INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Discrete Duration Model For Speech Synthesis

Bo Chen, Tianling Bian, Kai Yu

Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering
Brain Science and Technology Research Center
Shanghai Jiao Tong University, Shanghai, China

bobmilk@sjtu.edu.cn, bian_tianling@sjtu.edu.cn, kai.yu@sjtu.edu.cn

Abstract

The acoustic model and the duration model are the two major components in statistical parametric speech synthesis (SPSS) systems. The neural network based acoustic model makes it possible to model phoneme duration at phone-level instead of state-level in conventional hidden Markov model (HMM) based SPSS systems. Since the duration of phonemes is countable value, the distribution of the phone-level duration is discrete given the linguistic features, which means the Gaussian hypothesis is no longer necessary. This paper provides an investigation on the performance of LSTM-RNN duration model that directly models the probability of the countable duration values given linguistic features using cross entropy as criteria. The multi-task learning is also experimented at the same time, with a comparison to the standard LSTM-RNN duration model in objective and subjective measures. The result shows that directly modeling the discrete distribution has its benefit and multi-task model achieves better performance in phone-level duration modeling.
Index Terms: speech synthesis, duration model, multi-task learning, long short-term memory

1. Introduction

Statistical parametric speech synthesis (SPSS) system consists of two major components, acoustic model and duration model [1, 2]. In the speech synthesis stage, the duration model first generates the phoneme duration in frames. The generated phoneme duration is used in the acoustic model to predict the acoustic features at each frame. Finally, the acoustic features are vocoded into natural speech. In the conventional architecture of hidden Markov model (HMM) based SPSS system [1], the distribution of acoustic features at each state in an HMM is modeled individually. Therefore, the phoneme duration is usually modeled at each state of an HMM in the acoustic model.

In neural network based acoustic model [3, 2, 4, 5], the concept of "state" from HMM is gradually deprecated, since the pronunciation period of the phoneme can be described by position features in the input of neutral network [6, 2]. Compared to state-level duration, phone-level duration has the inherent advantage that the reference duration can be more accurate than that of state-level duration. Some research groups pay attention to the phone-level duration modeling. Zen et al. first proposes a unidirectional LSTM-RNN based SPSS framework which the phoneme duration is provided by a simple phone-level LSTM-RNN duration model [2]. Ronanki et.al proposed a frame-level duration model based on transition probability to predict phone duration in [7]. It tasks the advantage of discrete duration values and provides a LSTM-RNN duration model as baseline. Decha et.al proposed a duration prediction method using multiple Gaussian process experts with phone duration as the minimal modeling unit [8].

In previous works [2, 9], the distributions of acoustic model and duration model are usually modeled under the Gaussian hypothesis. However, the duration model is slightly different from the acoustic model that the modeling target of the duration model is countable value, which means the distribution of duration is discrete and the Gaussian hypothesis may not be appropriate. Recently, the state-of-art speech generation model WaveNet [10] introduces a novel method to directly model speech waveforms instead of spectral acoustic features, in which the 8-bit speech signal is modeled discretely. It attracts the interest that whether the discrete modeling is also a better way for the duration model. This paper will conduct an investigation on the performance of discrete modeling in phone-level duration model. The multi-task learning method is also experimented in this paper, in comparison to the standard LSTM-RNN duration model. The predicted distribution given the phone sequence will be examined and the final predicted duration in SPSS architecture will be compared with the standard LSTM-RNN duration model in objective and subjective measure.

The rest of the paper will be organized as follows: Section 2 describes the LSTM-RNN duration model architecture. Section 3 introduces the discrete modeling method for countable phoneme duration. Section 4 provides the experiments with objective and subjective evaluations. Section 5 gives a conclusion.

