F0 modeling using DNN for Arabic parametric speech synthesis

Imene Zangar, Zied Mnasri, Vincent Colotte, Denis Jouvet

F_0 modeling using DNN for Arabic parametric speech synthesis

Imene Zangar[1], Zied Mnasri[1,3], Vincent Colotte[2], and Denis Jouvet[2]

[1] University Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, Electrical Engineering Department, Tunis, Tunisia {imene.zangar,zied.mnasri}@enit.utm.tn
[2] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France {vincent.colotte,denis.jouvet}@loria.fr
[3] Università degli studi di Genova, DIBRIS, Genoa, Italy

Abstract. Deep neural networks (DNN) are gaining increasing interest in speech processing applications, especially in text-to-speech synthesis. Actually state-of-the-art speech generation tools, like MERLIN and WAVENET are totally DNN-based. However, every language has to be modeled on its own using DNN. One of the key components of speech synthesis modules is the prosodic parameters generation module from contextual input features, and more particularly the fundamental frequency (F_0) generation module. Actually F_0 is responsible for intonation, that is why it should be accurately modeled to provide intelligible and natural speech. However, F_0 modeling is highly dependent on the language. Therefore, language specific characteristics have to be taken into account. In this paper, we aim to model F_0 for Arabic speech synthesis with feedforward and recurrent DNN, and using specific characteristic features for Arabic like vowel quantity and gemination, in order to improve the quality of Arabic parametric speech synthesis.

Keywords: Arabic parametric speech synthesis · Fundamental frequency F_0 · Deep neural networks · Recurrent neural networks

1 Introduction

Speech processing systems have gained increasing interest since a few years. Actually, new technology advances have made possible to interact with machines, through speech recognition and speech synthesis. Speech synthesis could be performed from text (Text-to-speech synthesis). In text-to-speech synthesis, the text is processed to obtain a sequence of units (phonemes, diphones, syllables, etc). The parameters of which are predicted by dedicated modules. These parameters could be classified into prosodic parameters and acoustic parameters, mainly mel-cepstrum coefficients like Mel-Generalized Cepstral coeffcients (MGC), Mel-Frequency Cepstral Coefficients (MFCC), Linear Spectral Pair (LSP), etc. and their temporal derivatives (Δ and $\Delta\Delta$). Prosodic parameters include segment duration and fundamental frequency (F_0). Both parameters are jointly responsible for rhythm and intonation.

F_0 modeling has always been the cornerstone of any speech synthesis system. In fact the evolution of F_0 is the manifestation of complex and interdependent phonological phenomena like intonation, accentuation and voicing. Therefore, an accurate F_0 model is necessary to produce intelligible and naturally sounding synthesis speech. So far, there has been a variety of F_0 models, highly accurate and successfully applied to speech synthesis systems. However, most of them are language dependent, since each F_0 model was established for a specific language. Furthermore, when a model is applied for another language, it should undergo many adjustments to fit the new target language. F_0 models could be classified into phonological vs. phonetic models. Phonological models can also be divided into tone-sequence and perceptual models. In tone-sequence models like ToBI (Tone and break indices) [1] the intonation of an utterance is described as a sequence of phonologically opposite tones, High (H) and Low (L). The different combinations of both tones give a finite state grammar. Then F_0 is regarded as an intonation event (high/low/rising/falling); whereas in perceptual models like IPO (Institute of perception research) model [2], the F_0 contour is described by a sequence of the most relevant movements like prominence. In this model, the intonation contour consists of a linear sequence of discrete intonational elements, i.e. the most relevant F_0 movements (but not tones). On the other side, i.e. the phonetic models, F_0 is regarded as a physical quantity to be measured/predicted. These models try to establish an analytic formulation or approximation of F_0 and/or its variations. Tilt model [3, 4], and PaIntE, Parametric Intonation Event System model [5] are amongst the well known analytic F_0. The Tilt model attempts to label the utterance by one of four intonational events, i.e. pitch accents, boundary tones, connections and silence. The PaIntE model is similar to the Tilt model as it tries also to model the accents using a small set of parameters. Moreover, this model uses the sum of two sigmoid functions to represent F_0 contour locally.

