9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Duration modeling using DNN for Arabic speech synthesis

Imene Zangar[1], *Zied Mnasri*[1], *Vincent Colotte*[2], *Denis Jouvet*[2], *Amal Houidhek*[1, 2]

[1]University Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis,
Electrical Engineering Department, Tunisia
[2]Inria, Villers-lès-Nancy, F-54600, France
[2]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[2]CNRS, LORIA, UMR 750, Villers-lès-Nancy, F-54600, France
imene.zangar@enit.utm.tn, zied.mnasri@enit.utm.tn,
vincent.colotte@loria.fr, denis.jouvet@loria.fr, amal.houidhek@loria.fr

Abstract

Duration modeling is a key task for every parametric speech synthesis system. Though such parametric systems have been adapted to many languages, no special attention was paid to explicitly handling Arabic speech characteristics. Actually, in Arabic phoneme duration has a distinctive role, because of consonant gemination and vowel quantity. Therefore, a precise modeling of sound durations is critical. In this paper we compare several modeling of phoneme durations (including duration modeling by HTS and MERLIN toolkits), and we propose a new approach which relies on using a set of models, each one being optimal for a given phoneme class (e.g., simple consonants, geminated consonants, short vowels, and long vowels). An objective evaluation carried out on a set of test sentences shows that the proposed approach leads to a more accurate modeling of the phoneme durations.

Index Terms: Arabic TTS, phoneme duration modeling, HTS, MERLIN, DNN.

1. Introduction

Text-to-speech synthesis (TTS) has become a useful component in many voice applications, such as online translators and text message readers. Furthermore, TTS is nowadays available for most widely spoken languages all over the world on the main online services. Hence, it is important to have high quality TTS for Arabic language since it represents a large market with more than 300 million potential users.

Following the development of the PSOLA technique [1], concatenative TTS has been the dominant method as it made it possible to build automatically a good-quality synthetic voice using a relatively small corpus with an appropriate segmentation, mainly into diphones [1]. While concatenating the adjacent segments, spectrum smoothing is performed using time or frequency-domain pitch synchronous overlap and add (TD/FD-PSOLA).

Less than a decade after, a higher quality of TTS was obtained by unit selection [2]. Unit selection TTS is based on the selection and the concatenation of original speech segments, such as phonemes, diphones or demisyllables [3]. The selection and the concatenation of units are based on minimizing the weighted sum of two costs namely the target cost, that is the difference between a candidate unit and the target, and to the concatenation cost, which evaluates the quality of joining consecutive units [2]. Hence, unit selection does not require any prosodic or spectral modification of the selected units, which leads to a very high level of naturalness of the generated speech. However, a huge database is required to provide units covering most of the speech units, unless the TTS system is designed for a specific application.

In the last decade, taking advantage of the success of stochastic speech modeling, especially for automatic speech recognition [4], statistical parametric speech synthesis (SPSS) based on hidden Markov models (HMM) was successfully developed [5]. This approach has many advantages in comparison to the concatenative speech synthesis technique such as stable quality of speech synthesis, robustness for speaker adaptation [6] and the flexibility to change voice characteristics [7]. This technique is the backbone of HTS system [5], which has successfully been transformed into a multilingual TTS system [7].

Since a few years, deep learning and mainly deep neural network (DNN) technique has been growing so fast that it becomes the must-have in most data-driven systems. DNN allows modeling large data sets with high accuracy. Therefore, DNN have been recently used to design new generation TTS systems [8], such as MERLIN [9] and WAVENET [10].

As far as Arabic speech is concerned, Arabic TTS systems have been developed since the beginning of TTS technology. Unit selection TTS was successfully adapted to Arabic [11] as well as HTS [12]. Recently, a new set of linguistic features, that takes care of Arabic phonemes specificities, such as vowel quantity and consonant gemination, has been successfully introduced into the HTS system, to better fit it to Arabic language [13].

