G. Droua-Hamdani
Laboratoire TAP CRSTDLA, 1, Djamel Eddine Al Afghani Bouzaréah Alger, Algérie.
M. Guerti
Ecole Nationale Polytechnique BP 182-El Harrach 16200 Alger, Algérie.

Application of Model to correct the Phoneme's Duration produced by TTS System -case of Arabic Voiced phonemes -

Abstract

To approch a human voice, some modifications on speech signal produced by TTS systems are necessary. As we know, the relevant parameters of the vocal signal are: the fundamental frequency, energy and duration. In our work, we have studied the duration of segmental phonemes. In fact, we tried to reduce the different imperfections which depend on duration of voiced consonants having unvoiced oppositions of arabic language. For that, we measured first different coefficients of variation (position and context), and after, we constructed our model of correction. We tested this model to modify the duration of some phonemes (voiced) produced by Text-to-Speech system.

Key words: segmental duration, TTS, arabic language, voiced consonants

1. Introduction

In this article, we present a contribution of phonemics' duration modelling of the Arabic voiced consonants having unvoiced oppositions and its application to speech system synthesis. Our predictive model is based on that D. Klatt system [1].

Spoken is a complex process and its analysis remains a central element in the realization of the human language. In fact, the relevant parameters of the vocal signal are: the fundamental frequency, energy and duration. The phonemic duration depends on several factors of variability intra and interlocutors, like: the co-articulation (the influence of a sound on the contiguous sound), the sex (man/woman), the age, the regional accent, flow of elocution, the emotional state of the speaker, etc [2]. Several studies (models and techniques) in the duration domain were made for the Latin languages but for Arabic very few works were realised [3] [4] [5] [6]. But it is not the case for Arabic language where we noticed few works only [7] [8].The interest by studying this parameter for the Arabic is justified double because the duration has in important position in the Arabic linguistics. In fact, it is considered as being a main property of distinction between words (gemination and madd)

2. Effect of duration in Arabic sound system

The sound system of the Standard Arab consists of six vowels: three short vowels [a], [u] and [i] respectively [fetŸa], [‡amma] and [kasra], and three long ([ā], [ū] and [ī]), and twenty eight consonants but, in our work, we were interested only on the voiced consonants having unvoiced oppositions.

Among the characteristics of the Standard Arabic, we have in addition to the [madd] (long vowels), the emphasis and the geminating. The emphasis represents the pharyngalization which occurs at the time of the contraction of the higher part of the pharynx to produce a second place of articulation and then, four specified phonemes ([©], [®], [‡] and ¼]). The gemination, as for it, also called redoubling, corresponds to the phenomenon of reinforcement of the consonant articulation. Set a part [hamza], all Arabic consonants may be geminating.Tab.1 represents some examples for each characteristic.

قَتَلَ قَاتَلَ	[qatala] [qa:tala]
حَمَام	[Ÿama:m]
حَمَّام	[Ÿamma:m]
سَلَبَ	[salbun]
صَلَبَ	[©albun]

Tab.1: Examples of Arabic characteristics

3. Methodology

The corpus worked out for the study of voicing consists of 100 sentences whose flow of elocution is normal. The takings away were made on sequences of type [C1V1C2V2C3V3] with $C_i \equiv$ consonant and $V_i \equiv$ vowel. All consonants are taking from a vocalic context of the three short vowels with three different positions (Initial (I), Middle (M) and Final (F)). The material used for the recordings data is: the CSL (Computerized Speech Lab Model 4300B) of Kay and the technique used for the analysis is the sonnagraphic one. The selected sampling rate is 11025 Hz samples coded on 16 bits.

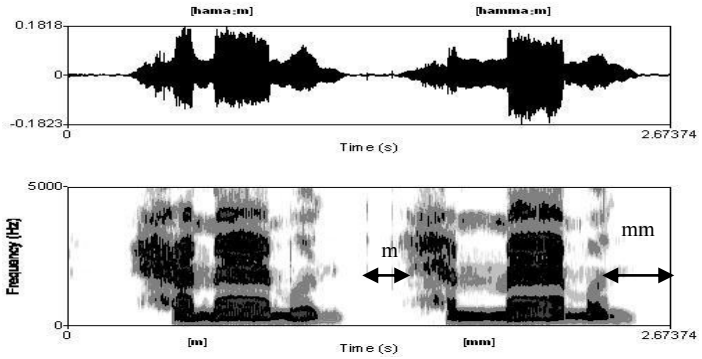

Figure.1: Representation of geminating ([Ÿama:m]/ [Ÿamma:m])

4. Analyze of segmental durations for voiced Arabic consonants

To measure the effect of voicing over the durations, we chose representative consonants i.e. which form minimal pairs. So we recorded a corpus of sentences based on the voiced oppositions voiced/ unvoiced. On the seven pairs of existing oppositions of voicing in the Standard Arab, we studied six of them which are: [z]/ [s], [d]/ [t], [‡]/ [®], [†]/ [±], [‰]/ [¡], [`]/ [Ÿ]. For that, we obtained a whole of 40 values per consonant voiced according to its position in the word and of the vocalic context. The study applied was repeated for corresponding unvoiced consonants.

