

Modeling sound duration in an Arabic text to speech system

S. Khabet, Z. Zemirli, M. Mosteghanemi
LCSI Laboratory, SCALA TEAM, INI, Algiers, Algeria

Keywords : Sound duration, Arabic TTS, Contextual model, Syllabic structure.

Abstract This article describes our method to model the sound duration of standard Arabic speech. The final aim is the prosodic generation for an Arabic Text-To-Speech (TTS) synthesis system. Several authors have already identified on isolated word corpora the effects of speakers, word structure, sound nature and Arabic dialects. In our work hypothesis is to verify these effects for standard Arabic on continuous speech. Firstly a continuous speech corpora is defined and recorded. Then, two automatic tools—the SYNTHAR+ grapheme-phoneme system and the MBROLIGN alignment—are used to label the speech database recorded. Finally several contextually analysis of phonetic unit duration are conducted. Our studies on continuous speech confirm the results already obtained on isolated words—both the consonantal and final pause context effect and the consonantal gemination effect on the vowel duration. We will report also on the effect of the syllabic number on the speech rate.

51

Introduction

The current systems of (TTS) synthesis for Latin languages such as French, English, German, etc. are able to produce an intelligible and an almost natural speech synthesis (Dutoit, 1997).

For the French language, these qualities are the result of many basic researches dedicated for decades to the study of the prosody (Di Cristo, 1998), (Malfrère, 1998) and (Mertens, 1999), to quote only those. This quality is essentially due to the use of high-performance automatic learning and labeling tools for prosodic process as well as the availability of speech databases. (Bartkova, 1987) has defined a set of rules which determines the segmental duration according to syntactic and prosodic marker relative to the word, the syllable position in the word, the phoneme position in the syllable, etc.

Many works have studied the variation of sound duration for the Arabic language according to several factors: speakers, phonological phenomena, syllabic structures, consonantal geminating and Arabic dialects, etc. Among them, (Ghazali, 1992a) has considered the phonological levels for the Arabic TTS; (Jomaa 1994) has treated the opposition between short and long vowels in Arabic language whereas (Amrouche, 1998) has studied the variation of vowel duration according to the syllabic structures. It is important to mention that all these studies were mainly based on isolated and non-sense word corpora. A neural-network based model of Arabic syllable duration was presented by (Chehab, 2000) for the naturalness amelioration of Arabic TTS and (Zaki, 2000) has proposed a set of rules to describe the variation of the stylized F0 curves for interrogative synthesized sentences. (Zemirli, 1998) underlined the major role of the duration of sounds in the intelligibility rate of the Multivoix system. Few works are still dedicated to the prosodic generation for the TTS of the Arabic language.

The prosodic representation system consists in three prosodic parameters: fundamental frequency, duration and intensity. The sound duration is one of the most difficult to model. As reported by the works described above on isolated words, the duration depends on the contextual realization of phonemes characterized both by the nature, size and structure of syllables and the stress, etc. The aim of this paper is to study the effects of these parameters on continuous speech corpora useful to elaborate prosodic scheme for new generations of TTS systems.

This present work describes the approach adopted to predict the sound duration useful for the automatic rate generation in standard Arabic TTS system. It aims to quantify the effects of the left/right consonant context, the elongation of geminated consonants as well as the influence of the syllabic structure of sound duration on continuous speech corpora. Firstly, we will proceed with the elaboration of representative continuous speech corpora including all the phonotactical constraints for the Arabic language. Secondly, we will describe briefly two automatic tools used to align the phonetic units on the speech. Several extracted criteria based on syllabic structure and phonetic characteristics of the Arabic language are defined to compute the phonetic unit duration. These criteria are quite similar to those described in (Dutoit, 1997) but they are adapted to phonological and syllabic phenomena of the Arabic language.

Various results reporting on the effects of the syllabic structure, the geminating and the position



of phonemes on the variations of phonetic unit duration are then commented. Implementation and preliminary evaluation of the automatic generation model of the rhythm are discussed. Finally, perspective works are evoked.