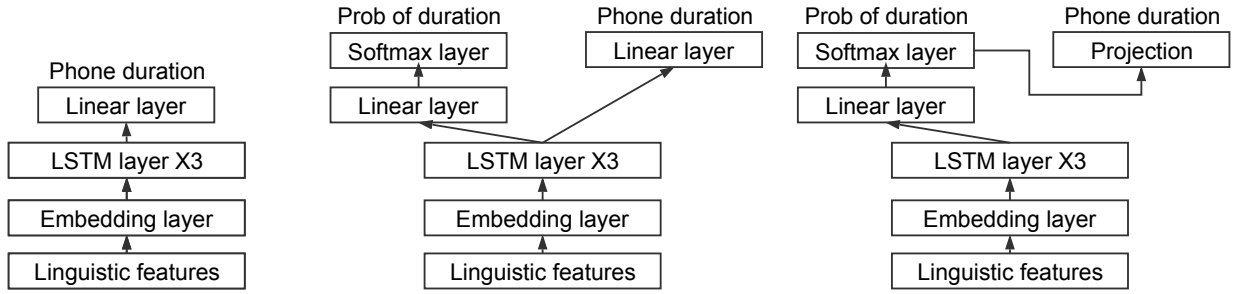
2. LSTM-RNN Duration Model

2.1. Duration Model

The duration is modeled either at state-level [9, 11] or at phone-level [2, 8] in statistical parameter speech synthesis. In conventional HMM based speech synthesis system, phoneme duration is modeled at state-level and the duration model is trained iteratively at the same time with acoustic model [1]. In neural network based duration model, the reference duration is pre-determined before training [9]. The reference duration is commonly provided by forced-alignment (FA) with a speaker-independent (SI) automatic speech recognition (ASR) system or a speaker-dependent (SD) speech synthesis system. It is believed that the majority of the reference duration is accurate, but it is inevitable that there are also phonemes with inaccurate duration. The neural network based duration models appeared in many works [2, 7, 9, 11] (including the baseline model) directly adopt mean square errors (MSE) as the optimization target to

This work was supported by the Shanghai Sailing Program No. 16YF1405300, the China NSFC projects (No. 61573241 and No. 61603252) and the Interdisciplinary Program (14JCZ03) of Shanghai Jiao Tong University in China. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

http://dx.doi.org/10.21437/Interspeech.2017-1144



(a) MSE (b) CE and MT (c) p-MT

Figure 1: *Network structures of phone-level duration models with different criteria.*

model the duration of phoneme or duration of state without an explicit probability distribution. The principle is to minimize the mean square errors between the predicted duration and the reference duration. For an utterance u_i with m_i phonemes in the corpus, the mean square error is defined as

$$\mathcal{L}_{mse} = \frac{1}{m_i}\sum_{j=1}^{m_i}(d_{i,j} - \hat{d}_{i,j})^2 \tag{1}$$

where $v_{i,1}, v_{i,2}, ..., v_{i,m_i}$ is the phoneme sequence of u_i, $d_{i,j}$ is the predicted duration of $v_{i,j}$ and $\hat{d}_{i,j}$ is the reference duration of $v_{i,j}$. In the predicting stage, the output $d_{i,j}$ is directly elected as the predicted duration.

2.2. Network structure

LSTM-RNN [12] is a powerful model in sequence modeling. It has shown its great advantages in speech synthesis [2, 5]. The LSTM-RNN duration model has already been adopted in many works [2, 7, 11]. In this paper, deep unidirectional LSTM-RNN is selected to model the relationship between linguistic features and phoneme duration. The baseline model consists of 3 concatenated LSTM-RNNs with mean square error as the criterion. The network structure is shown in Figure 1-a. The input features consist of the linguistic features, while the output features only consist of the phoneme duration.

3. Modeling discrete duration distribution

3.1. Discrete distribution

The task of duration model is different from acoustic model that the values of phoneme duration are positive integers[1], which means the distribution of duration can be described by discrete probability instead of continuous probability. Let $p(d|v_{i,1}, ..., v_{i,j})$ be the probability that the duration of $v_{i,j}$ is d frames given the phoneme sequence history $v_{i,1}, ..., v_{i,j}$. The distribution of p is modeled to minimize the cross entropy (CE) criterion that

$$\mathcal{L}_{ce} = -\log p(\hat{\boldsymbol{d}}_i|\boldsymbol{v}_i) = -\sum_{j=1}^{m_i}\log p(\hat{d}_{i,j}|v_{i,1}, ..., v_{i,j}) \tag{2}$$

