

ARABTALK®

An Implementation for Arabic Text To Speech System

Yasser Hifny¹ Shady Qurany² Salah Hamid³ Mohsen Rashwan⁴

¹Department of Computer Science, University of Sheffield, UK

²Department of Information Technology, Cairo University, Egypt

³Department of Electrical Engineering, Higher Technological Institute, Egypt

⁴Department of Electronics and Communication Engineering, Cairo University, Egypt

Other participants

Muhammad Atiyya⁵ Ahmad Ragheb⁵ Galaal Khallaaf⁵

⁵Research & Development International, RDI

Abstract

This paper describes the ARABTALK® Text-To-Speech (TTS) synthesis system, developed at RDI¹, for Arabic language. ARABTALK® is a state-of-the-art corpus based concatenative TTS system. The system employs Artificial Neural Networks (ANN) statistical prosody based models for duration, energy, and global pitch contour prediction. In addition, it has a real time synthesis by selection algorithm to explore large speech corpus. ARABTALK® has a hidden Markov models (HMMs) based procedure to automatically time-align new voices transcriptions to their acoustic phoneme boundaries. In this framework, a mature phonology framework has been developed and many perfect rule based models were utilized in the process of letter to sound conversion. The system is multi-user and safe-threaded enabled for server based applications. This research aims to advance the process of developing high quality Arabic TTS synthesis, which yields natural and human sounding Arabic voices.

1. Introduction

Corpus based unit-selection concatenative text to speech paradigms are the state-of-the-art high quality natural TTS systems. ARABTALK® is one of these systems, which is developed specially for Arabic Language. This paper describes the overall architecture, several components of the system, and linguistic concepts for Arabic. Many components of the system are corpus based like statistical prosody models and corpus preparation.

This paper is structured as follows. Section 2 describes the Arabic phonology developed to generate phrases targets and phonological features, which are utilized in the prosody prediction and unit selection process. The approach to prepare the speech-aligned corpus is presented in section 3. The description of the statistical prosody models is presented briefly in section 4. Section 5, describes the two kind of units used for concatenation, monophone and diphone. Section 6, describes the unit selection process. Finally, section 7 summaries our conclusions and an expected future work.

2. A framework for Arabic phonology

ARABTALK® has a mature Arabic phonology framework. Many problems have to be defined and solved in order to achieve automatic letter to sound conversion and provide all the necessary information for other components of the system like phonological to acoustic components mapping (duration, energy, and intonation models) and unit selection process.

Our vision for Arabic language suggested the following tasks to be solved in order to have a reasonable output:

Standard Arabic language has twenty-eight consonants and six vowels. The six vowels are divided into three short vowels and three long vowels. The long vowels have similar spectral properties like their short vowel version with

¹ Research and Development International (RDI) <http://www.rdi-eg.com/>

longer durations than the short vowel version. However, the current system has 41 phonetic letters by adding extra phonemes to consider the effect of the pharyngealized phonemes.

Morphological diacritics are the diacritics of a word characterized by word structure and it is one of the core tasks in order to have automatic letter to sound conversion. Arabic orthography does not consider short vowels within the word structure. RDI has a statistical solution developed by the NLP group to predict possible short vowel patterns for a sequence of words [1].

Syntactic diacritics are the short vowels assigned to the end of each word and they are assigned on the basis of syntactic analysis for the whole phrase. The prosody generation and the unit selection algorithms are affected directly by syntactic diacritics as the actual databases are recorded in a natural way. In order to avoid developing syntactic Arabic analyzer, we suggested and introduced a novel corpus-based approach as a workaround to predict the syntactic diacritics based on HMM Tagging methods. This approach will be developed by NLP group and integrated to the system in the future versions. Currently, we assign a blind default diacritic type for the syntactic diacritics during the automatic letter to sound conversion or they could be supplied manually.

Consonant clusters are eliminated as Arabic has a prosodic nature to remove heavy pronunciation. The consonant clusters are three adjacent consonants, which may result during the physical pronunciation, and are eliminated by inserting a short vowel between the first and second consonant. The type of the short vowel is selected by using simple rule based model.

Phonetic grammar validation is a procedure to ensure that a given phrase could be parsed correctly by the syllabification algorithm where an Arabic syllable must start with only one consonant and the syllabic structure prevents three consonants or two vowels to appear adjacently. This problem usually happens when mixing an Arabic text and non Arabic text (written in Arabic orthography) in one sentence.

