Design and Development of a Prosody Generator for Arabic TTS Systems

Zied Mnasri
Signal, Image and Pattern Recognition Research Unit -ENIT
Université Tunis El Manar
BP 37, le Belvédère, Tunis 1002

Fatouma Boukadida
Signal, Image and Pattern Recognition Research Unit-
ENIT Université Tunis El Manar
BP 37, le Belvédère, Tunis 1002

Noureddine Ellouze
Signal, Image and Pattern Recognition Research Unit-ENIT
Université Tunis El Manar
BP 37, le Belvédère, Tunis 1002

ABSTRACT
Prosody modeling has become the backbone of TTS synthesis systems. Amongst all the prosodic modeling approaches, phonetic methods aiming to predict duration and F0 contour are being very praised, thanks to the development of regression tools, such as neural networks (NN). Besides, parametric representations like Fujisaki model for F_0 contour generation help to reduce the problem into the approximation of parameters only. But, prior to the prediction process, text analysis should be carried out first, to select and encode the necessary input features. In our purpose to promote Arabic TTS synthesis, an Integrated Model of Arabic Prosody for Speech Synthesis (IMAPSS) tool has been designed to integrate our developed models for text analysis, NN-based phonemic duration prediction and Fujisaki-inspired F_0 contour. Hence, the yielding parameters provide a command file to be read by speech synthesis systems, like MBROLA.

General Terms
Signal processing, Speech synthesis, Prosody, Neural Networks.

Keywords
Arabic TTS, prosodic parameters, text analysis, phonemic duration, F0 contour, neural networks, Fujisaki model.

1. INTRODUCTION
Whereas TTS systems have been popularized since several years, especially for wide-spread languages such as English, French, Chinese...etc., Arabic is still awaiting more interest. Actually, though many systems are suggesting TTS tools for Arabic, a few of them are based on this language specifically-designed model. In fact, any language needs to be processed on its own, in order to extract its characteristics and to meet its requirements, and especially to model the dynamics of its prosodic features variations.

Phonologically speaking, prosody stands for the abstract phenomena incurring from speech, including accentuation, intonation and rhythm. The phonetic realization of these cognitive concepts is the physical notions of duration, pitch and intensity [1]. Whereas intensity deals with the speech loudness, i.e. the speech signal's energy, duration and pitch are the main transmitters of the acoustic information, whether linguistic, paralinguistic or non-linguistic [2]. On another side, the prosodic parameters could be interpreted physiologically as the duration and the frequency of the vibration of the vocal tract. Both viewpoints suggest that a quantitative processing of the prosodic parameters can provide a reliable model able to generate the melodic effect of speech.

Many approaches have been investigated to reach this goal. Whereas rule-based models aim to take profit of the linguists' know-how to establish computational model for prosodic parameters [3][4], the statistical models are based on the analysis of a dedicated corpus, to achieve a mapping between input features and output targets. Input data are generally extracted during corpus analysis through the selection of the most relevant features. In contrary, output targets depend on the model's structure. In fact, a variety of prosodic parameters could be predicted, such as syllabic or phonemic durations, raw F0 values [5] or underlying pitch parameters [6]. Therefore, a preliminary study of prosodic parameters should be conducted, to determine **(a)** which parameters to predict **(b)** at which level, based on the language's characteristics and the prosodic system components. Actually, though the modeling goal is the same, the parameterization of the model may reduce the task to the prediction of the parameters, provided they are phonologically significant. This is the cornerstone of Fujisaki model for F_0 contour generation, which suggests a superpositional representation of F_0 contour based on the physiological description of the vocal tract [7]. The advantage of this model is the phonological interpretation of its components and parameters, i.e. the baseline frequency, the phrase and the accents commands, both described by their timing and amplitude. Hence the F_0 contour yields from the superposition of these components in the logarithmic domain.

To increase the input/output mapping, supervised learning, i.e. specific and separate input and output sets, is highly recommended. Amongst the supervised statistical learning techniques, neural networks are famous for their ability to approximate non linear functions [8], thanks to their generalization power and their ability to capture the latent relationship between the input data and the output targets. However, special care should be taken whilst adjusting the neural schemes and especially, a clear strategy could help linking the predicted parameters in order to use some of them as input features to increase the prediction potential of other ones.

