www.ijaceeonline.com ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

Modeling and Optimization of Segmentation Unit Stable Point for Arabic Synthesis Speech

Mehdi Lachiheb
Assistant professor, Mathematics Department, FSG, University of Gabes, Tunisia.
Email: lachihebm@yahoo.ca

Abdelkader Chabchoub
Assistant professor, Electronic Department, MCT, Saudi Arabia
FST, University of Tunis El Manar, Tunisia.
Email : achabchoub@yahoo.fr

Abstract: *This paper presents a new approach for modeling the segmentation stable point for the elaboration of an Arabic speech analysis and synthesis database. This segmentation method minimizes human intervention and helps achieve higher quality synthetic speech. The proposed mathematical definition of the most stable point is the result of a combination of nonlinear optimization and statistical analysis. The statistical study determines the disagreement among the standard deviations of several recordings relative to the same phonetic unit. The optimization ensures the minimum variation of the prosodic parameters in the segmentation stable point of a single or multiple recording. It is also minimizing the error in identifying the segmentation point location. A preference test is conducted to demonstrate the efficiencies of the proposed method in reducing discontinuities in concatenation, thus improving the quality of synthesized speech.*

Keyword: *Arabic speech synthesis, Pitch, Formant, Segmentation, Optimization, Stable point*

1. INTRODUCTION

Synthesis speech plays an important role in modern life. Now a days computers are used in helping people to learn languages [1], to communicate with control devices [2-4], to make contact and to communicate very easily with robots that may understand human voice and give the right replies. Recent research has focused mainly on producing speech that seems more natural or more human, for many languages [3] the goal has been achieved. However, this is not the case for the Arabic language [5, 6], so our research must have an interesting Arabic synthesis speech. Focus on improving speech synthesis in Arabic without having a disadvantage. Currently, speech synthesis by concatenation of acoustic units is the most recent technique to automatically produce speech [7, 8]. A synthetic speech signal can be generated by this method using the concatenating obtained by segmentation of the natural signal [9–12].

The quality of the speech produced by this method being influenced by three main factors [13, 14], the first is the choice of the units, the second is the segmentation of this unit and finally, the con-catenation of those units. Theoretically, the first two parameters concerning the continuity of speech parameters in phase of concatenation are necessary for the third, to improve the heterogeneity of natural and intelligent synthetic voices, Hence the need is for a good choice of units and segmentation or stable point. It should be noted that many researches have proposed more than stable point is in the middle of consonants or vowel without defining any mathematical approach to model this stable point [9–14]. This research work focuses on a mathematical modelling of the most stable point. Also, it deals with the creation of acoustic database units for Arabic concatenation speech synthesis. The acoustic units used in our database are of variable size (diphone, di-diphone phoneme, triphone.) [9] [16] they must be segmented into previously proposed stable point. Our proposed mathematical definition of the most stable point has generated excellent results for the precision chosen and segmentation in Arabic concatenation of speech synthesis using combination of nonlinear optimization and statistical study. This paper is organized as follows: In Section 2, the new definition and modeling of stable point was described. Section 3 presents the results of our statistical and optimal study and results discussion. Section 4 proves the choice of the most efficient units in the most natural synthetic Arabic word. Finally, section5 is the conclusion and future work

Cite this paper:
Mehdi lachiheb, Abdelkader Chabchoub, "Modeling and Optimization of Segmentation Unit Stable Point for Arabic Synthesis Speech", International Journal of Advances in Computer and Electronics Engineering, Vol. 4, No. 4, pp. 1-9, April 2019.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

2. STABLE POINT

The stable point is cited by many researchers as a middle of a phone segmentation point without defining it mathematically. This segmentation was used in the selection of the concatenation unit in the speech synthesis process. The choice of this point in most researches is the middle of each consonant or vowel phoneme in an audio-visual way. This procedure influences speech performance and minimizes the discontinuity rate between two syllabus in the concatenation phase. But this raises many questions: The beginning what is the scientific definition of stable point which represents the optimal point in the concatenation procedure of many characters from different words. The second is, how to compare between the stability of two points and the middle ones. Then, does that stable point determined in view of one or more registration of one unit. In the following part we present several types of stable point that improve the segmentation and the concatenation of units for the synthesis of speech. The body text shall have the font size of 10 points, times new roman font and justified with single line spacing.

Definition 1

Let be σ a positive real number and RU a one registration of a given unit U, defined by a time interval T, and p functions g_j , j = 1, ..., p. A point t^* is called local σ-stable point of the registration RU if t_ is the optimal solution of the problem:

$$min_{t\in[\sigma,1-\sigma]}(max_{r\in[t-\sigma,t+\sigma]}\sum_{j=1}^{p}\frac{g_j(rT)}{\frac{1}{T}\int_0^T g_j(x)dx}-\min_{r\in[t-\sigma,t+\sigma]}\sum_{j=1}^{p}\frac{g_j(rT)}{\frac{1}{T}\int_0^T g_j(x)dx})$$

The first definition relies on measuring the range of the main characteristics of the voice in a certain period and then choose the middle point of the period of optimal range.