Letter to sound conversion for Arabic usually has simple one to one mapping between orthography and phonetic transcription for given correct diacritics. Some simple rule based methods are used to complement the generation of the phonetic transcription.

Syllabification for Arabic language as Arabic has only six syllable types (CV, CVC, CVV, CVVC, CVCC and CVVCC). The last three types usually appear at the end of a phrase only due to their heavy pronunciation. The durations of the consonants and the vowels within these three types are known to be longer than the other remaining types. The number of vowels and the number of syllables in an Arabic phrase must be equal. Hence, any stream of valid Arabic syllables could be accurately parsed according to these rules.

Morphological stress assignment typically is described as predictably falling on a particular location in the word, depending on the internal structure of the syllables making up the word [2]. So, Arabic stress is known directly from the word syllable structure of a word. Arabic stress assignment is different from English language, which uses the stress as a free phoneme. Hence, Arabic stress is a morphological stress and it is not a lexical stress. The stress patterns are derived from an implementation of the stress assignment procedure, which is a combination of the work that has been developed by the phoneticians [3, 4, and 5]. Further enhancements will be integrated to the current model when we develop Part Of Speech (POS) tagger for Arabic.

Currently, the system does not have any automatic procedure to assign different accents degrees to word sequence. The accent degree for a word could be assigned manually or could be ignored during the transcription process. The last word of a phrase has a higher accent degree by default in the current implementation. Moreover, an algorithm that

changes the accent degree for a sequence of words is implemented. The primary objective of this procedure is to assign different accent degrees for function words and content words. Data driven approaches for *prosodic phrasing* and *accent label* predictions will be integrated in next versions as we have suggested and developed the specifications for a new general-purpose text corpus for Arabic "AL-KHALIL" [6]. This corpus will have rich annotation tags for syntactic, prosodic phrasing, and accents. These tags are assigned to guide statistical models to discover some rules about Arabic grammar and Arabic semantics.

Parsing a given utterance results in a prosodic tree, which is constructed to represent the different levels of the phonological description and the relationship between these levels. The output of this linguistic component is utilized by prosody models and unit selection process.

3. Database preparation

The system has two databases one for male speaker at 22 kHz sampling rate (one hour) and the other database is for a female speaker at 16 kHz sampling rate (four hours). The speech is coded into 12 dimensional MFCCs plus log energy and their derivatives. The EGG signal is recorded with each utterance to support pitch synchronous analysis and prosodic modification if necessary during the synthesis process. We use HMMs based Viterbi alignment procedure that is developed at RDI for this purpose [7]. The Viterbi alignment procedure can be summarized as a problem of searching time boundaries for known sequence of HMM models for phonemes. Since the best state sequence, which is known to be the Viterbi path, is obtained during decoding process, time boundaries can be obtained directly. This process is illustrated in figure (1).

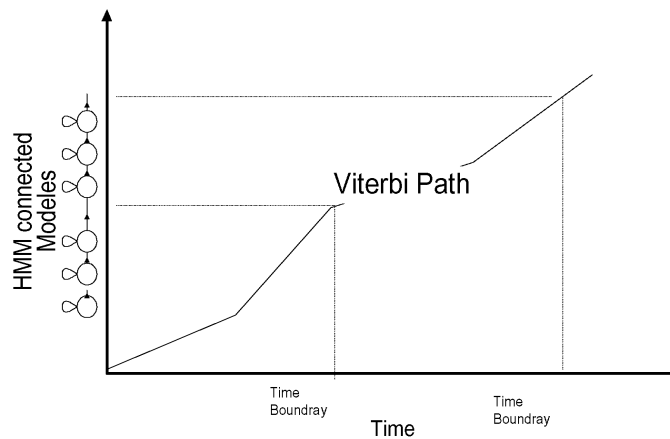


Figure (1) Viterbi Alignment

Actually, Viterbi (forced) alignment procedure results are reasonably well but the labels need to be more accurate for a synthesis database than for recognition. Hence, we did many manual corrections and we have developed many tools to correct and move boundaries for similar error patterns automatically.

