INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Duration modeling with global phoneme-duration vectors

Jinfu Ni, Yoshinori Shiga, Hisashi Kawai

Advanced Speech Technology Laboratory, ASTREC,
National Institute of Information and Communications Technology, Japan
{jinfu.ni, yoshinori.shiga, hisashi.kawai}@nict.go.jp

Abstract

A duration model is a major component in every parametric speech synthesis system. Conventional methods use full contextual labels as features to predict phoneme durations that require morphological analysis of text. By contrast, advances in bidirectional recurrent neural networks (BRNN) and global space vector models make it possible to perform grapheme-to-phoneme (G2P) conversion from plain text. In this paper, we investigate duration prediction from plain phonemes instead of using their full contextual labels. We propose a new approach that relies on both BRNN and global space vector representations of phonemes (GPV) and durations (GDV). GPVs represent the statistics of phonemes used in a language, whereas GDVs capture duration variations beyond linguistic features. They are essentially learned from a large-scale text corpus in an unsupervised manner where phonemes are converted by G2P.

We conducted experiments on two speech corpora in Korean and Chinese to train BRNN-based models in a supervised manner. An objective evaluation conducted on a set of test sentences demonstrated that the proposed method leads to more accurate modeling of phoneme durations than the baselines.

Index Terms: prosody modeling, speech synthesis, GloVe, bi-directional recurrent neural networks, sequence-to-sequence learning

1. Introduction

In text-to-speech (TTS) synthesis systems [1][2][3], the duration model predicts one of the prosodic parameters (duration, fundamental frequency, and intensity). The prosodic parameters determine speech rhythm and accentuation. For some languages, duration also plays a role in distinguishing the meaning of speech sounds [4]. Therefore, the accurate modeling and prediction of speech-sound duration is important for ensuring that synthetic speech is well perceived.

Conventional methods such as the widely used hidden Markov model-based speech synthesis system (HTS) [5] use hidden semi-Markov models (HSMMs) with Gaussian duration distributions. Recently, the deep neural network technique has grown so fast that it has become the core in most data-driven systems, including TTS systems [3]. Neural network approaches have also been widely adopted to model duration [6][7][4]. However, most approaches [8][3][6][7][4] predict phoneme duration using the full context labels that represent phonemes in context, including linguistic features, such as stress, and positional features, such as the relative positions of different segment levels (phoneme, syllable, and word) inside higher-level segments. A front-end tool is used to extract the contextual features from text and an embedding layer to represent the linguistic features along with duration model training [7]. By contrast, prosodic features represent paralinguistic information, such as the speaker's emotional state and speaking style. Generally, duration variations are not fully determined by linguistic features. Technically, with advances in learning a continuous space vector representation of words in an unsupervised manner [10], large amounts of unannotated data can be used to improve TTS components [11][12][13]. Moreover, a bidirectional recurrent neural network (BRNN) [14] [15] and bidirectional long short-term memory are well known to be powerful for sequence modeling [7] [16] [17]. This leads us to consider a basic approach for predicting duration using phonemes only but with advanced technology for sequence modeling and unsupervised learning.

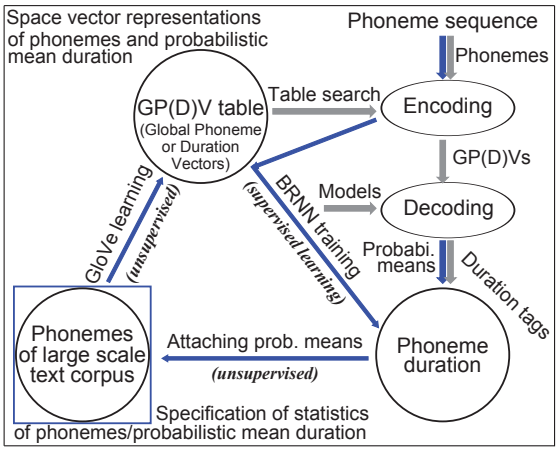

Figure 1: *Schematic diagram of BRNN-based duration models with global phoneme-duration vectors; the group of blue thin arrows indicate model building and that of thick gray arrows the duration prediction from plain phonemes.*

In this paper, we investigate a BRNN-based duration model combined with global space vector representations of phonemes (GPVs) and durations (GDVs). GPVs are used to replace the full contextual labels, whereas GDVs capture general duration variations caused by paralinguistic information and speaker factors. GPVs are learned offline by GloVe [10] from a large-scale text corpus with grapheme-to-phoneme (G2P) conversion [17] in an unsupervised manner. GDVs are learned in a similar manner but rely on a pre-trained GPV-BRNN model.

