Speech Synthesis System for the Holy Quran Recitation

Nadjla Bettayeb and Mhania Guerti
Department of Electronics, Signal and Communications Laboratory Ecole Nationale Polytechnique, Algeria

Abstract: *This paper aims to develop a Text-To-Speech (TTS) synthesis system for the holly Quran recitation, to properly helps reciters and facilitates its use. In this work, the unit selection method is adopted and improved to reach a good speech quality. The proposed approach consists mainly of two steps. In the first one, an Expert System (ES) module is integrated by employing Arabic, Quran language, phonetic and phonological features. This part was considered as a preselection to optimize the synthesis algorithm's speed. The second step is the final selection of units by minimizing a concatenation cost function and a forward-backward dynamic programming search. The system is evaluated by native and non-native Arabic speakers. The results show that the goal of a correct Quran recitation by respecting its reading rules was reached, with 97 % of speech intelligibility and 72.13% of naturalness.*

Keywords: *Speech synthesis, holy Quran, unit selection, expert system, Arabic language processing, tajweed rules.*

Received March 21, 2019; accepted April 10, 2020
https://doi.org/10.34028/iajit/18/1/2

1. Introduction

Text-To-Speech (TTS) synthesis is the process of artificial conversion of a digital text into speech. It is integrated as a module in many human-machine interaction systems, such as alarms, talking devices, or assistance systems for handicapped persons. Nowadays, digital Quran is a term adopted in several applications mostly when it comes to its reading, teaching its recitation rules, its security, or authentication [9, 20]. The dedicated systems for Quran recitation consist mainly of commercial devices, like Quran reading pen, mobiles, and web software [19]. Generally, these applications based on recording and storing all the Quran surahs, chapters, and Verses, and limit the user to listen by word or by a complete verse. Hence, our main goal is to overcome these limitations using speech synthesis technology. The Development of a Holy Quran Text-To-Speech (HQ_TTS) synthesis system may allow reducing the occupied memory space and gives users the freedom to choose the part they want to listen to.

In speech synthesis, Unit Selection (US) is one of the adopted methods for its high speech quality in which it, generally, needs only basic speech processing. It is based on the concatenation of natural sound segments, called units, after they were selected from a large database. The performance of this method depends then, on the database richness and the selection algorithm efficiency [8, 15]. In that context, our contribution consists of enhancing the unit selection process by employing the Arabic language and Quran phonetic and phonological characteristics as tuning parameters. This is because of their high influence on the units acoustic and prosodic features [1, 4, 13]. To the best use of these characteristics in the selection process, a rule-based Expert System (ES) is developed. This ES is integrated into a contextual preselection step. After that, a forward-backward dynamic programming search is applied for the best selection of units to concatenate.

2. Related Works

To the best of our knowledge, no full speech synthesis research addressed the Holy Quran. However, HQ_TTS is considered as an Arabic TTS system, because the Quran is written and read in Arabic. The only difference is that its recitation requires other reading rules, in addition to, the ordinary ones in Arabic, called the Tajweed rules. With these additional rules, new phonemes appear, like the vowels with double and triple duration as the usual ones (the Madd). Besides, new phonetic and phonological phenomena may present (e.g., the emphasis, the assimilation, etc.,).

TTS systems consist mainly of two parts: Natural Language Processing (NLP) and Digital Signal Processing (DSP). In NLP, the text is transcribed into phonetic writing, while DSP aims to generate speech from this phonetic representation. Arabic TTS systems have received considerable attention over the past two decades and research is still ongoing to improve them. Most of these works have been done in the NLP module that is considered a challenge in Arabic speech synthesis [4, 10]. This problem is caused by the lack of diacritical marks in the general used texts, in which they represent the vowels of this language. Despite the

natural speech quality that can give the Unit Selection Speech Synthesis (USSS) compared to the other methods. The research in the DSP module seems to focus on the Statistical Parametric Speech Synthesis (SPSS) [3, 12]. The lack of a standard and available Arabic database for speech synthesis may be a reason because it drives the researchers to make their ones like in [5, 18] which is not an easy task.

In other languages, like English, USSS has undergone significant progress since its development, twenty years ago, particularly in terms of selection algorithms. According to [2, 15], the unit features or their corresponding weights, involved in the selection process are some of the crucial factors for natural speech quality. in those works, several objectives, subjective or hybrid techniques were applied to tune the used unit features weights (e.g., Weight Space Search (WSS) and active interactive Genetic Algorithms (aiGAs)) [2]. Other studies rely on exploiting more unit features by the use of some machine learning techniques like Progressive neural networks [11].

