TECHNIQUES FOR HIGH QUALITY ARABIC SPEECH SYNTHESIS

Husni-Al-Muhtaseb[1], Moustafa Elshafei[1] and Mansour Al-Ghamdi[2]

[1]College of Computer Science and Engineering, King Fahd University of Petroleum and Minerals
Dhahran 31261 – Saudi Arabia
[2]King Fahd Military College, Riyadh, Saudi Arabia

ABSTRACT

The paper proposes a diphone/sub-syllable method for Arabic Text-to-speech systems. The proposed approach exploits the particular syllabic structure of the Arabic words. For good quality, the boundaries of the speech segments are chosen to occur only at the sustained portion of vowels. The speech segments consists of consonants-half vowels, half vowel-consonants, half vowels, middle portion of vowels, and suffix consonants. The minimum set consists of about 310 segments for classical Arabic.

1. INTRODUCTION

High quality speech synthesis from electronic form of text has been a focus of research activities during the last two decades, which lead to an increasing horizon of applications. To mention a few, commercial telephone response systems, natural language computer interface, reading machines for blinds and other aids for handicapped, language learning systems, multimedia applications, talking books and toys, and many others.

Figure 1 shows the functional diagram of the Arabic Text-to-Speech System ATTS. It comprises a Natural Language Processing engine (NLP), capable of producing a phonetic transcription of the text read, together with the desired intonation and stress pattern (prosody), and a Digital Signal Processing module (DSP), which transforms the symbolic information it receives into speech.

The pre-processing module organizes the input sentences into manageable lists of words or breathing groups. It also identifies numbers, abbreviations, acronyms and idiomatics and transforms them into full text when needed.

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 processor 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.

The morphological analysis module proposes all possible part of speech categories for each word taken individually, on the basis of their spelling. Inflected, derived, and compound words are decomposed into their elementery graphemic units (their *morphs*) by simple regular grammars exploiting lexicons of stems and affixes [1].

The NLP engine comprises a morpho-syntactic analyser, underlying the need for some syntactic processing in a high quality Text-To-Speech system. Indeed, being able to reduce a given sentence into something like the sequence of its parts-of-speech, and to further describe it in the form of a syntax tree, which unveils its internal structure, is required for at least two reasons :

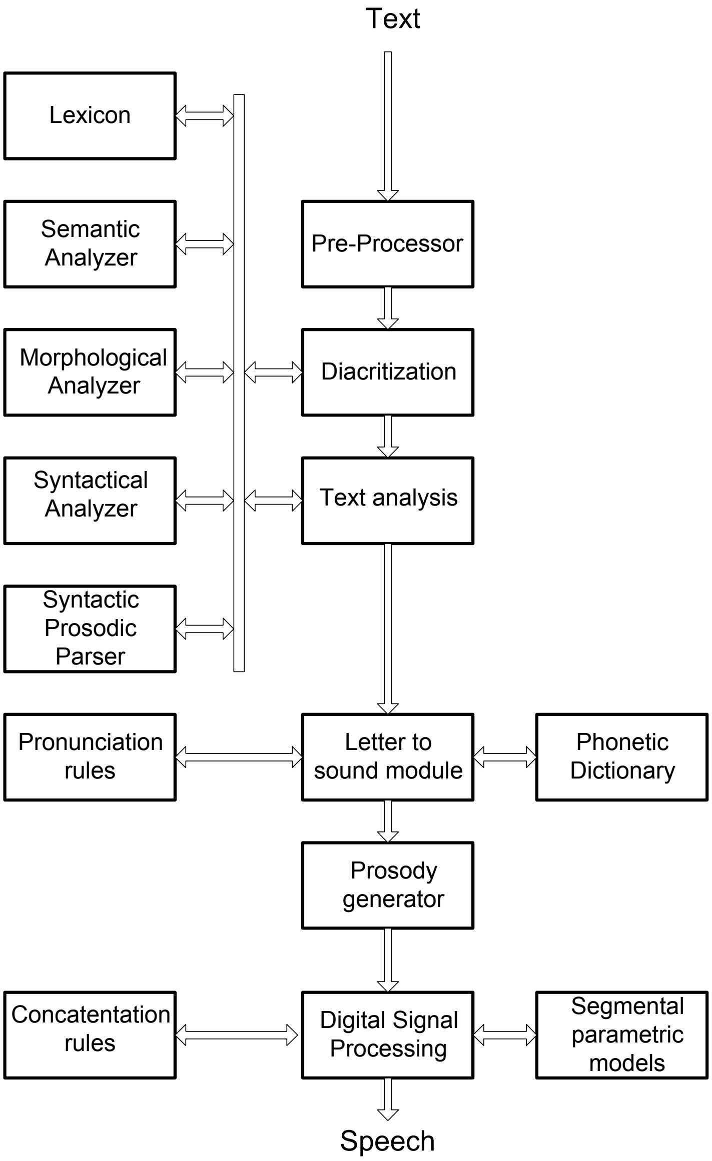

Figure 1. A general functional diagram of a TTS system.

