Synthesizing Intonation of Standard Arabic Language 
Using Neural Network

A. Zaki 1,2, A. Rajouani 2, M. Najim 1

(1) Equipe Signal et Image, ENSEIRB. B.P 99, F-33 402 Talence Cedex, France
(2) LEESA. Faculty of Sciences. B.P 1014 - Rabat, Morocco

Tel.: ++33 556 84 61 85 - Fax: ++33 556 84 84 06
{zaki, najim}@tsi.u-bordeaux.fr, rajouani@fsr.ac.ma

Abstract

In this paper, we propose a model to generate fundamental frequency (F0) contours using neural networks. A learning procedure is proposed as an alternative to synthesis-by-rules. The generation of correct fundamental frequency contour is one of the important issues in the naturalness of automatic text-to-speech conversion systems. The proposed approach is based on a standard feed-forward multi-layer network that produces global F0 contours of sentences, directly from encoded linguistic features of standard Arabic language. Our model does not need syntactic information to produce suitable declarative intonation. TD-PSOLA synthesizer is used for validation of our results.

1. Introduction

Generating acceptable prosody is currently one of the most challenging tasks in the development of text-to-speech synthesis systems. High quality is essential for intelligibility and acceptable synthesis, especially when considering read texts. Prosody helps the listeners to interpret an utterance and to distinguish its sense according to the speaker’s intonation.

The purpose of this work is to develop an automatically trainable model for the generation of F0 contours. The model can be incorporated into Arabic TTS synthesis systems, currently in development [1][2], in order to improve the naturalness and intelligibility of synthesis speech.

For Arabic language, there is a limited number of studies that deal with the prosody at the linguistic level. This is a major drawback for automatic speech processing systems such as recognition, synthesis, etc. Furthermore the absence of a general theory of intonation still prevents the correct derivation of this important feature in unlimited text applications.

We chose neural networks based approach as a model for automatic generation of F0 contours because it makes direct association of linguistic description and prosodic substance possible. This approach avoids the need of the knowledge of the transition from the higher levels of representation to the prosodic substance. Moreover, the use of neural networks in speech synthesis has been successfully applied to prosody generation for various languages: in the automatic generation of F0 contours in isolated English words [3], in the prediction of minor phrase in Japanese [4] and in the whole sentence prediction for German [5]. The neural network based approach has been also applied for the prediction of other prosodic parameters such as segmental duration [6] and the generation of F0 and the syllabic duration [7].

This paper deals with the generation of F0 contours of standard Arabic language starting from a linguistic description based on phonological, phonetic, and phonotactic informations. The neural network is trained for automatic generation of F0 values corresponding to the information presented at the entry for each syllable of incoming text, converted into equivalent string of phonemes. In the first part of this paper we briefly describe standard Arabic linguistic and prosody. In the second part we present the neural network model used for synthesizing F0 contours and the results obtained.

2. Background linguistic, aspects for standard Arabic language and prosody

Arabic is a Semitic language. The contemporary standard Arabic language, a modern version of classical Arabic, is the language commonly used in the Arab-speaking countries today. This is the language of science, learning and other domains different than specific dialects of the Arabic countries.

2.1. Vowels and consonants

There are three short vowels in Arabic, /i/, /u/, and /a/, which phonemically contrast with their long counterparts, /ii/, /uu/, and /aa/. The pharyngalization of some consonants has an influence on all six vowels. They undergo some modifications in formants frequencies [8]. According to [9], taking into account the pharyngalization phenomena, the vocalic system of standard Arabic language can be organized as vocal sub groups: six pharyngalized vowels and six non-pharyngalized vowels (3 short and 3 long in each group). Arabic contains 28 consonant phonemes, which are characterized by physiological speech parameters consisting of both horizontal and vertical places of articulation and also phonetic features, which distinguish voiced and unvoiced consonants.

2.2. Syllables

Isolated phonemes do not give any information about their linguistic function. They can be pronounced only on a syllable. Such a feature is very interesting for language study at phonologic level and for the research of the functionality of each phoneme.

Like other languages, Arabic includes syllable as linguistic unit. In Arabic, the structure of a syllable is based on the phonemic system. The peak, or nucleus, is always the most prominent element of the syllable. It must be composed of a vowel: long or short. The onset always consists of single consonant, and a coda consists of zero, one, or two consonants. From this description of the syllable structure the types found for the six possible syllables may be postulated as follows (table 1):
Table 1: types of syllables on standard Arabic language.

<table>
<thead>
<tr>
<th>Syllable types</th>
<th>Arabic example</th>
<th>English meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>/bi/</td>
<td>“in, at”</td>
</tr>
<tr>
<td>CVV</td>
<td>/maa/</td>
<td>“what”</td>
</tr>
<tr>
<td>CVC</td>
<td>/min/</td>
<td>“from”</td>
</tr>
<tr>
<td>CVCC</td>
<td>/dars/</td>
<td>“lesson”</td>
</tr>
<tr>
<td>CVVC</td>
<td>/baab/</td>
<td>“door”</td>
</tr>
<tr>
<td>CVVCC</td>
<td>/maarr/</td>
<td>“passer-by”</td>
</tr>
</tbody>
</table>

The first three types occur much more often than the last three. CV, the simple open syllable, is the most frequent one, and CVVCC is the least frequent one.