The obtained models will be the nucleus of the integrated prosodic model, which executes successively a series of modules responsible for:

- Text analysis, to extract the input features.
- Phonemic duration prediction

- F_0 contour generation module
- The command files generation, to be read by MBROLA multi-language speech synthesis system [9].

This paper starts by describing the speech corpus, and then duration prediction using neural networks is investigated. The following section shows the F_0 contour generation through the prediction of Fujisaki parameters. Finally, the integration of these components is described to evaluate the obtained results and to discuss the incurring issues.

2. CORPUS ANALYSIS

2.1 Speech material

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/ [10]. 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 targets. Then the main task while shaping the input space consists in classifying these features. Therefore, a twofold classification was suggested. The first part is processed linguistically, where segmented data are divided according to their linguistic, contextual, and phonological aspects, and the second is achieved statistically, 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 numeric classes (Cf. Annex2).

2.2 Fujisaki parameters extraction

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 [11]:

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

Where

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

$$G_a(t) = \begin{cases} min(1 - (1 + \beta t).e^{\beta t}, \gamma) & if \quad t > 0 \\ 0 & else \end{cases} \quad (3)$$

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.

Fujisaki constants, α, β and γ of the recorded voice were set at, respectively, 2/s, 20/s and 0.9 [12]. The variable Fujisaki parameters were obtained by Mixdorff's tool [13] 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 [14].

The first step consists in quadratic stylization 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 [14].

3. PROSODIC PARAMETERS MODELS

3.1 Duration analysis

This core is built upon a model which was developed using statistical learning and based on the analysis of a phonetically balanced Arabic speech corpus [15]. The first step consists of the analysis of speech material to optimize the learning process. Actually, many decisions have to be taken before presenting the input features to the statistical learning tool.

3.1.1 Analysis level

First, which level should be considered to predict duration? Or in other words, which segment should be considered as the duration unit? This controversial question was the topic of many researches. While Campbell [16] and Barbosa [17] have opted for the syllable, arguing that all phonemes within a syllable are stretched or shortened by the same factor and suggesting that it's crucial to determine the syllable's duration to decide about the phoneme's one, Van Santen [18] has shown that it's rather the phoneme's duration which influences the syllable's one, since the latter one is actually the sum of the earlier ones. Thus we opted for the phoneme as a duration unit.

3.1.2 Analysis domain

Also, while examining the phonemes durations available at the corpus, we noticed that their distribution is not normal. In fact, the duration distribution is deviated to the side of short durations, so that the mean duration and the standard deviation are also located at that side. This means one can easily obtain negative predicted duration values after learning. The solution is to switch to the logarithmic domain, which offers the following advantages:

1. It increases the resolution of small values, and since they are more frequent,
2. The logarithmic domain will help getting rid of negative values.

It normalizes the duration distribution by moving the mean value and the standard deviation to the center.

3.2 F_0 contour analysis

The analysis of the Fujisaki parameters allows a twofold study of their variation:

A. **The morphological side**: Which level should be considered as a framework for the parameters estimation, i.e. the phoneme, the syllable or the phrase group?

B. **The strategic side**: Which parameters should be determined first? and which ones could be highly correlated in order to define a stepwise strategy, where some predicted parameters could contribute as input features in the estimation process of other ones ?

Every Fujisaki parameter is predicted using NN by training the input features extracted from Text. Whereas all parameters could be sorted out together using a single NN, delivering many outputs simultaneously, a stepwise prediction based on a linked scheme could better the prediction performance. Hence:

1. Phrase and accent components are processed separately.
2. For every component, timing parameters are estimated firstly, to be used later as input features for magnitude parameters prediction.

Actually, this choice could be explained by:

1. Longer phrase groups require closer T_0 to the beginning of the phrase group and higher A_p to carry the phrase component as far as possible, to avoid its declination before reaching the end of the phrase group.
2. Longer accent groups, i.e. higher (T_2-T_1), are usually associated to the most accented syllables, i.e. which A_a is higher.

This strategy aims to increase the prediction power of NN by introducing extra-linguistic features, i.e. T_0, T_1 and T_2.

3.3 Selected input features

These features are extracted by the first module. Nevertheless, every prosodic parameter was trained with its own features set. This set is selected after the analysis stage, which is achieved by studying the distribution of phonemes duration according to every single feature. Those which effect on phoneme's duration seems important are kept. Nevertheless, the contribution of each feature should be weighted to check its relevance. Actually, some features may not be useful, or worse, they may struggle the learning performance. The final set of features was selected after such a contribution evaluation yielding the following set.