With the development of machine learning, speech synthesis has been taking benefit from data-driven models to predict prosodic parameters, including F_0, directly from input features, using machine learning. Amongst data-driven models, HMM (Hidden Markov Models) have been used for parametric speech synthesis, since 1999 [7]. More recently, deep neural networks (DNN) have taken benefit of the development of GPU to become the leading technique in machine learning. With the recent advances of DNN, parametric speech synthesis and more particularly prosodic parameters prediction modules have been migrating from HMM-based model to DNN-based ones, in the aim to increase accuracy, and especially to avoid the over-averaging problem, noticed in HMM-predicted parameters [15]. Therefore, feedforward and recurrent DNN have been integrated to parametric speech synthesis for many languages, such Japanese and English [8]. Also for Arabic, DNN have been recently used to predict phone duration for Arabic parametric speech synthesis [10], through a dedicated DNN model for each class of Arabic phones, i.e. short vs. long vowels. and simple vs. geminated consonants. Initially, DNN were inserted in HMM-based parametric speech synthesis as a replacement of CART (classification and regression trees) which are

used for HMM models clustering. In a second approach, DNN were used to predict raw prosodic parameters, that were injected to HTS models. For instance, DNN were proved to be more efficient than HMM to model segmental durations [9], [10]. Also, state-of-the-art TTS tools like MERLIN [6] and WAVENET [11] are fully based on DNN. In this work, the studied problem is F_0 modeling using DNN and adding specific features for Arabic, i.e. vowel quantity and gemination, to the standard parametric speech synthesis feature set [10].
This paper is organized as follows: Section 2 reviews, F_0 modeling for parametric speech synthesis, with a special focus on DNN models; Section 3 details the proposed DNN models used in this work; and Section 4 presents, the experiments and the results.

2 F_0 modeling for parametric speech synthesis

Parametric speech synthesis was originally designed using HMM models in HTS (HMM-based text-to-speech synthesis) system [7], including many updates regarding the use of multi space probability distributions (MSD) [12], and the hidden semi-markov models [13]. However, since a few years, DNN models are used for parametric speech synthesis, and lead to an enhance the quality of the generated speech.

2.1 F_0 modeling in HMM-based parametric speech synthesis

The parameters modeling in HMM-based parametric speech synthesis is based on a five-state left-to-right HMM modeling. Each state is modeled by a single Gaussian distribution with diagonal covariance. Decision trees are used to cluster HMM model according to the contextual input features. Finally, expectation-maximization (EM) algorithm is used to predict the HMM model output.
In HTS system, the prosodic parameters, including duration and F_0 were originally modeled using HMM. However for F_0, it is a special process, since F_0 is defined only in voiced regions of speech, where it is represented by one-dimensional continuous values. However, in unvoiced regions F_0 is replaced by a discrete label, so that it is impossible to apply the same HMM model for F_0 in voiced and unvoiced parts of the speech signal. Therefore, a multi-space probability distribution (MSD probability) was designed to model continuous and discrete F_0 respectively in voiced and unvoiced regions using HMM [14].
Nevertheless, F_0 modeling is still suffering from over-averaging, since the predicted Gaussian distributions tend to provide mean values of F_0. This phenomenon has been related in subjective tests of HTS vs. unit selection synthesis, where the latter technique was preferred [15], though it is older.

2.2 F_0 modeling in DNN-based parametric speech synthesis

In [8], $log(F_0)$ was modeled by DNN. In this work, F_0 was first interpolated in the unvoiced regions to provide continuous F_0 values as set by [16]; then $log(F_0)$