Eq (1) is equivalent to Eq (2) under the Gaussian hypothesis that $p(d|v_{i,1}, ..., v_{i,j}) \sim \mathcal{N}(\mu = d_{i,j}, \hat{\sigma})$,where $\hat{\sigma}$ is the global variance shared by all Gaussian distributions. From the perspective of probability distribution, minimizing the MSE criterion is equivalent to approximate the mean of the Gaussian distribution given the linguistic features. Therefore, Eq (2) is a more general criterion than Eq (1) and is convenient for mapping the probability directly to specific duration values instead of using Gaussian function. Meanwhile, the continuity of integer duration can be ignored since the duration values are treated as discrete labels. Also, the reference of duration generated from forced-alignment is not always accurate. Minor distortions commonly exist between reference duration and real duration. It is worthy to examine the performance of joining the MSE and CE criteria together as multi-task (MT) learning.

[1]The values of phoneme duration are usually at least 5 frames for Forced-Alignment from 5-states SD-HSMMs, or at least 6 frames for FA from 3-states SI-HMMs.

3.2. Multi-task learning

By interpolating Eq (2) and Eq (1), the criterion for multi-task learning is defined as:

$$\mathcal{L}_{mt} = \lambda\mathcal{L}_{mse} + \mathcal{L}_{ce} \tag{3}$$

where p and $d_{i,j}$ are both variables of $\mathcal{L}_{mt}$. The network structure of `CE` model and `MT` model is shown in Figure 1-b. When making predictions for duration, the output of the `MT` network is the distribution $p(d|v_{i,1}, ..., v_{i,j})$, while the output of the secondary task is ignored. The final predicted duration $d_{i,j}$ is defined as the expectation of d under distribution p:

$$d_{i,j} = \sum_{d=1}^{D} d \cdot p(d|v_{i,1}, ..., v_{i,j}) \tag{4}$$

where D is the maximum duration that appears in the training set. Eq 4 shows that $d_{i,j}$ is a function of p. Therefore, Eq (4) can be substituted into Eq (3) to get the new multi-task criterion:

$$\begin{aligned}\mathcal{L}_p &= \lambda\mathcal{L}_{mse} + \mathcal{L}_{ce} \\ &= \sum_{j=1}^{m_i}(\lambda(d_{i,j} - \hat{d}_{i,j})^2 - \log p(\hat{d}_{i,j}|v_{i,1}, ..., v_{i,j})) \\ &= \sum_{j=1}^{m_i} F(p; \hat{d}_{i,j}, v_{i,1}, ..., v_{i,j}, \lambda)\end{aligned} \tag{5}$$

where F is a function with p as the only variable. The procedure of calculating expectation can be easily implemented with a constant projection layer in network description language. The criterion is called `p-MT` and the corresponding network structure is shown in Figure 1-c.

4. Experiment

4.1. Experiment setting

A Chinese male corpus XIJUNM was used for experiment. It consisted of 14677 utterances (13977 train, 500 dev, 200 test) and about 13-hours human speech.

A speaker dependent hidden semi-Markov model (HSMM) system [1] was pre-trained by HTS-2.2 [13] for state-level forced-alignment to generate the reference duration of each phoneme. The acoustic features were extracted by STRAIGHT vocoder [14] at 5ms per frame-shift including 25 mcep, 5 bap, 1 lf0, 1 v/uv [15]. Therefore 1 frame distortion in the phoneme duration corresponded to 5ms divergence in speech. The phone-level duration models consisted of 3 stacked RNN-LSTM layers as shown in Figure 1-b and Figure 1-c, in which the input features consisted of the linguistic features and the output features consisted of phoneme duration or discrete probability distribution. The linguistic features consists of 519-dim binary and numerical features including phone indicators, position in word/phase/sentence, POS, CTobi, etc. The silence phonemes at the beginning and end of an utterance were removed before experiments, but the silence phonemes between two phonemes inside an utterance were retained since the duration of in-text silence could influence the rhythm. Each LSTM-RNN layers consisted of 256 LSTM units in the networks. The networks with different criteria were trained with mini-batch stochastic gradient descent (SGD) based back propagation through time (BPTT) [16] on GTX1080 with same training configuration and same random seed using Nerv toolkit [17].

