated in mismatched accent conditions.

Furthermore, compared to other languages, there are relatively limited speech recognition studies devoted to the Arabic language studies. Lack of spoken training and testing data is one of the main issues encountered by Arabic studies. The speech corpora, used for the Arabic recognition, usually are not designed for the purpose of ASR researches. Indeed, they are not based on phonetically balanced corpora, and they rarely include the regional variation. Therefore, previous works were principally conducted on Arabic alpha-digits recognition; isolated Arabic vowels and isolated Arabic word recognition; and more recently on the development of continuous speech recognition systems (Alotaibi et al. 2008; Elshafei et al. 2008; Elmahdy et al. 2009; Vergyri et al. 2004).

The present work aims to summarize the main steps of the development of Arabic speech continuous recognition system, for Modern Standard Arabic language that is based on phonetically rich and balanced corpora, so-called *ALGerian Arabic Speech Database* (ALGASD) (Droua-Hamdani et al. 2010). In addition, the paper intends also to reflect the effect of the regional variability on the Arabic ASR performance. The experimental approach tests first the performance of the system using recordings from six regions of ALGASD. Afterwards, details of the recognition results are performed to observe the deterioration of the performance of the ASR according to the regional variation included in the speech material.

The paper is organized as follows: Sect. 2 summarizes the main characteristics of the Arabic language. Section 3 gives an overview of the speech data used—ALGASD corpus. Section 4 gives explanation about train and test corpora. Section 5 describes the acoustic front-end of the ASR system. Section 6 gives an outline about the lexicon used in the development of the ASR. Section 7 describes the acoustic and language models. Section 8 shows the evaluation of the Arabic ASR. Section 9 concludes and indicates the perspective of this work.

2 Arabic language background

Arabic is an official language in more than 22 countries. The estimated number of Arabic speakers is about 300 million. Recent approaches in language and speech processing categorize the Arabic language as Modern Standard Arabic (MSA) and Modern Colloquial Arabic (MCA). MSA language is the standard variety shared by educated speakers throughout Arabic-speaking regions. It is the form that is used in education, media, and formal talks. Colloquial Arabic is what is spoken in everyday conversations and varies considerably not only across countries, but also within the same country.

The MSA basic phonological profile includes 28 consonants. Among these consonants, there are two distinctive classes, which are named pharyngeal and emphatic phonemes. MSA vowel phonemes are limited in number compared to English or French. There are three short vowels and three long vowels. However, there are many allophones to each of them depending on the consonantal context. In addition, MSA has two diphthongs /ay/ and /aw/. All Arabic syllables must contain at least one vowel (Watson 2007). Moreover, Arabic vowels cannot be initials and they can occur either between two consonants or be the final phoneme in a word.

3 Data collection

Speech corpus is an important requirement for developing any ASR system. This section describes the characteristics of the database, ALGASD, used in this study (Droua-Hamdani et al. 2010). The sound corpus is designed to train and test automatic speech recognition engines. Texts material used to record the speech corpus consists of 200 Arabic Phonetically Balanced (APB) sentences which included all Arabic phonemes with the respect of their actual distribution in MSA language.

Speakers are different from each other depending on the acoustics differences which are related to the vocal tract or to the pronunciation differences which are generally associated the geographic localities. In addition to acoustics differences between speakers, ALGASD corpus takes account of the main pronunciation variations of MSA due to regional differences in Algeria. Thus, the regional coverage corresponds to the major dialect groups.

Since most of the Algerian population is settled in the north of the country rather than in the south, the distribution of speakers according to their number and gender is proportional to the population of the regions. The texts to read are also proportional to the numbers of speakers in each region. Therefore, the number of recordings is not equal in the studied localities.

The ALGASD database is a good challenge for speech recognizers because of its diversity. In fact, it is enriched with varieties of 300 Algerian speakers taking into consideration the following characteristics:

1. Gender: females/males speakers.
2. Age: there are 3 age groups: younger speakers between 18 and 30, middle-aged speakers between 30 and 45, and speakers over 45.
3. Education level: three categories of education level are also considered: middle group (primary to secondary school), graduate group (university), and post-graduate group.
4. Speakers come from different socioeconomic backgrounds and their mastery of Arabic is different: doctors, teachers, students, unemployed persons, etc.
5. Regional variation: 11 regions scattered across the country.