is modeled using DNN. Besides, voicing decision labels were predicted.
In the preprocessing phase, 80% of silence data was removed from the training set. Input features were normalized to have a zero-mean and a unit-variance, whereas output targets were normalized to be in the interval $[0.01, 0.99]$. The DNN was trained using stochastic gradient descent (SGD), whereas the activation functions were all sigmoids, including the output layer. Different DNN architectures were trained on a speech corpus containing 33000 English utterances, covering nearly 13 hours. The best performing model was a feedforward DNN containing 5 hidden layers, each with 2048 nodes. Results show that DNN gave more accurate voicing decision prediction than HMM, whereas HMM gave the lowest $log(F_0)$ $RMSE$ error [8].
In [17], F_0 was interpolated in unvoiced regions using an exponential decay function [18]. To predict the voicing decision label and $log(F_0)$ values many DNN architectures were investigated based on feedforward and recurrent layers, i.e. LSTM (longs short-term memory) and BLSTM (Bidirectional LSTM). For instance, a 6-feedforward-hidden-layer DNN with 512 nodes at each hidden layer, a 3-feedforward-hidden-layer DNN with 1024 nodes at each layer, a hybrid 3-feedforward-layer and 1-LSTM-layer and a hybrid 2-feedforward-layer and 2-BLSTM-layer with 512 nodes at each layer, for both hybrid architectures, and using sigmoid activation function were used. 5000 Chinese utterances covering 5 hours of speech were used for training and development, and 200 utterances for test. Results show that in comparison to the baseline HMM model, used in HTS [16], the accuracy of voicing decision label prediction is nearly the same, whereas F_0 prediction $RMSE$ was lower for all DNN architectures, but without a significant difference between the different architectures, all varying between 15 to 16 Hz.
More recently, in [19], a hybrid 2-feedforward-layer and 2-LSTM-layer DNN with 512 nodes at each hidden layer, was trained on a 13-hour Chinese speech corpus. The $RMSE$ calculated on the test set was 12 Hz.
It should be noted that in all these architectures, the standard features set of HTS system [24] was used. No specific language features were added to model the language characteristics.

3 Proposed DNN models

F_0 modeling using DNN is a two-stage process, which requires first classifying speech segments into voiced and unvoiced regions, and then predicting F_0. Since both tasks are of different nature, i.e. classification and regression, we opted to train a different network for each.

3.1 Proposed architectures for discrete voiced/unvoiced decision classification and F_0 values prediction

Two different neural networks were used to predict voicing decision and F_0 values. For each DNN, many architectures were tried out, using dense layers only, dense layers with LSTM or BLSTM layers, and LSTM or BLSTM layers only (cf. Table 1). Actually, the intent from the choice of these architectures is to see whether recurrent neural network are more suitable than feedforward networks to model speech parameters, which are known to be highly recurrent.

Table 1. DNN architectures selected based on results on development set for voiced/unvoiced decision classification and for F_0 prediction.

DNN network	Number and types of hidden layers	Number of nodes in hidden layers
All-Dense	6 Dense layers	[512,512,512,512,512,512]
Dense-LSTM (1)	2 Dense layers + 4 LSTM layers	[1024,1024,512,512,256,256]
Dense-LSTM (2)	2 Dense layers + 4 LSTM layers	[512,512,512,512,512,512]
Dense-BLSTM (1)	2 Dense layers + 4 BLSTM layers	[1024,1024,512,512,256,256]
Dense-BLSTM (2)	2 Dense layers + 4 BLSTM layers	[512,512,512,512,512,512]
All-LSTM	6 LSTM layers	[512,512,512,512,512,512]
All-BLSTM	6 BLSTM layers	[512,512,512,512,512,512]

Also, it should be noted that after some experiments, all kept networks were implemented with *tanh* activation function for hidden layers, whereas linear output activation was used for F_0 and sigmoid output activation was used for voicing decision prediction. The *Adam* optimizer was used to train the voicing decision classification DNN whereas for the F_0 prediction DNN, the *Rmsprop* optimizer was used. The batch size was set to 100 in all the experiments.

3.2 Error minimization criteria

Since voicing decision prediction is a classification problem, cross-entropy loss function was used

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} v_i log(p(v_i)) \tag{1}$$

where N is the number of frames, v_i is the voicing decision value (0 or 1) of frame i, and $p(v_i)$ is its probability.
As far as continuous F_0 values prediction is concerned, MSE was selected as loss function, since it's a regression problem.

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(log(F_{0i}) - log(\hat{F}_{0i}))^2 \tag{2}$$

where N is the number of frames, $log(F_{0i})$ and $log(\hat{F}_{0i})$ are the target and the predicted values of frame i, respectively.
To avoid overfitting, early-stopping option was used. Thus, if L_{CE} (in case of voicing decision network) or L_{MSE} (in case of F_0 prediction network) evaluated on the development set, don't improve after a certain number of epochs, set to 20 in our case, the training process is stopped.

