

TOWARD AN ARABIC TEXT-TO-SPEECH SYSTEM

M. Elshafei Ahmed*

Department of Systems Engineering
King Fahd University of Petroleum & Minerals
Dhahran, Saudi Arabia

الخلاصة :

تحتوى هذه المقالة على وصف لنظام تجريبي لخوارزمي غايته نطق حروف الكتابة العربية باستعمال المقاطع الصوتية . وهذا النظام يستخدم فئة شاملة من الألفونيات العربية لتوليد الكلام طبقاً للغة العربية الحديثة والفصحى على السواء . وتحتوى هذه الفئة من الألفونيات على حوالي (١٥٠) ألفون عريباً مع بعض الثنائيات من الأصوات الصامتة والليننة لتسهيل نطق الكلام بمساعدة الحاسب . كما تصف هذه المقالة عدداً كبيراً من قواعد النطق طبقاً لهذه الفئة من الألفونيات . وتغطي هذه القواعد تنابع الاصوات المصنمته مع الليننة واستعمال الحركات القصيرة والطويلة وتأثير الأصوات المفخمة والحلقية على مخارج الأصوات ، وكذلك قواعد نطق أداة التعريف وقواعد الوقف والتنوين والغنة والإدغام والإقلاب والإخفاء وكذلك قواعد الحاق الكلمات ببعضها .

*Address for correspondence:
KFUPM Box 405
King Fahd University of Petroleum & Minerals
Dhahran 31261
Saudi Arabia

ABSTRACT

This paper describes an experimental system for algorithmic based segmental Arabic text-to-speech system. The system utilizes a comprehensive set of Arabic allophones for speech synthesis in accordance with both Modern Standard Arabic and the classical Arabic language. The proposed allophone set contains about 150 Arabic allophones and vowel/consonant combinations to simplify computer voice production. The paper also describes an extensive set of Arabic letter-to-sound rules from voweled text (with full diacritics) based on this allophone set. This set of rules governs consonant/vowel combinations, usage of short and long vowels, the coarticulation effects of emphatics and pharyngeals, pronunciation of the particularization, rules for word final aspiration, *Tanween* and *Ghonna* rules, rules for conversion, suppression, and assimilation, and rules for combining words.

TOWARD AN ARABIC TEXT-TO-SPEECH SYSTEM

1. INTRODUCTION

The last few years have witnessed an increasing interest in speech processing of the Arabic language [1–12]. However, a great deal of research still needs to be done in order to realize automated text to speech systems with comparable quality to the available commercial and laboratory systems for the English language. The progress in the English text-to-synthetic speech systems has been made possible by advances in linguistic theory, acoustic phonetics, computer characterization of English sound patterns, perceptual psychology, mathematical modeling of speech production, in addition to advances in computer hardware.

There are currently several approaches for Arabic speech synthesis; the direct storage of words and sentences [3, 23], the syllable method [5], sub-syllable method [5, 6], the diphone method [7], and the allophone method [1–3]. These approaches vary in complexity, memory requirements, and speech quality. In this paper we describe an algorithmic method for allophone synthesis of Arabic speech. The main advantages of this approach are the small memory requirement, the small number of the basic sound units required to synthesize an arbitrary Arabic word, and the simplicity of the hardware. However, the use of a limited number of fixed sound units may cause the resulting speech to sound “mechanical”. The direct concatenation of the allophones could cause unnatural abrupt transitions and neglects the coarticulation and interaction between consonants. To overcome this problem, the paper proposes over 150 allophone and consonant/vowel combinations to cover most of the prominent consonant interactions. Model parameter interpolation was also implemented over 2–5 msec to provide a smooth transition between the allophones. At a later stage, heuristic rules [29] will be developed to smooth the transition on a case-by-case bases.

An overview of a text-to-speech system is shown in Figure 1. The presence of the diacritic marks in the Arabic text is essential for the implementation of the automatic text-to-speech system. Unfortunately, most modern written Arabic, as in books and newspapers, is at the best partially vowelized. Hence a preprocessor for automatic vowelization of the text must be implemented before applying the text-to-speech rules. The automatic vowelization generator

requires integration of morphological, syntactical, and semantic information [13–18, 23].

The text with full diacritics is then passed through a second preprocessing step wherein a lexicon is used to replace numerals, abbreviations, special symbols as %, +, -, =, &, \$, *etc.*, and a few exceptions as *يس* *Yassen*, *هذا* *كجم*, *etc.*, by an equivalent phonemic spelling.