1. Accurate phonetic transcription can only be achieved provided the dependency relationship between successive words is known.
2. Natural prosody heavily relies on syntax. It also obviously has a lot to do with semantics and pragmatics, but since very few data is currently available on the generative aspects of this dependence, TTS systems merely concentrate on syntax. The contextual analysis module considers words in their context, which allows it to reduce the list of their possible part of speech categories to a very restricted number of highly probable hypotheses, given the corresponding possible parts of speech of neighboring words. This can be achieved, for example, by describing local syntactic dependence in the form of probabilistic finite state automata (i.e. as a Markov model), or by *mutli-layer perceptrons* (i.e., neural networks) trained to uncover contextual rewrite rules[2].
3. *et al.* 89], or with *local, non-stochastic grammars* provided by expert linguists or automatically inferred from a training data set with *classification and regression tree* (CART) techniques [Sproat *et al.* 92, Yarowsky 94].

Finally, a syntactic-prosodic parser, which examines the remaining search space and finds the text structure (i.e. its organization into clause and phrase-like constituents) which more closely relates to its expected prosodic realization (see below).

The Letter-To-Sound (LTS) module [3] is responsible for the automatic determination of the phonetic transcription of the incoming text. The task of the LTS module is usually classified into *dictionary-based* and *rule-based* strategies, although many intermediate solutions exist. *Dictionary-based* solutions consist of storing a maximum of phonological knowledge into a lexicon. In order to keep its size reasonably small, entries are generally restricted to morphemes, and the pronunciation of surface forms is accounted for by inflectional, derivational, and compounding morphophonemic rules which describe how the phonetic transcriptions of their morphemic constituents are modified when they are combined into words. Morphemes that cannot be found in the lexicon are transcribed by rule [1,2,4].

A rather different strategy is adopted in *rule-based* transcription systems [2], which transfer most of the phonological competence of dictionaries into a set of letter-to-sound (or *grapheme-to-phoneme*) rules. This time, only those words that are pronounced in such a particular way that they constitute a rule on their own are stored in an exceptions dictionary. A reasonably small exceptions dictionary can account for a large fraction of the words in a running text. In English, for instance, 2000 words typically suffice to cover 70% of the words in text. Clearly, some trade-off is inescapable. Besides, the compromise is language-dependent, given the obvious differences in the reliability of letter-to-sound correspondences for different languages.

2. SEGMENTATION OF SPEECH

Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units. With longer units high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems units used are usually words, syllables, demisyllables, phonemes, diphones, and sometimes even triphones.

Word is perhaps the most natural unit for some messaging systems with very limited vocabulary. Concatenation of words is relative easy to perform and coarticulation effects within a word are captured in the stored units. However, there is a great difference with words spoken in isolation and in continuos sentence which makes the continuous speech to sound very unnatural [1]. Because there are hundreds of thousands of different words and proper names in each language, word is not a suitable unit for any kind of unrestricted TTS system.

The number of different syllables in each language is considerably smaller than the number of words, but the size of unit database is usually still too large for TTS systems. For example, there are about 10,000 syllables in English. Unlike with words, the the inter-syllables coarticulation effect is not included in the stored units. So using syllables as a basic unit is not very reasonable unless various versions of each syllable is also included in the database, which would make the size of the database formidable. There is also no way to control prosodic contours over the sentence. At the moment, no word or syllable based full TTS system exists (English systems).

The current synthesis systems are mostly based on using phonemes, diphones, demisyllables or some kind of combinations of these. Demisyllables represent the initial and final parts of syllables. One advantage of demisyllables is that only about 1,000 of them is needed to construct the 10,000 syllables of English [5]. Using demisyllables, instead of for example phonemes and diphones, requires considerably less concatenation points. Demisyllables also take account of most transitions and then also a large number of coarticulation effects and also covers a large number of allophonic variations due to separation of initial and final consonant clusters. However, the memory requirements are still quite high, but tolerable. Compared to phonemes and diphones, the exact number of demisyllables in a language can not be defined. With purely demisyllable based system, all possible words can not be synthesized properly [6].

Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech. The inventory of basic units is usually between 40 and 50, which is clearly the smallest compared to other units [1]. Each phoneme is represented by a number of templates representing its contextual acoustic variations, called allophones. Using phonemes/allophones gives maximum flexibility with the rule-based systems. However, some phones that do not have a steady-state target position, such as plosives, are difficult to synthesize. The articulation must also be formulated as rules. Phonemes are sometimes used as an input for speech synthesizer to drive for example diphone-based synthesizer.

Diphones (or dyads) are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transitions between adjacent phones. That means that the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation points. Another advantage with diphones is that the coarticulation effect needs no more to be formulated as rules. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all combinations of phonemes are needed. The number of units is usually from 1500 to 2000, which increases the memory requirements and makes the data collection more difficult compared to phonemes. However, the number of data is still tolerable and with other advantages, diphone is a very suitable unit for sample-based text-to-speech synthesis. Longer segmental units, such as triphones or tetraphones, are quite rarely used. Triphones are like diphones, but contains one phoneme between steady-state points (half phoneme - phoneme - half phoneme). In other words, a triphone is a phoneme with a specific left and right context. For English, more than 10,000 units are required.

Table 1 gives a brief summary of the main features of some popular commercial and experimental TTS systems.

	System	**Technique**
1	Festival	diphone/non-uniform unit selectionn
2	AT&T Next Generation TTS	non-uniform unit selection synthesis with half-phones as units
3	CHATR ATR-ITL, Kyoto, Japan (English)	non-uniform unit selection
4	Laureate, British Telecom	diphone
5	Bell Lab TTS System	LPC diphone synthesis
6	YorkTalk	non-segmental (formant) synthesis
7	TTS3000 Lernout & Hauspie	diphone synthesis
8	Pavarobotti, National center for Voice and Speech	Articulatory speech synthesis (speech & singing)
9	Trainable Speech Synthesis	HMM-state sized subword uni, Rob Donovan
10	MBROLA, Le Mons Belgien	diphone synthesis
11	ProVerb, Speech Engine, Elan/CNET	diphone synthesis
10	infovox, Telia Promotor, KTH Stockholm	Formant synthesis
11	Multilingual TTS system, TI Uni Duisburg	Formant synthesis
12	DECTalk	a descendent of MITalk and Dennis Klatt's later work with Klattalk

Table 1. Features of Some popular TTS systems.

There are currently several approaches for Arabic speech synthesis; the allophone method [3], the syllable and sub-syllable method [7], and the diphone method [8]. These approaches vary in complexity, memory requirements, and speech quality. The combined Arabic/English allophone set in [3] contains about 150 allophones and consonant vowel combinations to simplify computer voice production.

3. DIPHONE DATABASE

The basic idea behind building diphone databases is to explicitly list all possible phone-phone transitions in a language. This makes the wrong, but practical, assumption that co-articulatory effects never go over more than two phones. In general the number of diphones in a language is the number of phones squared. In reality, additional sound segments and various allophonic variation may in some cases be also included. The basic idea is to define classes of diphones, for example: vowel-consonant, consonant- vowel, vowel-vowel, and consonant-consonant.

The syllabic structure of Arabic language is exploited here to simplify the required diphones database. The proposed sound segments may be considered as "sub-syllabic" units. For good quality, the diphones boundaries are taken from the middle portion of vowels. Because diphones need to be clearly articulated various techniques have been proposed to extracted them from subjects. One technique uses words within carrier sentences to ensure that the diphones are pronounced with acceptable duration and prosody (i.e. consistent). Ideally, the diphones should come from a middle syllable of nonsense words so it is fully articulated and minimize the articulatory effects at the start and end of the word.

4. PROPOSED METHOD

The Arabic phoneme set consists basically of 28 consonants, 3 short vowels, and three long vowels [9]. Several factors affect the pronunciation of phonemes. For example, the position of the phoneme in the syllable as initial, closing, intervocal, or suffix. The pronunciation of consonants may also be influenced by the interaction (co-articulation) with other phonemes in the same syllable. Among these coarticulation effects are the accentuation (pharyngealization) and the nasalization. Arabic Vowels are affected as well by the adjacent phonemes. Accordingly, each Arabic vowel has at least three allophones, the normal, the accentuated, and the nasalized allophone.

