F_0 Contour Modeling for Arabic Text-to-Speech Synthesis Using Fujisaki Parameters and Neural Networks

Zied Mnasri zied.mnasri@gmail.com
Ecole Nationale d'Ingénieurs de Tunis
Electrical Engineering Department
Signal, Image and Pattern Recognition Research Unit
University Tunis El Manar
Tunis, 1002, Tunisia

Fatouma Boukadida fatoumaboukadida@yahoo.fr
Institut Supérieur des Technologies Médicales
Electrical Engineering Department
University Tunis El Manar
Tunis, 1002, Tunisia

Noureddine Ellouze noureddine.ellouze@enit.rnu.tn
Ecole Nationale d'Ingénieurs de Tunis
Electrical Engineering Department
Signal, Image and Pattern Recognition Research Unit
University Tunis El Manar
Tunis, 1002, Tunisia

Abstract

Speech synthesis quality depends on its naturalness and intelligibility. These abstract concepts are the concern of phonology. In terms of phonetics, they are transmitted by prosodic components, mainly the fundamental frequency (F_0) contour. F_0 contour modeling is performed either by setting rules or by investigating databases, with or without parameters and following a timely sequential path or a parallel and super-positional scheme. In this study, we opted to model the F_0 contour for Arabic using the Fujisaki parameters to be trained by neural networks. Statistical evaluation was carried out to measure the predicted parameters accuracy and the synthesized F_0 contour closeness to the natural one. Findings concerning the adoption of Fujisaki parameters to Arabic F_0 contour modeling for text-to-speech synthesis were discussed.

Keywords: F_0 Contour, Arabic TTS, Fujisaki Parameters, Neural Networks, Phrase Command, Accent Command.

1. INTRODUCTION

TTS systems have known much improvement with a variety of techniques. However, naturalness is still a troublesome aspect, which needs to be looked after. In fact, naturalness is too large as a concept; it may be related to the speech synthesizer, which is required to produce an acoustic signal matching as closely as possible to the natural waveform, or to the listeners, who react perceptually to the sound they hear [1].

In both cases, it's prosody which is responsible of the naturalness quality. Prosody includes the underlying features spanning the speech segments. Again, there is a twofold definition of prosody, according to the adopted viewpoint. Thus, phonologically speaking, prosody stands for stress, rhythm and intonation. These terms describe the cognitive side of speech. Translated into phonetics, these abstract definitions are quantified by the signal's amplitude, duration and F_0 contour [1].

The latter feature, i.e. F_0 contour, is the physical transmitter of the acoustic information, whether linguistic, para-linguistic or non-linguistic. Linguistic information includes the lexical, syntactic and semantic data present in the text, while para-linguistic data describe the speaker's intention, attitude and style. Finally, non-linguistic information is related to his physical and emotional state [2].

All these sides, acting simultaneously to produce speech, need a model able to re-synthesize the F_0 contour transmitting them. Nevertheless, modeling F_0 contour is subject to many constraints, according to the approach to be used. Thus the way these data are dealt with is responsible of the modeling strategy, either in a timely manner, i.e. sequentially, or in a parallel and super-positional manner.

Besides, the presence/absence of parameters in the adopted model is a key index for its applicability. In fact, non-parametric modeling, in spite of its simplicity, is of lesser relevance than parametric modeling, which provides the opportunity of multi-language applicability.

For instance, the Fujisaki model is a super-positional model, inspired from the early works of Ohman [4], and developed to provide an analytical description of the phonatory control mechanism, through the introduction of three basic concepts:

1. The baseline frequency F_b
2. The phrase command
3. The accent command

Then the overall F_0 contour is calculated in the logarithmic domain as the superposition of the aforementioned concepts.

Since Fujisaki model is parametric, the main task is to measure its parameters. This can be done either by approximation, using the analysis-by-synthesis technique, [5] and [6], or by prediction, [22] and [23].

Amongst the prediction techniques, neural networks are famous for their generalization power, through capturing the latent relationship between an input set and the matching outputs, to be able to guess the value of any new coming sample. Nevertheless, supervised learning, i.e. specific and separate input and output sets, is highly recommended to ensure a faster convergence of the neural networks [7].

In the framework of Arabic TTS synthesis, we opted for a parametric tool, i.e. the Fujisaki model, to generate synthesized F_0 contours after the analysis of a phonetically balanced Arabic speech corpus.

Hence, we started by extracting the Fujisaki parameters from our corpus, using Mixdorff's tool [8]. Then neural networks were used to train a learning set, covering 80% of the corpus to predict the parameters related to the test set.

In this paper, we start by defining the different levels of intonation modeling, to locate the Fujisaki model and describe its components. Then, after a short description of our corpus and the extraction method, the selected neural architecture is explicitly shown, with the various involved phonological, contextual and linguistic features. Finally, synthesized F_0 contours and original ones are compared using statistical coefficients, and the synthetic parameters are discussed.

These instructions are for authors of submitting the research papers to the International Journal of Computer Science and Security (IJCSS). IJCSS is seeking research papers, technical reports, dissertation, letter etc for these interdisciplinary areas. The goal of the IJCSS is to publish the most recent results in the development of information technology.

2. INTONATION MODELING

2.1 Phonological vs. Phonetic Models

Phonological models attempt to explain the intonation system of a language through a set of principles of intonation and the relationship between intonation processes and the linguistic units, mainly syllables. This implies that phonological models are not directly related to the physical waveform, they are rather a symbolic representation.

In contrast, phonetic models focus more on the alignment of intonation movements with phonetic segments and their temporal location, which is a physical representation [1].

For example, the Tilt model is a phonetic intonation model using a continuous description of F_0 movements based on acoustic F_0 data [28] while ToBI model [3], is based on a linguistic survey which divides the speech tones into phonological categories having each its own linguistic function [9].

Then, F_0 pattern is inferred from the relationship between the phonological sets ad their assigned linguistic functions. However, phonological categories may linguistically interfere, causing mutual interaction between the linguistic functions. This phenomenon may seriously affect the model's results [10].