The remainder of the paper is organized as follows: In Section 2, we describe duration models with a combination of BRNNs and global phoneme-duration vectors. In Section 3, we explain the experimental setup. In Section 4, we present the experimental results and discussions. Finally, we present conclusions in Section 5.

2. Outline of the approach

Figure 1 shows a sketch of the diagram of BRNN-based duration models with GPVs, hereafter referred to as the one-

http://dx.doi.org/10.21437/Interspeech.2019-2126

level BRNN model, and GDVs, hereafter referred to as the two-level BRNN model. Figure 2 shows the relationship between the two types of BRNN models. The one-level BRNN model is language-dependent (LD) and the two-level BRNN model extends to one more BRNN level that can be language-independent. We use GloVe [10] to learn GPVs and GDVs in an unsupervised manner. Particularly, learning GDVs based on the probabilistic duration means that they are predicted using the one-level BRNN model. The two BRNNs are trained in a sequence-to-sequence supervised manner step by step given the alignment of the training data between phonemes and duration tags (as ground truth), which are quantized as an integer in milliseconds.

Model building comprises four steps that are performed in order. First, GPVs are learned from a large-scale text corpus using multilingual G2P conversion [17] and GloVe learning [10]. Second, the one-level BRNN model is trained on phonemes with GPVS as the input and corresponding duration tags as the output. Third, GDVs are learned using GloVe with sufficient probabilistic duration, which means that they are predicted using the one-level BRNN model. Forth, the second-level BRNNs are trained with GDVs as the input and the corresponding duration tags as the output.

2.1. Learning global phoneme vectors using GloVe

GloVe [10] is used to learn GPVs from a large-scale corpus of phonemes of sentences in an unsupervised manner. Using a multilingual G2P model [17], which converts plain text to phonemes with very high accuracy, we collect a phoneme corpus from a large-scale text corpus to gather various contexts of phonemes used in a target language. GloVe then uses the statistics of phoneme-phoneme co-occurrence frequencies counted from the large-scale phoneme corpus to learn GPVs based on a global log-bilinear regression model [10], specifically by minimizing the following cost function [10]:

$$J = \sum_{i,j=1}^{N} q(x_{ij})(\boldsymbol{c}_i^T \tilde{\boldsymbol{c}}_j + b_i + \tilde{b}_j - logx_{ij})^2, \quad (1)$$

where N is the number of unique phonemes, x_{ij} indicates the co-occurrence frequency of phonemes i and j, $q(x_{ij})$ is a weighting function applied to avoid frequent co-occurrence over-weighting [10], b_i and $\tilde{b}_j$ are biases, and $\boldsymbol{c}_i$ and $\tilde{\boldsymbol{c}}_j$ are space vectors to be learned, where $\boldsymbol{c}_i$ and $\tilde{\boldsymbol{c}}_i$ are equivalent in terms of representing the ith phoneme but have different initial values. As a result, $\boldsymbol{c}_i + \tilde{\boldsymbol{c}}_i$ provides the GPV of the ith phoneme. Note that, in this paper, we learn LD GPVs. As a result, there are two GPV tables, one for each language: Korean and Chinese.

2.2. Decoding GPVs to phoneme duration using BRNNs

First, BRNNs are trained in a supervised manner to learn relationships between the GPVs of phonemes and their duration tags that are represented by one-hot vectors. The standard multiclass cross-entropy is used as the objective function to train BRNNs with two hidden layers for this purpose. The following equations express the neural networks formally [16]:

$$\vec{h}_j^{(0)} = f(\vec{W}^{(0)} x_j + \vec{V}^{(0)} \vec{h}_{j-1}^{(0)} + \vec{b}^{(0)}) \quad (2)$$

$$\overleftarrow{h}_j^{(0)} = f(\overleftarrow{W}^{(0)} x_j + \overleftarrow{V}^{(0)} \overleftarrow{h}_{j+1}^{(0)} + \overleftarrow{b}^{(0)}) \quad (3)$$

$$\vec{h}_j^{(i)} = f(\vec{W}_{\rightarrow}^{(i)} \vec{h}_j^{(i-1)} + \vec{W}_{\leftarrow}^{(i)} \overleftarrow{h}_j^{(i-1)} + \vec{V}^{(i)} \vec{h}_{j-1}^{(i)} + \vec{b}^{(i)}) \quad (4)$$

$$\overleftarrow{h}_j^{(i)} = f(\overleftarrow{W}_{\rightarrow}^{(i)} \vec{h}_j^{(i-1)} + \overleftarrow{W}_{\leftarrow}^{(i)} \overleftarrow{h}_j^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{j+1}^{(i)} + \overleftarrow{b}^{(i)}) \quad (5)$$