Our work in this paper consists of developing a USSS system for the Quran recitation, named HQ_TTS. In this study we propose to employ the language phonetic and phonological features in the selection process, hence improving the speech quality. HQ_TTS is a software program developed using MATLAB. The block diagram of the HQ_TTS system is presented in Figure 1, while its main modules are detailed in the following sections.

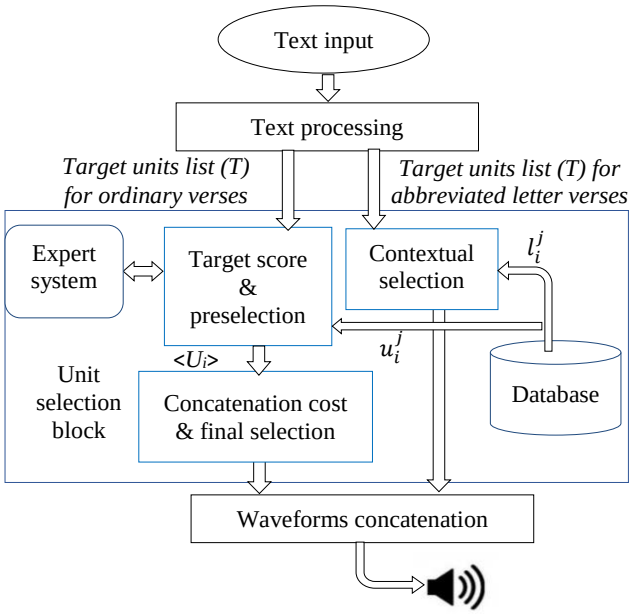


Figure 1. Block diagram of HQ_TTS system.

3. Text Processing

The first HQ_TTS module starts with a text analysis step, that consists of defining the number and the type of inserted verses. So, after converting the text into numeric code (using MATLAB), the phrases to be synthesized are distinguished by the character "," (a special character used for HQ_TTS system). Subsequently, the type of Quranic text is determined by defining "the Abbreviated letters" type that has a special reading way. These verses consist of a combination of one to five characters from 14 Arabic letters (just consonants) and appear at the beginning of 29 surahs, e.g.,: "طسم", "ألر ", etc., They are recited by concatenating the pronunciation of each composing letter out of context, and they have no linguistic meaning, e.g.,: "ألر " => [ʔalif laam raaʔ].

After this analysis step, the text is transcribed into phonetic writing with the rule-based method and depending on the verse's type [7]. This transcription method is also well known in translation systems like in [14]. The phonetic sequence is then analyzed and all its composing units are defined. Finally, a list of target units *T* is formed by coding and adding the contextual features of each unit as presented in Figures 2 and 3. The first two digits represent the word position in the sentence and the unit position in word respectively (beginning, in the middle or ending position). The remaining numbers are codes for the left and right phonemes respectively as detailed in [7].

" قُلْ أَعُوذُ بِرَبِّ ٱلفَلَق " <= *text input*

⇩ *Transcription*

[#qul !a3uu4ubirabbilfalaq_q#]

Defining the target units ⇩

{ #q ; #qu ; ul ; l ! ; … ; ira ; ... ; q_q# }

⇩ *Adding the contextual features codes*

T={ #q110031 ; … ; q_q#332900 }

Figure 2. Text transcription and target units list formation steps of surah: Elfalaq, verse: 1.

"ألم"

Transcription and defining the target units ⇩

{ ! ; l ; m}

⇩ *Adding the contextual features codes*

T={ !10023 ; l20124 ; m32300 }

Figure 3. Text transcription and target units list formation steps of surah : El-Baqara, verse :1.

4. Unit Selection

The unit selection block is the main part of the HQ_TTS system. It consists of searching in the database for the best units that suit the target list defined above. The database of HQ_TTS system contains 11070 different size sound units (diphones[1] and polyphones[2]) [7]. It was built from a real Quran recitation [16] by taking 3.85 % of the whole

[1]Sound unit that extends from the stable part of the first phoneme to the stable part of the second one, including the transition between them.

[2]A combination of three or more phonemes.

recordings. In this system, each verse type has its own selection process as follows:

4.1. Ordinary Verses Type

In this verse type, the selection algorithm starts with a search in the database for units that compose with the same phonemes as target ones. After that, they are structured as candidate units lists, and the best ones will be selected from these lists.

Contrary to previous works in USSS [2, 8, 15, 17], the selection process in this study is divided into two parts. The first one is a preselection step, based on the unit contextual features. Then, those chosen units become the new candidate units for the second step with a prosodic selection. Consequently, the algorithm time processing is reduced by minimizing the number of candidate units.