Definition 2

Let be $RU_i, i = 1,\dots,n$, n registrations of a given unit U, each is defined by a time interval T_i, and p characteristics functions $g_j^i, j = 1,\dots,p$. A point t^* is called the global stable point of $RU_i, i = 1,\dots,n$ if t^* is the optimal solution of the problem

$$\min_{t\in[0,1]} f(t) = \frac{1}{p}\sum_{j=1}^{p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{g_j^i(tT_i)}{\bar{g}_j(t)}-1\right)^2}$$

where

$$\bar{g}_j(t) = \frac{1}{n}\sum_{i=1}^{n} g_j^i(tT_i)$$

presents the mean of the characteristics functions $g_j^i, i = 1,\dots,n$ at *t*.

Definition 3

Let be $RU_i, i = 1,\dots,n$, n registrations of a given unit U, each is defined by a time interval T_i, and p characteristics functions $g_j^i, j = 1,\dots,p$. A point t^* is called the global stable point of order 1 of $RU_i, i = 1,\dots,n$ if t^* is the optimal solution of the problem

$$\min_{t\in[0,1]} f(t) = \frac{1}{p}\sum_{j=1}^{p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{g_j^i(tT_i)+\frac{\partial g_j^i(tT_i)}{\partial t}}{\bar{g}_j(t)+d\bar{g}_j(t)}-1\right)^2}$$

where

$$d\bar{g}_j(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_j^i}{\partial t}(tT_i)$$

presents the mean of the derived characteristics functions $\frac{\partial g_j^i}{\partial t}, i = 1,\dots,n$ at *t*.

They also set the point t_1 is more stable than t_2 if $f(t_1) \le f(t_2)$.

This definition is based on the measure of the function of the standard deviation of the main normalized functions of *n* records of a voice unit and choose the optimal point of the measured function.

Definition 4

Let be σ a positive real number and $RU_i, i = 1,\dots,n$, n registrations of a given unit U, each is defined by a time interval T_i, and p characteristics functions $g_j^i, j = 1,\dots,p$. A point $(t, t_1,\dots,t_n)^*$ is called the global σ-stable point of $RU_i, i = 1,\dots,n$ if $(t, t_1,\dots,t_n)^*$ is the optimal solution of the problem

$$\begin{cases}\min f(t,t_1,\dots,t_n) = \\ \frac{1}{p}\sum_{j=1}^{p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{g_j^i((t+t_i)T_i)}{\bar{g}_j(t,t_1,\dots,t_n)}-1\right)^2} \\ -\sigma \le t_i \le \sigma, \quad i = 1,\dots,n \\ \sigma \le t \le 1-\sigma\end{cases}$$

where

$$\bar{g}_j((t,t_1,\dots,t_n)) = \frac{1}{n}\sum_{i=1}^{n} g_j^i((t+t_i)T_i)$$

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

Definition 5

Let be σ a positive real number and $RU_i, i = 1, \dots, n$, n registrations of a given unit U, each is defined by a time interval T_i, and p characteristics functions $g_j^i, j = 1, \dots, p$. A point $(t, t_1, \dots, t_n)^*$ is called the global σ-stable point of $RU_i, i = 1, \dots, n$ of order 1 if $(t, t_1, \dots, t_n)^*$ is the optimal solution of the problem

$$\begin{cases} \min f(t, t_1, \dots, t_n) = \\ \frac{1}{p}\sum_{j=1}^{p}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{g_j^i((t+t_i)T_i) + \frac{\partial g_j^i}{\partial t}((t+t_i)T_i)}{\bar{g}_j(t, t_1, \dots, t_n) + d\bar{g}_j(t, t_1, \dots, t_n)} - 1\right)^2} \\ -\sigma \le t_i \le \sigma, \quad i = 1, \dots, n \\ \sigma \le t \le 1 - \sigma \end{cases}$$

where

$$d\bar{g}_j((t, t_1, \dots, t_n)) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_j^i}{\partial t}((t+t_i)T_i)$$

The last two definitions discuss the same point for the second definition, taking into consideration the mistakes that can be committed in the segmentation stage of each phoneme. Also, a good estimation of an unit require consideration of characteristic functions and its derivatives. This takes consideration in the formulation 4. The speech synthesis by concatenation is influenced by three main factors: the first is the choice of the unit, the second is the construction of the database by the segmentation of multiple units and finally the concatenation of a Word parts. In order to obtain a word similar to the reality, it is interested in these 3 phases which are directly interrelated. In fact the reduction of the difference of the discontinuity of main parameters between two units during the concatenation has a significant influence on the speech result. This passes through the first two steps. A local stable point ensures that a small variation of the main parameters is obtained if a small error is performed at the phase of segmentation. While a global stable point is often guaranteed that a minimization of the discontinuity gaps is obtained if we link several units obtained from different words. It remains to ensure that the local and global stable points are the same.