2.2 Rule-based vs. Data-driven Models

The rule-based models emphasize on the representations which capture maximal generality and focus on symbolic characterizations. This is completely compatible with the human way to produce speech, which variability depends on these abstract representations. But linguistics focus on the cognitive aspect of speech at the expense of data itself, which is, actually, the visible aspect.

On the opposite side, speech data-driven modeling doesn't require a close interaction with linguistics. Certainly some linguistic rules have to be considered or used to extract meaningful data, but do not interfere in processing, and do not impose major constraints on the output. This approach has proved that such models can explore areas that linguistics cannot reach, and give answers that they haven't found, thanks to its high-level computational processing, and also because of the difficulty to simulate the phonological and cognitive rules related to speech production.

Although these constraints can be modeled by using uncertainty, in the form of probability characterizations, it is still a wide and deep area to be explored, looking to the various and complicated interactions lying in [1].

In the case of intonation modeling, the rule-based modeling was used to generate F_0 by targeted interpolation [3]. Actually, linguistic rules are first set to define the target functions of every syllable structure. This function allows them to place the pitch target in the right range, between the top and base values [11].

Data-driven models are based on a phonetic and prosodic segmentation and labeling of the speech corpus. This data is used to predict either F_0 movements or F_0 values. For example the Tilt model is used for English pitch modeling using accent marks [11]. Besides, data-driven methods don't need a deep linguistic exploration and therefore are more adapted for statistical learning. Then they can be used to predict the pitch pattern using either a parametric representation or not.

2.3 Superpositional vs. Sequential Models

Superpositional models are built upon the idea that F_0 contour can be seen as the result of the interaction of many factors at different levels of the utterance, such as the phoneme, the syllable, the word...etc. Thus instead of processing the F_0 contour as a whole, the study can be split into many sub-models dealing each with a particular level, to be combined later to generate the desired F_0 contour.

Sequential models stand on the other edge. They aim to generate F_0 values or movements either directly or by means of parameters, but in both cases, they rely on a sole model moving from the

beginning to the end of the utterance. Hence, the components of the F_0 contour are generated together at any particular instant of speech [9].

2.4 Parametric vs. Non Parametric Models

In intonation modeling, parameterization consists in transforming the original F_0 values into some parametric forms. Hence, instead of predicting the F_0 values, it would be enough to predict the values of its parameters, to re-synthesize the F_0 contour.

In contrast, the non-parametric approach consists in estimating the F_0 values directly from a set of features. Though its simplicity and its direct scheme, the latter method provides equivalent results when compared to the first one. Actually, the F_0 values are considered as meaningful and intrinsic linguistic parameters, and thus, predictable from linguistic features.

However, non-parametric modeling proceeds directly, discarding the hierarchy and especially the interactions of the various input features. Hence, the components of the F_0 contour are ignored and its movements are neglected at the expense of its values. So forth, a post-processing is required to ensure the smoothness of the estimated F_0 contour.

This post-processing action is not required while parametric modeling, as the prediction of each parameter is made in a local level, e.g. the syllable, then smoothness is inherently processed. Nevertheless, care should be taken in the parameterization process, as too many parameters may provide better prediction accuracy, but on the other hand, may cause the loss of linguistic meaning [12].

3. FUJISAKI INTONATION MODEL

This is a phonetic data-driven superpositional and parametric model for intonation. It starts from a physiological interpretation of the intonation to provide an analytical description in the logarithmic domain through the introduction of some theoretical concepts such as the accent and the phrase commands [2].

Thus, the F_0 contour is considered as the response of the mechanism of the vocal cord vibration to the accent and phrase commands. So forth, the Fujisaki model gives a twofold description of the intonation:

1. A physiological and physical description of the phonatory control mechanisms through the shapes of phrase and accent components.
2. An analytical description through the magnitude, the timing and the superposition of the aforementioned commands.

3.1 Physiological Description

Fujisaki applied the Buchthal & Kaiser formulation of the relationship between the tension, *T*, and the elongation of skeletal muscles, *x*, to the vocal muscle

$$\begin{aligned} T &= a.(e^{bx} - 1) \\ &= a.e^{bx} \quad if \quad e^{bx} >> 1 \end{aligned} \tag{1}$$

and the relationship between the vibration frequency of elastic membranes and their tension

$$F_0 = C_0.T^{\frac{1}{2}} \tag{2}$$

to yield

$$Ln(F_0) = \frac{b}{2x} + ln(C_0.a^{\frac{1}{2}}) \tag{3}$$

Actually, the passage to the logarithmic domain is helpful to achieve a linear superposition, and so forth, a decomposition of the model into accent and phrase components. The constant $(C_0.a^{1/2})$ refers then to the baseline frequency $Ln(F_b)$ which is constant during each utterance.

As the elongation of the vocal muscle, x, is associated to the tension of the glottis, a further decomposition of the movement of the glottis into a rotation around the cricothyroid joint and a translation of the thyroid against the cricoids allows introducing the concept of the accent

component which refers to the rotation of the glottis, and the phrase component describing its translation [13].

3.2 Analytical Description

The Fujisaki model describes F_0 as a function of time in the logarithmic domain by achieving a linear superposition between:

1. The baseline frequency, which doesn't alter along the sentence
2. The phrase component
3. The accent component

The phrase and accent components are the outputs of 2 second-order linear systems, called the phrase and the accent commands [14]:

$$Ln(F_0(t)) = Ln(F_b) + \Sigma_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \Sigma_{j=1}^{J} A_{aj} [G_a(t - T_{1j}) - G_a(t - T_{2j})] \quad (4)$$

Where

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & if \quad t > 0 \\ 0 & else \end{cases} \quad (5)$$

$$G_a(t) = \begin{cases} min(1 - (1 + \beta t).e^{\beta t}, \gamma) & if \quad t > 0 \\ 0 & else \end{cases} \quad (6)$$

The parameters A_p, T_0, A_a, T_1, T_2, α, β and γ are called the Fujisaki parameters.
As inferred by the formulation of $Ln(F_0)$, F_b denotes the asymptotic value of F_0 in absence of accent commands. Furthermore, it is proved that F_b is highly correlated to the mode of the sentence. It has higher values in direct Yes/No questions than in declarative statements [15].