4.1.1. Contextual Preselection

The first selection step base on maximizing a target score function, calculated as the sum of four sub-scores (Equation (1)). The latter reflects the matching degree between each target unit t_i (from T) and its candidates u_i^j in the database. They are obtained by comparing the units contextual features (score for: the right context $S_r(t_i, u_i^j)$; left context $S_l(t_i, u_i^j)$; unit position in the word $S_{pw}(t_i, u_i^j)$ and word position in the sentence $S_{ps}(t_i, u_i^j)$). In the end, candidate units with the highest score value are selected to be the new candidate units, $< U_i >$, of the second step.

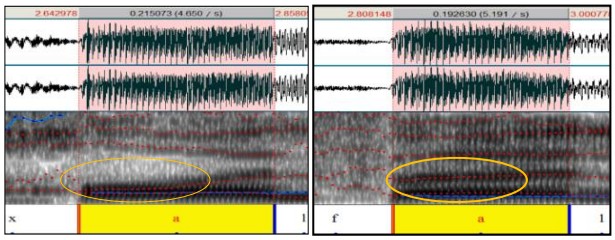
$$< U_i > = \underset{j}{argmax} \begin{Bmatrix} S_r(t_i, u_i^j) + S_l(t_i, u_i^j) \\ + S_{pw}(t_i, u_i^j) + S_{ps}(t_i, u_i^j) \end{Bmatrix} \quad (1)$$

Where: i= 1: n (the number of units in the sentence); j= 1: m (the number of candidate units for the target i).

The efficiency of the unit selection process mostly depends on the number or type of features (contextual or acoustic), used to calculate the target score, as well as, how to equilibrate between these features [2, 15]. In this study, the choice was made by looking for the minimum number of features with the most influence.

Quran language is characterized by its phonetic and phonological phenomena such us: emphasis, *Idgham* (assimilation), *Qalqalah*, etc., These language characteristics depend on the phoneme type (vowel, emphatic consonant, occlusive consonant, etc.,) and influence on adjacent phonemes acoustic and prosodic features [1, 4, 13]. Therefore, the adjacent left and right phonemes are selected as features. As an example, Figure 4 presents the analysis of the vowel [a]. The latter is extracted from two similar contexts, two ending words of surah "El-Falaq" verses 1 and 2: [falaq],(part (b)) and [xalaq] (part (a)). It is clearly noted that the second formant, F_2, of the phoneme [a] begins with low values when it is preceded by the emphatic phoneme [x] (Figure 4-a).

By taking advantage of these language characteristics, we developed a rule-based ES to deal with the score assigning process. This ES takes the features of both target and candidate units as inputs and deduces the score to assign to the candidate one. Figure 5 presents the main components of the ES, with taking one of the target units resulted in Figure 2 as an example. The ES knowledge base (its rules and facts) consists of rules that control the unit acoustic features by its context and other *Tajweed* rules that were not applied in the transcription step.



a) [a] preceded by an emphatic phoneme [x]. b) [a] preceded by non-emphatic phoneme [f].

Figure 4. Analysis by spectrogram and audiogram of vowel [a].

For each candidate unit feature, the inference engine of this ES starts with analyzing the unit neighbor phonemes and their characteristics (occlusive consonant, vowel in emphatic context, assimilated consonant, etc.,). Then, the scores assigned to the candidate unit depends on these characteristics influence on phonemes compared to the target one. In other words, if there is a full match in feature between the target and candidate unit, a maximum score is assigned. Otherwise, a lower or zero score is attributed. These scores depend if neighbor phonemes characteristics have the same, close, or no influence, as in the target unit.

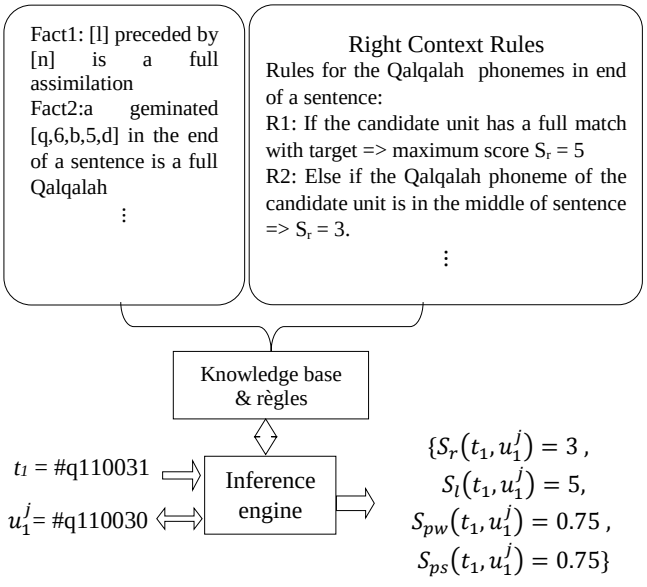


Figure 5. Block diagram of the developed expert system.