In classic Arabic, we can divide the Arabic consonants into three categories with respect to dilution and accentuation

- ***Always Diluted Consonants (ADC)***: The Arabic letters Hamza, Baa, Taa, Thaa, Jeem, Hhaa, Daal, Thaal, Zay, Seen, Sheen, Ain, Faa, Kaaf, Meem, Noon, Haa, Waaw (as a consonant), and Yaa (as a consonant) are always diluted. These are 19 consonants.
- ***Always Accentuated Consonants (AAC)***: This group consists of the emphatic phones { Saad /S/, Daad /D/, Zhaa /$\underline{\partial}$/, Thhaa /T/}, and the pharyngeal phonemes { Khaa /x/ , Ghain /γ/, and Qaaf /q/}, a total of 7 consonants.
- ***Context Dependent Consonants (CDC)***: The Arabic consonants Raa /r/ and Lam /l/.

A vowel after an Arabic consonant might be diluted or accentuated. The case of a vowel with respect to dilution and accentuation could be specified depending on the predecessor and successor consonants. The following sections elaborate more on these issues.

Arabic language has five syllable patterns: *cV*, *cW*, *cVc*, *cWc* and *cVcc*, where *c* represents a consonant, *V* represents a vowel and *W* represents a long vowel [9].

In the classical Arabic the following rules determine the case of a general vowel with respect to its predecessor/ successor consonants within a syllable:

- In *cV* and *cW* syllables, a vowel after an *ADC* consonant should be diluted.
- A vowel after an *AAC* consonant should be accentuated.
- A vowel after a *CDC* consonant follows context dependent rules as outlined in [Elshafei].

- A general vowel before an *ADC* consonant may be diluted.
- A general vowel before an *AAC* consonant should be accentuated.

Let us define a the *v-elements* to be *c*, $\underline{c}$, *x*, *x'*, *y*, *y'*, *z*, *z'*, *v*, *v'*, *w*, *w'* , where *c* represents a consonant, and $\underline{c}$ an intervocal consonant. *x* is an initial half vowel (HV), *x'* is an initial accentuated (pharyngealized) HV, *y* is a final HV, and *y'* is a final accentuated HV. *z* represents a sustained portion of a normal vowel, *z'* is a sustained portion of an accentuated vowel, *v* is a normal short vowel, *v'* is an accentuated short vowel, *w* is a long vowel, and *w'* is an accentuated long vowel.

Let us also define the *r-phone* set to be { *cx*, *cx'*, *y*, *y'*, *z*, *z'*, *yc*, *y'c*, *y*$\underline{c}$, *y'*$\underline{c}$, *c*, *cv*, *c'v'*, *cw*, *cw'*}. The syllable patterns could be mapped to the following patterns:

cV: *cx-y*, *cx'-y'*, *cv,* or *cv'*
cW: *cx-z-y*, *cx'-z'-y'*, *cw,* or *cw'*
cVc: *cx-yc*, *cx-y'c*, *cx-y'c*, or *cx'-yc*
cWc: *cx-z-yc*, *cx'-z'-y'c*, *cx-z'-y'c*, or *cx'-'z-yc*
cVcc: *cx-y*$\underline{c}$*-c*, *cx'-y'*$\underline{c}$*-c*, *cx-y'*$\underline{c}$*-c*, or *cx'-y*$\underline{c}$*-c*

A description of these sound patterns and their possible number of segments are discussed below.

Syllable initial consonant – HV (*cx*)
There are 21 consonants that would accept diluted vowels after them. 19 consonants are from *ADC* and 2 consonants are from *CDC*. Since, we have 3 different short vowels, this gives us *at most* 21 x 3 = 63 possible combinations.

Syllable initial consonant – Accentuated HV (*cx'*)
In classical Arabic, there are 9 consonants that would accept accentuated vowels after them. 7 consonants are from *AAC* and 2 consonants are from *CDC*. Therefore, there are 27 voice segments in this category.

HV – Syllable closing consonant (*yc*)
At most we have 63 possible combinations of this form.

HV – Intervocalic consonant (*y*$\underline{c}$)
At most we have 63 possible combinations of this form. This is similar to the *yc* case, except that the consonant here is shorter and unstressed.

Accentuated HV – Syllable closing consonant (*y'c*)
This form has *at most* 27 possible combinations.

Accentuated HV – Intervocalic consonant (*y'*$\underline{c}$)
This form has *at most* 27 possible combinations.

Suffix Consonant (*c*)
This is the final consonant in the *cVcc* syllables. The suffix consonants are usually longer than the voweled consonants. The number of such consonants is 28.

Syllable closing HV (*y* , *y'*)
The number of such units is only 6 combinations.

Steady state vowels (*z, z'*)
These are sustained vowel sounds of length double the half vowels. There are 6 of these sustained vowels.