4 Train and test corpora

For the purposes of the experiment, the data of 167 speakers corresponding to 592 recordings are extracted from 6 region subsets. From the total speech corpus, we build two subsets to train and test the Arabic recognizer. The system is trained using 434 sentences which represent two-thirds of the total sentences used for the development of the ASR system. The number of speakers used for the training is 118. Test corpus, composed of 158 sentences, includes 49 speakers representing all the regions devoted to the study. Speakers and corpus test have not been used in the training phase. The number

Table 1 Numbers of speakers and recordings in train and test corpora

Regions	Designation	Train		Test	
		Speakers	Recordings	Speakers	Recordings
Algiers	R1	56	198	24	78
Tizi-Ouzou	R2	24	93	10	30
Jijel	R3	12	45	4	12
Bechar	R4	5	19	2	8
El-Oued	R5	11	45	5	14
Ghardaia	R6	10	34	4	16

of sentences tested varies from region to region because the number of speakers is not constant across regions. Table 1 shows the number of speakers and recordings for each region in the in training and test phase.

5 Pre-processing

Characteristics of speech sounds vary substantially depending on the speaker and the acoustic environment. The widely-used statistical method for characterizing the spectral features of the speech frame is Hidden Markov Models (HMMs) (Jelinek 1999). Both the automatic speech recognition system (described below) and the parameterization method are designed by using the Hidden Markov Model Toolkit (HTK) which runs on a Windows platform (Young et al. 2006).

The same recognition system is trained and tested with identical settings of all relevant parameters. Acoustic front-ends in speech recognizers produce sequences of observation vectors which represent the short-term spectrum of the speech signal. The two most usually used parameterizations are Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) (Huang et al. 2003). However, the MFCC are commonly used as main features in speech recognition systems.

The conventional MFCC extraction method consists of several computational steps. The speech is first pre-emphasized with a pre-emphasis filter that is used to enhance the high frequency components of the spectrum. This is performed by applying the following formula:

$$x'_n = x_n - a x_{n-1} \tag{1}$$

where a is the pre-emphasis coefficient which should be in the range $0 \leq a < 1$.

The next, step in the processing, is to apply a window function on each individual frame of the signal, to reduce boundary effects. Typically the Hamming window is used. The impulse response of the Hamming window is defined as follows:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/N - 1) \\ \quad n = 0, 1, \ldots, N-1 \\ 0 \quad \text{otherwise} \end{cases} \tag{2}$$

where N is the total number of samples in the frame.

The MFC coefficients are calculated as a set of Discrete Cosine Transform (DCT) decorrelated parameters, which are computed through a transformation of the logarithmically filter-output energies. These energies are resulting through a perceptually spaced bank of equal height triangular filters $H_i(k)$ that are applied on the Discrete Fourier Transform (DFT) of a given speech signal. The frequency bands in the MFCC are equally spaced on the *mel* scale, which mimics the human auditory system response. We can calculate the Mel-Frequency cepstrum from the output power of the filter bank using the following equation:

$$c_j = \sum_{i=1}^{M} S[i] \cos\big(j\pi(i - 1/2)/M\big), \quad \text{with } j = 1, 2, \ldots, J \tag{3}$$

where M is the number of filters in the filter bank, J is the number of cepstral coefficients, $S[i]$ is the log-energy at the output of each filter:

$$S[i] = \log_{10}\left[\sum_{k=0}^{N-1} \big|X(k)\big| . H_i(k)\right] \quad i = 1, 2, \ldots, M \tag{4}$$

Temporal changes in the spectra play an important role in human perception. One way to capture this information is to use delta coefficients that measure the change in coefficients over the time. As a basic feature set, we extracted the (MFCCs) 1–12 along with energy and their first (δ) and second order ($\delta\delta$) regression coefficients. This greatly enhances the performance speech recognizers on based HMM. The derivatives of the MFCCs are calculated through the use of regression formula (3).

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{5}$$

where d_t is a δ coefficient at time t computed in terms of the corresponding static coefficients $c_{t-\theta}$ to $c_{t+\theta}$. The same formula is applied to the δ coefficients to obtain acceleration ($\delta\delta$) coefficients.

The principal usefulness of cepstral coefficients is that they are generally decorrelated and this allows diagonal covariances to be used in the HMMs. However, the higher order cepstra are numerically quite small compared to the lower order. So, it is appropriate to re-scale the cepstral coefficients to have similar magnitudes. Therefore, the last step in the processing is to apply a liftrage window. This is done according to the following formula:

$$c'_n = \left(1 + \frac{L}{2} \sin \frac{\pi n}{L}\right) c_n \tag{6}$$

where L is the size of liftrage window.

