Unit Selection Model in Arabic Speech Synthesis

Nedhal A. Al-Saiyd† and Mohammad Hijjawi†,

Computer Science Department, Applied Science Private University, Amman, Jordan

Summary

In text-to-speech synthesis the speech unit is segmented and extracted from natural speech, coded, classified, labelled and stored in the inventory of units. To generate a smooth synthesized speech the unit selection plays an important and crucial role. A multiple-instances approach of speech unit is suggested. A set of speech units is prepared for each speech unit type, taking into consideration target unit structure, the preceding and the succeeding syllables, and position in the utterance. In this paper, a diacritic Arabic Text-To-Speech system is described. The goal of this paper is to get intelligible, high quality speech synthesis based on unit selection using a syllables model.

Key words:

Arabic Speech, Unit Selection, Segmentation, Syllable, Speech Analysis.

1. Introduction

Text-to-speech system (TTS) is the production of intelligible speech waveform based on the phonetic transcription of written text [1]. TSS is an important field of artificial intelligence.

TTS is used in human-technology interaction and has many benefits for visually-impaired persons, navigation devices, e-learning systems, etc. In text-to-speech synthesis the speech unit is segmented and extracted from natural speech. The speech segments are coded, labelled and stored in the inventory of units to produce the phonetic transcription in written symbols (i.e. phonetic symbols). The coded segments are converted later to acoustic data [2], [3]. A general architecture of text-to-speech system consists of two main processes, which are the off-line that is performed to construct, the speech units and the on-line speech synthesis process that consists text analysis, searching and retrieving units from the inventory, prosody generation and speech synthesis [4]. Figure (l) shows the general architecture of Text-to-Speech system. The speech unit selection plays an important and crucial role to generate a smooth synthesized speech.

The first TTS system is done in Japan in 1968 and then some maintenance attempts is done on morpheme (i.e. unit of sound) database, letter-to-sound rules to convert English letter to phonemes, parser algorithms and phoneme-to-speech algorithm at Bell Laboratory, MIT and KTH [5].

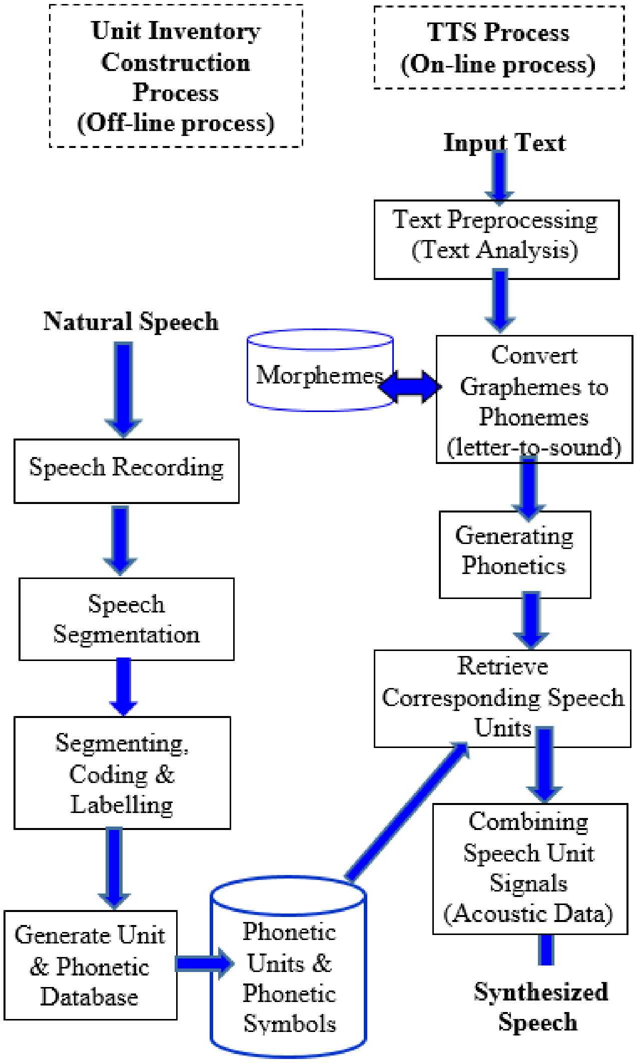
Unit Inventory Construction Process (Off-line process)
TTS Process (On-line process)
Input Text
Text Preprocessing (Text Analysis)
Natural Speech
Morphemes
Convert Graphemes to Phonemes (letter-to-sound)
Speech Recording
Generating Phonetics
Speech Segmentation
Retrieve Corresponding Speech Units
Segmenting, Coding & Labelling
Combining Speech Unit Signals (Acoustic Data)
Generate Unit & Phonetic Database
Phonetic Units & Phonetic Symbols
Synthesized Speech