3. NUMERICAL RESULTS AND DISCUS

3.1 Databases

The synthesis of speech by phoneme requires a very developed database that takes into consideration all the phenomena that can affect the speech. Indeed, speech depends on many factors such that the state of the recording space, the state of the person, the nature of the sentence, the position of the word in the sentence and the position of the unit in the word and the quality of the apparatus of recording... In this work we have chosen the room and the apparatus of recording and the speaker to avoid the external effects and we expect to record several versions of the same unit to study the main factors influencing the speech. The classified factors are those that have led to a significant increase in the standard deviation and many copies of one unit are stored in the database. They are selected according to the powerful effects. In addition, such copies shall be segmented according to one of the most stable points defined in the preceding paragraph. This requires taking test intervals for the bounded of each unit. Since the Arabic language as English and French is formed essentially by consonants and vowel, our study will be based on those two types of unity. The question we are going to answer which one of these two units is the most stable to choose it the extreme of our main unit in the database. Thus, the test sequence of the units will be formed by at least 4 phonemes so that one will have at least one vowel and one consonant with free extremes. All this would create a database containing all the versions that help to form all the words in their different positions with a minimum discontinuity between the units.

3.2 The effect of quality recording speech

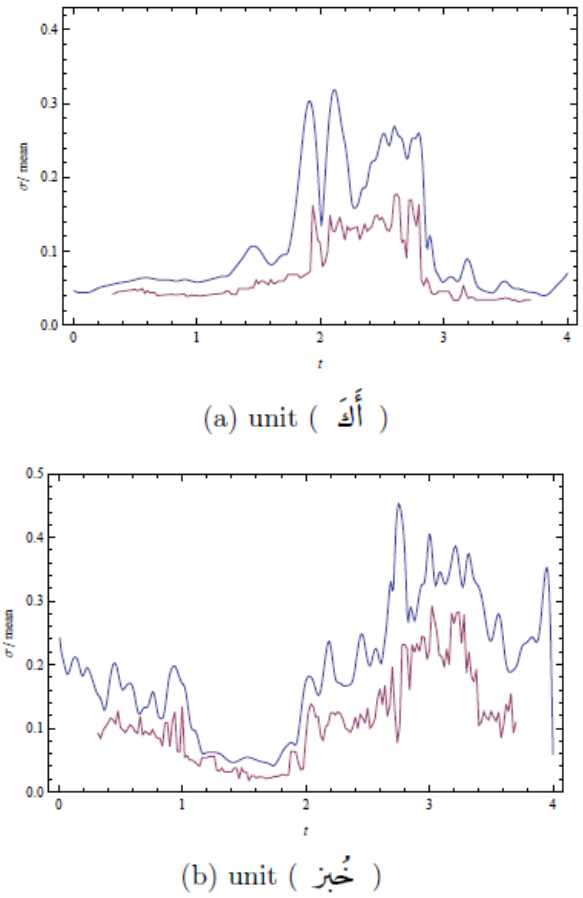

Figure 1. The normalized standard deviation for 12 recording of 2 units.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

Although we were limited to one speaker for the recording phase and we chose sophisticated tools and appropriate recording space to avoid noise, we noticed some differences between the data in some recordings. This is logical because the state and pronunciation of the registrar vary from moment to moment in his way of treating the breath. A diverse collection of units of four phonemes are chosen. Each contained consonants and vowels. Twelve recording by the same person for any unit are considered to analyse the function defined by the normalized standard deviation of pitch, four first frequencies and intensity. The units (أك) and (خُبزْ) was presented in *Figure*1. We notice that the difference is not negligible, and it is even more intensive for a consonant. These results show the order of magnitude of the normalized standard deviation from one record to another under the same conditions. To avoid side effects and copies influenced by rare states, the unit closest to the mean is chosen.

3.3 Position effect

The speech in a word passes through three essential stages : beginning when it increase, the medium where it is stationary and the end where it decrease. This point attracts our attention and we hope its influence in the speech synthesis is important. Therefore, we think that the position of the unit in the word affects the data of the parameters of a unit. To show the effect of the position, we try to segment several units formed by two phonemes, each unit is removed from several different words, from three positions: at the beginning, the interior and the end, and by several peoples including a woman and man. Table 1 presents the mean and standard deviation of length time of three units recorded by four peoples divided into four groups according to the position in the word. Note that for most units, the time mean varies from one position to another. The total standard deviation of all recording parts of each unit proves the effect of the phoneme position in the word. Indeed, the distribution of time length of the same unit is denser around the mean if it is obtained from different words and in the same position. In addition, the other speech parameters were targeted as the fundamental frequency F_0 and the following four frequencies F_1, F_2, F_3 and F_4 and the intensity.

Table 2,3,4,5 presents the percent of total standard deviation of six parameters pitch F_0, F_1, F_2, F_3,F_4 and intensity at five points of a unit in three positions. Similar to the last notes, the total standard deviation of all recording parts of each unit proves the effect of the phoneme position in the word. it should be noted that the second quartile point is the middle of the vowel has the lowest standard deviation. Then it is the stable point to respect the second definition. The question is however whether this value is the least of all the points that will be the goal for the next analysis.