3.4 Neural networks training

If the NN is fed with all features which may have a significant influence on phonemic duration, and if its architecture allows learning the effect of each factor, by matching it to the corresponding corpus-extracted duration, then it should be able to set a model able to estimate the outputs of new speech segments wherever they come from, i.e. the test base or the user's input text.

Therefore, focus was held to select the optimal architecture of the NN. Actually, many neural schemes were tested to check out their prediction performance. For every scheme, several architectures were tried by changing the network's parameters, such as the structure of hidden layers, the number of nodes and the transfer functions.

Neural networks rely on the total connection of nodes of every couple of successive layers. These weighted connections allow encoding features during their transfer through the network. The first layer, i.e. the input layer, contains features issued from the analysis stage. Thus, the number of nodes of this layer is not subject to any change, as it's equal to the number of input features. Furthermore, to ensure a quick convergence, or at least a finite-time convergence of the NN, input features were normalized into the interval [0,1].

The next stage is handling the NN black box, i.e. its hidden part, which includes the number of intermediate layers, the number of nodes and the transfer function of each one. Though this is mainly an empirical task, some rules have been followed to guide training trials [19]:

- If the input/output relationship is too complex, it would be better to increase the hidden elements, including the number of hidden layers.
- Hidden layers should be added only if the modeling process is separable into many stages. Otherwise the extra hidden layers will be used to memorize the network's output, including exceptions, instead of learning how to generate the suitable output. This may increase the risks to generalize the exceptional cases.
- A feed-forward NN is preferred to a recurrent NN as outputs are totally independent. In addition, a MLP having 2 hidden layers has been proved to be able to model any continuous function, provided the right inputs [8].
- The first hidden layer should include more nodes than the input layer to capture local features, which are more frequent; while the second hidden layer should focus on capturing global features by bearing fewer nodes.

Transfer functions are used to project features from one layer to another until they reach the last weighted sum stage at the output. Through this process, every feature contributes to the modeling process. Therefore, the transfer functions should be carefully selected to keep the features significance. In fact, the involved features have different types, and consequently, transfer functions should take care of this diversity. Hence the transfer functions should be:

1. Heterogeneous: So that when different transfer functions are aligned, they may be able to capture more types of features.
2. Symmetric: Such transfer functions help increasing the resolution and offer a larger range for the next layer's transfer function.
3. Non linear: This is necessary to model the real world phenomena.

Therefore, we opted for:

A. The Sigmoid and hyperbolic tangent transfer functions, respectively in the first and the second hidden layers for each NN. Both functions are continuous and strictly croissant, so that they are able to approximate any non linear function using MLP.

B. The Levenberg-Marquardt training algorithm to minimize the least square estimation error as it allows a

faster convergence than the gradient descendant algorithm.

C. The cross-validation technique, which helps early stopping of the training process, when the validation error starts to increase, even if the maximum epochs number is not reached yet, to avoid the generalization of the exceptions. Cross-validation could be achieved either in a regular way, by allocating half of the training set to it, or in a random way, using the bootstrapping technique. Both methods were tested, giving nearly the same performance.

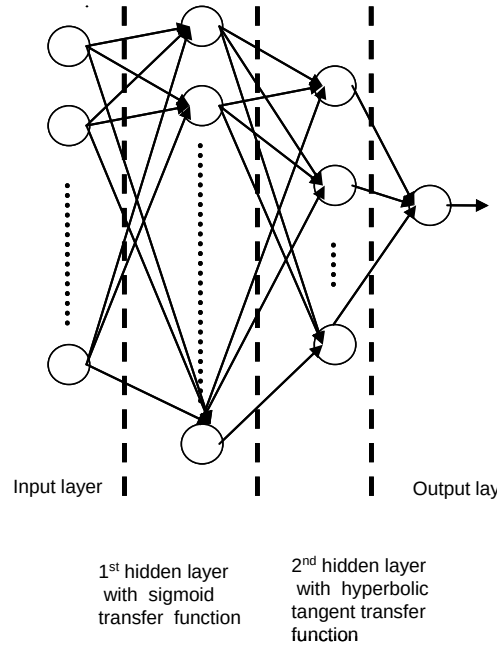

FIGURE 1. MLP scheme used for prosodic parameters prediction

4. EVALUATION

4.1 Statistical evaluation

For both prosodic features, statistical evaluation is carried out (a) to assess the generalization performance of the trained neural models (b) weigh the contribution of every single input feature (c) measure the impact of the generated F_0 contour.