4. Prosody modeling

As shown in figure (2), Phonology to prosody modeling is achieved via BP neural networks. The system utilizes three different neural networks to estimate the duration for each unit, the average energy per sample for a unit, and the global intonation contour for each phrase. The authors described the duration model of the system and its prediction

accuracy in details [8]. The global intonation contours used for training were extracted from the speech and each syllable was represented by eight values from the contour. The predicted phrase contour is a smoothing version of the concatenation of the predicted syllable pitch contours. During the unit selection the target costs are weighted scores between the predicted duration/pitch and the extracted values of duration/pitch for a unit. The training and testing procedures are based on the NN simulator that is developed for similar task [9].

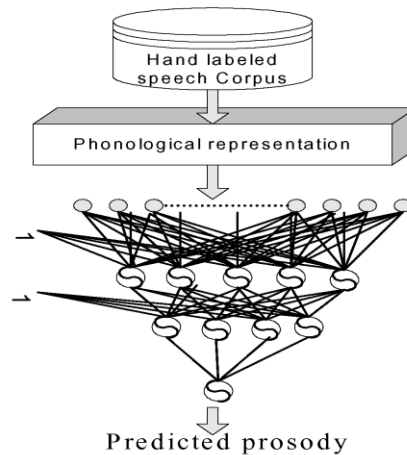


Figure 2: Phonology to prosody mapping

5. Database Unit

Current version of ARABTALK[®] is built so that the concatenation unit can be either monophone or diphone. In monophone case, the system was very sensitive to segmentation errors which degraded the intelligibility of the produced speech. Using human revision for the automatic segmentation -though tedious and time consuming- limited this problem although it wasn't completely solved.

As a solution for the system sensitivity to segmentation problems, the monophones were replaced by diphones as concatenation units. The diphone unit segmentation was made using the automatically segmented monophones and the boundaries of the diphone were taken from half the first phoneme to the half of the second phoneme. Segmentation can be done in two more ways: a) starting and ending at nearest pitch marks. b) Automatic segmentation of diphone units using HMM. This is left for future improvements.

The use of diphone units as concatenation units mainly solved the problem of system sensitivity to segmentation problems and the generated speech was much smoother at the concatenation points, this increased the system intelligibility. Also the direct use of automatically segmented speech was made possible.

For the context clustering of the diphone units, the same tree structure was used as the monophone case. The only difference is that the first part of the diphone was considered as the previous phoneme and the second part was considered the next phoneme.

6. Unit selection

Unit selection algorithms are developed to explore large databases in order to minimize prosodic and spectral modifications for high quality speech synthesis [10]. They aim to select the best sequence of units that match the required targets from a speaker database by Dynamic Programming (DP). The selection process is based on a

combination of target cost and continuity cost. Target costs measure how a unit in the database matches a target unit in the target phrase. The continuity cost is a distortion measure for coupling two neighboring units. In general, the unit selection algorithms are similar for the problem of searching the best state sequence in the HMMs using the Viterbi algorithm. The transition probabilities and observation scoring have the same role of the target and continuity costs in searching the best state sequence.