HTS prosody modeling is quite general. Indeed, though HTS has been successfully adapted for many languages, including Arabic [12], using almost the same set of features, the quality of the generated speech is usually less appreciated than unit-selection-generated speech. Actually, HTS suffers mainly from uniformly-distributed state/phoneme durations and over-smoothed F0 contours. To cope with such issues, it would be interesting to look for more accurate prosody models, possibly more language specific. Note that MERLIN [9], which is based on DNN modeling, relies also on generic approaches, whatever the language is.

Modern Standard Arabic (MSA), which is widely used among all Arab-speaking countries as the official and literary language, has 28 consonants and three vowels, /a/, /u/ and /i/ [14]. Most consonants could be geminated (doubled) which is indicated in writing through adding the specific diacritic sign (shadda) whereas each vowel has a short and a long version. Vowel quantity affects the meaning of the word, e.g.

10.21437/SpeechProsody.2018-121

ذهب/dahaba/ (he went) and ذهبا/dahaba:/ (they went (when concerning two people)) [15]. Consonant gemination is another important phenomenon in MSA, which may also modify the meaning, e.g. /darasa/ (he studied) and /darrasa/ (he taught). Both phenomena should be considered while modeling prosody, and particularly duration. Recently both phenomena were successfully taken into account in Arabic TTS using HTS, which was proved by subjective listening tests [13].

Therefore, this paper investigates the modeling of phoneme duration for Arabic language. Several modeling approaches based on DNN, and including recent advances such as LSTM and BLSTM architectures [16], are studied and analyzed on four important phoneme classes: short vowels, long vowels, simple consonants, and geminated consonants. Then a class-specific modeling is achieved by using for each class the approach that is performing the best in a given development set. This class-specific approach is compared to the duration modeling achieved by well-known TTS toolkits, such as HTS (HMM-based modeling) and MERLIN (DNN-based modeling). A fourth modeling, based on artificial neural networks (ANN), which was previously proposed for Arabic [17], is also included in the comparisons.

The rest of the paper is organized as follows. Section 2 describes duration modeling using HMM and DNN. Section 3 presents the LSTM and bidirectional-LSTM architectures developed for phoneme duration modeling for Arabic TTS. Section 4 details the experimental results with the associated objective evaluations. Finally, discussion and conclusion are presented in section 5.

2. Duration modeling

Prosodic parameters, i.e. duration, F0 and intensity are the physical manifestations of phonological phenomena of speech. Particularly, duration plays a major role (a) to define the speech rhythm and accentuation and (b) to preserve the speech meaning, as for some languages, such as Arabic, a long or a short vowel changes the meaning of the word. Therefore, an accurate duration model is needed to ensure that the synthesized speech is well perceived.

Duration modeling for speech synthesis has been the subject of many studies, such as in [18], where segment duration is determined by explicit formulas, considering some theoretical hypothesis, which assume the existence of an inherent duration for every phoneme and the existence of common compression/extension factor for all the phonemes within a syllable. Another explicit model was developed in [19] where the duration of a segment is calculated as the sum of products of some contextual features.

However, with the ability of machine learning algorithms to lead to accurate models provided the right features are used, recent phoneme duration models are now based on HMM and DNN approaches.

2.1. Duration modeling based on HMM

In HTS system, duration modeling is performed using a dedicated module, where state duration distributions are predicted separately. Then the phoneme duration is the sum of the predicted state durations. To model the set of state durations of each phoneme HMM, a multi-dimensional Gaussian distribution is used. The state duration distributions are clustered using a decision-tree-based clustering technique [20]. The state duration are determined in the synthesis stage by the state duration of the relevant HMM models. The decision trees are constructed by taking into account linguistic features, that include phoneme identity and class (vowel/consonant), phonological features such as stress and accentuation, and positional features like the relative positions of different segment levels (phoneme, syllable, word) inside higher level segments. Finally, the yielded model maps the phone segment linguistic features to the matching duration [5].