Hence 20 % of the corpus data is allocated for the test stage. Statistical coefficients are calculated involving original and predicted parameters:

- Mean absolute error

$$\mu = \frac{\sum_i | x_i - y_i |}{N}$$

- Standard deviation

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

- Correlation coefficient

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y}$$

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

X and Y are the actual and the predicted values, either for phonemic duration or for Fujisaki parameters (cf. Table 1).

In comparison to other languages, the difference could be explained by the difference of the corpus size. Nevertheless, results are very encouraging. Actually, far beyond the theoretic assessment, test data distribution can also tell about the accuracy of the model, since the normal distribution of data has been kept inside the same range, implying that the developed models have succeeded to **(a)** generalize the training data **(b)** capture the effect of input features on the variation of output parameters. In fact, this is primary goal of the neural modeling, which through the training stage, aims to detect the implicit relationship between features and targets, to be able to predict, as accurately as possible, the output of any new coming sample.

TABLE 1. Prosodic parametric statistical evaluation

Prosodic parameter	**Linear correlation γ**	**Mean absolute error μ**	**Standard deviation σ**	**Corpus mean value**
Phonemic duration	0.615	33.825 ms	43.938ms	114.484 ms
T_0[1]	0.431	0.072	0.508	- 0.409 ms
A_p	0.968	0.326	0.912	0.779
T_1	0.81	0.367	0.531	0.925 ms
T_2	0.784	0.415	0.584	1.074 ms
A_a	0.791	0.095	0.130	0.307

4.2 Input features' relevance evaluation

The feature selection stage is a key step in the prediction process, since the value and the interaction between the selected features may increase or decrease the model's performance. Actually, the selection of the input features is achieved in a preliminary study, based on linguistic rules and on previous modeling approaches of Arabic prosody modeling [15] [20] [21].

Also, the interaction between the output targets and input features may be used as a referee to include or exclude some features. Yet, the features selection is mainly theoretical, since their impact on the neural model couldn't be assessed before the relevance test. This test is made after the neural model is trained, yielding the relative variance of each input feature. Then this survey is necessary to **(a)** measure the contribution of analysis features **(b)** check the linguistic rules which had suggested them **(c)** optimize

[1] Negative values for T_0 are due to the fact that the corpus sentences were separate, as the phrase command is launched slightly before the beginning of each sentence.

the input space size, and consequently the NN size, by getting rid of the least relevant features.

5. SYSTEM INTEGRATION

IMAPSS is an integrated environment aiming to provide, from the SAMPA code transcription of Arabic, a command file readable by a synthesis system, containing the text phonemes, the duration and a F_0 sequence for each phoneme. This integrated system is presented as a Graphic User Interface (GUI) developed under MATLAB, to allow the following tasks (cf. Annex 1):

1. Open existing or type and save Arabic text transcribed in SAMPA code.
2. Select the prediction models for duration and F_0 contour and adjust their parameters
3. Generate a command file for a speech synthesis system.
4. Plot phonemes durations, the F_0 contour and its phrase and accent components.

This prosody generator is the result of the integration of 5 modules, having each a special task. Meanwhile, these modules are closely connected, as shown in Figure 2.

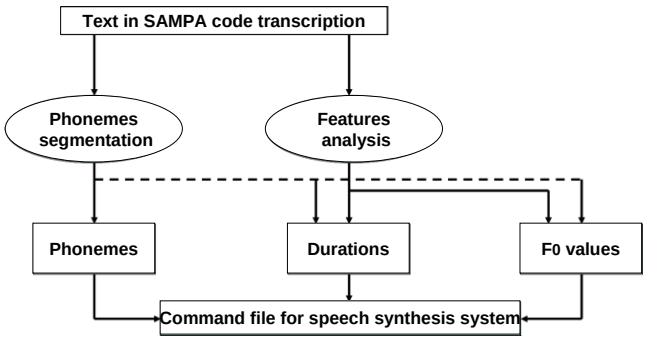

FIGURE 2. IMAPSS' block-diagram

5.1 Text analysis module

This module has a twofold task:

- It is used for the text segmentation into phonemes, to be included in the synthesis-ready command file.
- It performs a full-text analysis to provide the necessary features to the duration and F0 contour prediction models.