Next, the raw text produced from the preprocessing steps is partially parsed into breathing groups. This can be accomplished approximately in many ways. A simple rule that works most of the time as a natural speaker is obtained by locating the breathing group boundary at the following locations [26] whichever is encountered first: (1) at a punctuation mark; (2) preceding a conjunction; (3) before a noun phrase or a verbal phrase; (4) before a prepositional phrase; (5) after a fixed number of characters have appeared in the input.

In the next step, the text is processed by the first level of letter-to-sound rules. The objective here is to produce an abstract phonetic transcription of how the text is actually spoken. At this level the speech is also parsed into syllabic units, for example:

. مَر - رَ - ءَل - عَم - بَر is pronounced مَن رَأَى الْعُنْبَرِ

The next level is the translation of the abstract phonetic transcript to an allophonic transcription, wherein each phoneme sound is replaced by a context dependent acoustic version of the phoneme [26, 29]. In the mean time a companion description of the stress and inflection patterns of speech are generated [23, 29].

Finally, a parametric description of the articulation model, or the synthesis model, is used for generation of the speech sound. The parameters are modulated by the prosodic information as well as other heuristic rules for graded changes in the synthesis model parameters [29]. Several synthesis models are presently known:

1. Formant Synthesizer [29, 30, 32].
2. LPC Synthesizer [1, 2, 31].
3. Channel Synthesizer [30].
4. Articulatory Synthesizer [28, 30].

The synthesis model should allow the synthesized speech to be modified in fundamental frequency,

Figure 1.