TABLE 1: THE MEAN AND STANDARD DEVIATION (%) OF THREE UNITS IN THREE POSITIONS RECORDED BY 3 PERSONS

		unit (أ)	Unit(ب)	Unit (ك)
Person1	m_{be}	0.08522727	0.1143750	0.15
	m_{mi}	0.08874999	0.1546875	0.144375
	m_{en}	0.121875	0.14375	0.14453125
	m	0.09818548	0.13611111	0.14615384
	σ_{be}	3.52766841	23.8253872	4.1666666
	σ_{mi}	2.8169014	2.67247607	6.8310536
	σ_{en}	13.997275	17.9850342	1.4301358
	σ	19.549289	20.7747684	5.1802683
		unit (أ)	**Unit(ب)**	**Unit (ك)**
Person2	m_{be}	0.10340909	0.08562499	0.114375
	m_{mi}	0.08181818	0.12291666	0.1194444
	m_{en}	0.08625	0.11796875	0.1118055
	m	0.09062499	0.10763888	0.1151785
	σ_{be}	3.96214425	5.70091217	6.4887115
	σ_{mi}	16.1075187	3.38983050	4.5779115
	σ_{en}	9.61340518	3.17604736	9.2960431
	σ	14.6297954	16.2959108	7.4555752
		unit (أ)	**Unit(ب)**	**Unit (ك)**
Person3	m_{be}	0.09124999	0.110625	0.0857954
	m_{mi}	0.13680555	0.1656250	0.1242187
	m_{en}	0.206250	0.21328125	0.1902777
	m	0.14050925	0.15959821	0.1303571
	σ_{be}	28.603579	44.917919	19.421030
	σ_{mi}	39.1488720	15.490094	65.33316
	σ_{en}	21.6406922	20.2196475	7.912220
	σ	44.9133910	36.2178357	48.569806

3.4 Determination of stable points

For all parameter's functions F_0, F_1, F_2, F_3, F_4 and intensity a cubic splines approximations φ_i, i = 1, ..., 6 were used. This provides a simple function that coincide with the data and gives a good approximation of the derivatives. As they are at least 3 times differentiable that we can use the regular optimization methods based on the second derivatives. The test sequence of the units used in the next work will be formed by at least 4 phonemes so that one will have at least one vowel and one consonant with free extremes.

TABLE 2: THE STANDARD DEVIATION (%) OF 6 PARAMETERS OF UNIT (أ) IN THREE POSITIONS RECORDED BY PERSON 2

	be	Quad1	end	Quad2	mid
σ_{be}	12.266	7.2186	14.829	6.658	8.003
σ_{mi}	11.267	7.567	9.054	6.3532	7.873
σ_{en}	5.136	3.714	4.982	5.731	5.247
σ	11.676	8.04113	13.12480	7.840	8.657

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

TABLE 3: THE STANDARD DEVIATION (%) OF 6 PARAMETERS OF UNIT (بـ_) IN THREE POSITIONS RECORDED BY PERSON2

	be	Quad1	end	Quad2	mid
σ_{be}	11.29	8.2517	12.08	8.238	9.201
σ_{mi}	13.39	7.487	7.834	6.985	8.660
σ_{en}	9.867	8.555	8.456	6.703	7.844
σ	12.293	9.293	11.810	8.808	9.742

TABLE 4: THE STANDARD DEVIATION (%) OF 6 PARAMETERS OF UNIT (أ) IN THREE POSITIONS RECORDED BY PERSON 1

	be	Quad1	end	Quad2	mid
σ_{be}	19.113	4.3156	17.352	4.5732	4.645
σ_{mi}	13.138	4.808	17.632	3.4773	4.277
σ_{en}	24.679	11.375	23.578	6.9132	7.055
σ	27.633	10.501	22.526	10.3849	9.720

TABLE 5: THE STANDARD DEVIATION (%) OF 6 PARAMETERS OF UNIT (كَ) IN THREE POSITIONS RECORDED BY PERSON 1

	be	Quad1	end	Quad2	mid
σ_{be}	9.8218	12.821	17.184	7.972	8.692
σ_{mi}	10.482	12.227	19.977	7.588	13.46
σ_{en}	8.350	9.5674	9.298	7.228	8.15
σ	10.276	12.528	8.1450	8.145	11.69

3.4.1 Local σ- stable point

The optimization problem defined the local σ-stable point is

$$PLS\begin{cases} min(\max\limits_{r\in[t-\sigma,t+\sigma]}\sum\limits_{j=1}^{6}\dfrac{\varphi_j(r)}{\frac{1}{4}\int_0^4\varphi_j(x)dx}- \\ \min\limits_{r\in[t-\sigma,t+\sigma]}\sum\limits_{j=1}^{6}\dfrac{\varphi_j(r)}{\frac{1}{4}\int_0^4\varphi_j(x)dx}- \\ \sigma\le t\le 4-\sigma \end{cases}$$

where $\varphi_j, j=1,\dots,6$ present the cubic splines approximations of the frequencies F_0, F_1, F_2, F_3, F_4 and intensity functions.

One way to approach the solution of the problem P and avoid first and second derivative of the objective function of the main problem which is defined by an optimization problem, to consider for all $t_k = kh$ (h is a small positive number), the 2 following finite sequences of optimization problems

$$PSMAX_k\begin{cases} max\sum\limits_{j=1}^{6}\dfrac{\varphi_j(r)}{\frac{1}{4}\int_0^4\varphi_j(x)dx}- \\ t_k-\sigma\le r\le t_k+\sigma \end{cases}$$

with optimal value $psmax(t_k)$ and

$$PSMIN_k\begin{cases} min\sum\limits_{j=1}^{6}\dfrac{\varphi_j(r)}{\frac{1}{4}\int_0^4\varphi_j(x)dx}- \\ t_k-\sigma\le r\le t_k+\sigma \end{cases}$$

with optimal value $psmin(t_k)$.