4.1.2. Prosodic Selection

For each two successive candidate units from < U_i >, a concatenation cost at that junction is calculated by a weighted sum of three sub costs. With empirical weights tuning, these sub costs are calculated at the concatenation point by the Euclidean distance between the units energy $E\left(u_{i-1}^{m}, u_i^{j}\right)$, fundamental frequency F_0 $F\left(u_{i-1}^{m}, u_i^{j}\right)$, and 12 Mel Frequency Cepstral Coefficients (MFCC) $mfcc\left(u_{i-1}^{m}, u_i^{j}\right)$ values. Moreover, a penalty value $f_i^{\,j}$ that depends on the unit type (vowel, voiced or unvoiced consonant), is added to the F_0 distance in case of contradiction, e.g., at the concatenation point, one of the candidate units has an F_0 value while the other does not.

After the cost calculation, the best unit chain is determined by a dynamic programming search through the units network. In each order (target unit position in the sentence), the best path p_i^j for unit u_i^j is defined by looking for minimum cost C_i^j. As indicates Equation (2), this latter is calculated as the sum of the preceding order cost C_{i-1}^{j} and concatenation cost between the current and preceding order (for every two adjacent units in the stage) (2). After the final cost calculation C_{n+1}, the units chain "forward chain" is formulated by backtracking through the units trellis diagram and following the paths p_i^j , calculated as in Equation (3). This step is called the forward process because it starts from the beginning units <u_1> to the ending ones <u_n> as shows Figure 6 [5].

$$C_i^j = \begin{cases} 0 & if\ \ i = 1 \\ \min\limits_{m} \begin{cases} C_{i-1}^{1} + F\left(u_{i-1}^{1}, u_i^{j}\right) + f_i^{1} \\ C_{i-1}^{2} + F\left(u_{i-1}^{2}, u_i^{j}\right) + f_i^{2} \\ \vdots \\ C_{i-1}^{m} + F\left(u_{i-1}^{m}, u_i^{j}\right) + f_i^{m} \\ E\left(u_{i-1}^{1}, u_i^{j}\right) + mfcc\left(u_{i-1}^{1}, u_i^{j}\right) \\ E\left(u_{i-1}^{2}, u_i^{j}\right) + mfcc\left(u_{i-1}^{2}, u_i^{j}\right) \\ \vdots \\ E\left(u_{i-1}^{m}, u_i^{j}\right) + mfcc\left(u_{i-1}^{m}, u_i^{j}\right) \end{cases} \\ \min\limits_{k}\{C_n^k\} & if\ \ i = n+1 \end{cases} \qquad (2)$$

Where: *m* and *k* are the numbers of candidate units in the *i-1*th and *n*th order, respectively.

$$p_i^j = \arg\left(C_i^j\right) \qquad (3)$$

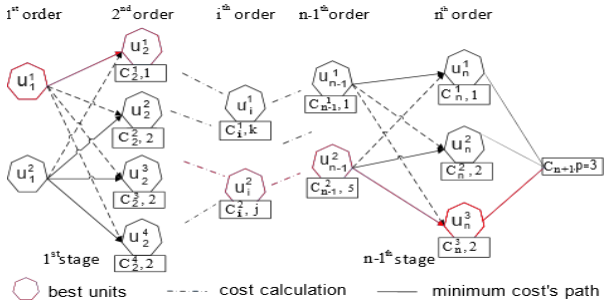


Figure 6. Forward search of the best units chain.

For a better selection, another process is applied with a backward search and a second chain is formulated "the backward chain". This second step follows the same way as the first one but in an opposite direction, i.e., starts with the n^{th} order units to the 1st order ones, and the cost calculation is from right to left. The final unit chain is decided after comparing the forward and backward chains. In each order, if the selected unit is the same in the two processes, that unit is chosen otherwise, the algorithm selects the unit with the minimum sum of its left cost (from the forward process) and right cost (from the backward process) [6].

4.2. Abbreviated Letters Verses

In the case of this exceptional verse type, a contextual selection is sufficient because all their letters combinations exist in our database (a total of 38 units). Therefore, the best units to be selected, < *L*>, are those composed with the same letter (phonetically words) as target units and maximize the comparison score of three contextual features, position in the verse, left and right letters: S_p, S_{ll}, S_{rl} respectively (Equation (4)). These scores get the value "1" if the candidate unit's feature match with the target one and "0" otherwise. As example, the scores for the second target unit t_2 = l20124 and the candidate l_i^2 = l20110 are $S_p\left(t_2, l_2^j\right)$ = 1, $S_{ll}\left(t_2, l_2^j\right)$=1, $S_{rl}\left(t_2, l_2^j\right)$=0.

$$< L > = \arg\max_{j} \begin{Bmatrix} S_{rl}\left(t_i, l_i^j\right) + S_{ll}\left(t_i, u_i^j\right) \\ +S_p\left(t_i, u_i^j\right) \end{Bmatrix} \qquad (4)$$

5. Waveforms Concatenation

In the sound generation phase, the selected units (waveforms samples) are concatenated without any specific processing. Although, because of the quality of the original sound used to build the database, just a little of intensity and sampling rate adjustment was made. We adopt this simple concatenation, to evaluate the performance of the selection process, without any sound processing. This speech modification can affect the speech quality, loss of some features, while the Quran recitation requires a good pronunciation of each sound (phoneme).