Fig. 1 General Architecture of TTS System.

Most text-to-speech systems have been related to Indo-European languages; English, French, German, Spanish, Italian, Russian, etc., but there is less-consideration given to Arabic Language [6].

Phonemes are divided into two major categories: vowels category, which is represented by ‘V’, and consonants category, which is represented by ‘C’. All vowels are

Manuscript received April 5, 2018
Manuscript revised April 20, 2018

voiced sounds while some consonants are voiced sounds and some are unvoiced sounds. Most of the languages have between 20 to 40 phonemes [7].

In the past few decades, Arabic speech synthesis has applied many approaches that were done mostly in academic research centers. The main approaches can be classified into two main categories, the rule-based formants (frequencies) synthesis systems; which is the most commonly used technique, and concatenation synthesis systems. Rule-based synthesis depends on speech production rules (letter-to-sound rules); therefore it eliminates extra space of storing speech segments and the produced speech is intelligible but not natural. Concatenation synthesis technique (data-driven technique) resolves the discontinuities at unit boundaries and enhance the produced speech through prosodic modification to speech units [8]. Speech units can be allophones, diphones, triphones, demisyllables, syllables, or words. Diphones are the most widely used units in concatenative synthesis [7]. The quality of the produced synthesized speech is influenced by the length of the unit. In general, Arabic language has no one-to-one mapping between letters and phonemes. A letter may give multiple phonemes and two letters may give one phoneme. With using diacritic mark can double the phoneme.

The rest of this paper is organized as follows. In Section 2, the Arabic grapheme-to-phoneme transcription is presented. Syllable unit selection is presented in Section 3. Speech segmentation, and labeling is explained in section 4. Arabic syllabication as a synthesis unit is presented in details in section 5. Lexical stress assignment is defined in section 6. Finally, conclusions is given in the last section.

2. Arabic Grapheme-To-Phoneme Transcription

Modern standard Arabic (MSA) has 28 consonant phonemes. Consonant sounds are differentiated according to the place of articulation, the way of travelling airflow from the lungs up and out of the mouth and nose and their voicing/un-voicing of the consonants. Arabic has few diacritical marks that appear at the end of the consonant letters. Sometimes the diacritics appear as a unique mark above or beneath the letter; as shown in table 1 and table 2 [9]. The existence if these diacritical marks increase the clarity of written text, present correct pronunciation and consequently enhance the intelligibility of the produced speech. Arabic language has three short vowels /i/, /u/, /a/, and three long vowels /ii/, /uu/, /aa/. Some of the Arabic grapheme letters are not pronounced, and affect the following grapheme; as in the lam-alshamsiya; the pronunciation will neglect the letter (lam); i.e. do not have the sound effect, and double the followed grapheme letter. Also, the pronunciation of long vowels is about twice as long as short vowels [10], [11].