2.2. Duration modeling based on DNN

Artificial neural networks (ANN) have been used to model prosodic parameters, particularly segment duration since many years, as in [21] for English, [22] for German, and [17] for Arabic. Nevertheless, these models were not too accurate in comparison to HMM, mainly because former ANN input dimension and number of hidden layers were limited.

Meanwhile, DNN have been gaining much interest for their ability to approximate any real continuous function. Therefore, DNN are now the state of the art for speech synthesis systems, such as MERLIN [9], and Wavenet [10].

Since segmental duration is a continuous value, DNN are used as a regression tool, which is trained to minimize the root mean squared prediction error (RMSE) [23]. Furthermore, recurrent networks architecture, like long short-term memory (LSTM) and Bidirectional-LSTM (BLSTM) are powerful models in sequential modeling. Therefore, deep LSTM and BLSTM are investigated to model the relationship between linguistic features and phoneme duration [24]. The input features could be the same as those used in HTS, since they cover most of the phonological, linguistic and contextual data with the addition of Arabic specific features regarding vowel quantity and consonant gemination. Input features need to undertake some preprocessing to reduce the data scattering, including normalization and/or saturation of some of the positional input values.

Output duration targets could be the state or the phoneme durations. To enhance the prediction quality of phoneme duration, it is recommended to normalize its distribution, using an adequate transform [25].

3. Experiment and result

3.1. Experimental environment

In order to train the duration model, a phonetically-balanced MSA corpus was used [26]. It consists of 1597 utterances corresponding to news bulletin read by a native-Arabic male speaker in a neutral style. The speech signals are sampled at 48 KHz with a16-bit precision.

The corpus is divided into three subsets, 1150 utterances for training, 287 for validation and 160 for test. The training set contains 37872 simple consonant occurrences, 23367 short vowel occurrences, 11565 long vowel occurrences, 4040 geminated consonant occurrences and 2458 pause segments. As mentioned before, input linguistic features and output duration targets have been pre-processed as required. Here we are using the same set of input features as proposed for HTS for Arabic synthesis in [13], for example the identity of the two previous phoneme and the two next phoneme of the actual phoneme, the position of actual phoneme in the syllable, word and phrase. Actually, the type of an input feature can be binary, like stressed/not-stressed, discrete like the phoneme

identity, or numeric like the phoneme position. All the input features are encoded into a 445-coefficient vector that includes all the binary and numerical features. Target duration outputs, i.e. phoneme durations have been analyzed to check their distributions, and a log-transform has been applied to normalize the values [25].

Table 1: *Description of the model architecture leading to the best accuracy on the development set, for each phoneme class, and for pauses.*

Phoneme Class	Model	Training set	Model description
Simple consonant	DNN-BLSTM	simple consonant	2 Dense layers with 512 units, activation function *tanh*, plus 2 BLSTM layers with 128 units.
Geminated consonant	DNN-BLSTM	geminated consonant	2 Dense layers with 16 units, activation function *tanh*, plus 2 BLSTM layers with 16 units.
Short vowel	DNN-BLSTM	short vowel	2 Dense layers with 512 units, activation function *tanh*, plus 2 BLSTM layers with 128 units.
Long vowel	DNN	long vowel	2 DNN layers with 512 and 256 units, activation function *tanh*.
Pauses	LSTM	all the phonemes	3 LSTM layers with 1024, 512 and 512 units.

3.2. Implemented duration models

Since our work is focused on modeling phoneme duration with DNN, several architectures have been implemented. This includes, feed-forward DNN using only dense layers, and recurrent DNNs based on LSTM and on BLSTM layers. For each model, various numbers of hidden layers, of nodes and of activation functions have been tried. The *RMSprop* optimizer was adopted in the experiments, as well as early stopping to avoid the over-fitting problem. But above all, the novelty of this work consists in determining the best model for each class of phonemes, considering two major characteristics of Arabic speech, i.e. vowel quantity and consonant gemination. Therefore, each type of neural network, i.e. DNN, LSTM and BLSTM has been trained on several subsets of data. For each class of sounds (e.g., simple consonants, consonants, …), a specific model is trained using only the segments corresponding to that class. Actually, 8 classes are considered: all phonemes including pauses between words, all phonemes only (i.e., without pauses), all consonants only, all vowels only, simple consonants, geminated consonants, short vowels, and long vowels. Hence, for each class, different models have been trained and evaluated, and the architecture leading to the most accurate prediction on the development set was selected, see Table 1. As the size of training corpus is different from one class to another one, this may explain why it is not the same model architecture that leads to the most accurate prediction of phoneme durations on the different classes of sounds. Consequently, in the following, we define the class-specific modeling as the fact of using, for predicting the duration of each sound, the model which is the most accurate for the corresponding class, as defined in Table 1.