6. Test and Evaluation

HQ_TTS system was evaluated by testing the quality of synthesized speech, the correct recitation of Holy Quran verses (the *Tajweed* rules verification) and the whole system performance as follows:

6.1. Speech Quality

Synthetic speech quality was assessed by applying some subjective tests, as we believe that human perception gives the best judgment for a concatenative speech synthesis system, especially when the system is designed for that purpose. These tests were conducted

on the basis of the most important criteria in synthetic speech quality: its intelligibility and naturalness [18].

This evaluation was taken by 16 native Arabic speakers and 8 non-native Arabic speakers. those evaluators were university students and teachers, aged between 25 and 50 from both genders (12 males and 12 females).

In the intelligibility test, five sentences (verses) and 10 words were used, in which the evaluators had to listen to each sentence/word and repeat what they heard. Meanwhile, we note intelligible if they pronounce correctly the targeted sentence/word, (verifying if they identify all the target phonemes), and not intelligible otherwise. The chosen verses compose of: the two main sentences in the Quran, (*Basmalah* and *Istiadhah*), a pair of short and similar part of verses with two different phonemes (العَزِيزُ الحَكِيم [ʔalʕaziizu lħakiim], العَزِيزُ الرَّحِيمْ [ʔalʕaziizu rraħiim]) and another longer verse composed of 30 phonemes. The used words in this test were chosen with different lengths and phonemes variety. After that, the intelligibility percentage was calculated as the number of intelligible sentences/words devised by the total number.

The results give 95.41% of total word intelligibility and 100% for the sentence. Most of the tested words were recognized and pronounced correctly, except the words: يَبْخَسْ [jabxas] and إِيتَاءْ [ʔiitaaʔ] that gave intelligibility of 87.5% and 66.67% respectively. The evaluators also commented that the synthesized verses were easy to detect and they don't show any doubt about them.

In the second test, the evaluators were asked to listen to 10 Quranic verses, then rated their satisfaction with the speech naturalness on a scale from 1 to 5 (very bad; bad; medium; good; very good). After that, the Mean Opinion Score (MOS) is calculated. Table 1 indicates six of the tested verses that were common to all evaluators. They were chosen from the two types mentioned in section 4, with different lengths, contain various unit contexts, and with different percentages of use in the database building (e.g., sentence 5 is one of the sound records used to build the database, i.e., use of 100%). The remaining four verses were left to the evaluators choice, resulting in a total number of 70 different verses or part of it.

This quality test gives a total naturalness result of 74.46%, MOS=3.72, for the native Arabic speakers, and 69.8%, MOS=3.49, for non-native Arabic speakers. The detailed result of the six common verses is presented as a box plot in Figure 7. It shows mixed results with good and medium speech naturalness. Table 2 shows that the free choice verses gave better results, in which more than 63.7% of the verses were rated "4" or higher.

Table 1. Common six verses used in the naturalness test, with their International Phonetic Alphabet (IPA) transcription.

N	Quranic verse	IPA code of the verse
1	وَمِنْ شَرِّ النَّفَّاثَاتِ فِي الْعُقَدِ	[wa min ʃarri nnaffaaɵaati fii lʕuqadi]
2	خَلَقَ الْإِنسَانَ مِنْ صَلْصَالٍ كَالْفَخَّارِ	[xalaqa lʔinsaana min ṣalṣaalin kalfaxxaar]
3	وَأَخَذْنَا مِنْهُمْ مِيثَاقًا غَلِيظًا	[waʔaxaɵnaa minhum miiɵaaqan galii%aa]
4	تَبَارَكَ الَّذِي بِيَدِهِ الْمُلْكُ وَهُوَ عَلَى كُلِّ شَيْءٍ قَدِيرٌ	[tabaaraka llaəii bijadihi lmulku wahuwa ʕalaa kulli ʃajʔin qadiir]
5	إِنَّ الَّذِينَ كَفَرُوا سَوَاءٌ عَلَيْهِمْ أَأَنذَرْتَهُمْ أَمْ لَمْ تُنذِرْهُمْ لَا يُؤْمِنُونَ	[ʔinna llaəiina kafaruu sawaaʔun ʕalajhim ʔaʔan4artahum ʔam lam tunəirhum laa juʔminuun]
6	كهيعص ذِكْرُ رَحْمَتِ رَبِّكَ عَبْدَهُ زَكَرِيَّا	[kaaf haa jaa ʕajn ṣaad əikru raħmati rabbika ʕabdahu zakarijjaa]