1. **The Phrase Component**
   The phrase control mechanism is a second-order linear system whose impulse response is stated in (5).Then the output impulses are defined by their magnitude Ap and onset time T0. The parameter α is constant during an utterance. Hence, Ap describes the declination degree in an utterance, and therefore cannot be subject of comparison between different types of utterances.

2. **The Accent Component**
   The accent command is also a second-order linear system whose step-response is stated in (6). Then, the accent command introduces a magnitude Aa, an onset time T1 and an offset time T2. Besides, the parameter β is constant during an utterance. The same for γ, which is fixed to a ceiling value of 0.9 to ensure that the accent component will converge to its maximum in a finite delay. As T2 is usually higher than T1, the variation of F0 is proportional to the accent component magnitude Aa, which was extended to the negative domain to be able to apply the model to many other languages [15].

3. **F0 Contour Analysis**
   In order to obtain an optimal approximation of the F0 contour, the analysis by synthesis of the natural F0 contour is applied [6]. This is done by modifying the input commands of the model until:
   - The F0 contour is approximated
   - The result is linguistically interpretable

These constraints are either linguistic, describing the relationship between linguistic units and structures, or paralinguistic, dealing with phrase and accent components.

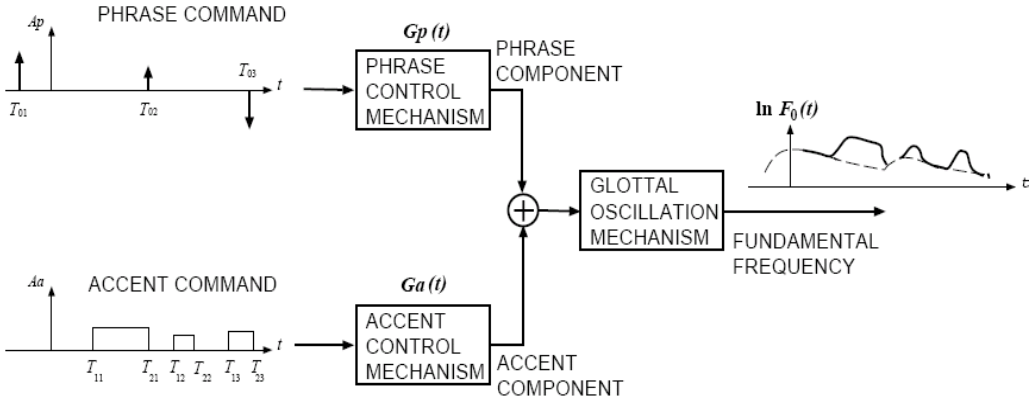

FIGURE 1: Fujisaki model components [4]

3.3 Motivation to use the Fujisaki model for Arabic

The Fujisaki model sound physiological background has been a great asset to its application to other languages. In fact, the model is based on the analytical formulation of the larynx movements, which are more related to the phonatory control system than to the linguistic rules. For instance, a polyglot speaker is able to produce, with his own larynx, the same intonation patterns as a native speaker [14].

Furthermore, the model can be modified to meet the specifications of the language to model. In fact, after its success in modeling, first Japanese [16], then English F_0 contours [36], it was adapted to many languages of different etymologies, Nordic (German and Swedish), Latin (Spanish and Portuguese) or south-east Asian (Chinese and Thai). This large use of the model brought many benefits to its primary formulation. For example, the falling accents can be modeled by introducing negative accent commands A_a.

Hence, the modified characteristics reveal the specific prosodic constraints of the language, while the invariable properties are mainly related to the speaker, regardless his language. Therefore, the Fujisaki model is potentially able to model the Arabic intonation.

4. SPEECH MATERIAL

4.1 Speech corpus

For this survey, we used a 200-Arabic-sentence corpus recorded by a male voice, with a 16-Khz sampling rate and 16-bit encoding, including the entire Arabic alphabet, composed by 28 consonants, 3 short vowels and 3 long vowels. In addition, amongst the 6 types of Arabic syllables, the most used ones are present in the corpus, i.e. /CV/, /CVV/, /CVC/ and /CVVC/ [29].

Corpus level	Type	Quantity
Sentence	Declarative	160
	Interrogative	20
	Imperative	20
Syllable	Opened and short /CV/	721
	Opened and long /CVV/	315
	Closed ans long /CVC/ and /CVVC/	519
Phoneme	Vowels	43%
	Consonants	57%

TABLE 1: Balanced Arabic corpus hierarchy and composition.

This corpus was first translated into phonetics, then segmented and labeled using spectrogram and waveform tools. The segmented data was stored in a database containing two levels: the predictors, i.e. the input features and the observations, i.e. the actual segmented durations. Then the main task while shaping the input space consists in classifying these features. Therefore, a twofold classification was suggested. The first part is linguistic, where segmented data are divided according to their contextual, positional and phonological aspects, and the second is statistical, as input data can be categorical or continuous. This classification generates a 2-dimension array where every factor is described according to its linguistic and numerical classes.

4.2 Fujisaki Parameters Extraction

Fujisaki constants, α, β and γ of the recorded voice were set at, respectively, 2/s, 20/s and 0.9 [17]. The Fujisaki parameters were obtained by Mixdorff's tool [18] which applies a multi-stage process called 'Analysis-by-Synthesis'. This process allows extracting the baseline frequency F_b, the phrase and the accent commands parameters through the minimization of the minimum square error between the optimal synthetic F_0 contour and the natural F_0 contour [6].The first step consists in quadratic stylization using the MOMEL algorithm [19] to interpolate the unvoiced segments and the short pauses within the F_0 curve, and to smooth the microprosodic variations due to sharp noises. Then, a high-pass filter is used to separate the phrase and the accent components through the subtraction of the filter output from the interpolated contour. This yields a low frequency contour containing the sum of phrase components and F_b. The third step consists in initializing the command parameters, i.e. A_p, T_0, A_a, T_1 and T_2. Finally, the synthesized contour is optimized, considering the interpolated contour as a target and the mean square error minimization as a criterion [18].