Table 1: Transcription of Arabic vowels (diacritics) to corresponding phonemes

Arabic short vowel Orthographical Representation	Pronounced as	Phonological Representation
◌َ	fatHa, front low unrounded	/a/
◌ُ	Damma Back high rounded	/u/
◌ِ	Kasra Front high unrounded	/i/
◌ْ	Sukuun Consonantal or vowelless	

Table 2: Transcription of one diacritic to more than one phoneme

Arabic Orthographical Representation	Pronounced as	Phonological Representation
◌ً	tanween alfatha	/an/
◌ٍ	tanween alkasra	/in/
◌ٌ	tanween alDamma	/un/
◌ّ	shadda	doubling the consonant

In our system, Arabic text with diacritic marks does not have a one-to-one correlation between written text and reading system. Arabic is considered relatively complex language because of its difficult linguistic structure and the amount of linguistic information [12]. To facilitate Arabic TTS, the written text have assign diacritics for the pronunciations needs.

Arabic phonetic representation defines the speech sounds of Arabic spoken language. They are represented by symbols to transcribe orthographic letters and diacritics sounds. Figure 2 shows the standard Arabic phonetic system corresponding to the Arabic orthographic letters.

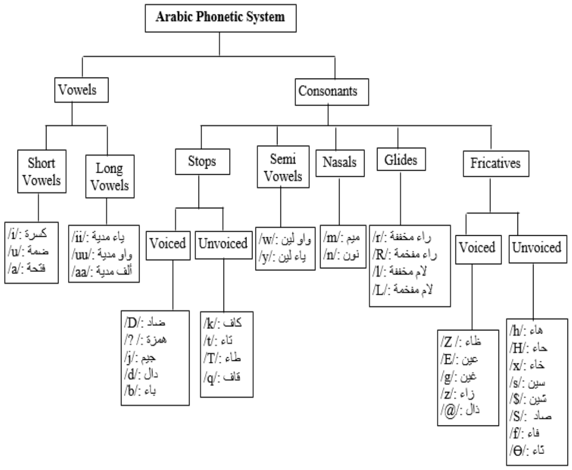

Fig. 2 Arabic phonetics and their corresponding Arabic orthographic letters and diacritics

Using diacritics helps in pronouncing each word correctly and therefore, will help in identifying the meaning of the utterance. But, sometimes the written Arabic text has not assigned the appropriate diacritics. This situation will make the process of generating the speech difficult [11].

The tri-literal roots <drs> can be pronounced as:

- <darasa> دَرَسَ which means learn a lesson,
- <darrasa> دَرَّسَ which means teaches,
- <duresa> دُرِسَ which means the lesson is learned,
- <durresa> دُرِّسَ which means the lesson is taught, and
- <dars> دَرْس which means lesson.

Some of the Arabic grapheme letters are pronounced in two ways depending on the context that consists this letter; such as the letter raa ر /r/ in the word ramz رمز, and the same letter in the word qurban قربان but it pronounced as /R/

The pronunciation of some input Arabic grapheme letters are represented by more than one phonemes.

The letter أ is represented by /?a/.

The letter أُ and ؤ is represented by /?u/.

The letter ئ and إ is represented by /?e/.

3. Syllable Unit Selection

One of the ongoing research area in speech synthesis is the selection of synthesis units and building the inventory that are used to produce very intelligible synthetic speech. Many researchers have discussed the challenges of selecting units to build a large inventory of phone-sized segments that depend on context [13]. They may have inconsistent in unit boundary concatenations that causes different degrees of recognized discontinuities (i.e. junctions) [6]. Context-dependent vowels have smooth concatenations, if the vowels are in the middle of the synthesis unit. Different vowels have different degrees of discontinuities when they are joined in the middle of the synthesis unit. This in turn will degrades the generated speech quality [6].

The method of unit selection plays an important role in building synthesized speech. It is suggested to select speech units and consider the Arabic diacritics and linguistic components of the TTS system to achieve high-quality speech. The process of segmentation and labelling is shown in figure 3.