4.1. Inherent duration of the voiced and unvoiced phonemes

The intrinsic (inherent) duration of the consonants is measured in initial position of the sequences. The aim set by this is the reduction of the phenomenon of the co-articulation which represents the influence of phoneme on other. Tab.2 presents the average of the inherent durations for each consonant voiced like that of its corresponding item unvoiced. We notice from the results that the segmental duration of the unvoiced phonemes is always higher compared to their correspondents. The Coefficient of Reduction of Voicing (CRV) represents the reduction of the voiced consonant length compared to unvoiced ones.

unV [C]	D_{inh}(ms)	V [C]	D_{inh}(ms)	CRV
[®] ط	118,16	[‡] ض	104,70	0,88
[s] س	137,89	[z] ز	101,45	0,73
[t] ت	110,98	[d] د	105,19	0,94
[¡] خ	122,63	[‰] غ	72,70	0,59
[Ÿ] ح	125,56	[‘] ع	66,70	0,53
[±] ث	126,21	[‡] ذ	100,11	0,79

Tab.2: Inherent durations for voiced/ invoiced consonants

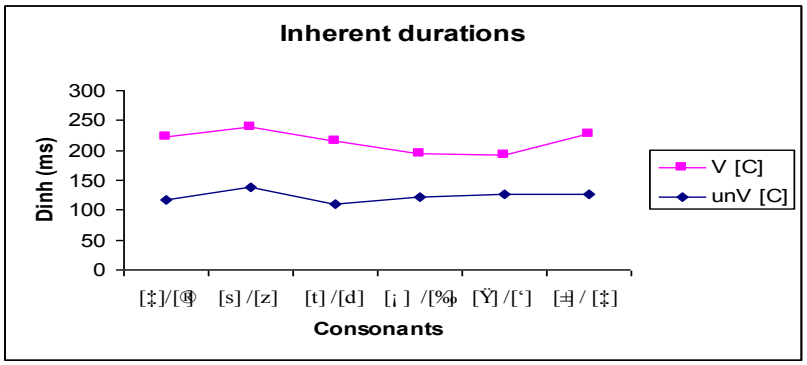

Figure.2: Representation of inherent duration

4.2. Variations of the consonant's duration of the according to the vocalic context and the position in the word

To study the variations of voicing according to the vocalic context right, we calculated the modifications of the durations generated by the change of the short vowels (3). The Coefficients of Reduction/Lengthening (CRV1 and CRV2) depending respectively on the vowel ([u] and [i]) are showing in Tab.3 (the vowel [a] is taking as the reference). For the position, we calculated Coefficients of Reduction/lengthening of Position (CRP1 and CRP2) which represent the modifications when we change the positions of the consonant in the sequence, for that we suppose three locations (Initial (I), Middle (M) and Final (F)). In this case, we suppose for reference the first position.

C	CRV_1	CRV_2	CRP_1	CRP_2
[z]	1,06	1,01	1,06	1,01
[‡]	1,06	1,04	1,06	1,04
[d]	1,04	1,02	1,04	1,02
[†]	1,04	1,02	1,04	1,02
[']	0,97	0,99	0,97	0,99
[‰]	0,96	0,87	0,96	0,87

Tab.3: Different coefficients of variation (vowel context and position)

5. Proposition of a modelling of voiced Arabic consonants

In our case, we are going to apply these ratios through a modelling to correct the values of the voiced consonant's duration generated by a speech system synthesis (TTS) Arphon conceived in the CRSTDLA [9]. In fact, we are going to predict the correct value by using a model of prediction inspired by that D. Klatt one []. By taking into account various analysed coefficients studied before, we built a system of rules. The general equation of the system is:

$$Dur_{cor} = CR(c_i) \times CRP(p_1, p_2) \times CRV(v_1, v_2) \times Dur_{inh}(c_i)$$

Dur_{cor} : corrected [C] duration;
Dur_{inh} : inhérent duration[C] ;
CR : coeff. of reduction of voicing
CRP: of reduction/lengthening depending on position;
CRV: coeff. of Reduction/lengthening depending on vocalic context.

6. Application of the model in an artificial speech

To verify the rates of variations obtained during the phase of analysis (the various coefficients), we proceeded to a comparison between phonemes duration's taken first time from a natural sentence and in second time from an artificial sentence (generated by a TTS system) (figure.3). The synthesized sentence was produced by an automatic Arabic TTS (Arphone2.0). We notice from sonagrams of both examples a net difference in the duration of the consonant [‡]. When, we used the rule of voicing to correct the duration of the consonant [‡] of the word [Ÿa‡ara], we obtain the results showed in figure 4.