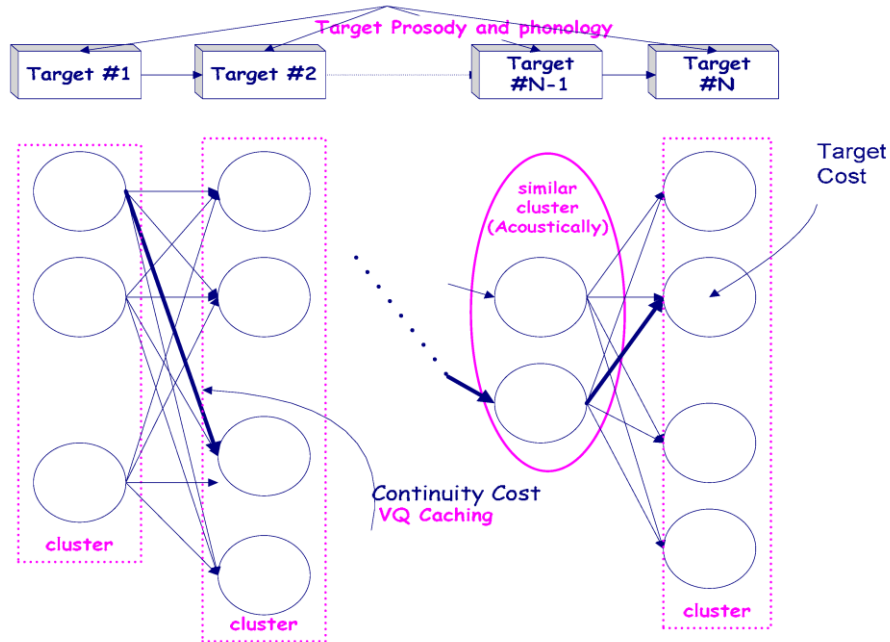


Figure 3: ARABTALK® Unit selection

The search space (the searched lattice) is considered large both in the horizontal and vertical directions in Arabic language because of two different factors. The horizontal direction, which is defined by the number of target units, is relatively large than English language since the phonetic letters per word are approximately doubled after adding the short vowels to each word. The vertical direction, which is defined by the available number of unit candidates, is also large as the actual number of vowels in a database is huge because Arabic language has only six different vowels and every syllable must have a vowel in the target phrase. ARABTALK® implementation of the unit selection process is optimized to achieve a real time performance. As shown in figure (3), the current implementation utilizes *Candidate Caching*, which reduces the search space and *Continuity Caching*, which reduces continuity cost calculations.

Candidate Caching aims to cluster the candidates of the same units [11]. Hence, the online search uses the units of the selected clusters to build the DP lattice. The clustering process is achieved by decision trees and only spectral similarity measure is used while splitting process to evaluate unit similarity. In this work, detailed in table (1), the clustering questions represent different phonological levels and wider context questions are used as they were more efficient to model context dependencies in our databases. The WAGON clustering program, output trees loader was available from EST tools [12].

Continuity Caching is achieved by Vector Quantizing (VQ) the spectral features (MFCC) of the coupling frames (first and the last frames) for each unit. The quantizer is based on Principle Component Analysis (PCA). In the current implementation the number of the centroids is 1024. A distance matrix between the centroids is saved and used during the synthesis process as a look up table to approximate the continuity costs.

During the synthesis process, ARABTALK® target costs imply only weighted prosodic cost between target pitches and durations with respect to candidates' values. The continuity costs are differently weighted at the coupling boundaries for syllables and words. These boundaries are defined during generating the targets for a phrase. The acoustic costs are not considered in the current implementation since the cluster units are very similar acoustically.

Phonology Level	Feature Description	Feature count	Possible range
Phoneme	Sound Type	11	1 to 13
	Voicing Type	11	1 to 5
	Consonant Type	11	1 to 9
	Type Of Articulation	11	1 to 13
	Place Of Articulation	11	0 to 15
	PhonemeID	11	0 to 41
	Fuzzy Emphatic	11	0 to 1
	Emphatic Type	11	0 to 1
	Shadda	11	0 to 1
	Tanween	11	0 to 1
Syllable	Phoneme Position	1	1 to 4
	Count of Phonemes	1	2 to 4
	Accent Degree	1	0 to 4
Foot	Syllable Position	1	1 to 10
	Count of Syllables	1	1 to 10
Phrase	Foot Position	1	0 to 3

Table 1: Clustering Questions Description

7. Summary

The overall architecture and general features of ARABTALK® Text-To-Speech system for Arabic language has been presented. An online demo is available at <http://www.rdi-eg.com/rdi/research/arabtalk.asp>. The system is corpus based system and has many statistical models. The system has real time unit selection with different caching methods. Our current research has many directions to improve the quality of the output speech. For example, different basic units will be investigated as we search better smooth continuity between the selected units. An automatic data reduction procedure, which offers flexibility in the database size, will be integrated in the next version for the handheld applications. Towards better intonation contours, the group will investigate methods based on ToBI labeling methods. Finally, parametric synthesis like H+N methods may be developed in order to have better coupling methods between units.

8. Acknowledgements

The Authors wish to express their thanks to members of RDI speech department for their support in constructing databases. We thank NLP group who made the necessary modifications for their work to match our vision and to be integrated with the current system. We thank Wael Hamza who developed the first speech synthesis system at RDI. We are also would like to thank Alan Black, Wael Hamza, Muhammad Afify and Christof Traber for their fruitful discussions. Many thank for Christof Traber who supplied us his PhD thesis hard copy.