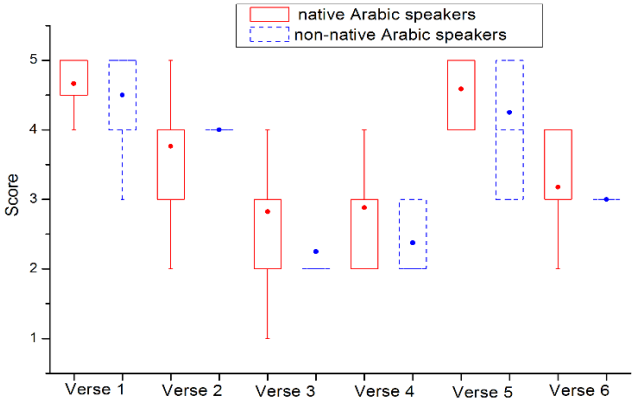


Figure 7. Naturalness results of the six common test verses.

Table 2. Naturalness results of the free choice verses.

Score	1	2	3	4	5
Sentence percentage (%)	1.93	7.67	26.7	33.5	30.2

6.2. System Performance and Tajweed Rules Realization

In the first test, we aimed to see the utility of dividing the selection process into two steps (contextual preselection+prosodic selection). In this context, we compared the synthesis time of the adopted approach with the traditional one presented in [2, 15, 17]. The latter is based on minimizing the sum of the target and concatenation costs together. The results presented in Figure 8 prove the necessity of this division in which the whole synthesis time was reduced. This time improvement does not depend on the length of the sentence but related to the number of candidate units in the final selection stage, which was dropped by the first one as shown in Figure 9. So, splitting the selection algorithm into two parts is very advantageous, especially with the double search application (forward-backward).

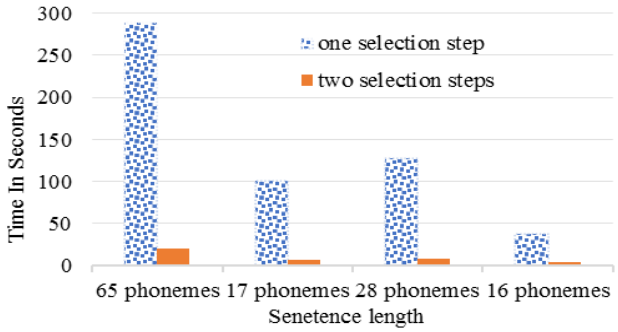


Figure 8. Synthesis time for some test sentences using one and two selection steps.

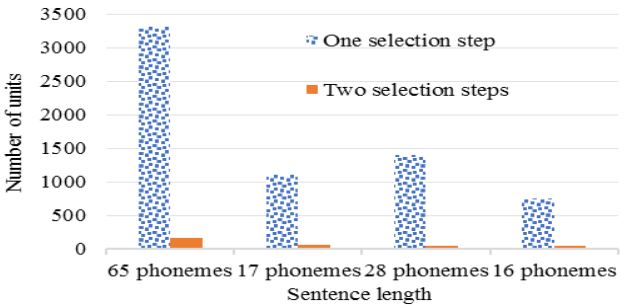


Figure 9. Last selection stage candidate units number for some test sentences using one and two selection steps.

Besides, another test was performed to check the effectiveness of our approach to integrate the ES in the unit selection process. In this evaluation, 20 other verses were synthesized using two different selection algorithms and the evaluators were asked to listen then choose the version with the best speech quality. The first algorithm uses the scores assigned by the developed ES. While in the second one, the unit selection base on trained scores using the active interactive Genetic Algorithm (aiGA), as proposed in [2]. Figure 10 presents the comparison results of this test, with a 54% preference of the selection algorithm that uses the score assigned by the ES.

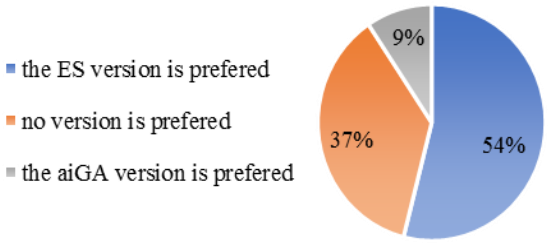


Figure 10. Comparison of speech quality between the ES and the aiGA scoring versions.