Phonetic classification and syllabic structure

The aim of this paragraph is to describe briefly the phonetic classification of the Arabic language as well as a brief description of its syllabic structure. Arabic language is composed of 28 consonants and 6 vowels (3 short vowels opposed to 3 long vowels). One of the characteristics of the Arabic language in regard to Indo-European languages is the emphatic phenomena present within the 5 consonants noted **s**, **d.**, **t.**, **z.** and **q** in MBROLA formalism (Dutoit, 1997). All the consonants can be geminated.

Syllables in Arabic language obey to two rules: R1). The syllabic nucleus is a vowel; R2) Two consonants can not follow themselves, except before a pause. The following syllabic combinations are: A short and an open syllable CV **kataba**, a long and an open syllable CVV **maata** and three long and closed syllables: CVC : **maktab**, CVVC : **salaam**, CVCC : **bint**. All the syllables start with a consonant followed by a vowel, all the syllables comprise more than one phoneme, all the syllables comprise only one vowel, syllable CVCC is only at the end of the word or isolated and all the other syllables can be at the beginning, the medium or at the end of the word. A vowel at the word beginning is not admitted; it is considered as the realization of the consonant hamza followed by a vowel. The accentual system was the object of several works (Lecomte, 1975) and (Zakaria, 1984).

Description of the method

To model any sound duration in an Arabic TTS, a corpus is needed. The corpus used consists of 625 sentences representing more than 20722 phonemes. The sentences are composed of two to fifteen words without abbreviation neither acronym. They were pronounced by An Arabic male speaker, without dialectal accent, according to three instructions of continuous speech production rate: slow (<10 phonemes per sec), normal (10-13) or rapid (>13). They represent also three types of intonative forms: affirmative, interrogative and explanative modes. The corpus was acquired at 16 kHz and cleaned of long silence before the speech beginning or between words.

Sentences	625
Phonemes	20722
Words	3630
Average number of phonemes per words	5,7
Average number of words per sentence	5,8
Number f verbs	643
Number of particles	1108
Number of nouns	1879

Table 1.
Description
of the speech
corpus

From a diacritized Arabic orthographic string, the grapheme-phoneme system SYNTHAR+ (Zemirli, 1998) produced the phonetic string. Its phonetic outputs were compatible with the MBROLA's entries. The output's evaluation was 100% success rate.

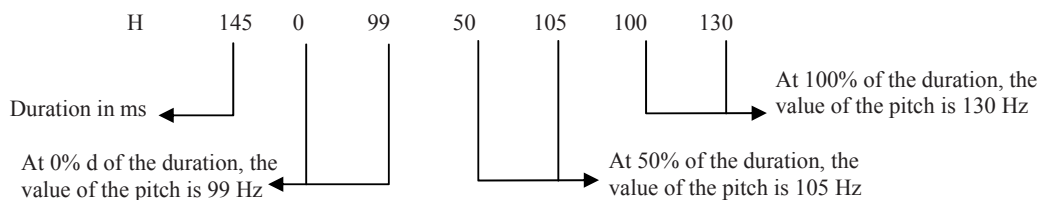
As next step, the phonetic alignment was made by means of the free tool MBROLIGN (available on the site www.tcts.fpms.ac.be/synthesis). The alignment was correct for all the sentences produced in the slow and normal modes of speech production. Adjustments of boundaries were realized for 5 % of phonemes, in several cases for long sentences (more than seven words) spoken in a rapid mode. Another checking concerns the geminated consonants where their durations are often reduced. Generation of the duration of the phonetic units

The word concept is different in the Arabic language compared to the French language or Indo-European languages. In Arabic, a word can be composed of numerous basic lexical entities (for more details, see (Zemirli, 1997). This phenomena is called agglutination. Affixed elements can be related to a word (suffixes and/or prefixes). With the exception of the particles any linguistic form can be analyzed in a root and a stem. Root and stem are the fundamental concepts of the Arabic language morphology. The decomposition of the word: **بِمَدْرَسَتِهِمْ** (in their school) produces (تَهُمْ | مَدْرَسَاتِ | بَ).

