Abstract
A system that uses short phonetic clusters, speech segments, or synthesis units to synthesize standard Arabic (SA) is described. The clusters are derived from the Arabic syllables. Basic and phonetic variants of the synthesis units are defined after qualitative and quantitative analyses of the language phonetics. A speech database of the synthesis units and their phonetic variations is created and the units are tested to control their segmental quality. A computer-based TTS system is developed using the method. Speech is synthesized by waveform concatenation. The intelligibility of synthesized speech is assessed by a standard intelligibility test method that is adapted to suit the Arabic phonetic characteristics.

Introduction
Concatenating units extracted from speech utterances is a popular method for segmental speech synthesis from discrete speech segments. The units are usually sub-words of fixed contexts and durations i.e. units derived from a given phonetic context, and having fixed lengths. Typical units used in the past are diphones, triphones, and demisyllables. For example, the polyphone approach (Bigorne, et al 1991) used on the multilingual PSOLA. Small sized units of speech are prone to a great deal of variations in phonetic quality depending on the context in which they appear. It is often important to conduct an objective study of the contextual phonetic variations of the units and enhance the basic units with some phonetic variants and perform waveform processing to synthesize speech of good quality.
Arabic has been synthesized using different types of fixed-length discrete synthesis units (El-Imam, 1990). The present method is yet an addition to the pool of synthesis algorithms for Arabic speech. It differs from the previous method in that it requires less number of synthesis units. The basic synthesis units are enhanced by few allophonic variants after analysis of the phonetic of Arabic. The allophones of the synthesis units are shown to be viable phonetic variants by perceptual and quantitative analysis. After segmentation, the speech database is tested and modified to improve its quality and that of the synthesized speech. Intelligibility tests were conducted on synthesized speech.

1-The Arabic sounds and their phonetic variations
In Arabic there are 28 consonant, 6 vowels, and two diphthongs. The consonants are (/f/, /bl/, /t/, /l/, /kl/, /fl/, /bl/, /fl/, /gl/, /d/, /k/, /l/, /m/, /n/, /h/, /w/, and /j/). The Arabic equivalent of the above symbols are respectively, ( , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , , ). The vowels are three short (/a/, /a/, and /a/) and three long, which are the counterparts of the short (/a/, /a/, and /a/). The diphthongs are (/ai/ and /aw/); formed when the glides /i/ or /u/ are preceded by short vowel /a/. The articulation of these basic sounds was presented in El-Imam, 1990. There are numerous contextual phonetic variations that can occur to basic Arabic sounds. Among the most prominent phonetic variations are those that are caused by the influences of pharyngealization, nasalization, anticipatory coarticulation, aspiration, and segment duration changes. The most important phonetic variations of Arabic sounds are caused by pharyngealization. Pharyngealization is both intrasyllabic and intersyllable phenomenon. It affects all Arabic phonemes but its effects are most prominent on the vowels (short and long), the two diphthongs (/ai/ and /aw/), the sonorant /l/, the trill /r/, and the counterparts (/l/, /l/, /l/, /l/, and /l/) of the Arabic emphatics (/t/, /d/, /l/, and /z/). The counterparts of the emphatics assimilate to the corresponding emphatic whenever they occur in a context that causes them to become pharyngealized. A sound that is likely to become pharyngealized will do so when it is in the same syllable or next to an emphatic or another sound that is heavily pharyngealized. Phonetic variations caused by nasalization affect vowels and diphthongs. All Arabic vowels and the two diphthongs are nasalized when they are followed by a nasal sound (/mv/ or /nv/). For diphthongs nasalization is an intersyllable process (because diphthongs always occur as syllable closing sounds). But for vowels, nasalization can be intersyllabic or intrasyllabic process. Phonetic variations caused by anticipatory coarticulation affect the voiceless stops /t/ and /k/. Either the regular places of articulation of these sounds change or their articulation overlaps with another neighboring vowel sound. The /t/ and /k/ can be coarticulated in anticipation of the following front or back vowel. In Arabic there is aspiration, or the production of an /h/ like sound, which is encountered following the release of the voiceless stops /t/ and /k/. The phoneme /t/ is aspirated when word final and /k/ is aspirated when followed by any vowel. Sound duration changes affect the Arabic consonants. A consonant can occur in initial, intervocalic, or syllable closing position. In Arabic, an initial or intervocalic consonant is shorter than the same consonant when it occurs as a syllable-closing consonant.