Input features types	Accent command input features
Phonological	• Sentence mode • Syllable type • Syllable accent level • Nucleus weight
Positional	• Accent command rank in sentence • Number of accent commands in sentence • Accented syllable position in sentence • Number of syllables in sentence • Nucleus position in accented syllable • Number of phonemes in accented syllable • Nucleus position in sentence • Number of phonemes in sentence
Contextual	• Nucleus duration • Previous phoneme's duration in accented syllable • Accented syllable's duration • Sentence duration • F_0 in beginning of accented syllable • F_0 at end of accented syllable • F0 movement in accented syllable
Extra features for A_a	• Predicted accent command duration (T_2-T_1)

TABLE 2: Accent command's input features

Input features types	Phrase command input features
Phonological	• Sentence mode • Phrase command syllable type • Phrase command syllable accent level • Nucleus weight
Positional	• Phrase command syllable position in sentence • Number of syllables in sentence • Nucleus position in phrase command syllable • Number of phonemes in phrase command's syllable • Nucleus position in sentence • Number of phonemes in sentence
Contextual	• Nucleus duration • Phrase command syllable duration • Sentence duration • Utterance's baseline frequency (F_b)
Extra features for T_0	• Ap predicted for phrase command

TABLE 3: Phrase command's input features

5. MODELING FUJISAKI PARAMETERS WITH NEURAL NETWORKS

5.1 General Neural Scheme

Neural networks are famous for their large ability to link the input to the output through a functional relationship. Therefore, they have been used in several intonation models, not only to predict F_0 values [20] or F_0 movements [21], but also F_0 parameters [22] and [23].

Though the modeling goal may differ, the approach is always the same. Neural networks are used to map a learning set, representing intonation-related features to a target set. Once learning is achieved, the model becomes able to predict the output from the test set inputs.
In most cases, the learning set includes 80% of the samples whereas the test set covers the remainder. The input features consist of a variety of speech characteristics, describing different aspects. However, these features should be selected looking to their correlation with the intonation. According to a previous survey we made about segmental duration using neural networks [30], the introduction of some features may give better prediction accuracy, whereas it's best to discard other features. Though the targets are different, as we are aiming to predict the Fujisaki parameters, it is still important to study the relevance of every feature or class of features.
Besides, the variety of the targets requires different implementation schemes for the neural networks. Hence, looking to our corpus composed of 200 separate sentences, we decided to assign a single phrase command for every sentence and at most one accent command for every syllable.
Then the task is to predict Fujisaki Parameters related to each aspect. For the phrase command, we need to predict A_p and T_0, and for the accent command, we need to predict A_a, T_1 and T_2. It looks obvious that we are looking for different types of targets at different levels, i.e. amplitudes and temporal locations for sentences and syllables. Therefore, the neural processing should be carried out in a parallel and distributed way to predict every single target on its own. Furthermore, we have noted that the targets themselves are well correlated with each other at each level, i.e. A_p and T_0, A_a and (T_2-T_1). Therefore we set a double tier strategy where, at every level, i.e. the sentence or the syllable level, the predicted parameters are used as inputs to predict the other ones. This strategy has been helpful to give better results, as it captures the latent relationship between the amplitude and the location of phrase and accent commands.
In addition, we opted for a 4-layer feed-forward neural network to model each parameter, i.e. using 2 hidden layers and a single-node output layer. Actually, it's been proved that a 2-hidden-layer neural network is able to model any continuous vector-valued function [24]. The first hidden layer is used to capture local features from the input, while the second hidden layer is required to capture the global features.
For the activation functions, we used a linear function at the output whereas they were non linear at the hidden layers, respectively the logistic sigmoid function and the tangent hyperbolic function.

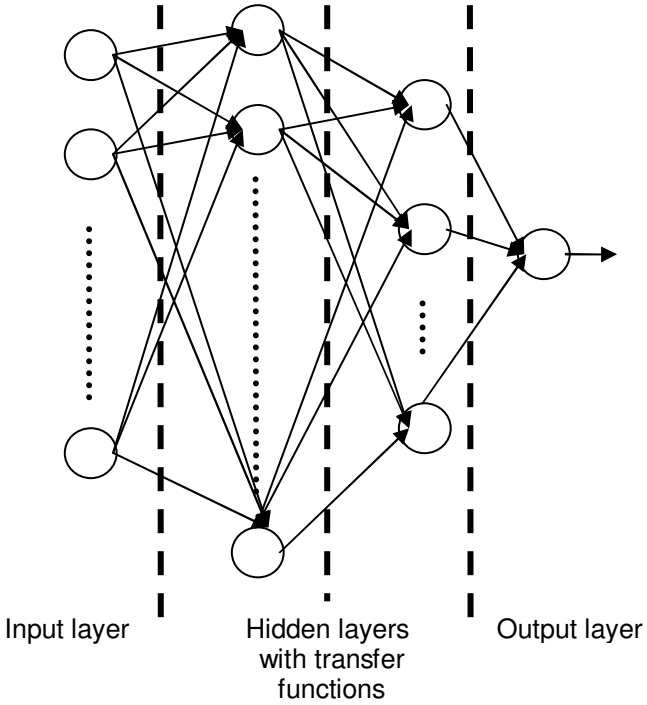

FIGURE 2: General neural network's scheme : a 2-hidden-layer FFNN with respectively, sigmoid and hyperbolic tangent transfer functions.

5.2 Modeling Phrase Command

In order to predict the phrase command parameters, i.e. A_p and T_0, we used an input set covering the contextual, phonological and positional features extracted from the corpus.

It's necessary to stress that the sentences of the corpus are totally distinct, which implies that contextual features are strictly internal to the sentence. Therefore we don't encounter any feature like 'previous phrase command amplitude' or 'previous phrase command timing' in the learning set.

However, looking to the high correlation between A_p and T_0, we start by predicting the A_p to be used later as input for T_0 prediction, but again, only within the same sentence. Also, to normalize the output, we opted to use the logarithm of the squared values for the temporal locations of the phrase commands.

5.3 Modeling Accent Command

In order to build the learning set to be mapped to the extracted targets, a certain number of constraints have to be considered:

1. Only accented syllables have accent commands, and one syllable have at most one accent command.
2. Each accent command is defined by its amplitude A_a, its onset time T_1 and offset time T_2. If the accented syllable lies in the beginning of the sentence, the accent command can start before the accent group, but in any case, the accent command cannot end beyond the accent group [25].