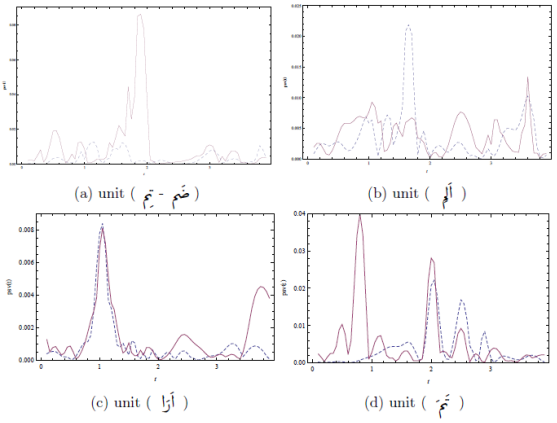

Figure 2 The function $psv(t)$ for 2 recording of 4 units.

All optimization problems deal with one dimension. The method was used to find the minimum of a smooth univariate function using both function evaluations and first derivative. This method uses a descent direction defined by either the secant direction or cubic interpolation. [17–20]. The objective function of *PLS*,

$$psv(t_k)=psmax(t_k)-psmin(t_k)$$

is plotted to determine the global minimum. A diverse collection containing many consonants and all the vowels to analyze the function $psv(t)$ with $h = 0.01$ is considered. Two recording for each unit were studied. A sample of 4 units was considered in graph 1.

We noted that in most cases the global in minimum is in the interval [0.25, 0.75] of a vowel phoneme and in general the $psv(t)$ function is more moderate in a vowel interval than in that of consonant.

3.4.2 Global stable and σ-point

The optimization problem defined the global stable point is

$$PGS\begin{cases}\min\ pgsv(t)\\ 0\le t\le 4\end{cases}$$

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

where $pgsv(t) = \frac{1}{6}\sum_{j=1}^{6}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{\varphi_j^i(t)}{\bar{\varphi}_j(t)} - 1)^2}$

and the optimization problem defined the global σ-stable point is

$$PGS_\sigma \begin{cases} \min\ pgsv_\sigma(t, t_1, \dots, t_n) \\ -\sigma \le t_i \le \sigma,\ \ i = 1, \dots, n \\ \sigma \le t \le 4 - \sigma \end{cases}$$

where

$$pgsv_\sigma(t, t_1, \dots, t_n) = \frac{1}{6}\sum_{j=1}^{6}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\frac{\varphi_j^i(t+t_i)}{\bar{\varphi}_j(t+t_i)} - 1)^2}$$

and

$$\bar{\varphi}_j(t) = \frac{1}{n}\sum_{i=1}^{n}\varphi_j^i(t)$$

If we use the standard deviation formula

$$sd = \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n^2}(\sum_{i=1}^{n}x_i)^2}$$

Then the normalized standard deviation

$$\frac{sd}{\frac{1}{n}\sum_{i=1}^{n}x_i} = \sqrt{\frac{n\sum_{i=1}^{n}x_i^2}{(\sum_{i=1}^{n}x_i)^2} - 1}$$

and the problem *PGS* is given by

$$PGS \begin{cases} \min \sqrt{\frac{n\sum_{i=1}^{n}(\varphi_j^i(t))^2}{(\sum_{i=1}^{n}\varphi_j^i(t))^2} - 1} \\ 0 \le t \le 4 \end{cases}$$

and the optimization problem defined the global σ-stable point is

$$PGS_\sigma \begin{cases} \min \sqrt{\frac{n\sum_{i=1}^{n}(\varphi_j^i(t+t_i))^2}{(\sum_{i=1}^{n}\varphi_j^i(t+t_i))^2} - 1} \\ -\sigma \le t_i \le \sigma,\ \ i = 1, \dots, n \\ \sigma \le t \le 4 - \sigma \end{cases}$$

In the last problems we have an optimization problem composed by a multivariate function subjects to bounds on the variables. To determine the global optimum, a graphical study of the function $pgsv\sigma(t_k)$ defined by the optimal values of the following sequence of optimization problems is used.

$$(PGS_\sigma)_k \begin{cases} \min\ pgsv_\sigma(t_k, t_1, \dots, t_n) \\ -\sigma \le t_i \le \sigma,\ \ i = 1, \dots, n \end{cases}$$

where $\sigma \le t_k = kh \le 4 - \sigma$ and h is a small real number. We use a quasi-Newton to solve this problem [20-22]. An active set *AS* is defined by the indices of the variables at their bounds. A free variable is a variable not in the active set. The feasible direction for the free variables is computed by the formula

$$d = H_k^{-1}g(x_k)$$

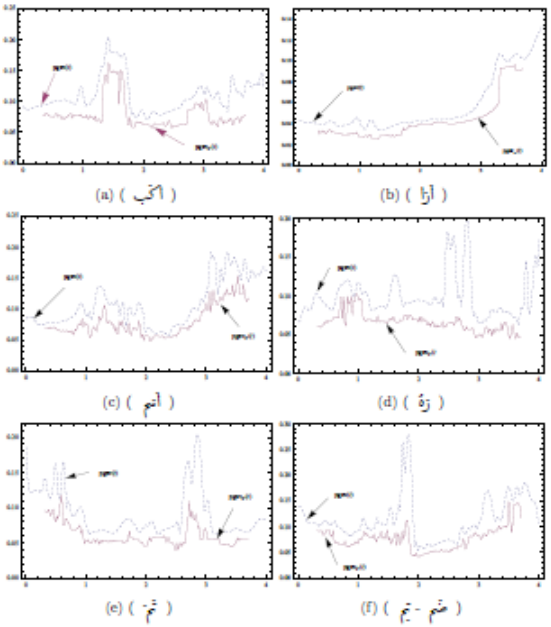

Figure 3 The function $pgsv(t)$and $pgsv\sigma(t)$ for 6 different units.