3.3. Objective evaluation

The objective evaluation consists in comparing the performance of the class-specific DNN modeling to state-of-art models, i.e. HMM model as used in HTS, DNN model as used in MERLIN, and a former ANN model developed for Arabic [17]. 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. The ANN model from [17], contains 2 hidden layers with 26 units each, and uses *sigmoid* and *tanh* as activation functions. This model does not use exactly the same set of linguistic features as HTS. Evaluation measures are reported for each class of sounds in Table 2, and globally in Table 3.

Table 2: *Comparison of RMSE, MAE and correlation between predicted durations and reference durations on test set, for each phoneme class and for the various modeling approaches.*

Phoneme class	Duration modeling	RMSE (ms)	MAE (ms)	Corr
Simple consonant	HMM from HTS	25	18	0.76
	Class-specific DNN	**25**	**17**	**0.77**
	DNN-MERLIN	26	18	0.75
	ANN from [17]	35	25	0.50
Geminated consonant	HMM from HTS	43	31	0.43
	Class-specific DNN	**42**	**32**	**0.51**
	DNN from MERLIN	54	40	0.15
	ANN from [17]	62	50	0.42
Short vowel	HMM from HTS	22	16	0.82
	Class-specific DNN	**22**	**16**	**0.84**
	DNN from MERLIN	26	19	0.81
	ANN from [17]	26	19	0.78
Long vowel	HMM from HTS	49	34	0.68
	Class-specific DNN	**40**	**28**	**0.77**
	DNN from MERLIN	54	38	0.66
	ANN from [17]	68	52	0.07
Pauses	HMM from HTS	109	73	0.54
	Class-specific DNN	**109**	**70**	**0.54**
	DNN from MERLIN	146	110	0.60
	ANN from [17]	188	158	0.56

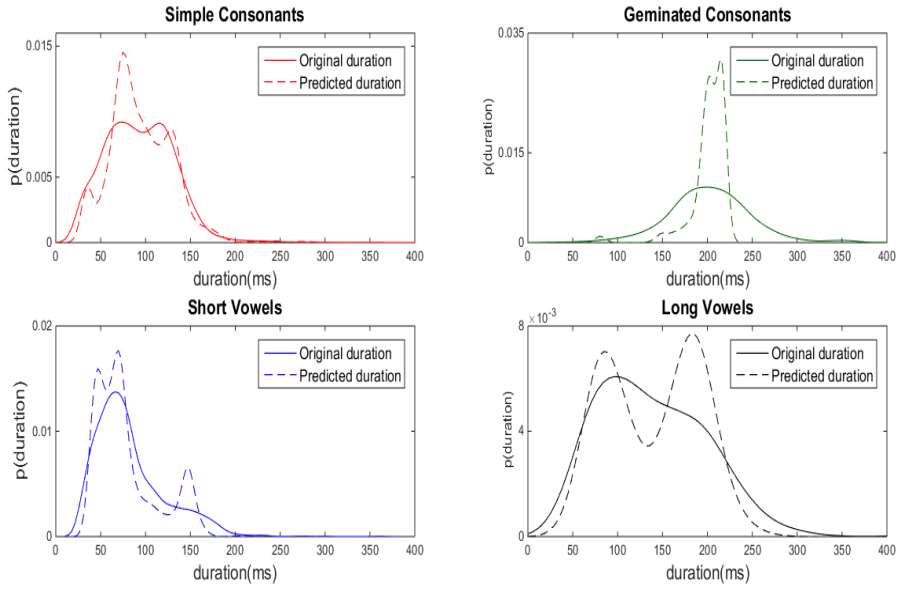

Figure 1: *Comparison of the distributions of the phoneme durations between original and predicted values.*

