DURATION MODELING FOR ARABIC TEXT TO SPEECH SYNTHESIS

Yasser Hifny¹ and Mohsen Rashwan¹, ²

¹Research & Development International (RDI)
²Department of Electronics and Communication Engineering, Cairo University, Egypt

(YHifny, MRashwan)@rdi-eg.com

ABSTRACT

Duration modeling is a fundamental task of prosody generation for Text To Speech (TTS) systems. The objective of this task is to predict the duration of a speech unit from its phonological representation. Duration modeling has a significant influence on the intelligibility and the naturalness of the synthesized speech. This paper presents a Neural Network (NN) based approach to predict the duration of Arabic phonemes. The developed model utilizes neural networks to map the relation between the phonological features and duration values.

1. INTRODUCTION

High quality speech synthesis is the ultimate goal of state-of-the-art TTS systems. Towards this objective, duration modeling is a well-known component that affects the quality of the synthesized speech. Duration modeling contributes implicitly during the calculation of the synthesized fundamental frequency contour. Moreover, synthesis by selection algorithms have a duration cost when exploring large databases to select the most likely sequence of occurrences which match the estimated targets from a given text [1]. Time scale prosodic modification algorithms utilize the predicted (target) durations to calculate the time scale modification factors for the selected speech units [2]. It is clear that duration modeling contributes to the most of the components that are usually utilized during the process of converting text to speech. Hence, accurate and robust duration modeling is one of the fundamental and interesting tasks of developing TTS systems.

Duration modeling has been addressed over many years. Klatt developed his known rule-based model to predict segmental durations for English; each segment is assigned a duration by a set of rules [3]. Rule based models have been developed for many TTS systems and for many languages. These rule systems were initially developed at a time when there was not much speech data and computer power available to analyze the data.

Corpus based approaches have recently seen increased interest in the field of language and speech processing. Corpus based approaches minimize the required human knowledge to describe a certain phenomena, provide reasonable solutions for difficult modeling problems, and outperform or have comparable results to rule based models if they exist. They also allow for rapid development cycles, and a unified approach to solving different problems. Hence, corpus based approaches have proved to be an effective and practical solutions for multilingual TTS synthesis. However, when perfect rule based solutions could be described, it will be the best to develop these solutions even if they are difficult since error sources could be reduced for large and complex systems. It is expected that hybrid approaches between rule based and statistical approaches will outperform each individual approach.

Non-parametric statistical modeling methods are attractive methods to predict the duration from text. Classification and Regression Trees (CART) models [4], Artificial Neural Networks (ANN) [6], and Multivariate Adaptive Regression Splines (MARS) modeling [8] have been successfully applied to predict duration from text [5, 7, 9]. MARS and ANN models have comparable prediction accuracy and outperformed CART models [9]. In this paper, a MultiLayer Perception (MLP) neural network is utilized to discover a relation between an Arabic phonology representation and the phoneme duration.

This work is developed in the framework of the ArabTalk TTS system for Arabic. ArabTalk is multi-user environment, phoneme based, large database, concatenative text to speech system. ArabTalk has a mature phonology framework. It has many perfect rule based models for automatic letter to sound, syllabification, morphological stress assignment, phonetic grammar validation, and consonants clusters elimination tasks. It has a statistical morphological analyzer to assign morphological diacritics and ANN-based prosody components for duration, energy, and global pitch contour prediction. In addition, it has a real time synthesis by selection algorithm to explore a large speech corpus. Moreover, ArabTalk has a MultiLayer Perception (MLP) based procedure to automatically align new voices to their phoneme boundaries level. Prosodic modification is a part of this large-scale project. All of this research is to advance the process of developing high quality Arabic TTS synthesis. Research is still going on to achieve state of the art quality.

In this paper, we will introduce the duration component of this system. The proposed model predicts the duration of a phoneme given its phonological representation. A brief introduction about the developed Arabic phonology framework will be introduced in section 2. The approach to prepare the speech-aligned corpus is presented in section 3. The description of the utilized features, neural network modeling and the experimental results for the proposed model are presented in section 4. Section 5, discusses briefly the developed model and introduces an application of the developed model to predict the average energy per sample for a phoneme given its phonological representation.
2. A FRAMEWORK FOR ARABIC PHONOLOGY

The phonological representation that has been developed contains all the information that is needed to drive the duration model. The details consist of a sequence of phones, accent/stress markers, and information about the boundaries of syllables, words and phrases.

The Arabic language has twenty-eight consonants and six vowels. The six vowels are divided into three short vowels and three long vowels. Each long vowel has similar spectral properties to those of their short vowel version but with longer durations. In general, Arabic orthography does not consider short vowels within the word structure. Hence, the phonetic transcription of Arabic orthography implies morphological and syntactic analysis to assign short vowels to a given word. ArabTalk uses statistical morphological analyzer to predict possible short vowel patterns for a sequence of words [10]. The possible generated patterns are based on the unified Arabic morphological language grammar. Arabic usually has simple one to one mapping between orthography and phonetic transcription for given correct diacritics. Some simple rule based methods are used to complement the generation of the phonetic transcription.