2-The basic synthesis units
Three types of synthesis units are defined, which are a consonant vowel cluster, CV, a vowel-consonant cluster, VC, and a stable consonant, C. Arabic syllables are very regular and are characterized by two important facts; the nucleus of every Arabic syllable is a vowel and the juncture between two closed or a closed followed by an open or closed syllables is always a point of minimal acoustic activity. Because of these two facts, it is possible to acoustically break the syllables into such clusters and to use them to synthesize Arabic speech. Besides, syllables are known to be basic units to carry information regarding the prosody of the speech. The synthesis process is following; a syllable of type CV or CV: is synthesized from a CV unit. A syllable of type CVC or CV:C is synthesized from a CV plus a VC unit. The uncommon syllables of type CVCC or CV:CC are synthesized from a CV, a VC and a lone C synthesis unit. The difference between this synthesis method and diphone synthesis is in the treatment of words of composite syllabic structure (words that are formed of closed syllables or a mixture of closed followed by closed or open syllables). The regularity of Arabic syllables and minimal acoustic activities across the juncture points between constituent syllables of such words
made it is possible to tag a VC unit to a CV unit at the syllable boundary. Minimal waveform smoothing is carried out across the juncture to remove any concatenation artifacts (amplitude and timbre mismatches), if such artifacts exit. Because there are 28 consonants and 6 vowels, there are 336 synthesis units of type CV and VC (168 units of each type) and 28 units of type C giving rise to a total of 364 basic synthesis units. In the developed TTS system, it was found that the 364 synthesis units are enough to synthesize intelligible Arabic speech of good quality. The merits of using this method to synthesize Arabic speech are generally those that apply to good choice of synthesis segments. 1) The synthesis segments account for as many coarticulation effects as possible. 2) They are easily concatenated. 3) They are short enough to reduce the size of the synthesis units’ database and long enough to reduce the density of concatenation points. In comparison to the partial syllable approach that was used previously to synthesize Arabic (El-Imam, 1990), the present method uses less than half the number of synthesis units (including 244 extra synthesis units phones that are used to improve the quality, next section). If only the basic synthesis units are used, the present method uses approximately one quarter of the synthesis units used previously to synthesize Arabic speech. Both methods are syllabic and therefore can easily be extended to produce naturally sounding speech. Because the present method is based on clustering of vowels and consonants at their steady-state points, it automatically takes care of transitional coarticulations between these sounds. Other forms of intersyllabic and intrasyllabic phonetic variations in Arabic speech will require further allophonic studies of the language. This will be dealt with in the next section.

3-Phonetic variants of the synthesis units

Pharyngealization of the vowels and diphthongs leads to six vowel \{a’, a”, u’, u”, i’, and i”\} and two diphthongs \{a:j/ and aw:/\} allophones. The two sonorants /l/ and /l/ can become pharyngealized leading to two pharyngealized sonorant allophones \{l’/ and r’/\}. The effects of these allophones on the basic synthesis units are easy to deduce. Whenever a vowel is pharyngealized, the synthesis unit(s) containing that vowel will also be pharyngealized. When we exclude the Arabic emphatics and their assimilated counterparts, there will be 20 x 6 = 120 pharyngealized CV units and 6 x 20 = 120 pharyngealized VC units. Examples of pharyngealized CV and VC units are found in the word “/fa:DDa/’ [fa: DDa’ /l] ‘he prefers’. The word is mode of two closed syllables of type CVC. In this words the emphatic /D/ is geminante. The presence of the first /D/ in the same syllable with /a:/ leads to the pharyngealization of the vowel /a:/ which, results in the pharyngealized CV unit /a:/ /l. Likewise the presence of the second emphatic /D/ in the second syllable with /a:/ causes the pharyngealization of the vowel /a:/ which, results in the pharyngealized VC unit /a:/’ /l. The /l/ can be exceptionally pharyngealized (without a presence of an emphatic or another heavily pharyngealized sound as in the name of ALLAH (God’s name)). This causes two extra synthesis unit of type CV /a:/ and VC /a:/’ /l. In conclusion, the influences of pharyngealization prompted the introduction of additional 240 units of type CV and VC.