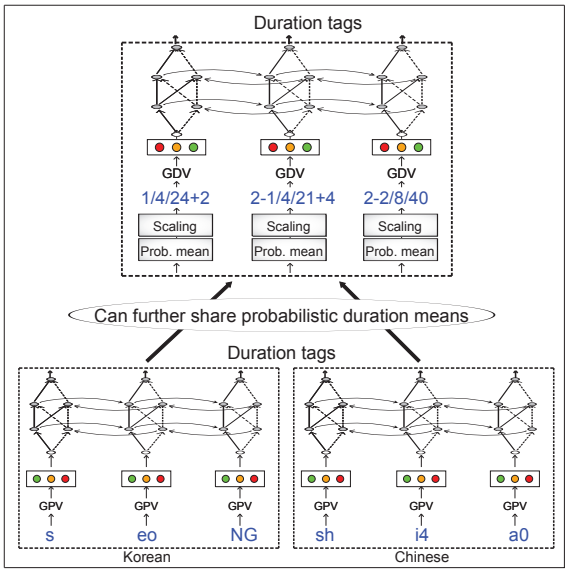

Figure 2: *Demonstration of two-level BRNN duration models, where symbol tags, such as* 2-1|4|21+4 *(discussed in Section 2.3), represent probabilistic duration means.*

$$\boldsymbol{y}_j = \text{softmax}(U_{\rightarrow} \vec{h}_j^{(2)} + U_{\leftarrow} \overleftarrow{h}_j^{(2)} + \boldsymbol{c}), \quad (6)$$

where $1 \leq i \leq 2$, and arrows $\rightarrow$ and $\leftarrow$ indicate forward and backward directions, respectively. $\boldsymbol{x}_j$ denotes the GPV of the jth phoneme, $j = 1, ..., n$ and n is the number of phonemes of a sentence; $\boldsymbol{h}$ denotes the hidden variables; $\boldsymbol{W}$, $\boldsymbol{V}$, and $\boldsymbol{U}$ denote the weight matrices; $\boldsymbol{b}$ and $\boldsymbol{c}$ denote the bias vectors; and $f(x)$ denotes the tanh function.

At the output layer, $\boldsymbol{y}_j$ provides the probabilities of the target duration for the jth phoneme. Consequently, the predicted duration of the jth phoneme $\hat{d}_j$ can be chosen using $argmax(\boldsymbol{y}_j)$, the top-ranked value, or by calculating the probabilistic duration mean:

$$\hat{d}_j = \boldsymbol{D}^T \boldsymbol{y}_j, \quad (7)$$

where $\boldsymbol{D}$ denotes the vectors of the duration tags.

2.3. Learning global duration vectors using GloVe

With the pre-trained one-level BRNN models and GPVs, we use Eq. (7) to calculate the probabilistic duration means of the large-scale corpus of phonemes. Consequently, a large-scale *duration* corpus is achieved; there is a *duration* per phoneme. Furthermore, we tag *duration* by quantizing it into four groups called $G1$, $G2$, $G3$, and $G4$ simply dividing it by 120 ms, 30 ms, 6 ms, and 50 ms, respectively. The purpose of the first three groups is to root the *duration* and the last one determines its context. Moreover, we use them with the phoneme itself for tagging *duration*. In this paper, we consider the following three tags, hereafter referred to as symbol tags:

$$T(ag)1: \quad G1|G2|G3|\text{phoneme}, \quad (8)$$

$$T2: \quad G1|G2|G3, \quad (9)$$

$$T3: \ G4_{\text{previous } duration}\text{-}G1|G2|G3\text{+}G4_{\text{next } duration}, \quad (10)$$

where symbols $|$, $-$, and $+$ represent the concatenation operation. A few examples are shown in Fig. 2. Thus, a large-scale corpus of symbol tags is obtained for learning GDVs.