The word edge short vowel marker implies syntactic analysis for Arabic (syntactic diacritics). It is not easy to develop a syntactic Arabic analyzer. Currently, we assign a blind default diacritic type for the syntactic diacritics. We noticed that the prosody generation and the unit selection algorithms were affected directly by this procedure during the online synthesis so we have a strong interest to solve this problem. Hence, we suggested and developed a novel corpus-based approach as a workaround to predict the syntactic diacritics. Of course, syntactic diacritics could be supplied manually.

Arabic language has only six syllable types (CV, CVC, CVV, CVVC, CVCC and CVVCC). The last three types usually appear at the end of a phrase only due to their heavy pronunciation. The durations of the consonants and the vowel within these three types are known to be longer than the other remaining types. The number of vowels and the number of syllables in an Arabic phrase must be equal. Any stream of Arabic syllables must follow a simple syntax that is known. An Arabic syllable must start with only one consonant. Arabic syllabic structure prevents three consonants or two vowels to appear adjacently (“VV” symbols describe an Arabic long vowel). Hence, syllabification is achieved using a rule based syllable boundary detection algorithm. As Arabic has a prosodic nature to remove heavy pronunciation, consonant clusters, which may result during pronunciation, are eliminated using simple rule based model. Any violation to these definitions results in invalid Arabic phrase structure while parsing.

The phonological literature typically describes Arabic stress as predictably falling on a particular location in the word, depending on the internal structure of the syllables making up the word [11]. So, Arabic stress is known directly from the word syllable structure of a word. Hence, a rule based model could fit well to assign stress for a word without any need for a lexicon. Arabic stress assignment is different from that in the English language, which uses the stress as a free phoneme. Hence, Arabic stress is a morphological stress and it is not a lexical stress. The stress patterns are derived from an implementation of the stress assignment procedure, which is a combination of the work that has been developed by the phoneticians [12, 13, 14]. Further enhancements will be integrated to the current model when we develop Part Of Speech (POS) tagger for Arabic.

Currently, ArabTalk does not have any automatic procedure to assign different accent degrees to a word sequence. The accent degree for a word could be assigned manually or could be ignored during the transcription process. The last word of a phrase has a higher accent degree by default in the current implementation. An algorithm that changes the accent degree for a sequence of words is implemented. The primary objective of this procedure is to assign different accent degrees for function words and content words.

ArabTalk will integrate data driven approaches for prosodic phrasing and accent label predictions. We have suggested and developed the specifications for a new general-purpose text corpus for Arabic "AL-KHALIL" [15]. This corpus will have rich annotation tags for syntactic, prosodic phrasing, and accents. These tags are assigned to guide statistical models to discover some rules about Arabic grammar and Arabic semantics.

Parsing a given utterance text results in a prosodic tree, which is constructed to represent the different levels of the phonological description and the relationship between these levels.

3. CORPUS GENERATION

A large database may cover enough phonetic and prosodic variation in the language so that most of the units that are required during the synthesis process may be contained within the database. Hence, large databases contribute directly to the quality of the synthesized speech. State of the art TTS systems attempt to minimize prosodic and spectral modifications for high quality speech synthesis. The large database approach is an attractive idea in this direction. Real time unit selection algorithms have been developed to explore large databases [1], a technique known as synthesis by selection.

ArabTalk is a large database-based system so we need an automatic procedure to align a speaker corpus with lower level labels such as phones. We use a HMM based viterbi alignment procedure that is developed for this purpose [16]. The viterbi alignment procedure can be summarized as a problem of searching time boundaries for known sequence of HMM models for phonemes. Since the best state sequence, which is known to be the viterbi path, is obtained while decoding process, time boundaries can be obtained directly. This process is illustrated in figure (1). Actually, viterbi (forced) alignment procedure results are reasonably good but the labels need to be more accurate for a synthesis database than for recognition. Hence, we did many manual corrections and we have developed many tools to correct and move boundaries for similar error patterns automatically.
4. DURATION MODELING

Based on the suggested and developed framework for Arabic many features (factors) are extracted to describe each phoneme in the speech corpus. Phonology to duration modeling is achieved via a MLP neural network.

4.1. Features selection

Phonological factors that affect the duration modeling have a long history of research since Klatt’s work [3]. Some of these factors are identities of the phonetic segments, syllabic stress, word importance or word accent, location of the syllable in the word, and the location of the syllable in the phrase [17]. In this work, detailed in table (1), some factors are specific for Arabic such as Emphatic, Shadda, and Tanween factors. The others are common features like those used in [7]. The possible values for these factors are selected based on the Arabic phonetics [12].