During the speech quality tests, the evaluators were also asked about the correct recitation of the synthetic speech. After listening and analyzing each verse, they had to choose between three choices as a final judgment for the correct recitation of the HQ_TTS: "no error is detected," "existence of some errors" or "the Tajweed rules are not well applied". Besides that, they were able to comment about the *Tajweed* rules realization in each verse.

Results of this test show that the HQ_TTS fulfill well these rules in its recitation, in which 85 % of the choices were for the "no error is detected". The remaining 15% was for "existence of some errors" in which few participants were confused about the exact *Madd* duration in some sentences, like verse 3.

7. Discussions

The comparison result in Figure 10 shows the clear preference of our proposed approach for score assigning over the aiGA technique. It proves that the use of the language phonetic and phonological features, via the ES, leads to a better selection of units, hence improving speech quality. the ES is also advantageous, as it does not require time to train its parameters, nor some evaluators to adjust it as the aiGA do. The percentage of preference (54%) is significantly good because sometimes the two compared techniques select almost the same units. Consequently, it becomes difficult to distinguish between the two synthesized sentences.

According to results presented in Figure 7, the HQ_TTS system synthesizes very well the verses used in the database building (such as verse 5). It also works better with short sentences. This is because the longer the sentence, the more concatenation effect between units is perceived. This constraint can be overcome by applying some speech enhancement techniques to those cases only. Compared to other works on Arabic speech synthesis [5, 12, 18], the total naturalness MOS of the HQ_TTS system is considered good and encouraging Especially with the addition of *Tajweed* rules that complicates the synthesis process (more sound variety to synthesize). Results, where the verse naturalness rated less than "3", are due to the small number of variants in the database of some rare Arabic units (e.g., [iiz̧]). The recording quality of the sounds in the database has also an important effect. Some essential sounds were originally bad and had a noise problem. To this day, they represent the only source with monotonous speech. So, by using a good database and well-recorded sounds the system's accuracy will be improved. Actually, those verses were chosen, for the test, to determine the lowest score that the system can obtain. Fortunately, these results are not general and only occur in a few examples, which can be proven by the speech naturalness results of the free choice verses, as illustrated in Table 2.

The variance in the naturalness scores shown in Figure 7 can be explained by some evaluators high expectations, especially the non-native Arabic speakers. First, they are not familiar with synthetic speech. Second, some of them were comparing the quality of the verses recited by HQ_TTS to the ones they used to hear. This latter is not only well recorded but also pronounced with an artistic recitation style, a pleasant melody compared to the monotonous style of

the HQ_TTS. From all the results we conclude that the speech quality can also slightly affect the *Tajweed* rules realization, as the duration problem pointed before, occurred in the sentence with a MOS less than "3" (verse 3).

HQ_TTS gives excellent sentence intelligibility results. We benefit from this advantage of the USSS method because we did not apply much signal modification that may affect the unit acoustic features, thus their correct pronunciation. From the word intelligibility results, it can be deduced that a little issue may occur while synthesizing a single word containing the phoneme [ʔ]. This is because of [ʔ] sound short duration that obstructs its identification. The sound [s] at the end of a word can be confused with [ʂ] as in the test word يَبْخَسْ [jabxas]. This is because the two sound has the same articulation point with some common features. The word was also unpredictable (rarely used by the evaluators). Fortunately, those issues do not much affect the sentence intelligibility because they are rare cases and a Quranic verse can still be predicted even with a no identified phoneme. Moreover, it can be concluded based on the participant comments that the synthesized sentences were clear and easy to detect.

8. Conclusions

In this paper, a speech synthesis system for the Holy Quran recitation was developed. Using the unit selection method, the synthesis algorithm was improved by integrating the ES technology to tune the units' features in the preselection step. After that, the best units are chosen through a final selection and forward-backward search.

The results conclude that the system intelligibility exceeds 97%. as well as a good speech naturalness of 72.13%. The correct recitation of the Holy Quran was achieved by the good realization of *Tajweed* rules. In addition, the occupied memory space was optimized, by using only 3.85% of the total Quran recording. This encourages us to integrate the system is small devices like phone applications.

Enriching the HQ_TTS system with other reciters and recitation styles will make it more interesting for users. This can be done by using a larger database or by adopting the latest hybrid USSS methods that stores speech models instead of sound units in the database, for less memory consumption. The developed system can also be used to synthesize an ordinary Arabic text by using the right database and modifying some transcription rules.

References

[1] Ahmed A., " فونولوجيا القرآن : دراسة لأحكام التجويد في ضوء علم الأصوات الحديث [Quran Phonology: Quran Reciting Rules Based on Modern Acoustic]," MSc Thesis, Ain Chems University, 2004.

[2] Alías F., Formiga L., and Llorá X., "Efficient and Reliable Perceptual Weight Tuning for Unit-Selection Text-To-Speech Synthesis Based on Active Interactive Genetic Algorithms: A Proof-of-Concept," *Speech Communication*, vol. 53, no. 5, pp. 786-800, 2011.

