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

Comparison of Modeling Target in LSTM-RNN Duration Model

Bo Chen, Jiahao Lai, 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, ljhao1993@sjtu.edu.cn, kai.yu@sjtu.edu.cn

Abstract

Speech duration is an important component in statistical parameter speech synthesis(SPSS). In LSTM-RNN based SPSS system, the speech duration affects the quality of synthesized speech in two aspects, the prosody of speech and the position features in acoustic model. This paper investigated the effects of duration in LSTM-RNN based SPSS system. The performance of the acoustic models with position features at different levels are compared. Also, duration models with different network architectures are presented. A method to utilize the priori knowledge that the sum of state duration of a phoneme should be equal to the phone duration is proposed and proved to have better performance in both state duration and phone duration modeling. The result shows that acoustic model with state-level position features has better performance in acoustic modeling (especially in voice/unvoice classification), which means state-level duration model still has its advantage and the duration models with the priori knowledge can result in better speech quality.

Index Terms: duration model, multi-task learning, statistical parameter speech synthesis, long short-term memory

1. Introduction

The neural network based statistical parameter speech synthesis system has outperformed the decision tree clustered hidden Markov model (HMM) system [1]. However, the concept of ”state”, which comes from HMM, still remains in some of the deep neural network based SPSS systems [2, 3, 4, 5, 6]. The typical neural network based SPSS system consists of a frame-level acoustic model and a duration model [7, 8, 3]. For the acoustic models with input components derived from ”state”, the duration is necessarily to be modeled at state-level (but the object measure is the distortion of the predicted phoneme duration [7]). However, ”state” is no longer necessary in LSTM-RNN based acoustic model since the sequence characteristic (which is modeled by HMM before) can be internally modeled in LSTM-RNN [8]. Thus, it seems better to model duration at phone-level since it has the advantage that the phone-level duration can be obtained from forced-alignment (FA) with automatic speech recognition(ASR) model or even human labelling, which leads to the more accurate reference duration than duration from FA of a speaker dependent text-to-speech model.

Works from different groups have modeled duration at different levels. Zen et.al provide a LSTM-RNN based SPSS framework with a simple phone-level LSTM-RNN duration model [8]. Logarithmic phone duration is modeled by bi-directional LSTM-RNN within the prosody contours in [9]. Gustav et.al introduce a robust method to deal with the inaccuracy of forced-alignment in state-level duration models [7]. A frame-level duration model based on transition probability to predict phone duration is introduced in [10]. A joint training method to model state-level duration with state index as a secondary task in acoustic model is introduced in [11]. This paper gives an investigation on the effectiveness of different methods to model duration in LSTM-RNN SPSS architecture. The acoustic models with position features derived from duration at different levels are compared to indicate that the state-level duration model still has its advantage. The duration models in sequences of different levels are also investigated with different training criteria. A state sequence based state-level duration model is proposed, with a tricky method to explicitly utilize the priori knowledge into the network training that the sum of state duration should be equal to the phone duration.

The rest of the paper will be organized as follows: Section 2 describes the effects of position features in acoustic model. Section 3 introduces the state-level duration models in phone sequence and state sequence, with the method to utilize the priori knowledge. The experiment setup and results are provided in Section 4. Section 5 gives the conclusion of this paper.

2. Position Feature in Acoustic Models

In the neural network based statistical parameter speech synthesis system, duration affects the overall quality of generated speech in two different aspects: the prosody of the speech and the position features in acoustic models.

In a typical neural network based acoustic model, the input features consist of linguistic features and position features [8, 3]. The position features are derived from duration including phone-level duration and state-level duration [3]. In the LSTM-RNN SPSS system with only phone-level duration models, state-level position features cannot be derived to form the input features of the acoustic model. Therefore, the LSTM acoustic models with position features at different levels are first investigated to examine whether state-level duration model still has its advantage. To avoid the possible distortions caused by the weak position features, duration-derived rich position features are included in the input features of the acoustic models. The state-level position features consist of

- forward and backward absolute position in the state
- forward and backward relative position in the state
- state duration, and the fraction in the phoneme duration
- 1-of-5 state index