In order to evaluate the performance of experiments with different training data sizes, two subsets were randomly sampled from the training set with about 1-hour speech and 5-hours speech. For each training set, 4 types of duration models were trained for evaluation, each corresponded to one of the criteria shown in Section 2 and 3.

A frame-level LSTM-RNN acoustic model was trained from the whole training set to synthesis speech based on the duration prediction from different duration models. The input features of acoustic model consisted of the linguistic features and position features which were derived from phone-level duration, while the position features derived from state-level durations were not included. Also, an English female corpus SLT in Arctic database [18] was also used for double checking in objective evaluation. 600 utterances in SLT corpus were used as Train set, while 300 utterances were used as Dev set, 200 utterances were used as Test set. The experiment setting on SLT corpus was slightly different from XIJUNM corpus that the phoneme duration in SLT corpus was directly calculated from the timestamps provided in the database instead of FA from SD-HMMs. The LSTM-RNN layer consisted of 128 LSTM units since the training data in SLT was much smaller.

4.2. Duration distribution

Figure 2 shows the histogram of reference phoneme duration in the Train set of XIJUNM corpus. 4 dashed lines indicate that $X\%(X = 50, 90, 95, 99)$ of samples lie on the left side of the lines. Figure 3 provides the distributions of some predicted phoneme duration of model `p-MT` in Test set. 2 vowels (`ao`, `ee`), 2 consonants (`g`, `r`) and 1 in-text silence phoneme (`sil`) are selected as examples[2]. The dashed lines in Figure 3 have the same meaning as the dashed lines in Figure 2. It can

[2] An example only corresponds to 1 phoneme in 1 specific utterance.

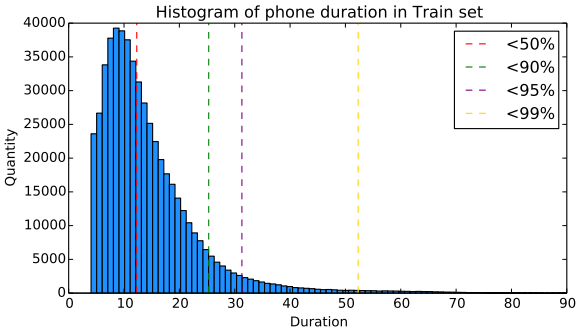


Figure 2: *Histogram of phoneme duration in Train set.*

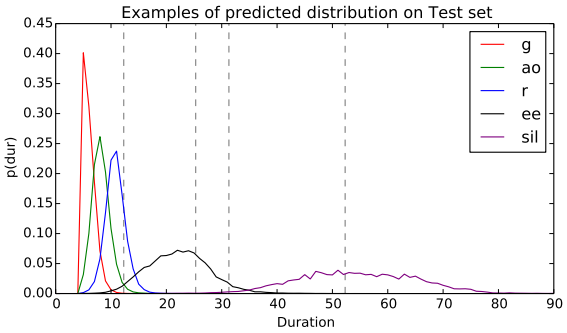


Figure 3: *Examples on the predicted discrete probability distribution of some phonemes in Test set.*

be observed in Figure 3 that the predicted distributions for each phoneme are very Gaussian-like, but the 'variance' get increase with the mean duration. As the predicted mean duration reaches the minority of the dataset, the 'variance' get much larger than the 'variance' of the phonemes with mean durations lying in the majority part of the dataset. This example shows that the discrete model has better modelling ability than simple Gaussian hypothesis with global variance.