Consonant-short vowels syllables (*cv, cv'*)
There is a total of 28 x 3 = 84 syllables. These are redundant and can be composed from cx-y and cx'-y'. However, having them completely stored could lead to a significant improvement in the sound quality.

Consonant-long vowels syllables (*cw, cw'*)
There is a total of 28 x 3 = 84 syllables. These are also redundant and can be composed from *cx-z-y* and *cx'-z'-y'*. However, having them completely stored could lead to a significant improvement in the sound quality.

5. IMPLEMENTATION ISSUES

1- The minimum set of sound segments consists of the r-phones {*cx*, *cx'*, *y*, *y*', *z, z'*, *yc*, *y'c*, *y<u>c</u>*, *y'<u>c</u>, c*}. The number of these diphones/allphones comes to only 310 segments.

2- About 55% percent of the Arabic syllables consist of the cV and cW syllables. By adding the extra 168 syllables { cv. cv'. cw, and cw'} appreciable improvement in the sound quality can be achieved by avoiding concatenation artifacts.

3- In modern standard Arabic significant coarticulation is noticeable. In particular, the effect of the AAC letters on the accentuation of the whole word can be clearly noticed. The effect is both inter-syllabic and intra-syllabic. At the syllable level, in the cWc, cVc, and cVcc, the entire vowel would be accentuated. To accommodate this effect, the set of sound segments {*cx', y'c* and *y'<u>c</u>* } should be extended to all consonants. The sizes of these sets would become now 84 instead of 27 for the classical Arabic alone.

4- In the cVcc syllables the final consonant cluster is approximated by the xc (or x'c) segment and a suffix consonant. That is, in general, a crude approximation. Some consonants in the consonant clusters can be significantly different from normal allphones. However, considering storing ycc and y'cc r-phones can be very costly. it would require about 2352 segments. Fortunately, the cVcc syllable is a rare syllable. it only occurs less that 1% of time. Any reasonable approximation would not lead to severe degradation of the quality.

5- Concatenative synthesizers however make a fixed decision and they cannot reasonably produce anything outside their pre-defined vocabulary. For example, in addition to lacking of foreign phones as v and p, it would not be able to generate cluster consonants as the word " string" or initial clusters as "cry". This is particularly a problem when it comes to pronouncing foreign names, and terminology. The proposed set needs to be further enhanced with a few English allophones or diphones to be able to accomplish its job.

CONCLUSION

The paper defines a set of Arabic diphones/sub-syllables for concatenative Arabic text-to-Speech synthesis. The minimum set covers the Classical Arabic where the co-articulation is minimal. The paper also proposes extension of the set to improve the quality of the speech and to incorporate the common co-articulation effects found in modern standard Arabic.

ACKNOWLEDGMENT

This work has been supported by KACST under contract number # AT-18-12. The Authors would like to acknowledge also the support of KFUPM.

REFERENCES

[1] J. Allen, S. Hunnicut, and D.Klatt, D. KLATT, *From Text To Speech, The MITTALK System*, Cambridge University Press, 1987.

[2] S. Furui and M.M. Sondhi (edrs), *Advances in Speech Signal Processing*, Marcel Dekker, Inc, 1992.

[3] M. Elshafei Ahmed, “Toward an Arabic Text-To-Speech System”, The Arabian Journal for Science and Engineering, Volume 16, Number 4, October 1991.

[4] S.E. [Levinson, J.P. Olive, J.S. Tschirgi, "Speech Synthesis in Telecommunications", *IEEE Communications Magazine*, November 1993, pp. 46-53.

[5] Donovan R.. *Trainable Speech Synthesis*. PhD. Thesis. Cambridge University Engineering Department, England, 1996.

[6] T. Dutoit , *An Introduction to Text-To-Speech Synthesis¸* **Kluwer Academic Publishers, 1996.**

[7] .M.F. Abu Alyazeed, M.M. R. Al-Ghoneimy, and M.F. Mohammad, Comparison of Syllable and Sub-syllable Methods for Speech Synthesis," Proceedings of the Second Conference on Arabic Computational Linguistics, Kuwait, 1989.

[8] Yousef A. El-Imam, "Unrestricted Vocabulary Arabic Speech Synthesis System," IEEE Trans. Acoustic, Speech and Signal Processing, ASSP-37, No. 12, pp. 1829-1845, 1989.

[9] Ibraheem Anees, Al-Aswat Al-Arabia, Arabic title, Anglo-Egyptian Publisher, Egypt, 1990.