The parameters of the ASR system are 16 kHz sampling rate with a 16 bit sample resolution, 25 millisecond Hamming window duration with a step size of 10 milliseconds, normalized energy, MFCC coefficients with 22 as the length of cepstral leftering and 26 filter bank channels of which 12 are as the number of MFCC coefficients, and of which 0.97 are as the pre-emphasis coefficients. In computation of δ and $\delta\delta$ coefficients, Θ is set to 3.

6 Arabic phonetic dictionary

The development of an effective ASR system requires a phonetic dictionary to training and test phases. The lexicon or pronunciation dictionary is used to map word sequences to phone sequences. The HMMs corresponding to the phone sequence may then be concatenated to form a composite model representing words and sentences. Each word in the lexicon may have several pronunciations, and in this case, there will be one branch in the network corresponding to each alternative pronunciation. Each pronunciation may consist either of a list of phones or a list of HMM names (Young et al. 2006).

Usually alphabet-to-sound conversion for Arabic has simple one-to-one mapping between orthography and phonetic transcription for given correct diacritics. However, at word level there are very few exceptions. Indeed, some words may have a slight difference between the pronunciation and the orthographic form as /haða/ which is pronounced /ha:ða/.

Texts of ALGASD are transcribed using *Speech Assessment Methods Phonetic Alphabet* (SAMPA), but some of the phonemes are renamed for machine convenience. Thus, to create the dictionary, we replace them by symbols that are proposed in the West Point Modern Standard Arabic database by LDC with minor modifications (LDC). Table 2 gives a sample of the lexicon used throughout our ASR experiments.

Table 2 Sample of the lexicon used throughout the ASR

ASR	SAMPA
d'a?ula	D_ ah Q_ uh l ah
d'aru:ratun	D_ ah r uw r ah t uh n
Gadan	G_ ah d ah n
Gadrahum	G_ ah d r ah h uh m
Gafala	G_ ah f ah l ah
Gafat	G_ ah f ah t
Gala:	G_ ah l ae

7 Acoustic and language models

Currently, the most popular and successful speech recognition systems use Hidden Markov Models in the acoustic modeling. HMM Toolkit is used to train the acoustic models of thirty-four MSA phonemes (28 consonants and 6 vowels) to which we add a model of silence (sil). A short pause (sp) model is created from the silence model and tied to it. All the models are context independent, 5-state HMM (first and fifth states were non-emitting) left to right without skip state, all with one Gaussian mixture (diagonal covariance) per state.

Baum-Welch re-estimations algorithm is used in order to estimate the transition probabilities of the context-independent HMMs. An alignment of speech data is done after the seventh re-estimation using the Viterbi algorithm (Rabiner and Juang 1993; Huang et al. 2003).

Language model is another important requirement for any ASR system. In practice, n-gram models have been shown to be extremely effective in modeling language data in speech recognition tasks. Current research mainly focuses on bigram topic models which are built from labels used in the training process computed according to the following formula:

$$p(i,j) = \begin{cases} (N(i,j) - D)/N(i) & \text{if } N(i,j) > t \\ b(i)p(j) & \text{otherwise} \end{cases} \tag{7}$$

where $N(i,j)$ is the number of times word j follows word i and $N(i)$ is the number of times that word i appears.

8 Evaluation and discussion

In this section, we discuss two experiments that gauge how well the system performs with respect to regional influence.

- The first experiment assesses the general performance of the model for the entire data speech (i.e. R1 to R6).
- The second experiment aims to detail the results obtained above in order to investigate the impact of regional variation on the general performance of the system. This is done by testing individually each region.
  - The speaker-independent ASR performance in the first experimentation is 91.7 % which is a satisfactory result. The accuracy is 90.6 %. Indeed, from the overall data test corpus (i.e. 158 recordings), 18 sentences are

not recognized correctly. These sentences are affected either by a deletion, a substitution or an insertion at the beginning or the ending of the sentence. The system fails in recognizing 38 tokens out of 458. The number of substitutions is 32 while the number of deletions is 6, and the number of insertions is 5.

- As shown in Table 3, the results of the second experiments indicate considerable variation among the 6 localities. The results show that northern regions R1, R2, and R3 have a reasonably high correct rate of recognition. On the other hand, the worst performance is encountered in southern regions R4, R5, and R6. Figure 1 illustrates numbers of recognized labels compared to the total number of the tokens given in the data set corpus of each region.