where H is a positive definite approximation of the Hessian matrix and $g(x_k)$ is the gradient evaluated at the current point x_k, both are computed with respect the free variables. The new point is defined by

$$x_{k+1} = x_k + \lambda\, d$$

λ is determined by a line search minimization of the objective function *f*. The matrix H_k is updated according the *BFGS* formula

$$H_{k+1} = H_k + \frac{H_k s s^t H_k}{s^t H_k s} + \frac{y y^t}{y^t s}$$

$$where\ s = x_{k+1} - x_k\ and\ y = g(x_{k+1}) - g(x_k)$$

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

The objective function $pgsv(t)$ and $pgsv\sigma(t)$are plotted with $\sigma = 0.25$ and $h = 0.01$ to determine the global minimum and the interval where the values closed to that. A diverse collection containing many consonants and all the vowels to analyze the function $pgsv(t)$ and $pgsv\sigma(t)$ is chosen Between 9 and 12 recording for any unit are considered. A sample of 6 units was considered in graph 2. We noted that in the almost case the global minimum is in the interval [0.25, 0.75] of a vowel phoneme and in general the $pgsv(t)$ and $pgsv\sigma(t)$ functions are more moderate in a vowel interval than in that of consonant. Rather, the variation of this functions is more important in the interval of a consonant. However, graph 2.b shows that the vowel" mad" is not stable and similar to consonant. On the other hand, the function $pgsv\sigma(t)$ presents a remarkable reduction of the function $pgsv(t)$ in all cases of graph 2.

This confirms our proposition that the fixing of the terminals of a unit is very sensitive and it may be subject to errors. Then the function $pgsv\sigma(t)$ helps to obtain the most stable area and to correct the errors that we encounter in the determination of the endpoints during the segmentation step.

3.4.3 Global stable and σ-point of order 1

Similar to the global stable point, to determine the stable point **σ-** stable point of order 1, a graphical studies of the functions

$$pgsfv(t) = \frac{1}{6}\sum_{j=1}^{6}\sqrt{\frac{n\sum_{i=1}^{n}(\varphi_j^i(t) + \frac{\partial\varphi_j^i}{\partial t}(t))^2}{(\sum_{i=1}^{n}\varphi_j^i(t) + \frac{\partial\varphi_j^i}{\partial t}(t))^2}} - 1$$

Where $0 \le t \le 4$ and $pgsfv_\sigma(t_k)$ defined by the optimal values of the following sequence of optimization problems is used.

$$(PGSF_\sigma)_k \quad \begin{cases} \min\ pgsfv\sigma(t_k, t_1, \dots, t_n) \\ -\sigma \le t_i \le \sigma,\ \ i = 1, \dots, n \end{cases}$$

Where

$$pgsfv\sigma(t_k, t_1, \dots, t_n) = \frac{1}{6}\sum_{j=1}^{6}\sqrt{\frac{n\sum_{i=1}^{n}(\varphi_j^i(t_k+t_i) + \frac{\partial\varphi_j^i}{\partial t}(t_k+t_i))^2}{(\sum_{i=1}^{n}\varphi_j^i(t_k+t_i) + \frac{\partial\varphi_j^i}{\partial t}(t_k+t_i))^2}} - 1$$

and $\sigma \le t_k = kh \le 4 - \sigma$ and h is a small real number. The same units used in the last section are tested according to the new definition. Among these units, four are represented in graph 3. As long as the values of the curves $pgsfv(t)$ and $pgsfv_\sigma(t)$ are largely different, each one is represented on a graph. We noted that the curve $pgsfv(t)$has several peaks due of the wide variation of the parameters used and the instability of the interpolation of the derivative function. A good reduction is obtained during the translation of each unit recording according to a well defined path which is given by the curve $pgsfv_\sigma(t)$. Despite the fact that in this case the curve contains several peaks, one can notice clearly in most units that the vowel zones are more stable than those of a consonant and the global optimum is generally equal or close to the middle of a vowel interval.