The other forms of phonetic variations of Arabic are, implicitly, taken care of by the basic synthesis units themselves. For example the anticipatory coarticulation of /t/ with the long vowel /a:/ as in the CVVC word “/h/ /u:/t/ (cherries) is taken care of by the CV unit /h/ /u/. This unit is a cluster of the /h/ the transition from the /h/ to the long vowel /a/, and a stable portion of the vowel /a/. It is not expected that anticipatory coarticulation of the /t/ on the vowel /a/ will last longer than the onset of the vowel. Likewise nasalization of the vowels is an intrasyllabic phenomenon and is inherently taken care of by the VC units themselves. For example, the nasalization of the long vowel /a:/ as in the CVVC word “/n/ /a:/” /kaw/ (has been) is taken care of by the VC unit /n/ /a:. This unit is a cluster of a stable part of the vowel /a:/, a transition from that vowel to the /n/, and the nasal /n/. Because diphthongs are syllable-closing sounds, their nasalization requires the introduction of two extra synthesis units. For example, in the CVVCVC word “ /zajnun/ [zaj’ nu:n] (very good) the diphthong /a:j/ falls in the first syllable but is nasalized to [aj’] due to the influence of the nasal /n/ in the second syllable.

Sound duration changes of initial, intervocalic, and syllable-closing consonants are inherently taken care of by the CV, VC, and the lone C synthesis units. The consonants are always parts of the synthesis unit clusters no matter where they occur in words. The above analyses show that the total number of extra phonetic variants of the synthesis units required is 242. 120 of these units are of type CV to take care of some forms of pharyngealization, anticipatory coarticulations, and some forms of sound duration changes. 122 are of type VC and they are needed to take care of other forms of pharyngealization, other forms of sound duration changes, and nasalization.

With the exception of the sound duration changes, use is made of neural methods to discover and verify the phonetic variations of the Arabic sounds of interest to the present synthesis method. A 3-layer multilayer perceptron, MLP is trained to recognize the phonetic differences between minimal pairs. For example, to distinguish between the pharyngealized and the non-pharyngealized /a:/ in the minimal pair “/t/” /sa:r/ (he walked) and “/t/” /sa;/ (he happened) /Sa:/ [Sa:’ r ’], numerous recordings of pharyngealized [a:] and non-pharyngealized /a:/ were used to train the MLP. The MLP is then used to recognize the difference between a pharyngealized and a non-pharyngealized sound by presenting it with test segments. Algorithms are developed and implemented by software to extract the synthesis units and their allophones from input text. The algorithms work on any input text to convert it to a string of phonemes and then to allophones and the respective synthesis units derived from the phonetic string. A synthesized utterance is obtained by stringing these synthesis units together after the synthesis units’ junctures are smoothed using a concatenative synthesis approach. Besides, the regular Arabic grapheme-to-phoneme rules, some additional text processing rules were defined to derive the allophonic variants of the Arabic phonemes on the basis of their phonetic contexts and to extract the syllables and the synthesis units.

4-Testing and Quality control of the synthesis’ units database
When segmenting the synthesis units and extracting them from carrier words, phrases, or sentences, errors and misjudgments can occur no matter how the segmentation process is automated. For this reason quality control of the synthesis units’ inventory is important Van Santen, (1993). Preliminary perceptual analysis of the present synthesis method using a limited test text that include samples of the synthesis units in different contexts, showed that there are certain concatenation problems that can be attributed to bad segmentation. These problems are not of the same caliber as the other concatenation artifacts (timbre mismatches and variations in amplitude distortions) that are usually treated by adequate smoothing methods on a parametric synthesis model.