These features are extracted from different levels, i.e. the phoneme, the syllable and the phrase. Besides, they describe different aspects of the text, such as its syntax, its context and its phonology. In fact, each prosodic parameter has its own strand of features, previously selected during the neural network's learning. Thus, these features are neither identical, nor homogeneous. In addition, some of them are extracted from the output of other modules, as this integrated prosody model works as a linked structure, where some predicted parameters are used as inputs for other ones.

Nevertheless, the feature extraction from text is performed in an optimal way, to avoid redundancy. Actually, some features are commonly used to estimate most prosodic parameters. Therefore, we have defined an extraction strategy, which takes care of the prediction order. Hence, duration features are extracted first for each phoneme, the phrase command features, and at last, the accent command features, which are related to syllables.

In addition, the features coding is related to their nature. Thus, discrete features mostly describing syntax and phonologic have categorical coding, while other ones, mainly contextual and positional, are considered as ordinal.

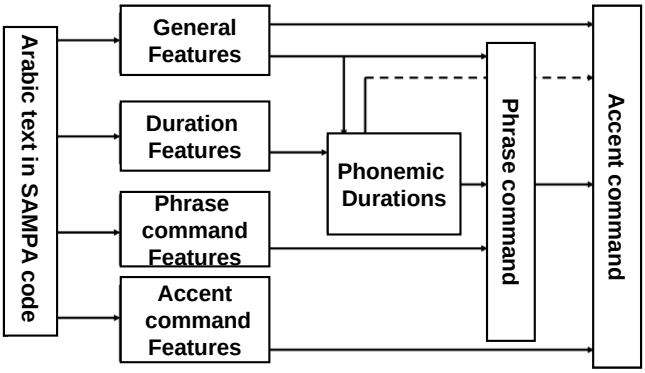

FIGURE 3. Features analysis strategy

5.2 Duration prediction module

This module uses the general features and those extracted at the phonemes level to be presented as inputs for the duration prediction model.

Then, the models yielding from the learning process of both methods were incorporated to the integrated prosody generator. Actually, one of the greatest advantages of using learning tools, is that once learning is over, whatever the time it takes, the prediction becomes an immediate operation, provided the right inputs are available.

Yet, the model selection is not specific to the duration prediction, as it will be shown later that it's also extended to the F_0 contour components. The importance of this module is not restricted to phonemes duration prediction. Actually, this step is crucial to localize syllables, and then pitch accents. Also, it calculates the phrase duration, which is necessary to estimate the F_0 contour components.

5.3 F_0 contour generation module

F_0 contour is generated from text using Fujisaki model and statistical learning. Fujisaki has relied on the earlier works of Ohman about the physiological description of speech production [11], to develop a parametric model describing F_0 as the superposition of two components, each being the response to a second order filter resulting from a impulsion command, i.e. the phrase command and a step wise command, i.e. the accent command.

Since both components are parametric, statistical learning techniques, such as NN, are used to predict their parameters, i.e. timing and magnitude, from text-extracted features, to generate the corresponding F_0 contour as a sum of both components in the logarithmic domain. Meanwhile, this integration requires the adjustment of some constants, too. Nevertheless, each component has its own model, and then its own feature set. Therefore, the

Fujisaki parameters prediction is executed as a step-wise process, starting by estimating the timing parameters for each component, i.e. T_0 and (T_1, T_2), to be used as extra features for the estimation of the magnitude parameters, respectively A_p and A_a

5.4 Command file generation module

This is the ultimate output of the integrated prosody system. All the previously mentioned modules are executed to generate a command file containing:

- The text segmentation into phonemes in SAMPA code transcription
- The duration of each segmented phoneme
- A sequence of F_0 value extracted from the contour for the voiced phonemes

Such a file should be read by a speech synthesis system to generate intelligible and natural sounding Arabic speech.