4. UNIT CHOICE

Note that the variance of the duration of the unit classified by its position in the word is very small compared to the one classed by the speaker who shows that the effect of the position[15]. The other

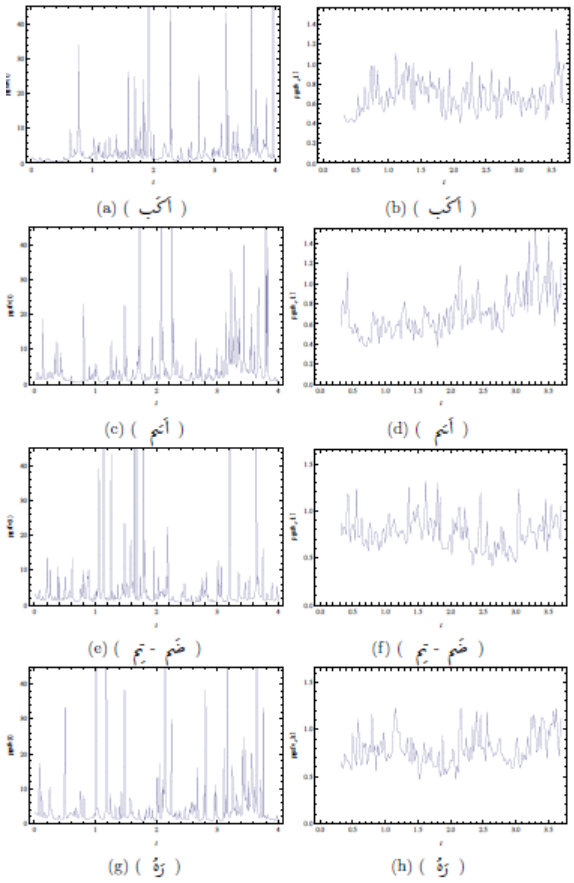

Figure 4 The functions $pgsfv(t)$ and $pgsfv_\sigma(t)$for 4 different units

fundamental result observed from these statistics is that in most cases the vowel interval is more stable than that of consonant and the midpoint of the vowel is the most stable point. Therefore, we consider units that end with vowels so that the concatenation will be at the most stable points. These units are formed by at least three phonemes. To process these points, the following type of words are chosen:

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 4, April 2019, pp. 1–9

Word *W1*
Space C1 V1 C2 V2 C3 V3 space
Word W*2*
Space C1 V1 C2 C3 V2 C4 space
Word *W*3
Space C1 V1 V2 C2 V3 V4 C4 space
where C presents a consonant and V presents a vowel.
From word W1 three different unit type are determined as follows:
Unit 1 : Space C V
which guarantees the beginning of a word. Since the Arabic language contains 28 consonants of which two can be in laminated and inflated forms, then 30 consonants are considered. The number of the first vowel is equal to three then we obtain 90 different units of unit 1 in the database.
Unit 2: VCV
which guarantees a middle part of a word and which can have 270 different units in the database,
Unit 3 : VCV space
which guarantees the middle part of a word and which can have the same number of the last unit in the database.
From word *W2* two different type of units are determined as follows:
Unit 4: VCCV
which guarantees a middle part of a word, and which can appear with 4020 different forms in the database [23].
Unit 5: VC space
appear with 90 different forms in the database
From word *W3*, we determined the following different units:
Unit 6: VVCV
which can take 90 different forms in the database.
Unit 7:VCVV space
which can take 270 forms
Unit 8: VVCVV space
which can take 270 forms.
With this unit we will have a data base that includes 6270 units all of which are obtained by segmentation at the level of a vowel

5. CONCLUSION

This work aims at conducting a study on Arabic speech synthesis by concatenation in a large database. We were particularly interested in the study of concatenation unit choices and the segmentation points of these units. The best choice of the beginning and the end of a unit directly influences the concatenation step and therefore the quality of synthetic speech. We have first made a statistical study of the factors that affect the quality of segmented speech units (recording, unit location in the word and speaker conversion effect).
This study is based on the variance function, which indicates the difference rate between several samples of the same unit. It shows the influence of corpus recording on the unit parameters. To minimize this influence, it is proposed to record the same unit of the same word several times and to choose the closest recording of a unit to the average using the least square method.

The position of a unit in the word (beginning, middle or end) has been taken into consideration as this affects the quality of speech synthesis. Therefore, the units in the database have been divided into three classes (start, middle and end). Furthermore, we have done a mathematical modulation of the stable points based on the minimum of the variation of the prosodic parameters along a unit over a fixed period of time and between several records. A multivariate optimization problem has been considered to correct the error that can appear at the ends of a unit. A statistical study on several units formed by two vowels and two consonants shows that, in most cases, the vowel area is more stable than that of the consonant and that the most stable point is in the range [0.25 T, 0.75T] of a vowel.

The correction applied to the ends of a modulated phoneme, through an optimization problem which defines the global stable sigma point, shows a good reduction of the variance. This work gives rise to a large database and suggests further study of the stable points of all consonants and vowels to particularly look for classes of phonemes where consonants are more stable than vowels in order to reduce the size of the database.

REFERENCES

[1] Heiga Zen a,b, Keiichi Tokuda a , Alan W. Black c.,(2009), ”Statistical parametric speech synthesis” *Speech Communication Elsevier*, Volume 51, Issue 11, pp. 1039-1064

[2] Arzu Mutlu, Betil Eros-Tuga, (2013), “the role of computer-assisted language learning (call) in promoting learner autonomy” *Eurasian .Journal of Educational Research*, issue 51,spring, pp. 107-122.

[3] Md Fahad Jahan, Md Taslim Arefin, (2017), "Design and Implementation of an Automated Home Security Robot", *International Journal of Advances in Computer and Electronics Engineering*, Vol. 2, No. 12, pp. 1-7.