In addition to these constraints, we adopted 2 major modifications to the Fujisaki model, which were suggested to allow its extension to other languages, i.e. German and Swedish:

1. For the interrogative sentences, a final rise component is added. It's represented by an extra accent command [26].
2. Accent command magnitude, A_a, can be negative [27].

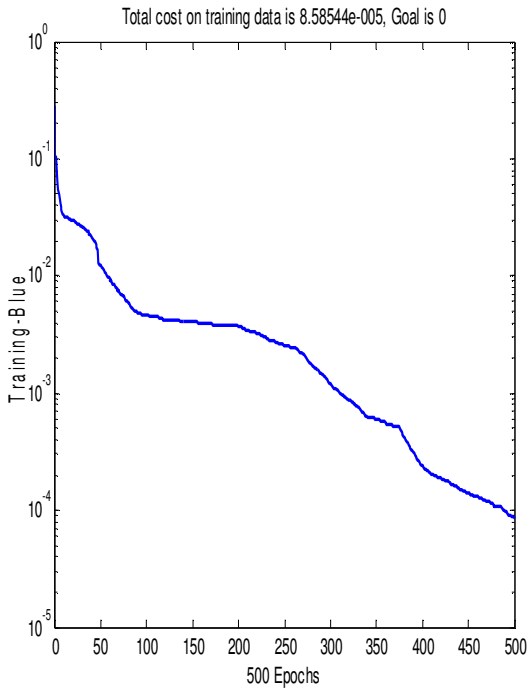

FIGURE 3: A_p training error

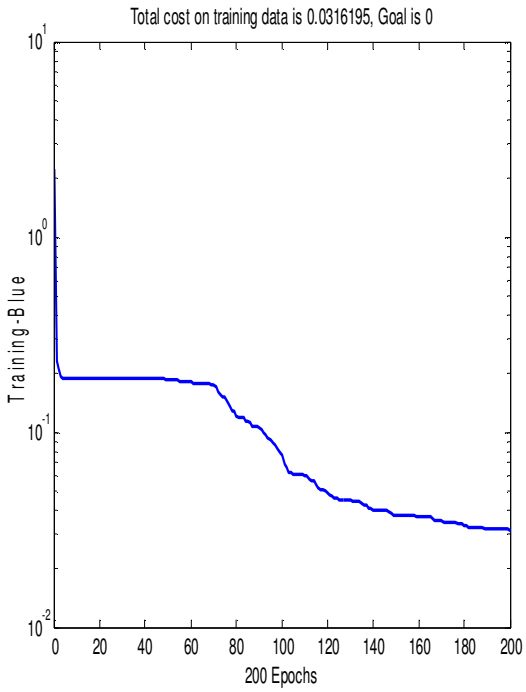

FIGURE 4: T_0 training error

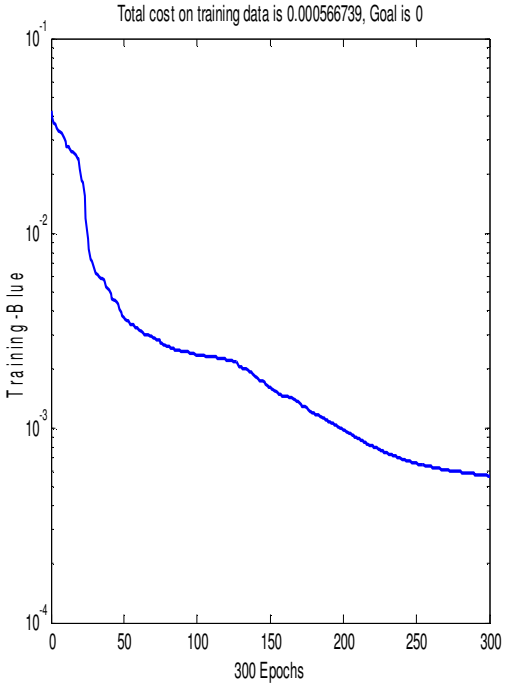

FIGURE 5: A_a training error

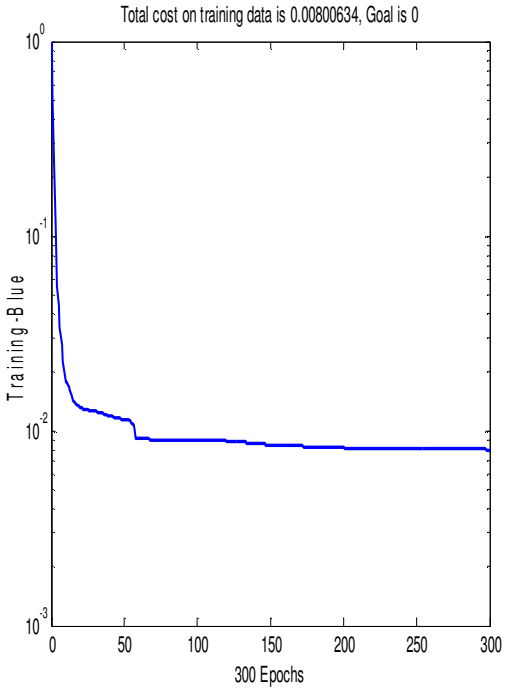

FIGURE 6: (T_1,T_2) training error

5.4 Selection of Input Features

For both levels, phrase and accent commands, the input features are treated according to their hierarchical class, i.e. the phoneme, the syllable or the sentence. It's important to note that for Arabic, the word is the acoustic unit between 2 successive pauses. However, the phonetic transcription doesn't always match with the text, because of the frequent presence of word-

boundary-linking syllables. To cope with this problem, we opted for the syllable as the acoustic unit.
The other criteria of features selection are firstly, the type of data values, whether discrete or continuous, and secondly their classes, i.e. contextual, phonological or positional.
Actually, such a broad classification is widely used while dealing with neural networks. As learning is supervised, these classes have to be defined to reduce the scarcity of data incurring a high-dimension input space, and to get rid of unnecessary data which may reduce the learning performance. This pre-processing is also useful to avoid the over-learning problem. Actually, too many inputs may yield generalizing the learning exceptions.