4.3. Objective Evaluation

To evaluate the performance of different duration models, an objective measure using root mean square error (RMSE) and mean absolute error (MAE) [7] is conducted in frames per phone in relation to the reference durations. The results shown in Table 1 are evaluated on the experiments on the whole Train set. The RMSE and MAE of 3 systems with discrete probability (`CE`, `MT`, `p-MT`) are significantly better than the system with `MSE` as the only criterion. Within the 3 systems, the RMSE(90%↓) and MAE(90%↓) [9] are close to each other (the divergence is less than 0.01), which means the multi-task training has minor affect on majority part of the distributions. Meanwhile, the `p-MT` system has slightly better RMSE(100%) and MAE(100%) than `CE` and `MT`, which means the `p-MT` is more robust than the rest.

The RMSE evaluation of experiments on training set with different training data sizes in XIJUNM corpus is also shown in Table 2. Result shows that in all the comparison, the performance of the `p-MT` system is better than the `MSE` system, which indicates that the discrete modeling duration model is capable for different data sizes.

Table 3 shows the objective result of the same experiments

Table 1: *RMSE and MAE between predicted duration and reference duration on Test set of XIJUNM corpus. The result is reported for all data (100%), 90% of the data with smallest error (90%↓), 10% of the data with largest error (90%↑)*

Crit.	RMSE			MAE		
	100%	90%↓	90%↑	100%	90%↓	90% ↑
MSE	4.112	2.393	10.84	2.684	1.910	9.655
CE	4.029	2.219	10.86	2.526	1.743	9.565
MT	4.000	**2.209**	10.77	2.513	**1.734**	9.524
p-MT	**3.959**	2.219	**10.60**	**2.509**	1.741	**9.422**

Table 2: *RMSE and MAE between predicted duration and reference duration on Test set of XIJUNM corpus. The result is reported for all data (100%), 90% of the data with smallest error (90%↓), 10% of the data with largest error (90%↑)*

Data Size	Crit.	RMSE		
		100%	90%↓	90%↑
1h	MSE	4.594	2.664	12.13
	p-MT	4.485	2.550	11.94
5h	MSE	4.345	2.574	11.36
	p-MT	4.150	2.361	11.05
13h	MSE	4.112	2.393	10.84
	p-MT	3.959	2.219	10.60

on SLT corpus. The performances within the 3 discrete models are close. In terms of MAE, the discrete models significantly outperform the `MSE` model. The RMSE of discrete models are significantly better than `MSE` model on 90% data with least errors. The result is similar as in the XIJUNM corpus. The result shows that the discrete modeling works both for Chinese and English. Meanwhile, the RMSE of `CE` model is slightly worse than `MSE` model on 10% data with the largest error.

Since the difference between synthesised speech which is caused by minor duration distortion is hard to be distinguished by human listeners, we propose "correct rate" as an objective measure. The prediction of phone duration is considered correct if the distortion between predicted duration and reference duration is within a specific threshold. The correct rates of the predicted duration are listed in Table 4. Result shows that the `p-MT` model can predict more accurate duration than the `MSE` model in all the data sizes and with threshold from 1 frame to 4 frames. It can also be observed that the predicted duration of over 80% of phonemes has at most 4 frames (20ms) distortion, which means the duration of most of the phonemes in an utterance can be considered as correct.

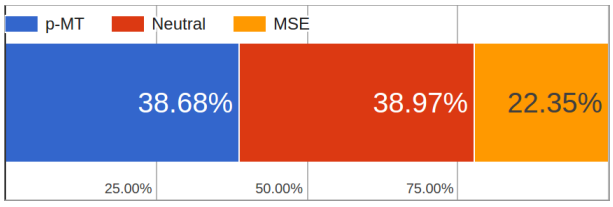


Figure 4: *Preference Score between p-MT and MSE duration model in XIJUNM corpus.*

Table 3: *RMSE and MAE between predicted duration and reference duration on Test set of SLT corpus. The result is reported for all data (100%), 90% of the data with smallest error (90%↓), 10% of the data with largest error (90%↑)*

Crit.	RMSE			MAE		
	100%	90%↓	90%↑	100%	90%↓	90% ↑
MSE	5.757	3.737	14.34	4.047	3.009	14.38
CE	5.733	**3.551**	14.66	3.899	**2.818**	13.62
MT	5.665	3.601	14.29	3.917	2.875	13.28
p-MT	**5.612**	3.576	**14.13**	**3.890**	2.859	**13.16**