[3] Al-Radhi M., Abdo O., Csapó T., Abdou S., Németh G., and Fashal M., "A Continuous Vocoder for Statistical Parametric Speech Synthesis and its Evaluation Using an Audio-visual Phonetically Annotated Arabic Corpus," *Computer Speech and Language*, vol. 60, 2020.

[4] Alsharif B., Tahboub R., and Arafeh L., "Arabic Text To Speech Synthesis Using Quran Based Natural Language Processing Module," *Journal of Theoretical and Applied Information Technology*, vol. 83, no. 1, pp. 148-155, 2016.

[5] Amrouche A., Falek L., and Teffahi H., "Design and Implementation of a Diacritic Arabic Text-To-Speech System," *The International Arab Journal of Information Technology*, vol. 14, no. 4, pp. 488-494, 2017.

[6] Bettayeb N., Guerti M., and Ramzan N., "A Forward-Backward Dynamic Programming Search for Arabic Unit Selection Speech Synthesis," *in Proceedings of 1st International Conference on Embedded and Distributed Systems*, Oran, pp. 73-77, 2017.

[7] Bettayeb N. and Guerti M., "A Study to Build a Holy Quran Text-To-Speech System," *International Journal on Islamic Applications in Computer Science and Technology*, vol. 7, no. 4, pp. 1-10, 2019.

[8] Dutoit T., *Springer Handbook of Speech Processing*, Springer Berlin Heidelberg, 2008.

[9] Elsayed E. and Fathy D., "Evaluation of Quran Recitation via OWL Ontology Based System," *The International Arab Journal of Information Technology*, vol. 16, no. 6, pp. 970-977, 2019.

[10] Elshafei M., Al-Muhtaseb H., and Al-Ghamdi M., "Techniques for High Quality Arabic Speech Synthesis," *Information Sciences*, vol. 140, no. 3, pp. 255-267, 2002.

[11] Fu R., Tao J., and Wen Z., "Progressive Neural Networks Based Features Prediction for The Target Cost in Unit-Selection Speech Synthesizer," *in Proceedings of 14th International Conference on Signal Processing*, Beijing, pp. 504-509, 2018.

[12] Houidhek A., Colotte V., Mnasri Z., and Jouvet D., "DNN-Based Speech Synthesis for Arabic: Modelling and Evaluation," *in Proceedings of 6th International Conference on Statistical Language and Speech Processing*, Mons, pp. 9-20, 2018.

[13] Jongman A., Herd W., and Al-Masri M., "Acoustic Correlates of Emphatic Consonants in

Arabic," *The Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 3169-3169, 2007.
[14] Kharb S., Kumar H., Kumar M., and Chaturvedi A., "Efficiency of a Machine Translation System," *in Proceedings of International Conference of Electronics, Communication and Aerospace Technology*, Coimbatore, pp. 140-148, 2017.
[15] Narendra N. and Rao K., "Optimal Weight Tuning Method for Unit Selection Cost Functions in Syllable Based Text-To-Speech Synthesis," *Applied Soft Computing*, vol. 13, no. 2, pp. 773-781, 2013.
[16] Sweed A., "Islamway: the Quran Teacher," Available from: https://ar.islamway.net/collection/11899/-المصحف المعلم , Last Visited, 2019.
[17] Taylor P., *Text-to-Speech Synthesis*, Cambridge University Press, 2009.
[18] Tebbi H., Hamadouche M., and Azzoune H., "A New Hybrid Approach for Speech Synthesis: Application to the Arabic Language," *International Journal of Speech Technology*, vol. 22, no. 4, pp. 629-637, 2018.
[19] "The Noble Quran," https://quran.com/, Last Visited, 2019.
[20] Zakariah M., Khan M., Tayan O., and Salah K., "Digital Quran Computing: Review, Classification, and Trend Analysis," *Arabian Journal for Science and Engineering*, vol. 42, no. 8, pp. 3077-3102, 2017.

Nadjla Bettayeb Ph.D student in the Electronics Department of the National Polytechnique school of Algiers, Algeria (ENP). She received her Engineer and Master degree in 2013 from the Electronics Department, ENP. Her current interests include Arabic and Holy Quran language processing and speech synthesis.

Mhania Guerti Professor in the Signal and communication laboratory, Department of Electronics, National Polytechnique school of Algiers, Algeria (ENP). She received her MSc in 1984, from the ILP Algiers in collaboration with the CNET-Lannion (France). She got her Ph.D from ICP-INPG (Grenoble France), in 1993. She is specialized in Speech and Language Processing. Her main research interests include the areas of speech processing, acoustics, and audio-visual systems.