A MBROLIGN function computes prosodic information stored in a result file (it is a “.pho” file

which can be read by the system MBROLA). Each record of this file is associated to a phoneme (including the pause). Each phoneme (including the pause) is associated its wording, its duration and possibly a whole of couples of numbers representing the position of F0 expressed as a percentage compared to the total duration of the phoneme and the value of F0 to this position.

Modeling sound duration in an Arabic text to speech system



53

The following example is the prosodic output file for the sentence: “كَانَ فِيهَا سَرِيرَانِ كَبِيرَانِ”. In MBROLA formalism: /kaanafehaasariraanikabiraani / “There were two double beds inside.

Figure 1. Example of a line of the file “.pho” Generated with Mbro-lign

```

_ 200
k 74
aa 160 9 115 39 135 64 147 79 152 99 152
n 82 8 150 37 145 46 143 56 139 66 136 75 134 85 129 95 126
a 112 3 120 32 111 39 111 46 109 75 105 82 101 89 98 96 89
f 107
ii 150 11 137 16 139 32 145 48 146 53 149 69 149 74 148 80 146 85 146 90 145 96 143
h 70 2 141 14 139 25 135 37 134 48 130 60 129 71 125 82 124 94 122
aa 145 2 118 8 118 19 114 30 111 41 111 52 110 63 108 74 107 85 103 90 103 96 103
s 104
a 70 2 113 14 112 25 112 37 113 48 111 60 111 71 111 82 110 94 109
r 110 3 109 18 108 25 108 40 106 54 95 61 100 69 102 76 103 83 104 90 104 98 105
ii 190 11 107 20 108 36 105 45 105 53 104 62 102 70 100 78 100 91 99 95 98 100 98
r 70 11 98 22 97 34 94 45 90 57 99 68 100 80 102 91 106
aa 208 12 120 20 126 35 140 50 146 58 149 66 150 77 148 85 145 93 143 96 142
n 61 2 132 15 133 28 133 41 131 54 128 67 124 80 122 93 119
i 70 5 116 17 114 28 109 40 110 51 108 62 107 74 105 85 102 97 93
k 107
a 89 3 101 12 101 21 97 29 97 38 95 47 94 56 93 65 92 74 91 83 89 92 88
b 54 2 87 17 86 32 86 46 86 61 86 76 87 90 88
ii 199 13 93 17 94 25 95 37 95 49 94 61 92 65 91 77 90 81 90 85 90 89 90 93 89 97 89
r 100 4 90 12 90 20 89 28 89 36 88 44 72 92 100 100 103
aa 212 11 108 15 110 22 114 37 118 48 118 60 114 71 107 82 95 90 92 93 90 97 86
n 97 3 86 11 81 19 82 27 82 36 81 44 79 52 79 60 77 69 78 77 76 85 72 93 73
i 90 2 72 11 72

```

From these data structure obtained for all the sentences of the study corpora described above, a base of phonetic unit duration was computed. With each phoneme are associated, its contexts left and right, the position and the value of the first pitch (F0), the position and the value of the last pitch, the number of pitch (NP) and the value of the slope of F0 (+,-).

#ka\74\0,0\0,0\NP=0\+0\
kaan\160\9,115\99,152\NP=20\+38\
aana\82\8,150\95,126\NP=10\+24\
naʔ\112\3,120\96,89\NP=14\+31\
afii\107\0,0\0,0\NP=0\+0\
fiih\150\11,137\96,143\NP=18\+8\
iihaa\70\0,0\0,0\NP=0\+0\
haas\145\2,118\96,103\NP=18\+15\

aasa\104\0,0\0,0\NP=0\+0\
sar\70\2,113\94,109\NP=9\+4\
arii\110\3,109\98,105\NP=14\+4\
riir\190\11,107\100,98\NP=24\+7\
iiraa\70\11,98\91,106\NP=8\+8\
raan\208\12,120\96,142\NP=26\+31\
aani\61\2,132\93,119\NP=8\+13\
ni#\70\5,116\97,93\NP=9\+23\

From the same corpus and for each word, we built another base of information which contains: the nature of the phoneme (Consonant or Vowel), the slope of the phoneme, the grammatical category of the word (Noun, Verb or particle) and its Arabic writing and couples of initial and final values of the fundamental frequency of each phoneme.