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.2017-1152

The phone-level position features consist of

- forward and backward absolute position in the phoneme
- forward and backward relative position in the phoneme
- phoneme duration and logarithm of phoneme duration
- 1-of-3 hard position index in phone (begin, middle, end)

3. Neural network based Duration Model

3.1. Phone Sequence Model

In LSTM-RNN, the duration of states within an utterance is modeled as a sequence. In the duraiton model of phone sequences, the basic unit is a phoneme in an utterance. The input features consist of the linguistic features of the phoneme, while the output labels are the state durations of the phoneme. The cost function is defined as follows:

$$\mathcal{L}_S = \sum_{i=1}^{N}\sum_{s=1}^{5}(dur(p_i, s) - \hat{dur}(p_i, s))^2 \quad (1)$$

where N is the number of phonemes, $dur(p_i, s)$ is the predicted duration of state s in phone p_i and $\hat{dur}(p_i, s)$ is the reference duration.

3.2. State Sequence Model

In the duration model of state sequences, the basic unit is a state of a phoneme. The input features consist of the linguistic features of the phoneme and a 1-of-5 state index, while the output label is the state duration. In the network, all the bottom layers are shared by different states, but each of the 5 states has its own final projection layer. Hence, the states are still modeled in different streams as in the phone-sequence model.

3.3. Explicit constraint

The phone duration is used as a secondary task of state-level duration model in [7], which achieve a slight improvement in objective measure in DNN duration model scene. The cost function of phone duration is defined as follows:

$$\mathcal{L}_P = \sum_{i=1}^{N}(dur(p_i) - \hat{dur}(p_i))^2 \quad (2)$$

where $dur(p_i)$ is the predicted duration of phone p_i and $\hat{dur}(p_i)$ is the corresponding label. Instead of a secondary task, we have a priori knowledge that the sum of the state durations within a phoneme should be equal to the phoneme duration. Therefore, a explicit constraint can be utilized in the network:

$$\mathcal{L}_C = \sum_{i=1}^{N}((\sum_{s=1}^{5} dur(p_i, s)) - \hat{dur}(p_i))^2 \quad (3)$$

The constraint is adopted to the network simply by adding some layers. The state duration is first denormalized by a constant projection layer after the output layer. In the phone sequence model, as shown in Fig.1(c), a constant linear projection layer with weight $w_{sum} = (1, 1, 1, 1, 1)^T$ is concatenated after the denormalizing layer. The output is supposed to be the predicted phone duration (ignoring the distortion caused by the float value of network output). A normalizing layer is appended after that. Thus, the constraint can be represented by a minimal square error (MSE) layer between the output and the labeled phone duration. In state sequence model, the method to apply the constraint is a bit tricky. As shown in Fig.1(d). the predicted state duration at each timestamp is delayed for 0, 1, 2, 3, 4 timestamps after denormalization. At timestamp t, the predicted state durations arrived from the $t-4$, $t-3$, $t-2$, $t-1$, t timestamps are summed up by w_{sum}. The rest procedure follows the same way as that in the phone sequence. After that, the output is gated by the state index that unless the current state is the 5th state of a phone (i.e. the summation of 5 state durations is within a phone), the cost is ignored.

4. Experiment

The experiments were conducted on a Chinese female corpus ANONYF and a Chinese male corpus XIJUNM. ANONYF corpus consists of 5400 utterances (5100 for Train, 150 for Dev, 150 for Test) with 5-hours nature speech. XIJUNM corpus consists of 14677 utterances(13977 Train, 500 Dev, 200 Test) with 13-hours speech. Both objective measures and preference test were examined on ANONYF corpus, while XIJUNM corpus were only used for double checking the performance on large dataset in objective evaluations.