GloVe [10] is further used to learn GDVs as mentioned in Section 2.1 by simply replacing phonemes by symbol tags

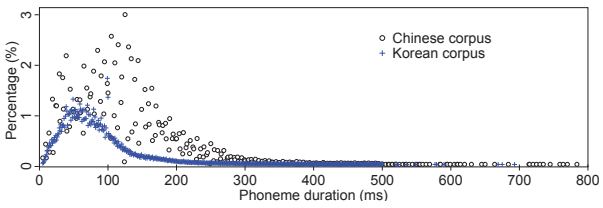

Figure 3: *Duration distribution of the two speech corpora.*

of the large-scale corpus. Because the last two symbol tags are phoneme-independent, their GDVs can be learned in a language-independent manner.

2.4. Decoding GDVs to phoneme duration using BRNNs

We further train BRNNs in a supervised manner to learn relationships between GDVs and (ground truth) duration tags that are also represented by one-hot vectors. The algorithm for decoding GDVs to duration tags is the same as that described by Eqs. (2)–(6). Duration tags are chosen by *argmax*.

3. Experimental setup

We evaluated the proposed method for Korean and Chinese, focusing particularly on the following three aspects:

- effectiveness of using GPVs for the neural duration model;
- effect of using GDVs on improving model performance;
- effect on model performance of learning GDVs on a large-scale phoneme-duration corpus.

Table 1 presents the datasets from two speakers (each in one language) used for supervised training and evaluation. The training, development, and test sets were disjoint. The ground truth duration of phonemes (a total of 9 hours of female speech) in Korean were manually labeled in terms of the Korean tone and break index (ToBI). The Chinese data (7 hours of female speech) were forced aligned using HTK [18], but with the manual correction of obvious mismatches. Figure 3 shows the distribution of phoneme duration in the datasets.

Table 1: *List of the training, development, and test sets.*

Lang.	Training (sentences)	Development	Test
Korean	3490	535	355
Chinese	5640	627	522

3.1. Learning GPVs and GDVs

We used a large-scale text corpus that consisted of approximately 0.5 million sentences in total from the two languages to learn GPVs and GDVs, with a size of 300 dimensions and a fixed 20-phoneme/tag window. Multilingual G2P [17] was used to convert text to phonemes. Punctuation in text was basically assigned as "pau." For comparison, we also learned GDVs by only using the sentences in the datasets for supervised training (4,380 sentences in Korean and 6,789 sentences in Chinese).

3.2. Network training

We trained BRNNs with the following hyperparameters [17]:

- number of units of input layers: 300 (GPV/GDV size)
- number of output units: number of duration tags (508 in the Korean corpus and 247 in the Chinese corpus)
- number of hidden layers: 2
- number of hidden units per hidden layer: 50
- stochastic gradient descent with a fixed momentum (0.9) with a small learning rate (0.0001)
- size of a mini-batch: 20 sentences
- maximum epoch: 2000.

We used the models with the best performance for the development set as the final models to be evaluated.

3.3. HTS duration model as a baseline for comparison

HTS [5] is a widely used method for building multilingual TTS systems. For comparison, we used the same datasets listed in Table 1 to build LD HTS speech systems with in-house front ends for extracting full contextual labels. To achieve a good duration model, we use pre-defined initial decision trees with "pau"-related questions to constrain decision-tree building. This is particularly effective for a speech corpus that has long utterances that involve large variations of pauses, such as the Korean corpus (average 7 seconds of speech per utterance).

3.4. List of duration model symbols

We trained the following duration models in the experiments:

- HTS: version of an HTS duration model [5]
- LD-GPV: version of a one-level BRNN model with LD GPVs
- LD-GDV(T1): version of a two-level BRNN model with LD-GPVs and LD-GDVs with T(ag)1 in Eq. (8)
- LD-GDV(T2): version of a two-level BRNN model with LD-GPVs and LD-GDVs with T(ag)2 in Eq. (9)
- LI-GDV(T2): version of a two-level BRNN model with LD-GPVs and language-independent (LI) GDVs with T(ag)2 in Eq. (9)
- LI-GDV(T3): version of a two-level BRNN model with LD-GPVs and LI-GDVs with T(ag)3 in Eq. (10).

In the last two versions, the two languages shared one two-level BRNN model.

4. Results

We evaluated the duration models using mean absolute errors (MAE) and root mean square errors (RMSE) in milliseconds between the predicted phoneme duration of test sentences and their ground truth duration. The experimental results are shown in Tables 2–4, where Table 2 excludes pauses and Table 3 includes pauses, and Fig. 4 with the MAE and standard deviation (SD) of individual phonemes by version of LD-GDV(T2). Additionally, the MAE of "pau" was 128.5 ms (SD: 131. 9 ms) in the Chinese corpus and 236.3 ms (SD: 138.3 ms) in the Korean corpus. We make several observations from the results.