(C VV+ C- V-) (VERB) (kaana) (كأن) (0,0) (114,152) (150,126) (120,89)
(C VV- C- VV-) (PART) (fiihaa) (فيها) (0,0) (135,143) (141,122) (118,103)
(C V- C- VV- C+ VV+ C- V-) (NOUN) (sariiraani) (سريران) (0,0) (113,109) (109,105) (105,98) (98,106) (111,142) (132,119) (116,93)
(C- V- C VV- C+ VV- C- V) (NOUN) (kabiiraani) (كبيران) (0,0) (101,88) (87,88) (93,89) (90,103) (108,86) (86,73) (72,72)

Study of the phonemes duration and preliminary interpretation

Our state of art had pointed that the phonetic unit duration is contextually and syllabically dependant. Our purposes are:

1. To study the effect of the left and right immediate context on the phoneme duration;
2. To measure the elongation duration of geminated phonemes;
3. To study the influence of the phoneme position in the syllable on its duration;
4. To establish a ratio between the size of the syllable and the duration of phonemes.

So, several extracting criteria were identified: the right and the left contexts defined in phonetic classes, the syllabic structure and the phoneme location in the syllable, and the accentual information.

We developed a model containing rule for the automatic prediction of the duration of the phonemes of the Arabic language. This model is a multiplicative type, we proceed in two stages: initially, the intrinsic durations of the all the phonemes were extracted starting from the corpus. In the second time, coefficients of reduction/elongation were calculated for each phoneme according to a certain number of criteria present below. The rules are form:

$$D_{Phon} = CRE \times D_{Int}$$

D_{Phon} represents the duration of the phoneme in context, CRE the coefficient of reduction / Elongation and D_{Int} the intrinsic duration of the phoneme.

The table 2 (without consideration of neither contextual information nor syllabic structure) consists of two columns: Column 2 corresponds to the minimal average (**MinAV**) duration computed on the not geminated phonemes which **duration is lower than 100 ms**; Column 3 shows the maximal average (**MaxAV**) duration of phonemes which are geminated phonemes e.g. a long vowel or a particular phoneme elongation. The next step was to verify the adequacy between the geminating information

given by SYNTHAR + and the phoneme durations (see table1). We can observe that geminated consonants have their duration doubled at least. Some preliminary perception tests have also confirmed the necessity of this elongation to distinguish between consonants and its corresponding geminating for listeners.

Modeling
sound duration
in an
Arabic text
to speech
system

Phoneme	MinAV	MaxAV
a	95	*
I	85	*
L	64	132
U	94	*
N	69	172
Aa	*	160
M	68	166
T	73	147
R	68	156
J	63	120
B	72	145
W	60	125
H	87	180
?	71	135
k	78	143
s	94	163
h	78	150
d	66	125
ii	*	200
f	70	135
q	80	151
X	97	159
Z	86	130
a.	74	165
D	72	150
s.	77	149
z	80	144
d.	80	160
G	82	160
i.	63	154
x	88	141
uu	*	160
z.	76	156
S	87	175
T	92	170
t.	85	130
u.	71	151

* The duration of short vowels is always lower than 100 ms. The duration of the long vowels is always greater than 100 ms. This explains the not appearance of their value in these cases.

Contextual interpretation

We defined two coefficients of reduction CR1 and CR2 of the intrinsic duration (Dint) of the vowel in the syllable types, CVC (**maktab**) and CVCG (**sabbaqa**), where CG is a geminated consonant, with regard to the duration of the vowel in the syllable of type CV. The table 2 allows noticing that the coefficient of reduction CR2 is greater than the coefficient of reduction CR1. New duration Dphon is equal to Dint* (CR1 or CR2) according to the type of syllable. The vowel durations are so more reduced in the CVCG syllables than in the CVC syllables. The values of CR2 obtained on our corpus confirm those of (Amrouche, 1998), who mentioned the reduction of vowel duration in front of a geminated consonant. Nevertheless this coefficient of reduction is more important for isolated words than for continuous speech. It is also necessary to underline that CR1 and CR2 depends on the nature of the vowel.