Table 2 shows that for each phoneme class, a class specific DNN modeling enhances the prediction accuracy of the phoneme duration, on the test set data, as measured by the various criteria: root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient between original and predicted duration. Results show that for each class of sounds, the novel class-specific DNN modeling performs better than the HMM modeling from HTS, the DNN modeling from MERLIN, and the former ANN model: the root mean square error and the mean absolute error are lower, and the correlation between predicted and reference duration values is higher.

Table 3: *Comparison of RMSE, MAE and correlation between predicted durations and reference durations on test set, for the various modeling approaches.*

Phoneme Class	Duration modeling	RMSE (ms)	MAE (ms)	Corr
All phonemes	HMM from HTS	30	20	0.83
	Class-specific DNN	**28**	**19**	**0.85**
	DNN from MERLIN	33	22	0.80
	ANN from [17]	41	28	0.66
All phonemes +pauses	HMM from HTS	40	24	0.93
	Class-specific DNN	**39**	**22**	**0.93**
	DNN from MERLIN	50	28	0.92
	ANN from [17]	63	37	0.87

When the accuracy are computed globally, i.e., on all the classes of sounds, as in Table 3, results show that using a class-specific modeling leads to a global improvement, compared to state of the art models, when all the phonemes are considered together in the evaluation, and also when considering all the phonemes and the pauses. This confirms that optimizing the duration modeling on the development set, for each class separately, allows achieving the overall best performance on the test set.

Figure 1 shows the distributions of phoneme durations for the original and predicted values on the test set, for each class of phonemes. Results show a good match for simple consonants, short vowels and long vowels. For the geminated consonants, the distribution of the predicted values is sharper, i.e. has a lower standard deviation, and slightly shifted towards higher durations. This means that the model slightly over-estimates the durations of long vowels.

4. Discussion and conclusions

This paper has investigated the modeling of the duration of the Arabic sounds for text-to-speech synthesis. Various DNN-based architectures have been developed and evaluated. Each model have been trained on various subsets of the training data corresponding to classes of sounds, as for example training on all phoneme segments, training on vowel segments only, training on short vowel segments only, etc. The various modeling architectures trained on various subsets of sounds have been compared on the development data. It appears that it is not the same model and training data, which leads to the best prediction accuracy on the various classes of phonemes (short vowels, long vowels, simple consonants and geminated consonants). This led us to define a class-specific modeling approach, which for each sound, uses the model that performs the best on the validation set. This class-specific modeling approach has been compared on the Arabic test set, to several state of the art modeling approaches, as the HMM-based modeling from the HTS toolkit and the DNN-based modeling from the MERLIN toolkit. Objective evaluations show that the proposed approach leads to a better prediction of the sound durations. The next steps will consist in integrating this class-specific duration modeling into state of the art TTS toolkits, in order to produce corresponding speech signals necessary for subjective evaluation through listening tests.

5. Acknowledgements

This research work was conducted under PHC-Utique Program, grant N°15G1405.

6. References

[1] E. Moulines, and F. Charpentier. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5–6, pp. 453–467, 1990.

[2] A. J. Hunt, and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP 1996 – 21st IEEE International Conference on Acoustics, Speech, and Signal Processing, May 7–10, Atlanta, Georgia, Proceedings*, 1996, pp. 373–376.

[3] A.W. Black, and K. Lenzo, "Optimal utterance selection for unit selection speech synthesis databases," *International Journal of speech technology*, vol. 6, no. 4, pp. 357–363, 2003.