6. EVALUATION AND DISCUSSION

After training the neural networks for each parameter on its own, the test phase is carried out jointly with statistical evaluation. Thus we used the following statistical coefficients to measure the accuracy of the model:

- Mean absolute error

$$\mu = \frac{\sum_i | x_i - y_i |}{N} \quad (7)$$

- Standard deviation

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \; d_i = e_i - e_{mean}, \; e_i = x_i - y_i \quad (8)$$

- Correlation coefficient

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y} \quad (9)$$

$$V_{X,Y} = \frac{\sum_i |(x_i - x_{mean})| . |(y_i - y_{mean})|}{N} \quad (10)$$

X and Y are the actual and the predicted F0 values.

Then we used the test results to build synthetic F_0 contours to be compared with the original ones, which voiced parts were extracted by SHR algorithm [31] whereas we used Praat software [32] to locate the voiced and unvoiced parts of speech.

Statistical coefficients	With extracted F_0 contour by SHR
Mean absolute error	38.59 (29.75%)
Correlation in voiced parts	0.78
Standard deviation	43.12

TABLE 4: Statistical evaluation of synthesized F0 contours with extracted F0 contours in the test set.

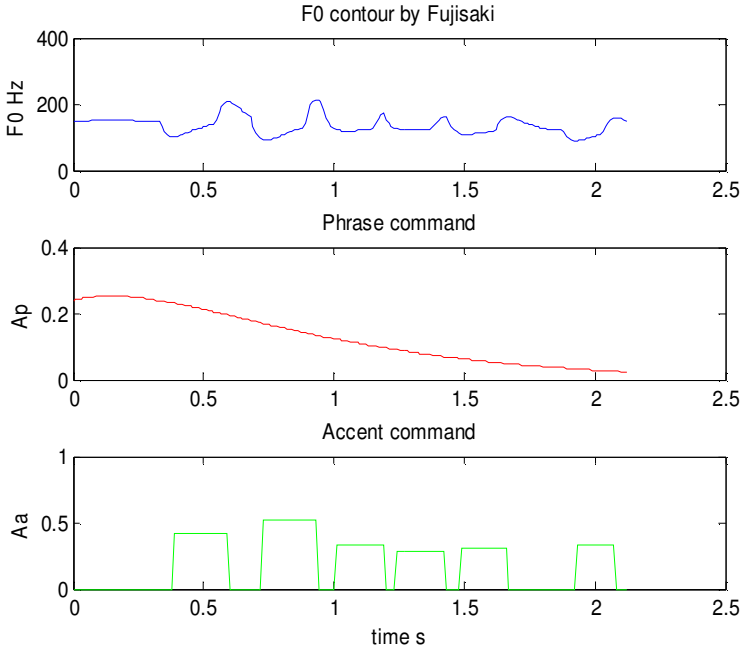

FIGURE 7: Synthesized F_0 contour, AC and PC of the Arabic sentence``hal ka:na juqabilukuma: ?" ( ``Was he meeting with both of you?")

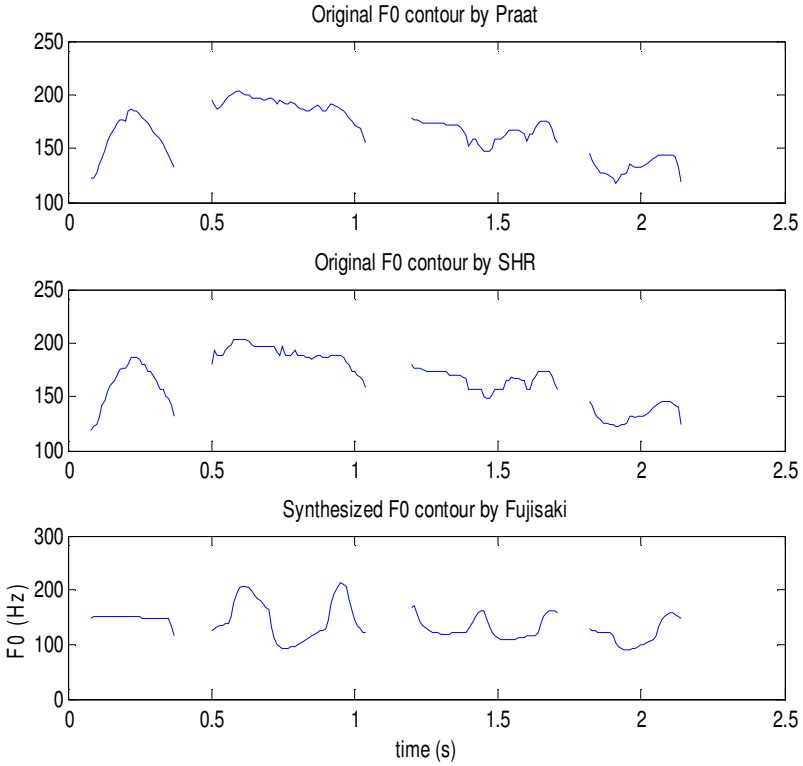

FIGURE 8: Extracted and synthesized F_0 contours in voiced parts of the Arabic sentence ``hal ka:na juqabilukuma: ?" ( ``Was he meeting with both of you?")

Though we opted to process the Fujisaki parameters separately, we have noticed a certain number of dependencies between them, especially within each level, i.e. the phrase and the accent groups. Actually, the single-node-output-layer scheme has been advantageous since it allowed the use of some already predicted parameters as inputs for the others.
Although it's still difficult to appreciate the quality of speech relatively to the synthesized F_0 contour before perceptual tests, the correlation between the natural and synthetic F_0 contour can tell about the accuracy of the model.
In addition to the statistical evaluation, these experiments revealed a certain number of notes about, not only the output parameters, but also about the predictors, i.e. the input features:

6.1 The Phrase Command's Evaluation

- **The Phrase Command' Amplitude A_p:** It's highly correlated with the duration of the sentence. It has higher values for longer sentences

- **The Phrase Command's Timing T_0:** This parameter is particularly critical, since it describes the very beginning of the phrase component, which precedes the onset of the utterance. Looking to our corpus, which is composed of separate sentences, T0 should usually be negative, unless there is a silence in the beginning of the utterance.