Phoneme	CV	CVC	CR1	CVCG	CR2
a	97	83	0.86	72	0.74
u	98	79	0.81	69	0.70
i	87	68	0.78	60	0.68

Table 2.
Average
phoneme
durations of
the corpora
(DMi, i is the
indice which
identifies the
phoneme;
this notation
will be reused
above)

Table 3.
Average
duration of
brief vowels
according to
the type of
syllable

The table 4 shows an elongation of the duration (D) of phonemes (in comparison to those of the Table 1) placed at the end of word and before a pause, even for vowels when they are voluntarily pronounced (vowel is usually omitted at the word end before a pause). This increase is due to an effort made by the speaker for the pronunciation of the vowel in front of a break. The phoneme n is a particular case; its duration is more than doubled. After a finer study on this phoneme we noticed that this strong elongation is verified when it marks the tanwin mark). CEi represents the ratio of elongation of the phoneme in front of a break.

Table 4.
Average duration and elongation ratio for some phonemes at the end of word and in front of a break

Phoneme	CEi	D	Phoneme	CEi	D
a	1.25	120	m	2.5	170
u	1.22	115	h	2.0	150
i	1.29	110	q	1.37	110
s	1.8	170	n	2.9	200
f	2	160	Z	1.5	129

The Table 5 illustrates some examples of relation between the number of syllables and the duration of the word phonemes. SC (syllabic Coefficient) is the average for all SCi (SCi is the reduction factor associated to each phoneme).

Table 5.
Reduction factor of the duration for some phonemes according to the number of syllables of the word

Number of syllables	1	2	3	> 3
SC (syllabic Coefficient)	1	0.96	0.92	0.87
Phoneme l	70	67	64	60
Phoneme q	90	86	83	78
Phoneme m	72	69	66	62
Phoneme Z	86	83	79	75

The vowels represent the core of the syllables. Table 6 represents the duration of the short vowels and the short vowels in stressed syllables and not accentuated. The standard deviation between the stressed syllables and non stressed syllables in the short vowels is 1.41 and that between the long accentuated and not accentuated vowels is 2.12. These weak variations make it possible to affirm that the duration of the vowels does not seem to be affected by the stress.

For the perception of the lexical stress of Arabic, the fundamental frequency is thus much more relevant than the duration. The accent will be characterized by the maximum of the intonate contour (F0) of the stressed syllable.

Table 6.
Duration of vowels in stressed syllables and not stressed syllables

	Non stressed syllables	Stressed syllable	Standard deviation
a	93	96	1,41
u	95	93	1,41
i	87	84	1,41
aa	188	185	2,12
uu	187	184	2,12
ii	201	197	2,12

Processing and evaluation

To evaluate the effectiveness of the SC (syllabic Coefficient), CR1 or CR2 and CE parameters, we proceeded to the automatic generation of the sound duration on a new corpus (EVAL0: 40 sentences, pronounced by the same Arabic speaker, which contains 1200 phonemes) according to the following steps:

1. Automatic transcription of the corpus EVAL0 by means of the SYNTHAR+ system and calculation of the following parameters, SC for every word and CRi or CEi for phonemes (by default SC=CRi=CEi=1);
2. Automatic generation of sound duration (Di) as follows $D_i = DM_i * SC * (CR_i \text{ or } CE_i)$;
3. Automatic generation of the sound duration by means of MBROLIGN;
4. Comparison of duration generated by MBROLIGN (step 3) with those produced at the step (2). The analysis of the comparison shows that the difference is around distance 12.02 ms for vowels and 7.78 ms for consonants.

The values of the short vowels are lower than those proposed by (Mrayati, 84) which locates the interval of the short vowels between 100 and 150 ms and with those of (El-Ani, 1970) which presents values going up to 300 ms. They are on the other hand close to those provided by (Amrouche, 1998) and (Chehab, 2000).

We note that the durations of the vowels in a context CV are higher than their equivalent in a context CVC. This phenomenon was observed in several Arabic works (Ghazali, 1992b) (Amrouche, 1998).