4 Experiments and results

4.1 Speech corpus

To train the DNN-based F_0 model, an Arabic speech corpus of 1597 utterances was used [20]. The utterances correspond to news bulletin in Modern Standard Arabic (MSA), read by a native-Arabic male speaker. The signals were recorded at a sampling rate of 48 KHz and a precision of 16 bits. 70% of the utterances were used for training, 20% for development and 10% for test. To extract original voicing decision labels and F_0, SWIPE algorithm [23], included in SPTK toolkit [21] was used. Since the speaker is male, the F_0 tracking algorithm was bound between 80 Hz and 320 Hz. A single value of F_0 was extracted at each 5-ms frame. Voicing decision labels were deduced automatically using the fact that null F_0 values corresponds to unvoiced frames.

4.2 Features selection and preprocessing for F_0 modeling

In addition to the standard HTS model input features set [24], classically used for parametric speech synthesis, two Arabic specific features, namely vowel quantity and gemination were added. Actually both features have been recently proved to enhance the quality of Arabic speech synthesis using DNN [22]. The input features were coded in different ways, with respect to their natures. Thus three types of feature encoding were used: yes/no features like stressed/non-stressed syllables were coded into binary values; class-wise features, such as phoneme identity, were coded with one-hot vectors and for unlimited-value features coarse coding was used. For example, with coarse coding, the relative position of phonemes in the syllable, i.e. beginning, middle or end, is encoded respectively, (1,0,0), (0,1,0) or (0,0,1) whatever the number of phonemes in the syllable. Thus, the input feature vector contains 439 coefficients. Moreover, the coefficients of the input features vector were normalized to have a zero-mean and a unit-variance. On the other side, a linear interpolation was applied to the output $log(F_0)$ targets in the unvoiced parts of speech. Then, interpolated values were normalized to be within [0.01, 0.99] interval. However, output voicing decision labels were not normalized, since they are already binary.

4.3 Selected DNN models

Various network parameters were experimented, i.e. various number and type of hidden layers, number of nodes in hidden layers and activation functions were empirically modified. The models were evaluated on the development set, using $RMSE(Hz)$ for predicted F_0 values and voicing decision error (VDE). The best models on the development set (cf. Table 2) were kept to be evaluated on the test set.

4.4 Objective evaluation

To assess the quality of the predicted parameters, i.e. $log(F_0)$ and voicing decision label, four measures were performed first on the development set, to select the best performing model, and then on the test set:

- **Root mean square error ($RMSE$)** between target and predicted F_0 values (cf.(3)) on voiced frames:

$$RMSE(Hz) = \sqrt{\frac{1}{N_v}\sum_{i=1}^{N_v}(F_{0i} - \hat{F}_{0i})^2} \tag{3}$$

  where N_v is the number of originally voiced frames and F_{0i} and $\hat{F}_{0i}$ are respectively the original and the predicted values of F_0 values at the originally voiced frame i.
- **Voicing Decision Error (VDE)**, which represents the proportion of frames for which a wrong voiced/unvoiced prediction is made. Actually VDE is the percentage of bad-predicted labels, i.e. false positives and false negatives:

$$VDE(\%) = \frac{N_{V\to U} + N_{U\to V}}{N} \times 100 \tag{4}$$

  where $N_{V\to U}$, $N_{U\to V}$ and N are respectively the number of voiced frames predicted as unvoiced, the number of unvoiced frames predicted as voiced and the total number of frames.
- **Gross pitch error (GPE)**, which is a measure which combines both F_0 prediction and voicing prediction classification. Actually GPE is the percentage of frames for which the F_0 relative error is higher than a certain threshold, set to 20% among the frames that are originally voiced and also predicted as voiced:

$$GPE(\%) = \frac{N_{GPE}}{N_{V\to V}} \times 100 \tag{5}$$

  where N_{GPE} and $N_{V\to V}$ are respectively the numbers of frames originally voiced and predicted as voiced for which $|F_0 - \hat{F}_0| > F_0 \times 0.2$ and $N_{V\to V}$ is the total number of frames originally voiced and predicted as voiced.
- **F0 Frame Error (FFE)**, is a measure which combines Gross Pitch Error (GPE) and Voicing Decision Error (VDE). Actually FFE is the percentage of frames for which an error is made, either according to the VDE or GPE criteria:

$$FFE(\%) = \frac{N_{V\to U} + N_{U\to V} + N_{GPE}}{N} \times 100 \tag{6}$$

The objective evaluation consists in comparing the performance of DNN modeling to state-of-art models, i.e. HMM model as used in HTS [7], DNN model as used in MERLIN [6] (cf. Table 3). The DNN model as used in MERLIN is composed by 6 hidden layers with 1024 units each and *tanh* as activation transfer function. This model relies on the same set of features as HTS. It should be emphasized that the acoustic parameters are generated from HTS or MERLIN using the original duration to compare the original and predicted vector F_0.