This module receives the outputs of the other ones to organize tasks in a structured way, as follows:

- Selection of the statistical learning method according to the user's choice, whether NN or CART.
- Segmentation of the input text into phonemes.
- Extraction of the general features.
- Extraction of duration's specific features.
- Estimation of the duration of every single phoneme.
- Extraction of the phrase command's specific features.
- Estimation of the phrase command's timing and magnitude
- Extraction of the accent command's features
- Selection of accented syllables.
- Estimation of accent command's timing and magnitude.
- Generation of F_0 contour by the application of Fujisaki's formula.
- Generation of the speech synthesis command file.

6. CONCLUSION AND DISCUSSION

During the implementation of IMAPSS system, different prosodic aspects of Arabic were investigated, since prosody is the phonetic realization of phonological concepts, i.e. accentuation, intonation and rhythm. Therefore, duration and pitch were analyzed to determine which way could provide the best prediction of their values. Neural networks were selected amongst many other regression tools, since they offer a generalized approximation of such non linear problems, especially by using MLP.

Both prosodic parameters' models were integrated to the IMAPSS system, in order to use the input features issued from the text analysis module. The final result is a command file ready for speech synthesis, providing the text segmentation in SAMPA code, the duration and a set of F_0 values, predicted for every phoneme.

Though tests have shown satisfactory results, in terms of statistical assessment, as good correlation have been noticed between generated and original parameters, some issues were revealed by this study, mainly:

- **The relevance of some input features:** This is still a problematic issue, since these features were selected upon previous studies of Arabic prosody modeling and according to their interaction with the target parameters. Though relevance evaluation could help erasing the less contributory features, focus should be held to investigate other unknown features which could be more correlated with outputs parameters, including non-linguistic features.
- **Fujisaki parameters analysis:** Output targets are provided by Fujisaki analysis tool, FujiParaEditor [13]. However, this tool hasn't been specifically designed for Arabic speech. Though some adjustments were introduced, such as for the Fujisaki constants α and β, and the extension of accent magnitude A_a to the negative domain, and also despite the good approximation of the generated contour with the original one, we believe that a dedicated tool for Arabic could provide better targets' values, henceforth more accurate training models.
- **Speech rate influence:** In addition to the phonological, contextual and positional features, some non-linguistic features, which cannot be extracted from text, may increase the model accuracy if included to the neural network. Amongst them, the speech rate could provide a better personalized speech quality. Our concern is how to extract such features in standalone real time applications.

7. REFERENCES

[1] Moebius, B. 1997. Synthesizing German F0 contours. In J. Van Santen, J. Sproat, R., Olive, J. and Hirschberg, J., Progress in speech synthesis, Chapter 32, pp 401-416, Springer Verlag, New York.

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

[3] Klatt, D. 1976. The linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of the acoustical society of America, JASA. No 59. pp1208-1221.

[4] Pierhumbert, J. 1980. The phonology and phonetics of English intonation, Ph. D. Thesis, MIT, Cambridge. USA.

[5] Rao, S. and Yegnanarayana, B. 2009. Intonation modeling for Indian languages, Computer speech and language Journal, Volume 23, pp 240-256, Elsevier.

[6] Sun, X. 2002. F0 Generation for speech synthesis using a multi-tier approach, in proceedings of ICSLP'02. pp 2077-2080. Denver, USA.

[7] Fujisaki, H. and Hirose, K. 1984. 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.

[8] Haykin, S. 1999. Neural Networks: a comprehensive foundation. 2nd edition. Engelwood Cliffs. Prentice Hall.

[9] MBROLA speech synthesis system, available at http://tcts.fpms.ac.be/Synthesis/mbrola.html.

[10] Boukadida, F. and Ellouze, N. 2004. Arabic intonative speech database. IEEE International Conference on Industrial Technologies. Tunis, Tunisia.

[11] Fujisaki, H and Ohno, S. 1996. 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.

[12] Mixdorff, H. and Jockisch, O. 2001. Building an integrated prosodic model of German. In proceedings of Eurospeech, vol2, pp 947-950. Aaloborg, Denmark.

[13] Mixdorff, H., Fujisaki, H. Chen, G. and Hu, Y. 2003. Towards the automatic extraction of Fujisaki model parameters for Mandarin. In Proceedings of Eurospeech. pp 973-976, Geneva. Switzerland.

[14] Mixdorff, H. 2002 An integrated approach to modeling German prosody. Habilitation Thesis. Technical University of Dresden., Germany

[15] Boukadida, F. 2006. 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, Tunisia.