Concatenation problems due to bad segmentation occur if the lengths of both or either the CV and VC units on opposite sides of the juncture are not adequate. For example the same CV unit can occur at the end of the word or adjoining a VC unit. Its length has to be optimally chosen to satisfy both situations. As an example, consider the CVCCV word “/katab/ (a child’s name for father), the word is synthesized from the following unit: CVVCVCV. The two CV units are the same (/ba/), the length of the first /ba/ is affected by the adjoining VC unit (/a:b/) but the last /ba/ unit is next to silence. Although the contexts of the same CV unit are different its length has to be optimally chosen to satisfy both situations.

Duration problems due to bad segmentation can alternatively manifest themselves as a prolongation of a medial consonant (the consonant appears like geminate. In reality it is not). For example, the CVCVCV word “/ba ta b/ (he wrote). The word is synthesized as CV+VC+CV+VC. The C in the underlined synthesis units is the consonant /t/ and if the lengths of the synthesis units are not appropriate, the /t/ will appear as geminate leading to a different word meaning.

To test for the quality of the synthesis units’ inventory against the segmentation problems described above; a minimal test text is formed. The text is made of words that contain all the medial consonants (all VC contexts) and all the closed syllable followed by closed or open syllable contexts (CVCCVC or CVCCVCV etc.). In total 1684 words are formed (1008 contain VCVs and 676 contain CC clusters). Although not essential, the words are placed into meaningful phrases to form a minimal test text. It is not essential to place the test words into meaningful sentences because the words are polysyllabic and meaningful. A native Arabic speaker can easily comprehend them.

The text-to-allophone component of the TTS system was used to derive the constituent syllables and synthesis units of each word in the test text. Every word in the test text is also synthesized by the TTS system. We selected two groups of 8 listeners each and trained them to hear the natural words in the test text (some words were played for the listeners, syllable by syllable to improve their discriminative judgements). The synthesized test words are then played for the trained listeners (again in whole and sometimes syllable by syllable). The listeners are asked to spot and record any word that they think is not good enough and to distinguish which syllables and, whenever possible, which synthesis unit(s) in the word is/are causing problems. Listeners were informed to pay attention the segmentation problems discussed above. Using this scheme and after analysis of the listeners’ responses, it was found that: 84 synthesis units (or about 14%) out of the 606 synthesis units (basic and phonetic variants), were in error due to bad segmentation. Misconceptions due to segmentation faults that are attributed to durations of the synthesis units account to the majority of the segmentation faults or around 70%. Misconceptions due to germination, sound confusion, and lack of emphasis account to the remaining 30% of the segmentation faults.

5-Intelligibility of synthesized speech
The previous quality control tests carried out on the synthesis units’ inventory measured a variety of segmental flaws but those tests are not specifically suited for testing segmental intelligibility. To test the segmental intelligibility of the synthesized speech using the present method, we have carried out a diagnostic rhyme test (DRT) (Voiers, 1983). The same distinctive features (voicing, nasality, suspension, sibilation, graveness, and compactness) used by Voiers were used. However, the test is enhanced to suit the phonetic characteristics of Arabic. The phenomenon of pharyngealization is somehow unique to Arabic and its influences extend beyond the syllable boundary. The quality of synthesized Arabic has to be assessed against this phenomenon. Other reasons for modifying the test are that: The syllables of Arabic, from which the present synthesis units were derived, are different from English or French. The vowels of Arabic are susceptible to great deal of phonetic variability, especially with respect to pharyngealization and nasalization. The consonants of Arabic vary in context. It is not enough to test for initial consonants; we must consider all the contexts in which a consonant appears (initial, medial, and final). In fact it was due to the phonetic variability of Arabic speech that we were motivated to modify and enhance the present synthesis method to take into account such variability. It is the author’s own belief that any other segmental synthesis method for Arabic speech has to be uniquely adapted by enhancing it with relevant phonetic variants.