- The proposed methods outperformed the HTS duration models in terms of MAE and RMSE (see Tables 2–3) except for the HTS-based duration model, which won for RMSE for counting pauses in the Korean corpus (see Table 3). The improvement in normal phonemes (excluding "pau") was significant: MAE decreased from 21.78 ms to 17.76 ms (RMSE from 30.4 ms to 28.96 ms) in the Korean corpus, and decreased from 30.21 ms to 23.18 ms (RMSE from 38.69 ms to 34.87 ms) in the Chinese corpus.

Table 2: *Results excluding pauses for a comparison of BRNN-based duration models with GPVs/GDVs and HSMM-based models for MAE and RMSE, both in milliseconds.*

Version	Korean corpus		Chinese corpus	
	MAE	**RMSE**	**MAE**	**RMSE**
HTS	21.78	30.40	30.21	38.69
LD-GPV	17.95	28.97	23.21	35.24
LD-GDV(T1)	17.97	29.02	23.37	35.41
LD-GDV(T2)	**17.76**	**28.96**	**23.18**	**34.87**
LI-GDV(T2)	18.20	29.51	29.60	43.75
LI-GDV(T3)	17.78	29.15	23.95	35.97

Table 3: *Results including pauses for a comparison of BRNN-based duration models with GPVs/GDVs and HSMM-based models for MAE and RMSE, both in milliseconds.*

Version	Korean corpus		Chinese corpus	
	MAE	**RMSE**	**MAE**	**RMSE**
HTS	25.73	**43.21**	33.71	49.47
LD-GPV	22.92	54.98	23.95	47.15
LD-GDV(T2)	22.59	54.03	**23.73**	**46.30**
LI-GDV(T3)	**22.06**	47.92	24.79	46.82

- The BRNN-based duration model with GPVs (i.e., one-level BRNN model) achieved good accuracy for predicting phoneme duration directly from the phoneme sequence instead of the full contextual linguistic features that are conventionally used.
- The use of GDVs, that is, the two-level BRNN model, improved accuracy for both MAE and RMSE in the version of LD-GDV(T2), that is, LD GDVs with tag $G1|G2|G3$ in Eq. (9). For language-independent GDVs, that is, LI-GDV(T2), contextual duration information is needed, that is, LI-GDV(T3), to achieve similar results to those achieved by LD-GDV(T2) (see Table 2).
- GDVs that combined *duration* with phonemes, that is, LD-GDV(T1), did not improve the performance of LD-GPVs (i.e., one-level model) (Table 2). This implies that only freely using GDVs can determine a deal of information that GPVs fail to determine; that is, GPVs were sufficiently powerful to represent the linguistic features.
- Comparing Table 3 (including pauses) with Table 2 (excluding pauses), the two-level BRNN model (using both GPVs and GDVs) performed better than the one-level BRNN model (using only GPVs). In Table 3, the RMSE decreased from 54.98 ms (LD-GPV) to 54.03 ms (LD-GDVs(T2)) to 47.92 ms (LI-GDV(T3)) in the Korean corpus. This indicates that GDVs worked well for phonemes with large duration variations, such as "pau," because more data were available for model training.
- Comparing LD-GDV(T2) in Table 4 (learning GPVs and GDVs only from the sentences in the speech corpora) and LD-GDV(T2) in Table 2 (learning GPVs and GDVs from a large-scale text corpus), the use of a large-scale corpus of phonemes improved the robustness of the duration model because the space vector models were learned by the statistics of the phonemes observed in various contexts.

Based on the experimental results, we conclude that the use of pre-learned GPVs instead of conventional full contextual

Table 4: *Results for learning GPVs and GDVs from the datasets for supervised training rather than a large-scale text corpus.*

Version	Korean corpus		Chinese corpus	
	MAE	**RMSE**	**MAE**	**RMSE**
LD-GDV(T2)	17.79	29.53	23.22	34.88