Like syllabic types distribution, the first three patterns operate freely in the language, i.e., they occur in an utterance without restriction: initially, medially and finally. Depending on the context they are occurring, the syllables CV and CVV can be realized phonetically and phonologically as CVC. Such a phenomenon appears in 2 cases:

- whenever a liaison between two words is done. If the second word contains 2 consonants in utterance CV, (CVV) is changed in CVC e.g., /daraba lwaladul/ “the child hits”;
- whenever the second word begins with “hamza al-wasl”, e.g. /fii-?alkitaab/, “in the book”, becomes /fii-kitaab/.

However the last three types of syllables have constraints on their distribution. There are very few instances in which the CVCC; and CVVC patterns do not occur in the final position in pause forms. The last syllable CVVCC; usually occurs only as the final syllable of utterances or of one words in pause form. Several researchers do not take into account this last syllable. Such conclusion is confirmed by the study reported in [9].

2.3. Lexical stress

Syllables of an utterance are not produced with the same level of loudness. Syllable has three lexical stress levels: primary (1), secondary (2), and weak (3). As they were reported in [8], the rules that govern the lexical item stress are:

- whenever a word is made up of a string of the CV type syllables, the first syllable receives the primary stress and the remaining syllables receive weak stresses, e.g. /kat(1)tib(3)/CV(1)CV(1)CV(3) “he wrote”;
- when a word contains only one long syllable, the long syllable receives the primary stress and the rest of the syllables go unmarked, receiving weak stresses. The final long syllable never receives a primary stress. /kaat(1)tib(3)/CVV(1)CV(3) “writer”;
- for polysyllabic word, a stress is placed on the first long syllable counting from the penultimate. The nearest long syllable to the beginning of the word receives the secondary stress. /mus(3)taw(2)da(3)>a(1)t(3)hum(3)/

2.4. Prosody of standard Arabic language

In this study, our aim is to characterize the prosody of standard Arabic language specially variations of fundamental frequency, by taking into account the linguistic description previously presented.

The prosody, as defined in [10], is a description (phonetic aspect) and formal representation (phonologic aspect) of the oral elements of utterances such as stress, tons, intonation and quantity, of which concrete manifestation in production of speech is associated with the variations of the fundamental frequency (F0), the duration and the intensity (physical prosodic parameters).

The syllable is considered to be essential for language learning and for correct pronunciation of the intonation of speech. The realization of stress and ton on syllables gives it a prosodic characteristic.

The study presented in [11] and [12] shows that fundamental frequency is a pertinent parameter for perception of lexical stress in Arabic. The place of principal lexical stress is preserved on the sentence and the maximum of F0 curve corresponds to the stressed syllable [13]. This correlation of the realization of stress and variations of fundamental frequency is useful for automatic processing of prosody.

2.5. Extension of lexical stress on sentence level

Our perceptive study of interrogative [14] and affirmative intonation brings some modifications for the realization of lexical stress on the sentence level:

- taking into account variations of intonation levels and stress realization, we noticed that in the case of the liaison between two words, the last syllable of the first word, which has a weak stress in isolated case, receives, a second stress;
- monosyllabic prepositions (/fii/, /min/, /an/, /maa/) receives a second stress instead of a principal one;
- realization of stress for different levels, correlated with fundamental frequency variations, takes place with respect to declination from the beginning to the end of sentences, for affirmative utterances.

3. Neural Network Model

The major problem when using neural networks for automatic prosody generation is the selection of input parameters. Typically a neural approach uses information about the content and the relative position of linguistic unit as input parameters [15].

Our application is based on the linguistic description previously presented, which depends on the syllabic unit for prediction of F0 contours. For each syllable, we assign prosodic features correlated with phonological, phonetic and phonotactic description as information.

The neural network model presented here is based on the same principle used by the NET-talk System for prediction of English language pronunciation [16]. The network consists of a number of input units completely interconnected with a number of hidden units, which are in turn completely interconnected with a number of output units. Such a network can be trained by standard back-propagation algorithm of an error measure in order to develop an optimum value for the weights of each connection [17]. At the prediction stage, the
connection weights are used to test the inputs. The nonlinear sigmoid function is only used in the middle layer.

The purpose of the network is to learn how to relate F0 values to the target syllable, by taking into account the contextual information of one past and one future syllable of the input sequence. Figure 1 shows the architecture of the network.

3.1. Database.

We use a corpus, consisting of 112 utterances of declarative sentences with lengths from 1 to 10 words offering a large variety of lexical, syntactic and semantic forms. The corpus was read by a single native Arabic male. 100 sentences were used at learning step and 12 were kept for the generalization test.