[4] S. J. Young, "The HTK HMM toolkit: Design and philosophy," *Cambridge Univ. Eng. Dept. Tech. Rpt. CUED/F-INFENG/TR*, vol. 152, 1993.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *Proceedings of the EUROSPEECH*, vol. 5, pp. 2347–2350, 1999.

[6] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[7] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[8] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP 2013 – 38th IEEE International Conference on Acoustics, Speech and Signal Processing, May 26–31, Vancouver, Canada, Proceedings*, 2013, pp. 7962–7966.

[9] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *9th ISCA Workshop on Speech Synthesis, September 13–15, Sunnyvale, CA, USA*, 2016.

[10] A.V.D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, …, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv: 1609.03499*, 2016.

[11] R. Abdelmalek, and Z. Mnasri, "High quality Arabic text-to-speech synthesis using unit selection," in *SSD 2016 – 13th IEEE International Multi-Conference on Systems, Signals and Devices, March 21–24, Leipzig, Germany*, 2016, pp. 1–5.

[12] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving Arabic HMM based speech synthesis quality," in *INTERSPEECH 2006 – 9th Annual Conference of the International Speech Communication Association, September 17–21, Pittsburgh, Pennsylvania, USA, Proceedings*, 2006, pp. 1332–1335.

[13] A. Houidhek, V. Colotte, Z. Mnasri, D. Jouvet, and I. Zangar, "Statistical modelling of speech units in HMM-based speech synthesis for Arabic," in *LTC 2017 – 8th Language & Technology Conference, November 17–19, Poznań, Poland*, 2017, pp. 1-5.

[14] D. Newman, "The phonetics of Arabic," *Journal of the American Oriental Society*, vol. 46, no. 1984, pp. 1–6, 1986.

[15] S. Baloul, *Développement d'un système automatique de synthèse de la parole à partir du texte arabe standard voyellé*. Le Mans : Doctoral dissertation, 2003.

[16] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *ICASSP 2015 – 40th IEEE International Conference on Acoustics, Speech and Signal Processing, April 19–24, Brisbane, Australia, Proceedings*, 2015, pp. 4470–4474.

[17] Z. Mnasri, F. Boukadida and N. Ellouze, "Modeling Segmental Duration by Statistical Learning for an Arabic Text-to-Speech System," *International Review on Computers and Software,* vol. 4, no. 5, 2009.

[18] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1221, 1976.

[19] J. P. Van Santen, "Prosodic modelling in text-to-speech synthesis," in *Eurospeech 1997 – 5th European Conference on Speech Communication and Technology, September 22–25, Rhodes, Greece, Proceedings*, 1997, KN. 19–28.

[20] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration Modeling in HMM based Speech Synthesis System", *Proceedings of ICSLP*, vol. 2, pp. 29–32, 1998.

[21] W. N. Campbell, *Syllable-based segmental duration – Talking machines: Theories, models, and designs*. North Holland: Elsevier, 1992.

[22] H. Mixdorff, and O. Jockisch, "Building an integrated prosodic model of German," in *Eurospeech 2001 – 7th European Conference on Speech Communication and Technology, September 3–7, Aalborg, Denmark, Proceedings*, 2001, pp. 947–950.

[23] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust tts duration modelling using dnns," in *ICASSP 2016 – 41st IEEE International Conference on Acoustics, Speech and Signal Processing, March 20–25, Shanghai, China, Proceedings,* 2016, pp. 5130–5134.

[24] B. Chen, B. Tianling, and Y. Kai, "Discrete duration model for speech synthesis," in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association, August 20–24, Stockholm, Sweden, Proceedings*, 2017, pp. 789–793.

[25] K. M. Rosen, "Analysis of speech segmental duration with the lognormal distribution: A basis for unification and comparison," *Journal of Phonetics*, vol. 33, no. 4, pp. 411–426, 2005.

[26] N. Halabi, W. Wald, "Phonetic inventory for an Arabic speech corpus," in *LREC 2016 – 10th International Conference on Language Resources and Evaluation, Slovenia*, *May 23-28, Proceedings*, 2016, pp. 734–738.