The main inventory is decomposed into five data files according to the five consonant categories (stops, nasals, semi-vowels, glides and fricatives) to make the inventory more manageable and to make search process for a specific unit faster. The data structure of each record in the unit inventory is described as:

{Current Syllabus Unit Code, Preceding Syllabus Unit Code, Succeeding Syllabus Unit Code, Syllabus label, syllable structure, stress type, position of 1st vowel, the length of vowel samples, position of 1st consonant (sample), the length of consonant samples, Intrinsic duration, Pitch period, power of the last phoneme in the syllable, the power of the last phoneme in the syllable, Speaking tempo}.

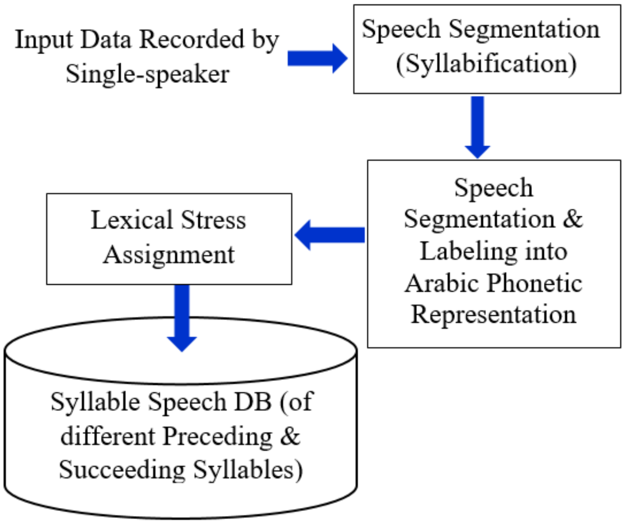
Input Data Recorded by Single-speaker
Speech Segmentation (Syllabification)
Speech Segmentation & Labeling into Arabic Phonetic Representation
Lexical Stress Assignment
Syllable Speech DB (of different Preceding & Succeeding Syllables)

Fig. 3 Arabic Syllable Unit Selection

There is a correlation between the acoustic variety of the units and the unit inventory size. If the unit is decreased in size, the inventory size is increased. The unit inventory have to store at least 3 instances of each unit, and taking into the consideration all possible concatenations states and various modifications in prosodic data. Classification and clustering the speech units is another challenging issue.

4. Speech Segmentation, and Labeling

In the inventory of speech synthesis at least one instance of each speech unit is required. But in our work, multiple-instances approach of speech unit is suggested. A set of speech units is identified for each speech unit type, taking into consideration unit structure and position in the utterance. Each unit is influenced by the predecessor and successor syllables of the uttered words, as it is shown in figure 4. The syllables are then coded and stored in the inventory, with other important data; as it is explained previously in section 3.

Fig. 4 The neighboring of the current Syllable

This will help to overcome the problem of mismatching between the selecting of single speech unit and the target and reduce extra signal processing that is needed to enhance produced speech quality. The phonetic inventory will have large number of speech units that reflects phonetic and prosodic features of the context.

The unit selection process, however, will select the speech unit that fits a given phonetic and prosodic context. It requires more complex unit-selection search algorithm and maintains relatively large number of speech units in the database [14].

5. Arabic Syllabication As A Synthesis Unit

In concatenation process, the speech unit plays a crucial role in the quality of speech to be more natural, where the amount of concatenation is preferred to be minimum.

It is preferred to select the Arabic syllable as a variable-size speech unit, which combines consonants 'C' and vowels 'V'; and use it as a sound unit. Syllable is a segment of phonological words that depends on pronounceable segments of words. Arabic language has very large of monosyllabic (one syllable) or bi-syllabic words. The syllable can have stress in the words and this will reduce the discontinuities in the inter-units, which in turn will enhance the quality of the produced speed. They have all possible phone transitions [15].

The Arabic syllables are always starts with one consonant. They are categorized into open syllabuses that ends with vowel ‘V’, the closed syllables that are ends with single consonant ‘C’ and the double closed syllables that are ends with two consonant ‘CC’.