[4] A. Hunt and A. Black, (1996), “Unit selection in a concatenative speech synthesis system using a large speech database” *ICASSP*, vol. 1, pp. 373-376.

[5] M. Chu, H. Peng, H. Yang, E. Chang, (2001),” Selecting nonuniform units from a very large corpus for concatenative speech synthesizer” In: Proc. *ICASSP*, Salt Lake City, USA, pp.785-788.

[6] Keiichi Tokuda ; Yoshihiko Nankaku ; Tomoki Toda , (2013), “Speech Synthesis Based on Hidden Markov Models” *Proceedings of the IEEE*, Vol 101 ,Issue 5, pp. 1234-1252.

[7] T. Toda, A. W. Black, and K. Tokuda, (2007), “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory” *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 8, pp. 2222-2235.

[8] G. Demenko, S. Grocholewski, A. Wagner, M. Szymansk, . (2006)” Prosody Annotation for Corpus Based Speech Synthesis "*In Proceedings of the Eleventh Australasian*

ISSN: 2456 - 3935

International Journal of Advances in Computer and Electronics Engineering

Volume 4, Issue 4, April 2019, pp. 1–9

International Conference on Speech Science and Technology. Auckland, New Zealand, pp. 460-465.

[9] M. Brown, A.P. Salverda, L.C. Dilley, M.K. Tanenhaus, (2011) "Expectations from preceding prosody influence segmentation in online sentence processing", Psychon. Bull. Review 18, pp. 1189-1196.

[10] A. chabchoub, W.barkouti, S.Alahmadi, A.cherif, (2012), "Di-Diphone Arabic Speech Synthesis Concatenation", International Journal of Computers & Technology, Volume 3, pp. 218-222.

[11] Raja Abdelmalek, Zied Mnasri, (2016), " High quality Arabic text-to-speech synthesis using unit selection" 13th International Multi-Conference on Systems, Signals and Devices (SSD), pp. 1-5.

[12] K. Khalil, C. Adnan, (2013), " Arabic hmm-based speech synthesis" *In International conference on electrical engineering and software applications (ICEESA).*

[13] M. Z. Rashad, H. M. El-Bakry, I. R. Ismail, (2010), "Diphone speech synthesis system for Arabic using mary tts», International *Journal of Computer Science and Information Technology (IJCSIT),* 2(4), pp. 18-26.

[14] J.v. Santen, B. Moebius, (2000), "A quantitative model of F0 generation and alignment, In: Botinis, A. (Ed.), *Intonation Analysis Modeling and Technology. Springer*, pp. 269-288.

[15] I. Abu Doush, F. Alkhatib, A.R. Bsoul, (2016), "What we have and what is needed, how to evaluate Arabic Speech Synthesizer" *International Journal of Speech Technology* 19, pp. 415-432.

[16] A. Chabchoub, A. Cherif, (2011)," An Automatic mbrola tool for high quality Arabic speech synthesis" *IJCA* volume 36-No-1, pp. 27-36.

[17] J. Nocedal, S. Wright, (1999), "Numerical Optimization", *Springer-Verlag*.

[18] R.K. Beatson, M.J.D. Powell, , (1992), " Univariate multiquadric approximation: quasi-interpolation to scattered data", Constructive Approximation, pp. 275-288.

[19] M.J.D. Powell, ,(1964) , "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives*", Computer J.* , pp. 7155-162.

[20] P.E. Gill, W. Murray and M.H. Wright, , (1981), "Practical Optimization", *Academic Press, New York*, pp. 92.

[21] J.E. Dennis, R.B. Schnabel, (1983), "numerical methods for unconstrained optimization and nonlinear equations", *Published by Englewood Cliffs, N.J., Prentice Hall.*

[22] P.E. Gill, W. Murray, (1972), " Quazi Newton method for unconstrained optimization», *J. Inst. Maths. Applics.*9, pp. 91-108.

[23] O.J.A. Rasoul, (2010)," تطوير التعرف الالي على الحروف العربية من خلال الية لغوية" Proceding, Sixth international computing conference in Arabic, *Yasmine Hammemet Tunisia*, pp. 365-378.

Authors Biography

Dr. Mehdi Lachiheb is an Assistant professor in the Mathematics Department, FSG, University of Gabes, Tunisia. and member Research unit ATSSEE, FST Al-Manar University Tunisia Company. He completed his Ph.D. at ENIT Tunisia in Applied Mathematics. His research interests are Applied mathematics, optimization signal processing and the structure of the system engineering.

Dr. Abdelkader Chabchoub is an Assistant professor in the electronic Department of in Madinah College of Technology KSA and member Research unit ATSSEE, FST Al-Manar University Tunisia Company. He completed his Ph.D. at FST, ENIT Tunisia in Electronic and signal. His research interests are artificial intelligence, embedded systems, robotic and smart system, Signal and image processing.

Cite this paper:

Mehdi lachiheb, Abdelkader Chabchoub, "Modeling and Optimization of Segmentation Unit Stable Point for Arabic Synthesis Speech", International Journal of Advances in Computer and Electronics Engineering, Vol. 4, No. 4, pp. 1-9, April 2019.