To test the segmental intelligibility of the synthesized speech produced by the current synthesis method, a word list containing multi-syllable words is formed. The list covers the important domains discussed above.

The Arabic word list is similar to Voiers DRT in that the words used are minimal pairs. I.e., they differ in only one phoneme (the phoneme under test). Second, the minimal pairs were selected such that their phoneme differences cover a particular set of phonetic features. The phonetic features tested are the same as those used by Voirs, 1983 and somehow enhanced by Van Santen, 1993 but we added the emphasis feature which is the root of pharyngealization in Arabic. We have classified the Arabic phonemes according to each feature and its counterpart. Attempts were made to include as many SA vowel environments as possible and to include each feature tested as many times as possible in each vowel environment. The DRT is a two-choice test. Each test item is a pair of rhyming words. The pair has consonants that differ in a single distinctive feature. The listener task is to judge which of the two words the synthesis system has uttered. This indicated that whether or not the listener has apprehended the tested distinctive feature. We have selected ten listeners. All the listeners are adult males with no hearing disorders. The test was administered as a pencil-and-paper choice. Five word-pairs appear at a time but each pair is individually spoken by the synthesis system. The listener’s task is to mark the word he has heard from each pair.
DRT can be scored in several ways depending on the interest of the investigator. For the purpose of the present speech synthesis system, we have used the seven major diagnostic scores (one corresponding to each distinctive feature tested). An overall score is obtained by averaging the seven major diagnostic scores.

Two versions of the present synthesis method (one version uses an inventory of the basic synthesis units and the other version uses the inventory of the basic synthesis units enhanced by allophonic variants, section 3, are compared to each other. For each speech synthesis version, attempts were made to compare synthesis by waveform concatenation to two other speech generation methods, which are LPC coded version of the waveform concatenation and naturally spoken test words.

For the three speech generation methods, the results show that nasality feature is highly distinctive with average scores of 95% (synthesis enhanced by allophonic variants) and 92% (synthesis using the basic synthesis units) across the three speech generation methods. The graveness is the hardest to differentiate with an average scores of 87% (synthesis enhanced by allophonic variants) and 81% (synthesis using the basic synthesis units) across the three methods. The other features fall in between the above two extremes. The feature average for the direct waveform concatenation using the present synthesis method when the basic synthesis units are enhanced by allophonic variants is around 91%. The feature average for the direct waveform concatenation using the present speech synthesis method and the basic synthesis units is around 85%. The results show that there is a reasonable improvement to the quality of synthesized speech to justify the effort spent on the phonetic variations of the synthesis units and the quality testing of the synthesis units’ inventory.

Listeners were then trained on machine voice by exposing them to scripts of synthesized speech, using the current synthesis method and direct waveform concatenation. The training text adopted is not broad enough to cover all phonemes’ contexts, intelligibility rose by about 3% on all features tested. Besides training people on machine voice, intelligibility of synthesized speech is better when the words are significant parts of phrases and sentences. Sentence-level intelligibility test were attempted, in the manner described by Van Santen (1993). It was found that when the same words used in the present test are inserted into meaningful sentences, the average intelligibility scores rose by approximately 2%.

6-Conclusions
A speech synthesis method for Arabic is presented. The method uses discrete fixed length segments of speech, which are stable demisyllabic clusters. To improve the quality of synthesized speech, phonetic variations of the synthesis units are studied and the most important context-sensitive allophones of the clusters are defined. The inventory of synthesis units is tested for its quality and modified on the basis of the quality test results. The intelligibility test revealed that the introduction of context-sensitive allophones of the cluster improved the quality of the resulting artificial speech.

Arabic has, previously, been synthesized by another method. The present synthesis method uses less number of synthesis units than the previous method. To demonstrate the suitability of the method for the synthesis of SA, a personal computer-based version of the synthesis method and the text-to-speech conversion was implemented. The quality of the synthesized speech is highly intelligible. The article focused on improving the quality of segmental synthesis of Arabic. Producing naturally sounding synthesized Arabic is an open research area.

References