Generally, the phrase command is higher, only in the beginning, to decay along the sentence, and therefore the accent command is more relevant in the global shape of F_0 contour.

Phrase command parameter	Mean absolute error	Mean Value	Correlation
A_p	0.279	0.779	0.913
T_0	0.072	-0.409	0.508

TABLE 5: Phrase command parameters statistical evaluation in the test set

6.2 The Accent Command's Evaluation

- Accent command's onset and offset (T_1, T_2): Unlike the phrase component, the accent component is characterized by a variable range of temporal locations. If we consider (T2-T1) as the accent group duration, and according to a previous study we made about the segmental durations [30], then the accent command onset T_1 and offset T_2 are mainly related to the phonological, contextual and positional features of the speech. Yet, a good result of the accent group prediction has been helpful to predict its amplitude A_a.

- Accent command's magnitude A_a: A good prediction of A_a requires, not only information about the speech characteristics, but also about the neighboring accent groups, i.e. A_a and (T2-T1) of the previous accent group, and (T2-T1) of the actual accent group. In fact, the accent component is responsible of the overall shape of the F_0 contour after the decay of the phrase component, and so forth, the amplitude, the temporal locations of the previous accent group, coupled with the F_0 movement, provide a relevant indication about the accent command amplitude.

The assumption made in the beginning, stating that A_a values can be expanded to the negative domain, was verified in the test set. However, synthetic and natural A_a values, for the same accent group, don't have always the same sign. This is due, from one side, to the prediction accuracy, and from the other side, to the interaction of the phrase and accent commands, which are superposed to generate a natural looking F_0 contour. Besides, the contextual input features, such as (T2-T1), A_a and F_0 movement of the previous accent group, act jointly as correction agents to keep watching the fall/rise evolution of the synthesized contour.

Accent command parameter	Mean absolute error	Mean Value	Correlation
A_a	0.205	0.438	0.595
T_1	0.136	1.134	0.903
T_2	0.141	1.332	0.918

TABLE 6: Accent command parameters statistical evaluation in the test set

6.3 Discussion

- **Fujisaki Constants (α, β):** In our model, we used average values of the angular frequencies α=2/s and β=20/s. The variation of these factors doesn't alter significantly the results. Actually, the tool we used to extract the Fujisaki parameters, [8], allows changing their values, but without a great influence on the resulting contour. This confirms many previous researches concluding that α and β are mainly speaker-dependent, such as English [36], German [13] and Japanese [4] since they characterize the dynamic properties of the glottal control mechanism. [4].

- **Input Features:** A good approximation of F_0 contour needs, not only a good model, but mainly a representative learning set, able to describe as closely as possible, the various interactions and synergies between different contributory types of data. Hence, after the feature selection phase, pre-processing is required either to normalize input data or to broad-classifying them. In fact, a heterogeneous input set, including different data classes with different value ranges may fall in over-fitting, where exceptions could be generalized, or may require a high calculation time before the training error falls down to an acceptable minimum. Back to the obtained results, the pre-processing phase was of great help to tune the output values, improve accuracy and minimize training error. It also allowed capturing the inherent relationship between some input features and their relative outputs, and therefore guided us to build a personalized predictors set for each following parameter.

7. CONCLUSION

In the general framework of developing an Arabic TTS system, spotlight was focused on modeling F_0 contour. This task was performed using the Fujisaki model, which parameters were extracted by Mixdorff's analysis-by-synthesis-based tool and trained by neural networks.
Therefore, a phonetically-balanced Arabic corpus was analyzed, firstly to extract the Fujisaki parameters and secondly to select and pre-process the input features used as predictors for the neural network. Neural networks were hired for their large ability to capture the hidden functional mapping between input and output sets. Many neural schemes were suggested during the elaboration of this work, to select those which performed best in the try-and-error test. Then, statistical coefficients were calculated between actual and predicted Fujisaki parameters.
This study revealed also that Fujisaki parameters are dependent at each of the phrase and accent levels. Therefore, some of them were used as inputs to predict the other ones. In fact, an interaction was noted between the phrase and accent component to keep the overall shape of F_0 contour. Thus, after the decay of the phrase command, the accent command rises, and so forth, the F_0 contour becomes more sensitive to the accent variations. This note was checked out by the introduction of correction agents to the input feature while training. Furthermore, some of Fujisaki's assumptions were verified in this study. Thus negative accent commands were necessary to model the variations of the accent group; and the variation of some parameter

values such as α and β didn't have a relevant impact on the results, confirming that they are rather speaker-dependant.
As a future projection, this model can be used along with our previous study on segmental duration modeling [30], to build an integrated model of Arabic prosody, able to generate automatically the duration and the F_0 contour of an input text. Also, this study can be expanded to a paragraph-composed corpus in different Arabic dialects.

8. REFERENCES

1. M. Tatham, K. Morton, *"Developments in speech synthesis"*, John Wiley & Sons Inc. (2005)

2. H. Fujisaki, *"Prosody, information and modeling with emphasis on tonal features of speech",* in Proceedings of Workshop on spoken language processing, ISCA-supported event, Mumbai, India, January 9-11, 2003

3. J. B. Pierrehumbert, *"The phonology and phonetics of English intonation",* Ph. D. Thesis, MIT, Cambridge, 1980

4. H. Fujisaki, *"Dynamic characteristics of voice fundamental frequency in speech and singing. Acoustical analysis and physiological interpretations"*.STL-QPSR,1981, Vol. 22(1), pp 1-20, KTH, Sweden

5. S. Narusawa, N. Minematsu, K. Hirose and H. Fujisaki, *"Automatic extraction of model parameters from fundamental frequency contours of English utterances",* in Proceedings of ICSP'2000, pp 1725-1728, Denver, Colorado, USA

6. H. Mixdorff, H. Fujisaki, G. P. Chen and Y. Hu, *"Towards the automatic extraction of Fujisaki model parameters for Mandarin"*, in Proceedings of Eurospeech'03, pp 873-976, Geneva, 2003