<table>
<thead>
<tr>
<th>Phonology Level</th>
<th>Feature Description</th>
<th>Feature Count</th>
<th>Possible Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoneme</td>
<td>Type of articulation</td>
<td>3</td>
<td>1 to 13</td>
</tr>
<tr>
<td></td>
<td>Place of articulation</td>
<td>3</td>
<td>1 to 15</td>
</tr>
<tr>
<td></td>
<td>Emphatic type</td>
<td>1</td>
<td>0 to 1</td>
</tr>
<tr>
<td></td>
<td>Shadda</td>
<td>1</td>
<td>0 to 1</td>
</tr>
<tr>
<td></td>
<td>Tanween</td>
<td>1</td>
<td>0 to 1</td>
</tr>
<tr>
<td>Syllable</td>
<td>Phoneme position</td>
<td>1</td>
<td>1 to 4</td>
</tr>
<tr>
<td></td>
<td>Count of phonemes</td>
<td>1</td>
<td>2 to 4</td>
</tr>
<tr>
<td></td>
<td>Accent degree</td>
<td>1</td>
<td>0 to 4</td>
</tr>
<tr>
<td>Foot</td>
<td>Syllable position</td>
<td>1</td>
<td>1 to 10</td>
</tr>
<tr>
<td></td>
<td>Count of syllables</td>
<td>1</td>
<td>1 to 10</td>
</tr>
<tr>
<td>Phrase</td>
<td>Foot position</td>
<td>1</td>
<td>0 to 3</td>
</tr>
</tbody>
</table>

Table (1) Phonological feature description

4.2. Phonology to duration mapping

As shown figure (2), phonology to duration modeling is achieved via a BP neural network. The network has one output, which is the duration value in milliseconds encoded between 0.1 to 0.9. The duration value is normalized to this range by extracting the maximum and minimum duration values from the aligned corpus. The extracted features are encoded to be valid inputs for the NN. The input layer has 111 nodes, which results from the possible values for the input features. The two hidden layers have 15 and 5 nodes respectively. This structure is assumed to discover a relationship between the input factors and the corresponding duration value. For example, it is known that the syllables at the end of a phrase in English and Arabic have longer lengths, so it is expected that the model will capture this rule.

The aligned speech corpus is divided into three sets. The first set is for training the neural network; the second set is used for cross validation. Cross validation is known to be one of the best methods to avoid over fitting and provide better generality models. The training is stopped when the Mean Square Error (MSE) on the validation set stops decreasing significantly or starts increasing. The last set is used for testing and evaluating the prediction accuracy. The training and testing procedures are based on the NN simulator that was developed for similar task [18]. During synthesis, the predicted duration values are re-mapped to the corresponding millisecond values.
4.3. Experimental results

The training sessions are done using one-hour labeled database. The aligned database contained 43541 phonemes (= 12 phonemes per second). These phonemes were divided into three sets. The training set contained 32239 phonemes (= 74% of the data). The test set contained 8458 phonemes (= 20% of the data). The validation set, which was used for cross validation, contains 2844 phonemes (= 6% of the database). In order to evaluate the prediction accuracy, a correlation coefficient ($p$) and Relative Mean Square Error (RMSE) were computed between the predicted values ($y$) and actual duration values ($x$) for the test set as shown in Table (2). Moreover, we have integrated this model into the ArabTalk system to evaluate the developed model practically. The output-synthesized speech using this model provides very good and promising results.

<table>
<thead>
<tr>
<th>Evaluation Method</th>
<th>Evaluation equation</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>$RMSE = \sum_{i} \left( \frac{x_i - \mu_x}{\sigma_x} \right)^2$</td>
<td><em>0.26</em></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>$\rho_{xy} = \frac{\sum_{i} (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i} (x_i - \mu_x)^2 \sum_{i} (y_i - \mu_y)^2}}$</td>
<td><em>0.87</em></td>
</tr>
</tbody>
</table>

Table (2) Prediction accuracy measurements

5. DISCUSSION AND CONCLUSION

An application of artificial neural networks for duration modeling for Arabic Text To Speech synthesis is presented. NN usually provide a flexible and simple approach for function approximation from an available corpus. Moreover, reported results provide positive feedback about our suggestions and the implementation for an Arabic phonology framework. Hence, the suggested framework for Arabic phonology has been utilized for intonation prediction and synthesis by selection of ArabTalk system. This model is used to predict the average energy per sample of a phoneme from its phonological representation. This is achieved by changing the output value of the neural network from the duration to the average energy per sample for a phoneme. The energy model achieved correlation coefficient accuracy of 0.91. As expected, corpus based approach provides a generic and fast method to develop new models, new voices, and new languages without devoting a great deal of effort to describe the relationship between the input and output of the modeled process.

ACKNOWLEDGEMENT

This work has been developed at Research & Development International (RDI) research labs. The authors would like to thank all the members of the Speech Department for their support. This research work has been supported by RDI research grants SD-TTS-00-001 and SD-TTS-00-002.

REFERENCES