Table 2. Objective evaluation results for the proposed models

Proposed DNN model	Development set				Test set			
	RMSE (Hz)	VDE (%)	GPE (%)	FFE (%)	RMSE (Hz)	VDE (%)	GPE (%)	FFE (%)
All-Dense	33.41	3.51	22.14	14.14	37.86	3.35	26.98	17.48
Dense-LSTM (1)	17.65	3.08	5.05	5.50	19.23	2.78	6.82	6.38
Dense-LSTM (2)	**15.70**	2.92	**2.41**	**4.09**	18.14	2.71	4.43	5.06
Dense-BLSTM (1)	19.72	2.88	5.85	5.71	20.14	2.38	6.33	5.74
Dense-BLSTM (2)	19.93	**2.59**	4.98	5.04	21.39	2.51	6.39	5.93
All-LSTM	16.70	2.95	2.62	4.22	**17.71**	2.79	**3.38**	**4.59**
All-BLSTM	15.95	2.64	3.16	4.18	18.65	**2.32**	6.57	5.82

Table 3. Comparison of objective evaluation results on test set for the best selected models on the development set and for the various modeling approaches

models	RMSE(Hz)	VDE(%)	GPE (%)	FFE(%)
HMM model from HTS [7]	30.52	7.19	8.90	11.93
All-Dense DNN model from MERLIN [6]	36.93	4.73	9.11	9.46
Dense-LSTM (2)	**18.14**	2.71	**4.43**	**5.06**
Dense-BLSTM (2)	21.39	**2.51**	6.39	5.93

4.5 Discussion

In comparison to the state-of-the-art results (cf. section 2.2), and taking into account the small size, only 3 hours, of the used corpus, the $RMSE$ results are quite near of the 15-Hz-$RMSE$ obtained using only feedforward DNN trained on a 13-hour English corpus [8], or the 12-Hz-$RMSE$ given by a 13-hour Chinese corpus, trained with a DNN containing dense and LSTM layers [17].

In Table 2, looking to VDE values, it looks obvious that using recurrent networks gives better voicing decision prediction results. The Dense-BLSTM (2) model leads to the best performance on the development set, with a VDE of 2.59%. The corresponding 95% confidence interval is ±0.05%; which make this result significantly better than that of the other models. This model leads to a VDE of 2.51% (±0.07%) on the test set.

In Table 3, the results show that using Dense-LSTM and Dense-BLSTM to model respectively F_0 and voicing decision label, leads to better performance compared to state of the art models, i.e. HMM model as used in HTS [7], DNN model as used in [6]. Finally the combined F_0 and voicing decision measures like GPE and FFE show that using recurrent networks is more fitted to the problem of F_0 modeling. This could be explained by the recurrent nature of speech, where consonants, mostly unvoiced or partly voiced, are followed by voiced vowels.

5 Conclusion

In this paper, F_0 modeling was achieved using DNN for Arabic parametric speech synthesis. F_0 modeling is a two-fold process, which requires the prediction of the voicing nature (voiced or unvoiced) of speech segments, and then the prediction of F_0 values in voiced parts. An Arabic speech corpus was used to train two DNN models, one for voicing classification and one for F_0 prediction, using the standard features set for parametric speech synthesis, in addition to two Arabic-specific features, vowel quantity and gemination. Several architectures were tried out for both networks, using (i) feedforward (dense) layers only, (ii) a combination of feedforward and recurrent layers (LSTM or BLSTM) and (iii) recurrent layers only. Objective results using standard metrics for F_0 prediction quality, like $RMSE$, VDE, GPE and FFE show that the best F_0 prediction results are obtained with recurrent networks.

Acknowledgement

This research work was conducted in the framework of PHC-Utique Program, financed by CMCU (Comité mixte de coopération universitaire), grant No15G1405.