The successive elements within Arabic syllable boundaries are made up of the segmental phonemes of Arabic language [16]. The five different structure patterns of Arabic syllables are:

- CV- consonant followed by short vowel, which has light stress; as /bi/ " بِ" .
- CVV- consonant followed by long vowel; as /maa/ "ما " , /kuu/ "كو", which has heavy stress.
- CVC- consonant followed by short vowel followed by consonant; as /ber/ "بِر", /kur/ "كُر", which has light stress.
- CVVC- consonant followed by long vowel followed by consonant; as /reem/ "ريم", /noor/ "نور", /naar/ "نار", which has heavy stress.
- CVCC- consonant followed by short vowel followed by consonant followed by consonant; as /sabt/ "نَهر", /fajr/ "فَجر", which has heavy stress.

The last two structures are called super-strong or super-heavy syllables and are appeared in the end of word. CV is considered as short syllable while the others are long syllables. CV is the most common in Arabic language and CVVC is the least common. The first four patterns occur initially, medially and finally in the word.

The syllable structure consists of three parts, A, B, C. B is considered as a main part and has the prominence and referred as the nucleus of the syllable. The remaining parts A, and C are referred to as marginal. Acoustically, the nucleus part has more intensity than marginal. The marginal part can be either initiation or the termination of the syllable. The consonant is always in the initiation, while the termination can be consonant, two consonant, or zero consonant factors [16]. Each syllable units is classified and labelled according to the Arabic phonetic representations.

6. Lexical Stress Assignment

The stress or accentuation is the amount of power that is distributed over the syllables of each utterance. The stressed vowel will have longer duration and higher fundamental frequency than the unstressed [17], [18], [19] [20]. It is characterized by temporal distinctions between articulatory gestures.

Stress is predictable prosodic information rather than phonemic. In Arabic, stress is dependent on Arabic syllable structure. It does not fall on the last syllables of the word. If the word has two syllables, the stress will occur on the first syllables. The stressed vowels has longer duration and higher fundamental frequency

After assigning lexical stress, the Arabic syllables of different length, structure and lexical stress is stored in syllable’s database with the corresponding coded speech signals. Building a large-size speech database that covers all unit variants is not an easy task.

The primary stress has relatively higher pitch and the secondary stress has soft and constant pitch. The following rules are applied to assign primary and secondary stress for the syllable(s) per each word in the utterance:

Rule 1: For each extracted word from written sentence do
Find each syllable length and syllabus occurrences per word

If syllable length=2 and No-of-syllables per word >=1 then the first syllable get the primary stress, and the rest syllables get weak stresses.
For example: /kataba/ is syllabified with stress as:
/ka´- ta´´- ba´´/ for the word: /كَتَبَ/ / CV´ - CV´´ - CV´´/

Rule 2: For each extracted word from written sentence do
Find each syllable length and syllabus occurrences per word
If word has only one longer-syllabus than other syllables then the long syllabus gets primary stress and the rest gets secondary stresses.
For example: /kaatib/ is syllabified with stress as:
/kaa´ - tib´´/ for the word: /كاتِب/
/CVV´ - CV´´C/

Rule 3: For each extracted word from written sentence do
Find the syllable length and syllabus occurrences
If word has two or more longer-syllable then the longer-syllables that is closest to the word-end (not the last syllable) gets the primary stress and the other long syllables get secondary stress.
For example: /ra?iisahunna/ is syllabified with stress as:
/ra - ?ii´´ - sa - hu´n - na/ for the word: /رَئيسَهُنَّ/
/ CV - CVV´´ - CV - CV´C - CV /

Rule 4: For each extracted word from written sentence do
Find the syllable length and syllabus occurrences
If word has long syllable=4 and the syllable structure is either CVVC or CVCC then this syllable gets primary stress.
For example: /kitaab/ is syllabified with stress as:
/ki - taab´/ for the word: /كِتاب/
/ CV - CVV´C/
Or
/katabat/ is syllabified with stress as:
/ka - tabt/ for the word: /كِتاب/
/ CV - CV´CC/