[16] Campbell, N. 1992. Syllable duration modeling by neural networks, PHD thesis, University of Essex, UK.

[17] Barbosa, P. 2004. Caractérisation et génération de la structuration rythmique du Français. Thèse de doctorat. Institut polytechnique de Grenoble. France.

[18] Van Santen, J. 1994. Assignment of segmental durations in Text-To-Speech synthesis, Computer Speech and language 8, pp 265-273.

[19] Sutton, R. 1998. Reinforcement learning, MIT Press, Cambridge, MA, USA.

[20] Baloul, S. 2003. Développement d'un système automatique de synthèse de la parole à partir du texte Arabe voyellé. Thèse de doctorat. Académie de Nantes. Université du Maine. France.

[21] Zaki, A. 2004. Modélisation de la prosodie pour la synthèse de parole Arabe standard à partir du texte, Thèse de Doctorat, Université Bordeaux I, France.

8. Annex 1: IMAPSS interface

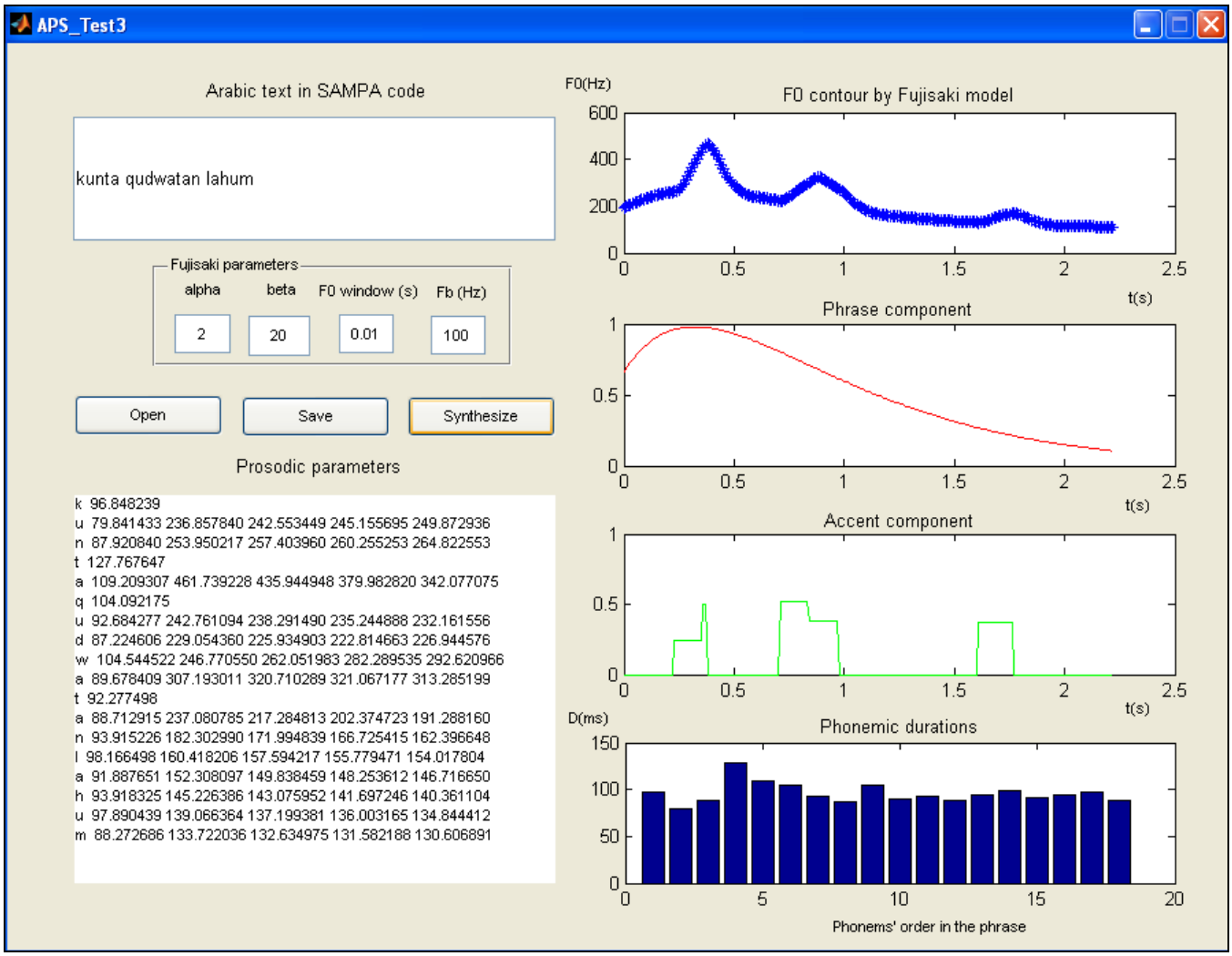

FIGURE 4. The IMAPSS interface: Fujisaki parameters, command file generated from text, phonemic durations, the phrase and the accent components and the resulting F0 contour for the Arabic utterance ``Kunta qudwatan lahum'' (``You were their leader'')

9. Annex 2: INPUT FEATURES INVENTORY

Parameter	Feature	Class	Coding
All	Phrase Mode	Contextual	Categorical
Duration	Previous phoneme's class	Phonological	Categorical
	Actual phoneme's class	Phonological	Categorical
	Following phoneme's class	Phonological	Categorical
	Actual phoneme's position in the syllable	Contextual	Ordinal
	Number of remaining phonemes in the syllable	Contextual	Ordinal
	Total number of phonemes in the syllable	Contextual	Ordinal
	Actual phonemes position in the phrase	Contextual	Ordinal
	Number of remaining phonemes in the phrase	Contextual	Ordinal
	Total number of phonemes in the phrase	Contextual	Ordinal
Duration, Aa, T1 & T2	Actual syllable's accentuation level	Phonological	Categorical
	Actual syllable's position in the phrase	Contextual	Ordinal
	Number of remaining syllables in the phrase	Contextual	Ordinal
	Total number of syllables in the phrase	Contextual	Ordinal
T0 and Ap	1st syllable's accentuation level	Phonological	Categorical
	1st syllable's type	Linguistic	Categorical