Table 4 and Fig. 2 show the deletions, substitutions and insertions of tokens of each region. We notice that the southern regions present the higher numbers of substitutions as in R5 and R6 compared to the total number of tokens to recognize.

Table 3 Word and sentence recognition rates by regions

	Designation	Word		Sentences
		%Corr	%Acc	%Correct
North	R1	94.34	93.87	88.89
	R2	98.98	97.96	93.33
	R3	100	100	100
South	R4	78.26	78.26	71.43
	R5	80.85	76.60	71.43
	R6	74.42	67.44	64.29

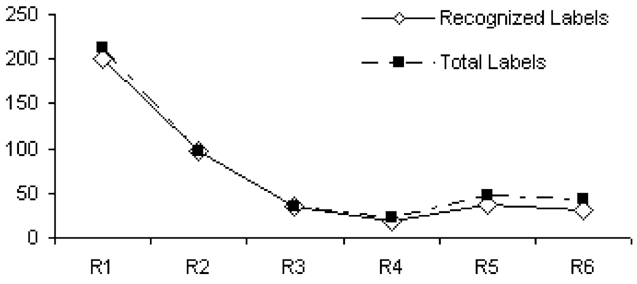

Fig. 1 Comparison between recognized and total words by region

In this experiment, we also investigate the number of sentences not correctly recognized according to the gender of the speakers. The participants of test phase have not been used in the training phase. We notice that the distribution of speakers that produced sentences which are not correctly recognized by the ASR (deletion or missed tokens) is generally equal for both kinds of speakers for each region. However, their number is higher in southern regions than northern ones compared to the total number of speakers of the test. The dividing of both genders (female/male) is reported in Table 5. Figure 3 plots the distribution of speakers that have pronounced unrecognized sentences.

The purpose of the experiment is to design a comprehensive system that can recognize all the phonemes of Arabic MSA, based on phonetically balanced corpora, regardless of the speaker's accent. The recognition result shows a satisfactory performance rate. From this perspective, the expected target is relatively achieved as long as Jijel (R3) with even smaller numbers of speakers and recordings, compared to Algiers and Tizi Ouzou (respectively R1 and R2), reaches a maximum recognition performance. Although it is well-known that certain phonemes pronounced with the accent of Jijel have distinctive phonetic features. Regarding the region R5 even with numbers of speakers and recordings almost similar to R3, we get a fairly low rate.

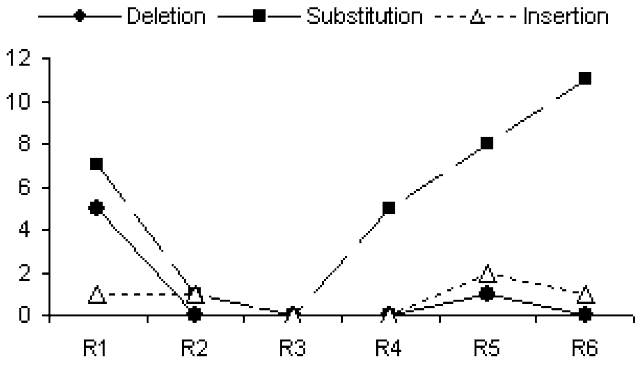

Fig. 2 Deletions, substitutions and insertions words for each region

Table 4 ASR performance analysis

	Regions	Total words	Recognized words	Deletion	Substitution	Insertion
North	R1	212	200	5	7	1
	R2	98	97	0	1	1
	R3	35	35	0	0	0
South	R4	23	18	0	5	0
	R5	47	38	1	8	2
	R6	43	32	0	11	1

Table 5 Speakers according to their gender: speaker's number that have produced unrecognized sentences are reported between parentheses

Regions	Males	Females	Total speakers
R1	8 (2)	16 (4)	24 (6)
R2	5 (1)	5 (1)	10 (2)
R3	1 (0)	3 (0)	4 (0)
R4	–	2 (1)	2 (1)
R5	2 (1)	3 (0)	5 (1)
R6	1 (1)	3 (1)	4 (2)