The corpus is recorded in a soundproof room and digitized at 16 KHz and 16 bits/sample. There are 1280 syllables in our database. The intonative contour of each sentence is extracted from natural speech by using interactive software graciously provided by Elan-informatique\(^1\). This ensures an automatic stylization of F0 curves. The errors occurred when computing F0 values imply a manual tuning for correcting and normalizing data of F0 without changing of intonation of sentences at perceptive level. Also, the gap of unvoiced phonemes was automatically filled.

3.2. Input/output parameters.

We use phonologic, phonetic (acoustic, articaltory) and phonotactic descriptions as input parameters.

For the phonologic parameters we assign to each syllable the lexical accent with the precision of its level by taking into account the modifications of the rules of accentuation as presented in (§2.5) and the type of each syllable.

The phonetic parameters represent acoustic features of consonants and vowels (voiced/invoiced, short/long, pharyngized) and articularutory features (nasalization, friction, occlusion, liquid, etc.).

The phonotactic parameters include the position of each principal accent in sentences and the position of each syllable in the text and its boundary. Those parameters make it possible to distinguish the prosodical features of syllables and strengthen their classification.

The set of parameters is automatically extracted from the phonetic string, by using hierarchical modules: syllabification, classification of types of syllables, accentuation, phonetic classification and positional parameters. These parameters are encoded into an 8 bit digital representation. Each syllable accepts up to 10 symbols, resulting from the prosodic labeling.

Four analog output values are assigned to each syllable. The output layer is ranged between 0 and 1, which represents a linear mapping of the 60 Hz to 150 Hz interval appropriate with our speech material. We assign two F0 values for each phoneme of the syllable. Each phoneme has beginning and end frequencies; the least frequency is at the beginning of the next phoneme, if exists. CV and CVV syllables have three F0 values; but we add a fourth identical value to the third in order to keep the same size in the output layer. The CVCC syllable does not appear in our corpus.

3.3. Experiments and results interpretation

In order to generate F0 contours for arbitrarily sentences, we use a windowing technique for the input like the one adopted in NET-talk system. We propose for each input, a window that includes linguistic information of target syllable and those of the right and left syllables. The window represents a set of symbols of prosodic features of each syllable and has a size of 180 bits. Therefore the input layer consists of 180 neurons. In order to predict the F0 values for each syllable, the window is shifted along the string of symbols (Figure 1).

We test two neural architectures with different numbers of hidden layers. The first network has one hidden layer and consists of 90 neurons. The second network has two hidden layers and consists respectively of 90 and 45 neurons. The use of two hidden layers does not bring any improvement on the performance of neural network in comparison with the architecture with a single hidden layer. The training of network was carried out on a PIII processor in one hour and its optimization was achieved with a 10\(^2\) error degree. The performance of the network is also influenced by the choice of the size and the composition of the training corpus.

We tried various networks and input sizes. For the hidden layer level, the number of neurons varied from 40 to 100 nodes and the size of the window varied according to the parameters used. For each experience we omitted some positional parameters in order to evaluate the network behavior.

The prediction of intonative contours for sentences from the database is achieved with a correlation coefficient average of 0.96. This result explains that such a network is able to learn a small number of F0 contours by heart and to capture the prosodic patterns of syllables in different contexts. The networks trained without taking into account the positional parameters had a poorer behavior when generating the intonative contours of sentences that appear in the learning database.

\(^{1}\) Prosel software: Alignment of speech signal with phonetic string of text.
In the generalization phase, most sentences synthesized using intonative contours computed with the proposed neural network model are found to be equivalent from a perceptive point of view of professional listeners. Some samples of synthesis speech are available in: http://www.tsi.u-bordeaux.fr/zaki/arabic-synthesis-demo.htm

The comparison between the shapes of both natural and synthesized contours shows that synthesized contours follow the general trends of natural ones. The lexical accent takes place correctly with respect of declination phenomena, as can be seen in figure 2.

The poorer quality at perceptive level for the synthesized sentences can be explained by the fact that the network used is unable to acknowledge some prosodic features of some syllabic realization particularly at positional level. Despite this behavior, which is due to the small size of learning database, the intonative contours preserve in general the global shape of intonation realization.

![Figure 2: Natural and Synthetic F0 contours of sentence /kataba lwaladu risaalatan/ “the child writes a letter”](image)

4. Conclusion

In this paper we have formulated an approach consisting of a set of procedures based on linguistic and multi-layer feed forward neural networks, for automatically learning the prosody in Arabic speech. Our experiments show the feasibility of the approach and the applicability of such a network as a module for speech synthesis from text. The model is tested with synthesis system based on TD-PSOLA technique for Arabic language integrated in Prosel software by using natural segmental duration. Several comparisons between the utterances with synthesized F0 and natural sentences demonstrated clearly the ability of neural network to preserve the essence of intonation in F0 variations of whole sentence and various prosodic patterns of syllable unit. In our model, the syntactic information is not needed for automatic computing of intonative curve but positional parameters are essential for good classification of syllabic features. This characteristic and the encouraging results allow a perspective to generalize this model for the generation F0 contours for other modality of sentences in Arabic language: interrogative, exclamatory, call sentences, and imperative sentences.

5. References