The duration of a vowel is less low in a closed syllable than in an open syllable. This is truer if the following phoneme is geminated.

We note that the duration of the consonants C1 in a context C2 is higher than that of its equivalent in a context C1V.

Conclusion and perspectives

At the end of this study, we constructed a sound duration database. We have added three parameters CS (coefficient of syllabic reduction), CR (coefficient of phonemic reduction) and CE (coefficient of phonemic elongation) for all standard Arabic phonemes.

We noted a standard deviation of 12.02ms for the vowels and 7.78ms for the consonants between the durations calculated by our model and the real durations of the corpus. This weak variation seems to confirm the robustness of our model of generation of the phonemic durations.

Actually this module of generation of phonemic duration is used efficiently in an Arabic text to Speech System called ARAVOICE (Zemirli, 2004).

We are trying to complete this module who will supplement the intonate model already developed in (Zemirli, 2004) for the generation of many intonates contours: joy, anger, sadness, fear, perseverance, Astonishment.

References

- Amrouche, A., Boudraa, B., Rouvaen, J.M. (1998), "Organisation temporelle des voyelles dans les structures CVCVCV, CVCCVCV et CVCCCV de l'Arabe standard", JEP'98, pp. 91-94.
- Bartkova, K., Sorin, C. (1987), "A model of segmental duration for speech synthesis in French, Speech Communication", pp. 245-260.
- Chehab, A., Zaki, A., Rajouani, A. (2000), "Un Modèle Neuronal pour la Prédiction de la Durée des Syllabes de la Langue Arabe", Actes des XXIIIèmes Journées d'Etude sur la Parole, Aussois, pp. 95-98, juin 2000.
- Di Cristo, A., Di Cristo, P., Véronis, J. (1998), "Optimisation d'un modèle prosodique pour la synthèse par règles à partir du texte du Français", JEP'98, pp. 135-138.
- Dutoit, T. (1997), High-Quality Text-to-Speech Synthesis, Kluwer Academic Publishers.
- El-Ani S. (1970), Arabic Phonology : An Acoustical and Physiological Investigation, The Hague, Netherlands : Mouton, 1970, "Janua Linguarum" Series Practica 61. Translated into Arabic, 1983.
- Jomaa, M. (1994), "L'opposition de durée vocalique en arabe : Essai de typologie", JEP'94, pp. 395-400.
- Ghazali, S., Znagui, M., Benmiled, Z., Jemni, H. (1992a), "Synthèse de l'Arabe standard à partir du texte par TD_PSOLA : Le traitement des processus phonologiques", JEP'92, pp. 89-93.
- Ghazali, S., Brahim, A. (1992b), « Voyelles longues et voyelles brèves en arabe standard : Organisation temporelle », Actes des 19ème Journées d'études sur la parole, Bruxelles, pp. 153-154.
- Lecomte, G. (1975), La Grammaire de l'arabe, P.U.F, Paris.
- Malfrière, F., Dutoit, T., Mertens, P., (1998) "Un générateur de prosodie Tout Automatique", JEP'98, pp. 147-150.
- Mertens, P. (1999), "Un algorithme pour la génération de l'intonation dans la parole de synthèse", TALN'99, Cargèse, pp. 233-242.
- Zakaria, A. (1984), "L'accent Arabe Soudanais", Thèse de 3ème Cycle, Université de Franche-Comté, Besançon.
- Zaki, A., Rajouani A., Najim M. (2000), "Contours intonatifs de la phrase interrogative en arabe", JEP'00, pp. 249-252.
- Zemirli, Z. (1997), "Modélisation des règles phonologiques dans un système de traitement automatique de la langue arabe", JST'97, Avignon 15-17, pp. 361-368.
- Zemirli, Z. (1998), SYNTHAR+ : Synthèse vocale arabe sous Multivoix, TSI Vol17, N°6/98, pp. 741-761.
- Zemirli, Z., S. Khabet, (2004), « Synthèse Vocale Arabe sous MBROLA », CARI'04, 6ième Colloque Africain sur la Recherche en Informatique – Tunis, 22-25 Novembre 2004.