References

1. Pierrehumbert, J.: The phonology and phonetics of English intonation. Ph.D. Thesis, Massachusetts Institute of Technology (1980)
2. Hart, J., Collier, R., Cohen, A.: A perceptual study of intonation. Cambridge University Press (1990)
3. Dusterhoff k. and Black A.: Generating F_0 contour for speech synthesis using the Tilt Intonation Theory. In: 3rd ESCA workshop on Intonation: Theory Models and Applications, pp.107-110. Athens, Greece (1997)
4. Taylor, P.: Analysis and synthesis of intonation using the tilt model. Journal of the Acoustical Society of America **107**(3), 1697–1714 (2000)
5. Moehler, G., Conkie, A.: Parametric modeling of Intonation using vector quantization. In: 3rd ESCA workshop on speech Synthesis, pp.311–316. Jenolan Caves, Australia (1998)
6. Wu, Z., Watts, O., King, S.: Merlin: An open source neural network speech synthesis system. In: 9th ISCA Workshop on Speech Synthesis, pp.202–207. Sunnyvale, USA (2016)
7. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In: 6th European Conference on Speech Communication and Technology, pp.2347-2350. Budapest, Hungary (1999)
8. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: 38th International Conference on Acoustics, Speech, and Signal Processing, pp. 7962-7966. IEEE, Vancouver, Canada (2013)
9. Chen, B., Bian, T., Yu, K.: Discrete duration model for speech synthesis. In: 18th Annual Conference of the International Speech Communication Association, pp. 789-793. Stockholm, Sweden (2017)

10. Zangar, I., Mnasri, Z., Colotte, V., Jouvet, D., Houidhek, A.: Duration modeling using DNN for Arabic speech synthesis. In: 9th International Conference on Speech Prosody, pp. 597-601. Poznan, Poland (2018)
11. A.V.D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, , K. Kavukcuoglu: Wavenet: A generative model for raw audio. arXiv preprint arXiv: 1609.03499 (2016)
12. Yoshimura, T.: Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-to-Speech systems. Ph.D. Thesis, Department of Electrical and Computer Engineering, Nagoya Institute of Technology (2002)
13. Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Hidden semi-Markov model based speech synthesis. In: 8th International Conference on Spoken Language Processing, pp. 1393-1396. Jeju Island, Korea (2004)
14. Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T.: Multi-space probability distribution HMM. IEICE TRANSACTIONS on Information and Systems **85**(3), pp. 455–464 (2002)
15. H. Zen, K. Tokuda, A.W. Black: Statistical parametric speech synthesis. In: Speech Communication 2009, vol. 51, pp. 1093–1064. ELSEVIER (2009). https://doi.org/10.1016/j.specom.2009.04.004
16. Yu, K., Young, S.: Continuous F_0 modeling for HMM based statistical parametric speech synthesis. IEICE TRANSACTIONS on Information and Systems **19**(5), pp. 1071–1079 (2011)
17. Fan, Y., Qian, Y., Xie, F. L., Soong, F. K.: TTS synthesis with bidirectional LSTM based recurrent neural networks. In: 15th Annual Conference of the International Speech Communication Association, pp. 1964-1968. Singapore (2014)
18. Chen, C. J., Gopinath, R. A., Monkowski, M. D., Picheny, M. A., Shen, K.: New methods in continuous Mandarin speech recognition. In: 5th European Conference on Speech Communication and Technology, pp. 1543-1546. Rhodes, Greece (1997)
19. Chen, B., Lai, J., Yu, K.: Comparison of modeling target in LSTM-RNN duration model. In: 18th Annual Conference of the International Speech Communication Association, pp. 794-798. Stockholm, Sweden (2017)
20. Halabi, N., Wald, M.: Phonetic inventory for an Arabic speech corpus. In: 10th International Conference on Language Resources and Evaluation, pp. 734-738. Slovenia (2016)
21. Speech Signal Processing Toolkit (SPTK), http://sp-tk.sourceforge.net/
22. Houidhek, A., Colotte, V., Mnasri, Z., Jouvet, D.: DNN-Based Speech Synthesis for Arabic: Modelling and Evaluation. In: 6th International Conference on Statistical Language and Speech Processing, pp. 9-20. Mons, Belgium (2018)
23. Camacho, A., Harris, J. G.: A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America **124**(3), pp. 1638–1652 (2008)
24. Zen, H.: An example of context-dependent label format for HMM-based speech synthesis in English. The HTS CMUARCTIC demo (2006)