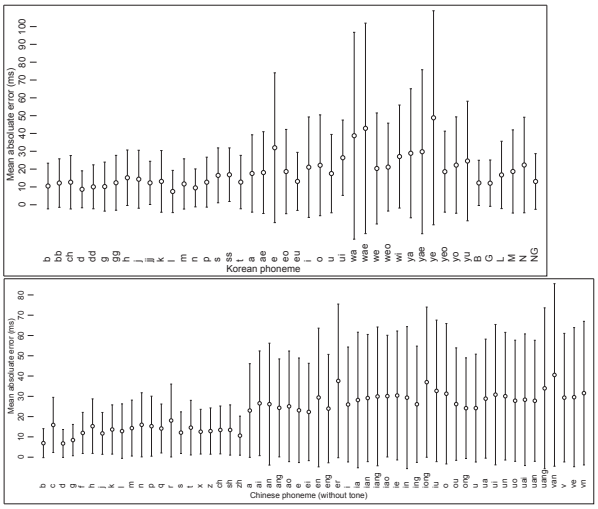

Figure 4: *MAE (open circles) and SD (bars) between the predicted and ground truth duration.*

labels performs well in terms of modeling phoneme duration. Because a front end is not required for extracting the full contextual features, the proposed method makes it relatively easy to develop multilingual speech synthesis systems [19], particularly in the case of low-resource languages. Using GDVs, it is possible to further improve the GPV-based duration model to consider factors of speakers and speaking styles that are beyond the scope of linguistic features. Moreover, GDVs can be language-independent; therefore, it is vital to easily use a large number of available speech corpus, such as huge speech data for automatic speech recognition, to improve the duration model.

There are some shortcomings in this study. We should compare the proposed method with other neural network-based methods. Additionally, no speech synthesis experiment was conducted in this study. We should investigate why some phonemes had large MAE and SD, such as /e/, /wa/, /wae/, and /ye/ in Korean, and /er/, /iong/, and /van/ in Chinese, as indicated in Fig. 4. A great deal of further work is required.

5. Conclusions

We have shown that using GloVe-based GPVs as features, instead of conventional full context labels, successfully achieves duration models using deep learning in multiple languages. We also demonstrated that the use of GDVs further improves the performance of duration models affected by factors of speakers or speaking styles that may fail to be represented by linguistic features. Both GPVs and GDVs can be learned from plain phoneme and duration corpora in an unsupervised manner. The proposed method outperformed HTS-based baselines and had the flexibility to learn GDVs across languages, thus enabling duration models to use more training samples not limited to individual speakers or languages. This study was motivated by a desire to develop multilingual TTS systems [19] relatively easily. Future work will evaluate the proposed method on common data for comparison with other approaches.

6. References

[1] H. Kawai, *et al.*, "XIMERA: A concatenative speech synthesis system with large scale corpora," *IEICE Trans. Inf. & Syst.,*, vol. J89-D, no. 12, pp. 2688–2698, 2006.

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

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, pp. 7962–7966, 2013.

[4] I. Zangar, Z. Mnsri, V. Colotte, D. Jouvet, A. Houidhek, "Duration modeling using dnn for Arabic speech synthesis", in *Proc. Speech Prosody*, Poznan, Poland, 2018.

[5] "HMM/DNN-based speech synthesis system (HTS)", http://hts.sp.nitech.ac.jp

[6] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust TTS duration modelling using DNNs," in *Proc. of ICASSP*, pp. 5130-5134, Shanghai, China, 2016.

[7] B. Chen, B. Tianling, and Y. Kai, "Discrete duration model for speech synthesis," in *Proc. of INTERSPEECH 2017*, pp. 789–793, 2017.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Duration modeling in HMM based speech synthesis system", in *Proc. of ICSLP*, vol.2 pp. 29-32, 1998.

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

[10] J. Pennington, R. Socher, and C. D. Manning, 2014, "GloVe: Global vectors for word representation," *http://nlp.stanford.edu/projects/glove/*.

[11] O. Watts, "Unsupervised learning for text-to-speech synthesis," *Ph.D. dissertation*, University of Edinburgh, 2012.

[12] H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech," in *Proc. the 8th ISCA Speech Synthesis Workshop (SSW)*, pp. 281-285, 2013.

[13] J. Ni, Y. Shiga, and H. Kawai, "Global syllable vectors for building tts front-end with deep learning," in *Proc. of INTERSPEECH*, 2017.

[14] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, 45(11), pp. 2673–2681, 1997.

[15] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in *Advances in Neural Information Processing Systems*, pp. 190–198, 2013.

[16] O. Irsory and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 720-728, 2014.

[17] J. Ni, Y. Shiga, and H. Kawai, "Multilingual grapheme to phoneme conversion with global character vectors," in *Proc. of INTERSPEECH*, pp. 2823–2827, 2018.

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

[19] Y. Shiga and H. Kawai, "Multilingual speech synthesis system," *Journal of the National Institute of Information and Communications Technology*, vol. 59, nos 3/4, 2012.