Rule 5: For each extracted word from written sentence do
Find the syllable length and syllabus occurrences
If word has long syllable=3 and the syllable structure is either CVV or CVC then the closes syllable to the beginning of the word gets primary stress and the one that is closest the end of the word gets the secondary stress.
For example: /kitaabuhu/ is syllabified with stress as:
/ki - taab´/ for the word: /كِتابُاتُهُ/
/ CV - CV´V-CV´´V-CV-CV/

In unit selection process, the proper unit is found and retrieved, when the selected unit matches phonetically and prosodically the target units. The search algorithm impacts unit selection process, when the search algorithm weights the suitability of extracted the desired units. The database has a set of candidate units for each syllable and covers a large collection of phonetics and prosodic variants of context dependent units. The size of the speech database is related to variable length of unit synthetic speech units and to various syllables that precede and succeed them, which will increase the searching time for the appropriate selected unit in concatenation process.

7. Conclusion

- In our system, the phonetic string is generated from Arabic orthography with diacritics. Firstly, a system was applied to the verbal database so that when we chose a word to be pronounced it automatically divided each word, transcribed phonetically, into syllables according to the specific rules of the Arabic language and taking into account the syllable location in the sentence.
- The syllable is selected and segmented from spoken utterance, taking into consideration different instances of the preceding and succeeding syllables. Each syllable units are classified and labelled according to the Arabic phonetic representations.
- The Arabic syllables of different length, structure and lexical stress is stored in syllable's database with the corresponding speech signals. A large female single-speaker speech database is constructed.
- The Arabic text is also syllabified and labelled according to the Arabic phonetic transcription that represent speech sounds.
- The produced speech is natural enough and the intelligible with some exceptions. The search algorithm impacts unit selection process, when the search algorithm weights the suitability of extracted desired units. The size of the speech database is related to variable length of unit synthetic speech units and to various syllables that proceed and succeed them, which will increase the searching time for the appropriate selected unit in concatenation process.

Acknowledgment

The authors are grateful to the applied science private university, Amman-Jordan, for the full financial support granted to cover the publication fee of this research article.

References

[1] S. Furui, Digital Speech Processing, Synthesis and Recognition. Marcel Dekker, 2001.

[2] C. Demiro ˇ glu* and E. Guner, Hybrid statistical/unit-selection Turkish speech synthesis using suffix units, EURASIP Journal on Audio, Speech, and Music Processing, Vol 2016 Issue 1, Dec. 2016. DOI 10.1186/s13636-016-0082-0. Also available at: https://link.springer.com/article/10.1186/s13636-016-0082-0

[3] J. Pribil, A. Pribilova, J. Matousek, Automatic Text-Independent Artifact Detection, Localization and Classification In Synthetic Speech, Radio-engineering, Vol. 26, No. 4, pp. 1151-1160, Dec. 2017.

[4] M. Dong1, K.-T. Lua, H, Li1, A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese, Journal of Chinese Language and Computing 16 (3): 135-144 , 2006.

[5] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. IEEE Transactions on Speech and Audio Processing, 9(1):39—51, 2001.

[6] A. Amrouche, L. Falek, and H. Teffahi, Design and Implementation of a Diacritic Arabic Text-To-Speech System, The International Arab Journal of Information Technology, Vol. 14, No. 4, July 2017.

[7] M. Z. Rashad, Hazem M. El-Bakry, Islam R. Isma'il, Nikos Mastorakis, An Overview of Text-To-Speech Synthesis Techniques, CIT'10 Proceedings of the 4th international conference on Communications and information technology, pp. 84-89, Greece , July 22 - 25, 2010.