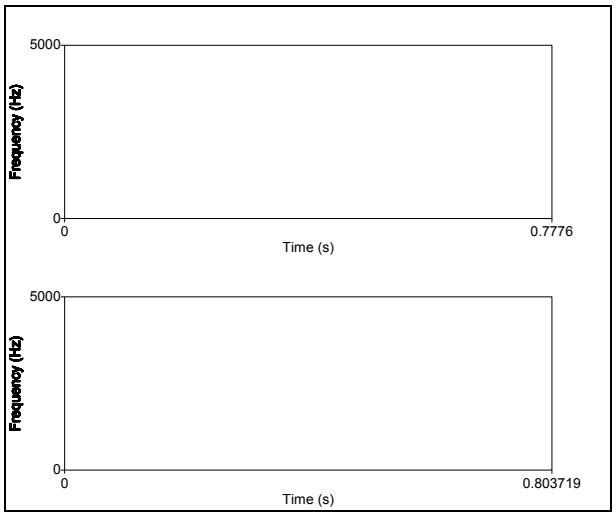

Figure.3 : Duration of [‡] produced par TTS system and naturally in the word [‡araba]

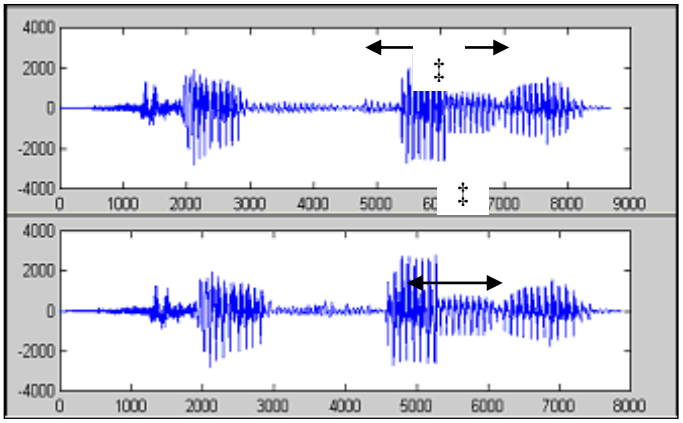

Figure.4: Correction of duration of the voiced consonant [Ÿa‡ara] produced by Speech System Synthesis

7. Conclusion

To correct the duration of voiced consonants produced by a Arphon system (TTS system), we measered first the different coefficients of variation of duration. In fact, by taking into account the position in the word (initial, middle and final) and the vocalic context ([a], [u] and [i]), we calculated different ratios which permitted to us to construct a model of prediction. By using this model we correct the duration of voiced phonemes (reduction or lengthening) and then approach the reel value of duration (human voice).

References

[7] Abbas, M., Benbellil, K., Droua-Hamdani, G., Ferrat, K.: 2002, *Lecture automatique de textes et chiffres en arabe* Standard. Symposium international sur le traitement automatique de l'arabe, Université de Manouba, Tunisis, Tunisie, 18-20 avril 2002

[5] Barbosa, P.A.: 1994, *Caractérisation et génération automatique de la structuration rythmique du français par apprentissage automatique*. Thèse de doctorat 1994.

[3] Barktova, K, Sorin, C.: 1987, *A model of segmantal duration for speech synthesis in French***.** Speech communication 6, pp.245-260, 1987.

[2] Calliope: 1989*,* La parole est son traitement automatique*, Masson, Paris France.*

[8] Chenfour, N., Benabbou, A., Mouradi, A..: 2000, *Elaboration du dictionnaire de di-syllabe pour un système TTS arabe par concaténation*, Conférence Maghrébine MCSEA l'2000 sur les sciences informatiques, 1-3 novembre, 2000.

[7] Colotte, V., Laprie, Y.: 2002, *Amélioration de la précision de la synthèse avec TD-PSOLA*, XXIV^{ème} journées d'études sur la parole, Nancy (France), 24-27 juin 2002.

[1] Klatt, D.H.: 1976, Linguistic *uses of segmental duration rules in real-speech data base*. Phonetica N°**5**, may, 1976.

[4] Santen, van, R.P.H.:1997, Prososdic Modeling in text to speech synthesis, Proceedings of Eurospecch 97, pp.2455-2458, Rhodes, Greece.

[6] Stylianou, Y.: 2001, *Applying the harmonic plus noise model in concatenative speech synthesis*, Transaction on speech and audio processing, vol. **9**, No. 1, IEEE, January 2001.

[6] Zemirli, Z., Vigouroux, N.: 2000, *Vers une modélisation de la durée des sons pour la génération automatique du rythme dans la synthèse de la langue arabe*, XXIIIèmes journées d'Etudes sur la parole, Aussois, 19-23 juin, 2000.