Table 4: *Correct rate of predicted duration. The prediction is considered as correct if the difference between predicted duration and forced-aligned duration is less than a threshold.*

Data	Crit.	≤ 1	≤ 2	≤ 3	≤ 4
1h	MSE	35.21%	55.32%	69.68%	79.49%
	p-MT	38.90%	58.56%	72.30%	81.25%
5h	MSE	39.67%	59.18%	73.28%	82.11%
	p-MT	44.06%	64.17%	76.38%	84.25%
13h	MSE	41.39%	62.26%	75.94%	84.33%
	p-MT	46.23%	66.32%	78.48%	85.91%

4.4. Subjective Evaluation

A preference test was conducted to evaluate the performance between MSE model and p-MT model trained on 13-hours dataset. It is believed that the difference between speech which is caused by minor duration distortion is hard to be detected by listeners. A simple method is adopted to select the speech that differs most in the test set. For an utterance u, the distance of u given 2 duration models (M_1, M_2) is defined as

$$Dist(u|M_1, M_2) = \max_{1 \leq i \leq m} |d_{1,i} - d_{2,i}| \quad (6)$$

where $v_1, ..., v_m$ is the phoneme sequence in u, $d_{x,i}$ is the duration of v_i predicted from $M_x (x = 1, 2)$. In the preference test, 20 utterances were selected from the test set with the largest distance given `p-MT` and `MSE` duration model. It ensures that there is at least 1 phoneme with considerable large duration difference in each pair of the testing speech. 17 Chinese listeners were asked to give preference to the 20 test utterances in phoneme duration. Each pair of speech appeared twice in random orders to eliminate the occasionality. The result in Figure 4 shows that `p-MT` duration model significantly (p-value$<10^{-7}$) outperform the `MSE` duration model in the synthesis speech with LSTM-RNN acoustic model in the subjective measure.

5. Conclusion

This paper conducts an investigation on the discrete modeling in phone duration in SPSS architecture. The predicted distribution shows that discrete modeling can predict more robust distribution than MSE criteria under the Gaussian hypothesis. The performance of multi-task learning is also examined and is compared with LSTM-RNN duration model with MSE as the only criteria. Objective measures indicate that the models with CE criteria outperform the MSE model. And p-MT model has slightly better performance among the models with CE criteria. The preference test shows that the discrete duration modeling significantly outperformed the simple MSE criteria in LSTM-RNN SPSS architecture.

6. References

[1] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-markov model based speech synthesis." in *Interspeech*, 2004.

[2] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[3] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[4] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4460–4464.

[5] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks." in *Interspeech*, 2014, pp. 1964–1968.

[6] O. Watts, G. E. Henter, T. Merritt, Z. Wu, and S. King, "From hmms to dnns: where do the improvements come from?" in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5505–5509.

[7] S. Ronanki, O. Watts, S. King, and G. E. Henter, "Median-based generation of synthetic speech durations using a non-parametric approach," *arXiv preprint arXiv:1608.06134*, 2016.

[8] D. Moungsri, T. Koriyama, and K. Takao, "Duration prediction using multiple gaussian process experts for gpr-based speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5495–5499.

[9] G. E. Henter, S. Ronanki, O. Watts, M. Wester, Z. Wu, and S. King, "Robust tts duration modelling using dnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5130–5134.

[10] S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu *et al.*, "Wavenet: A generative model for raw audio," 2016.

[11] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks." in *Interspeech*, 2014, pp. 2268–2272.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0." in *SSW*. Citeseer, 2007, pp. 294–299.

[14] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[15] K. Yu and S. Young, "Continuous f0 modeling for hmm based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1071–1079, 2011.

[16] R. J. Williams and J. Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computation*, vol. 2, no. 4, pp. 490–501, 1990.

[17] "Nerv toolkit," https://speechlab.sjtu.edu.cn/gitlab/nerv-dev/nerv.

[18] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.