9. References

- [1] Muhammad Atiyya, "A large-scale computational processor of the Arabic morphology", MSc thesis, Cairo University, Egypt, 2000.
- [2] Kenneth de Jong and bushra Adnan, "Stress, duration, and intonation in Arabic word-level prosody", *Journal of phonetic*, Vol. 27,3-22, 1999.
- [3] Ibraheem Anis, "The Sounds of Language", Dar Al Nahda Al `rabia Press, Cairo, Egypt, 1961.
- [4] Tammam Hassan, " Research Methods in Language", Al Risala Press, Cairo, Egypt, 1955.
- [5] Salman h. al-ani, "Arabic Phonology: An Acoustical and Physiological Investigation". The Hague, Netherlands: Mouton and Co., 1970. "Janua Linguarum" series practica 61. Translated into Arabic, 1983.
- [6] Ahmid Raghieb, Yasser Hifny, Mohsen Rashwan, "ALKHALIL, General purpose Arabic text corpus for the applications of Text To Speech synthesis", Technical report, Research and Development International (RDI) company, Cairo, Egypt, 2001.
- [7] Wael Hamza and Mohsen Rashwan, "Concatenative Arabic speech synthesis using large database", *In Proceedings of ICSLP2000*, vol. 2, pages 182-185, Beijing, China. 2000.
- [8] Yasser Hifny, Mohsen Rashwan, "Duration Modeling for Arabic Text to Speech Synthesis", *In Proceedings of ICSLP2002*, pages 1773-1776, Denver, Colorado, USA, 2002.
- [9] Yasser Hifny, "Online Arabic Handwriting Character Recognition ", MSc thesis, Cairo University, Egypt, 2000.
- [10] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database". In *ICASSP-96*, volume 1, pages 373--376, Atlanta, Georgia, 1996.
- [11] Black, A, and Taylor, P. "Automatically clustering similar units for unit selection in speech synthesis", *Eurospeech97*, Rhodes, Greece, 1997.
- [12] Paul Taylor, Richard Caley, Alan W. Black, Simon King, "Edinburgh Speech Tools Library", version 1.2, June 1999.

About the Authors

Yasser Hifny: - received the B.Sc. and M.Sc. degrees in electrical engineering from the Department of Electronics and Communication Engineering, Cairo University, in 1995 and 2001, respectively.

Now he is a Ph.D. student at Department of Computer Science, University of Sheffield, UK, and can be mailed at y.hifny@dcs.shef.ac.uk, or yHifny@hotmail.com.

Shady Qurany: - received the B.Sc. and M.Sc. degrees in Information Technology from Department of Information Technology, Cairo University, in 2000 and 2003, respectively. Now He is a teaching assistant at Department of Information Technology, Cairo University; Shady@RDI-eg.com

Salah Hamid : - received the B.Sc. and M.Sc. degrees in Electrical engineering from Department of computer and automatic control , Ain-Shams University, and Department of system Engineering, Al-Azhar university, in 1991 and 1997, respectively

Now He is a Ph.D. Student at Department of Electronics and Communication Engineering, Cairo University. He is now head of speech technologies team at RDI and can be mailed at Salah@RDI-eg.com.

Mohsen Rashwan: - Received the B.Sc. and M.Sc. degrees from Department of Electronics and Electrical Communications, Faculty of Engineering, Cairo University, 1977 and 1980, respectively. In 1985 he received M.Sc. degree from the Systems and Computer Engineering Dept., Carleton University, Ottawa, Ont., Canada. In 1987 he received Ph.D. in Electrical Engineering, Queen's University, Kingston, Not, Canada.

Now He is a Professor of Electrical Communications, in the dept. of Electronics and Electrical Communications, Faculty of Engineering, Cairo Univ., Gaza, Egypt. He is also a Cofounder and CEO of Research & Development Int'l; RDI, and can be mailed at mRashwan@RDI-eg.com.

Muhammad Atiyya :- received the B.Sc. and M.Sc. degrees in Electrical engineering from Department of Electronics and Communication Engineering, and Department of Computer Engineering , Cairo University, in 1995 and 2000, respectively

Now He is a Ph.D. Student at Department of Electronics and Communication Engineering, Cairo University. He is now head of Engineering Department at RDI, and can be mailed at m Atteya@RDI-eg.com.

Ahmed Ragheb: received the B.Sc. degree in Arabic and Islamic sciences from Faculty of Dar Aleloom, Cairo University, in 2000.

Now he is a M.Sc. student at Faculty of Arts. Ain-Shams University. He is a team leader of linguistic researchers at RDI, and can be mailed at Ragheb@RDI-eg.com.

Galaal Khallaaf: received the B.Sc. degree in Electrical engineering from Department of Electronics and Communication Engineering, Cairo University, in 2001.

Now he is Senior NLP researchers at RDI and can be mailed at Galal@RDI-eg.com.