[8] N. K. Bakhsh, S. Alshomrani, I. Khan, A Comparative Study of Arabic Text-to-Speech Synthesis Systems, International Journal of Information Engineering and Electronic Business, No. 4, pp. 27-31, 2014. DOI: 10.5815/ijieeb.2014.04.04

[9] http://sites.middlebury.edu/arabiclingusitics2014/files/2014/02/class6_phonetics_1.pdf

[10] https://www.lebanesearabicinstitute.com/arabic-alphabet/#Arabic_Vowels

[11] M. Bebah, C. Amine, M. Azzeddine, and L. Abdelhak, Hybrid Approaches for Automatic Vowelization of Arabic Texts, International Journal on Natural Language Computing (IJNLC) Vol. 3, No.4, Aug. 2014. DOI: 10.5121/ijnlc.2014.3404 53

[12] M. Maamouri, A. Bies, and S. Kulick, Diacritization; A Challenge To Arabic Treebank Annotation and Parsing, in Proceeding of the British Computer Society Arabic NLP/MT Conference, England, pp. 35-47, 2006. Available at: https://pdfs.semanticscholar.org/de08/84da761ea20e7c081d001bea2f8ef3acc755.pdf.

[13] Z. Elberrichi and K. Abidi, Arabic Text Categorization: a Comparative Study of Different Representation Modes, The International Arab Journal of Information Technology, vol. 9, no. 5, pp. 465-470, 2012.

[14] M. Lee, D. P. Lopresti, and J. P. Olive, A Text-to-Speech Platform for Variable Length Optimal Unit Searching Using Perceptual Cost Functions, International Journal of Speech Technology, Vol. 6, NO. 4, pp. 347-356, 2001.

[15] E. Oancea, A. Badulescu, Stressed Syllable Determination for RomanianWords within Speech Synthesis Applications, INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY, Vol. 5, Issue 3, pp. 237–246, September 2002.

[16] Al-Ani S. H. Arabic Phonology-An Acoustical and Phonological Investigation, Mouton & Co. N. V, Publishers, The Hague, The Netherlands, 1970.

[17] Taylor P. A., Black A., and Caley R. The Architecture of The Festival Speech Synthesis in Third ESCA Workshop in Speech Synthesis, CSTR, PP. 147-151M UKM,UKM 1998, Ghazali, S., Elements of Arabic Phonetics, Applied Arabic Linguistics, Signal and Information Processing, First Fall Session, Morocco, Oct. 1983.

[18] J. Matouˇsek, R. Skarnitzl, D. Tihelka, and P. Machaˇc, Removing Pre-glottalization from Unit-Selection Synthesis: Towards the Linguistic Naturalness of Synthetic Czech Speech, IAENG International Journal of Computer Science, 39:1, pp. 123-130, 2012.

[19] J. Matouˇsek, R. Skarnitzl, D.l Tihelka, and P. Machaˇ, Towards Linguistic Naturalness of Synthetic Speech, Proceedings of the World Congress on Engineering and Computer Science 2011 Vol I

[20] WCECS 2011, October 19-21, 2011, San Francisco, USA. Also available at: http://www.iaeng.org/publication/WCECS2011/WCECS2011_pp561-566.pdf

[21] https://www.slideshare.net/fawz/arabic-syllable-structure-and-stress

Nedhal A. Al-Saiyd She got her B.Sc. degree in Computer Science from University of Mosul-Iraq in 1981, M.Sc. and PhD degrees from University of Technology, Baghdad-Iraq in 1989 and 2000 respectively. She is a Prof. at Computer Science Dept., Faculty of Information Technology, in Applied Science University, Amman, Jordan. She has got more than 26 years of teaching experience. She has published several papers in major international journals and peer-reviewed international conference proceedings. Her research interests include: Software Engineering, Ontology Engineering, Intelligent Systems, User Authentication, Security, Image Processing and Speech Processing.

Mohammed Hijjawi He got his B.Sc. degree in computer Science from Applied Science Private University, Amman-Jordan in 2004, M.Sc. degree in 2006 and PhD from Metropolitan University, UK, in 2011. He is an associate prof at Computer Science Dept., Faculty of Information Technology, in Applied Science University, Amman, Jordan. He is a dean of faculty of Information Technology since 2016. He has published several papers in major international journals and peer-reviewed international conference proceedings. His research interests include: Arabic Machine Translation, Social media computing services,Data mining, Conversational Agents, and Artificial Intelligence.