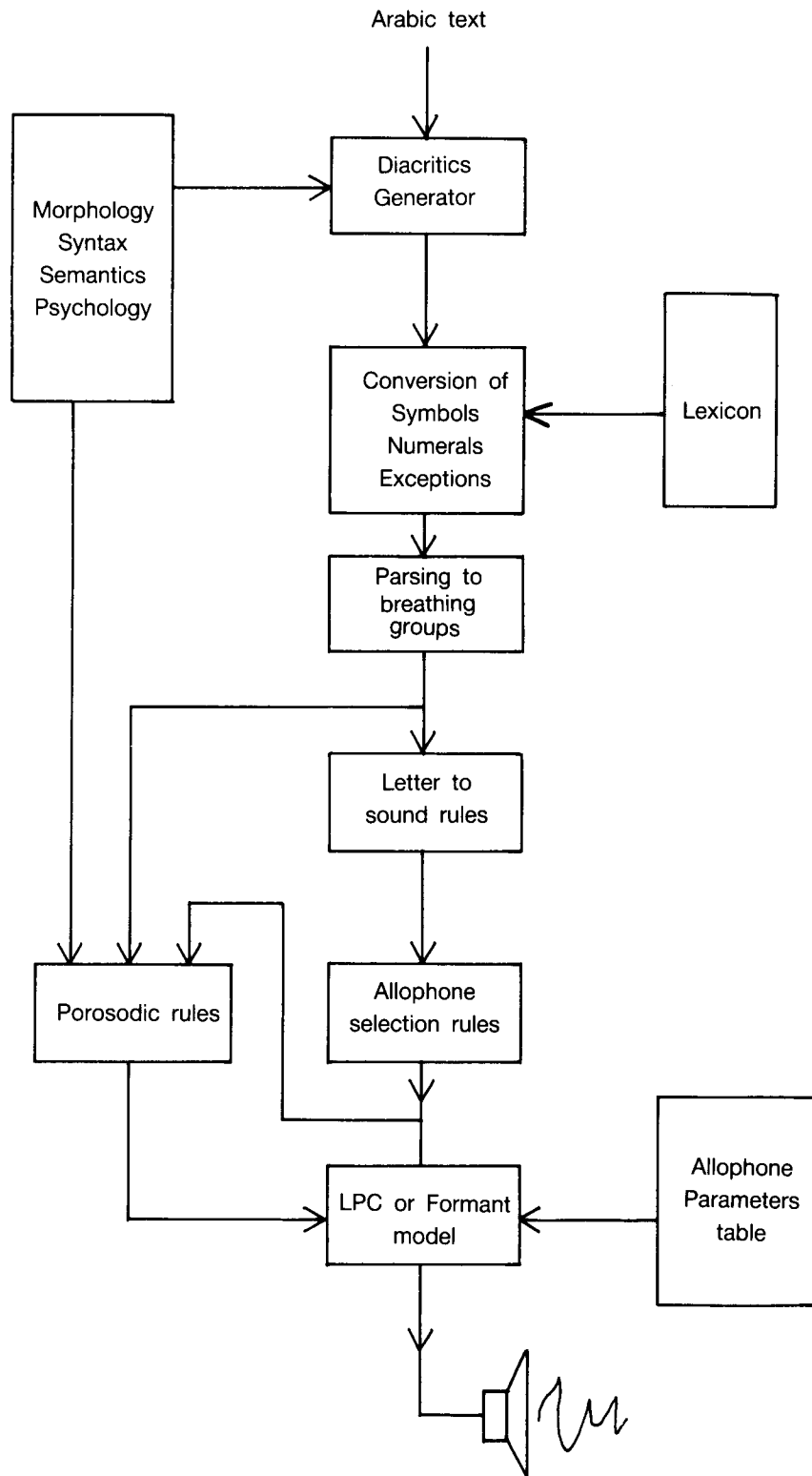


Figure 1.

amplitude, and duration, as well as performing some sort of parameter smoothing at boundaries between waveform pieces.

In this paper we describe an experimental system for rule based segmental text-to-speech system. Specifically, the paper addresses the following issues:

1. proposes an allophone set for speech synthesis;
2. presents a set of rules for conversion of fully diacritized text to an abstract phonematic transcription;
3. provides an extensive set of rules for conversion of the abstract phonematic transcription to allophone strings.

The development work was carried out on a commercial speech editing laboratory. The synthesis model is based on linear prediction coding [1, 31] using 10 filter coefficients and pitch synchronous analysis/synthesis. A detailed description of the model was given in [1]. Although the model was flexible enough to allow changing of duration, stress, and fundamental frequency, the speech quality produced by this method was still unsatisfactory for many applications because of the difficulty in smoothing of the filter coefficients without disturbing the formants trajectories, and because of the limited source excitations model, in addition to the limited bandwidth (4 kHz) which causes difficulty in pronunciation of fricative sounds. A formant model [32] is currently being investigated and recommended for future investigations. Nevertheless, speech synthesis by concatenating strings of Arabic allophones was displayed in [2]. The early version of the text-to-allophone program was also demonstrated separately in [4]. The main advantage of the allophone method is its simplicity and small memory requirements.

In the next section we introduce the proposed allophone set and in Section 3 an extensive set of Arabic pronunciation rules will be presented.

2. ARABIC ALLOPHONES

It is impressive to mention here the first phonetic study of the Arabic language (probably the first published study in phonology of all) of Ibn Sina (370–428 A.H.) [21]. Ibn Sina's treatise explains the mechanism of production of each letter as well as the points of articulation with reference to a fairly accurate anatomy of the vocal tract.

Among the recent studies is [19] which presents a comprehensive introduction to the modern classification of the Arabic sounds, their points of articulation, as well as some of the Arabic allophones. Al Ani [10] investigated the modern dialectal Arabic of Iraq using spectrographic techniques. El-Emam and Dannan [8] studied the allophones of the "Modern Standard Arabic MSA". In [2] about 100 basic allophones of classical Arabic were isolated and used for allophone-based speech synthesis system.

Table 1 shows the Arabic letters and their equivalent International Phonemic Alphabet IPA [3, 9].

The proposed Arabic allophone set contains about 150 allophones and consonant/vowel combinations to simplify computer voice production. Table 2 shows the Arabic set combined with a set of English allophones [25] for comparison, as well as to give the system some ability to pronounce English words, abbreviation, and names within the Arabic text.

The investigated set of Arabic allophones incorporates a number of allophones that appear in MSA [8], the basic allophones that characterize the classical Arabic and Quranic recitation [1, 2], as well as some vowel/consonant clusters to simplify computer speech synthesis.

The proposed allophone set is capable of producing synthetic speech for Quran, classical Arabic, and Modern Arabic with a relatively high degree of naturalness. However, it should be emphasized here that it is not the purpose of this study to define a minimal set of Arabic allophones from the linguistic point of view, but rather to identify a group of speech segments for computer generation of synthetic speech with relatively high naturalness.

The notations used to label the newly introduced allophones follow those used in the literature to describe the English allophones, see for example [26].

The allophones were isolated and digitally encoded and verified by inspection of the formant trajectories, as well as by using analysis-by-synthesis technique.

The Arabic language is characterized by a relatively large number of back consonants (produced in the back of the oral cavity). The back consonants cause a complex coarticulation phenomenon in Arabic speech. The back consonants that characterize the Arabic language are the pharyngealized set shown in Table 1. If we take another careful look at these consonants we can divide them further into the following groups; the emphatic consonants such as *Dhad* /D/, *Sad* /S/, *Tah* /T/, and *Zah* /ð/; the pharyn-