	1st syllable's nucleus class	Phonological	Categorical
	1st syllable's relative position in the phrase	Contextual	Ordinal
	Number of remaining syllables in the phrase	Contextual	Ordinal
	Total number of syllables in the phrase	Contextual	Ordinal
	1st syllable's nucleus relative position in the phrase	Contextual	Ordinal
	Number of remaining phonemes in the phrase	Contextual	Ordinal
	Total number of phonemes in the phrase	Contextual	Ordinal
	Nucleus relative position in the 1st syllable	Contextual	Ordinal
	Number of remaining phonemes in the 1st syllable	Contextual	Ordinal
	Total number of phonemes in the 1st syllable	Contextual	Ordinal
	1st phoneme's predicted duration	Contextual	Ordinal
	1st nucleus predicted duration	Contextual	Ordinal
	1st syllable's predicted duration	Contextual	Ordinal
	Phrase predicted duration	Contextual	Ordinal
	Phrase baseline frequency (Fb)	Non-linguistic	Ordinal
Ap	T0 value	Non-linguistic	Ordinal
T1, T2 and Aa	Previous syllable's type	Linguistic	Categorical
	Actual syllable's type	Linguistic	Categorical
	Following syllable's type	Linguistic	Categorical
	Class of syllable nucleus' previous phoneme	Phonological	Categorical
	Actual syllables' nucleus class	Phonological	Categorical
	Class of syllable nucleus' following phoneme	Phonological	Categorical
	Previous syllable's accentuation level	Phonological	Categorical
	Actual syllable's accentuation level	Phonological	Categorical
	Following syllable's accentuation level	Phonological	Categorical
	Number of primary (strong) accents in the sentence	Phonological	Ordinal
	Number of secondary (medium) accents in the sentence	Phonological	Ordinal
	Number of tertiary (weak) accents in the sentence	Phonological	Ordinal
	Actual syllable's position in the phrase	Contextual	Ordinal
	Number of remaining syllables in the sentence	Contextual	Ordinal
	Number of syllables in the phrase	Contextual	Ordinal
	Nucleus position the phrase	Contextual	Ordinal
	Number of remaining phonemes in syllable	Contextual	Ordinal
	Number of phonemes in the syllable	Contextual	Ordinal
	Last syllable in the sentence (Yes/No)	Contextual	Categorical
	Last phoneme in the syllable (Yes/No)	Contextual	Categorical
Aa	T1 & T2 values	Non-linguistic	Ordinal