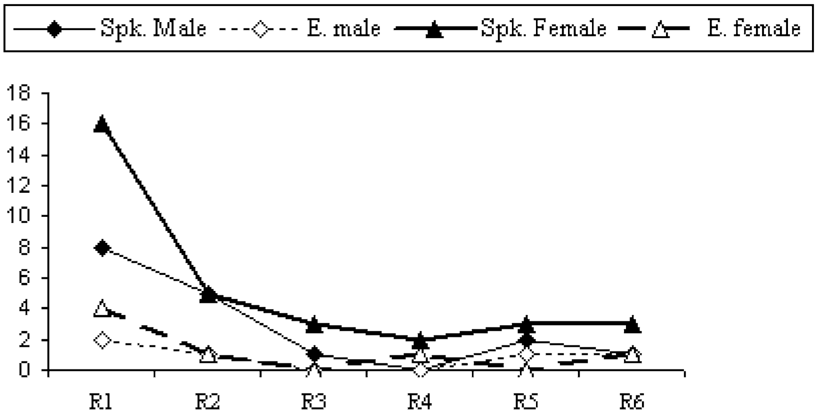

Fig. 3 Speakers that produced unrecognized by the ASR

We have also investigated the performance of ASR according to the gender of the speakers. The results show that the speakers of the southern regions produce more unrecognized sentences than those of the northern ones. The ASR rates for each region are generally equal between speakers be they female or male.

In a previous work (Droua-Hamdani et al. 2010), we have assumed that these regions of ALGASD have more homogeneous phonetic features and fewer phonologic differences to achieve a high recognition rate. But, the present study shows variability in the recognition rates between northern and southern localities—about 23 %. This shows that recognition differences are not due to random error but may be due to differences in pronunciation.

We suppose that this difference is caused by substantial differences in production of certain phonemes of the corpus (allophones). More investigations are needed to reveal which phones produce the deviation between recognition rates.

9 Conclusion

The present study concerns the development of a speech continuous recognizer of MSA. Data set is taken from six regions (3 from the north and 3 from the south) of ALGASD sound corpus. This voice bank mirrors different sources of speech variability related to regional and social variations of Algerian speakers. The speaker-independent system is based on HMM strategy carried out by HTK tools.

Two sets of experiments are conducted to test the ASR performance. The first one uses the overall test data of the six regions, while the second details the performance rates related to each region.

The results of monophone speech recognition models, for the first experiment, are successful for the purposes of ASR and might constitute a useful baseline model for further studies using complex ASR systems dedicated to MSA. However, in the second experiment, we observe unbalanced recognition rates between northern regions and southern ones. The results show a reasonably high correct rate of recognition for R1, R2, and R3 where R3 shows the higher recognition rates. The worst performance is encountered in southern regions; R6 is characterized by a lower accuracy.

Adaptive techniques (Maximum a Posteriori MAP, Maximum Likelihood Linear Regression MLLR) can be used to modify system parameters to better match variations related to the speaker's accent. This is our primary objective subsequently to improve the performance of the Arabic ASR.

References

Alotaibi, Y. A., Selouani, S. A., & O'Shaughnessy, D. (2008). Experiments on automatic recognition of non-native Arabic speech. *EURASIP Journal on Audio Speech and Music Processing*, *2008*, 679831. 9 pages, doi:10.1155/2008/679831

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: a review. *Speech Communication*, *49*(10–11), 763–786.

Droua-Hamdani, G., 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), 157–166.

Elmahdy, M., Gruhn, R., Minker, W., & Abdennadher, S. (2009). Modern Standard Arabic based multilingual approach for dialectal Arabic speech recognition. In *8th international symposium on natural language processing. SNLP'09* (pp. 165–174).

Elshafei, M., Al-Muhtaseb, H., & Al-Ghamdi, M. (2008). Speaker-independent natural Arabic speech recognition system. In *The international conference on intelligent systems ICIS 2008*, Bahrain.

Huang, X., Acero, H. A., & Hon, H.-W. (2003). *Spoken language processing. A guide to theory, algorithm and system development*. Upper Saddle River: Microsoft Research, Prentice Hall.

Jelinek, F. (1999). *Statistical methods for speech recognition* (2nd ed.). Cambridge: MIT.

Linguistic Data Consortium (LDC). Catalog Number LDC2002S02, 2002. http://www.ldc.upenn.edu.

Rabiner, L., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs: Prentice Hall.

Speech Assessment Methods Phonetic Alphabet (SAMPA): http://www.phon.ucl.ac.uk/home/sampa/arabic.htm.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2006). The HTK book (for HTK Version. 3.4). Cambridge University Engineering Department. http:///htk.eng.cam.ac.uk/prot-doc/ktkbook.pdf.

Vergyri, D., Kirchhoff, K., Duh, K., & Stolcke, A. (2004). Morphology-based language modeling for Arabic speech recognition. In *Proceeding of ICSLP* (pp. 2245–2248).

Watson, J. C. E. (2007). *The phonology and morphology of Arabic*. New York: Oxford University Press.