7. M.Vainio, *"Artificial Neural networks based prosody models for Finnish text-to-speech synthesis",* PhD. Thesis, Helsinki University of Technology, Finland, 2001

8. H.J. Mixdorff, *"FujiParaEditor program"*, Available at http://www.tfh-berlin.de/~mixdorff/

9. J. Buhmann, H. Vereecken, J. Fackrell, J. P. Martens and B. Van Coile, *"Data driven intonation modeling of 6 languages"*, in Proceedings of International conference on spoken language processing, October 2000, Beijing, China, Vol. 3, pp 179-183

10. K. S. Rao and B. Yegnanarayana, *"Intonation modeling for Indian languages"*, Computer speech and language Journal, Volume 23, pp 240-256, Elsevier, 2009

11. G. P. Giannopoulos and A. E. Chalamandaris, *"An innovative F0 modeling approach for emphatic affirmative speech, applied to the Greek language"*, in Speech Prosody 2006, Dresden, Germany

12. X. Sun, *"F0 Generation for speech synthesis using a multi-tier approach",* in *Proceedings of ICSLP'02*, Denver, 2002, pp 2077-2080

13. H. Mixdorff, *"An integrated approach to modeling German prosody",* Habilitation Thesis, Technical University of Dresden, Germany, 2002

14. H. Fujisaki and S. Ohno, *"Prosodic parameterization of spoken Japanese based on a model of the generation process of F_0 contours",* in Proceedings of ICSLP'96, vol 4, pp 2439-2442, Philadelphia, PA, USA, Oct. 1996

15. B. Moebius, *"Synthesizing German F0 contours",* in J. Van Santen, R. Spraot, J. Olive and J. Hirschberg, *Progressin speech synthesis,* Chapter 32, pp 401-416, Springer Verlag, New York, 1997

16. H. Fujisaki and K. Hirose, *"Analysis of voice fundamental frequency contours for declarative sentences of Japanese"*, in Journal of the acoustic society of Japan (E), 5(4), pp 233-241, 1984

17. H. Mixdorff and O. Jokisch, *"Building an integrated prosodic model of German",* in Proceedings of Eurospeech 2001, Aaloborg, Denmark, vo2, pp 947-950

18. H. Mixdorff and O. Jokisch, *"Evaluating the quality of an integrated model of German prosody",* International journal of speech technology, Vol 6, pp 45-55, 2003

19. D. Hirst, A. Di Cristo and R. Espesser , *"Levels of representation and levels of analysis for inronation in M.Horne, Prosody: Theory and experiment",* Kluwer editions, Dortrecht, 2000

20. K. S. Rao and B. Yegnanarayana, *"Intonation modeling for Indian languages"*, in Proccedings of Interspeech'04,Jeju Island, K0rea, 4-8 October 2004, pp733-736

21. G. Sonntag, T. Portele and B. Heuft, *"Prosody generation with a neural network: Weighing the importance of input parameters"*, in Proceedings of ICASSP, pp 931-934, Munich, Germany, April 1997

22. J. P. Teixieira, D. Freitas and H. Fujisaki, *"Prediction of Fujisaki model's phrase commands"*, in Proceedings of Eurospeech 2003, Geneva, pp 397-400

23. J. P. Teixiera, D. Freitas and H. Fujisaki, *"Prediction of accent commands for the Fujisaki intonation model"*, in Proceeding of Speech Prosody 2004, Nara, Japan, March 23-26, 2004, pp 451-454

24. J. J. Hopfield, *"Neural networks and physical systems with emergent collective computational abilities"*, Proceedings of the National Academy of Sciences of the USA, vol. 79 no. 8 pp. 2554-2558, April 1982

25. E. Navas, I. Hernaez, A. Armenta, B. Etxebarria and J. Salaberria, *"Modeling Basque intonation using Fujisaki's model and CARTS"* , in Proccedings of ICSLP 2002, Denver, USA, pp 2409-2412

26. H. Mixdorff, *"Intonation patterns of German-model-based quantitative analysis and synthesis of F_0 contours"*, Ph. D. Thesis, TU Dresden, 1998

27. H. Fujisaki, S. Ohno and S. Luksaneeyanawin, *"Analysis and synthesis of F_0 contours of Thai utterances based on the command-response model"*, in Proceeding of 15th ICPhS, Barcelona, Spain, 2003, pp 1129- 1132

28. P. Taylor, *"Analysis and synthesis of intonation using the Tilt model",* Journal of Accoustic society of America, No 107, pp 1697-1714, 2000

29. F. Boukadida, *"Etude de la prosodie pour un système de synthèse de la parole Arabe standard à partir du texte",* Thèse de doctorat, Université Tunis El Manar, 2006

30. Z. Mnasri, F. Boukadida and N. Ellouze, *"Modelling segmen1tal durations by statistical learning for an Arabic TTS system",* International Revue on Computer and Software, September 2009

31. X. Sun, *"SHR program"* , available at http://mel.speech.nwu.edu/sunxj/pda.htm, Copyrightht © 2001, X.Sun, Department of communication sciences and disorders, Northwestern University, USA

32. P. Boersma and D. Weenink, *"Praat: Doing phonetics by computer, version 4.4",* available at http://www.praat.org

33. A. Black and A. Hunt, *"Generating F0 contours from ToBI labels using linear regression"* ,in Proceedings of ICSLP, Philadelphia, Pennsylvania, 1996

34. K. Dusterhoff, A. Black and P. Taylor, *"Using decision trees within the tilt intonation model to predict F0 contours",* in Proceedings of Eurospeech, Budapest, Hungary, 1999

35. S. Sakai and J. Glass, *"Fundamental frequency modeling for corpus-based speech synthesis based on a statistical learning technique",* in Proceedings of IEEE ASRU 2003, Nov. 30-Dec. 4, 2003, St. Thomas, US Virgin Islands, pp 712-717

36. H. Fujisaki and S. Ohno, *"Analysis and modelling of fundamental frequency contours of English utterances"*, in Proceedings of Eurospeech'95, pp 985-988, Madrid, Sep. 1995