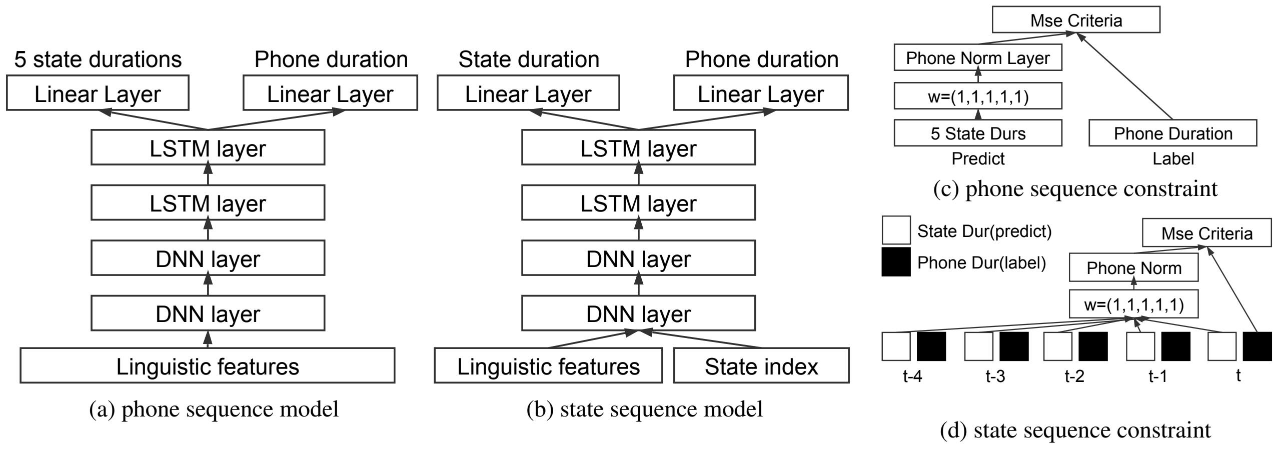
Full context labels were automatically generated from transcriptions with human labeled C-Tobi. A hidden semi-markov model(HSMM) synthesis system [12] was trained for state-level forced-alignment before neural network training. The STRAIGHT vocoder [13] was employed to extracted acoustic features at 5ms shift per frame including 25 mcep, 5 bap, 1 lf0, 1 v/uv. The lf0 was interpolated as continuous values [14] using SPLINE interpolation. The linguistic features consisted of 519-dim contextual features including phone indicators, position in word/phase/sentence, POS, C-Tobi, etc. The input features of acoustic model consisted of linguistic features and rich position features at different levels. All the networks were trained with truncated back propagation through time (BPTT) using Nerv toolkit [15] with mini-batch stochastic gradient decent(SGD) optimizer. The acoustic models were unrolled for 20 frames in truncated BPTT, while the duration models were unrolled for the whole sequence. Early stop was employed to avoid over-fitting. The input and output features were normalized to zero mean and unity variance for each dimension before training[1]. The silence phones at the beginning and the end of the utterances were removed in duration model training [2], but remained in acoustic model training. In synthesis stage, the silence phonemes of average duration was added to the beginning and the end of the input sequence.

4.1. Acoustic Model

4 acoustic models were trained in the experiments with different position features and acoustic features. The network architecture had 2 feed-forward layers concatenated with 2 LSTM-RNN layers and a final projection layer to output acoustic features. The number of nodes and cells for each hidden layer was set to 512. The objective results are reported in Table.1 for ANONYF and Table.2 for XIJUNM including mel-cepstral distortion (MCD), F0 root mean square error (F0 RMSE) and voice/unvoice classification error rate(V/U ER). ”Phone”/”State” indicates that the position features were derived from phone-level/state-level duration respectively. Acous-

[1] In the state sequence duration model, the state duration were normalized individually for each state, and each state had its own projection layer. Hence, the state durations were still modeled in different streams.

[2] It is observed that removing two-side silence phonemes can achieve extreme better objective measure in duration modeling.



(a) phone sequence model

(b) state sequence model

(c) phone sequence constraint

(d) state sequence constraint

Figure 1: *Architectures of duration models at different sequence levels, and the structure to import the explicit constraint into the network. The denormalizing layers after the output layer of state durations are omitted in the figure.*

Table 1: *Objective comparison between the LSTM acoustic models with different position features and output features (ANONYF corpus).*

Position Feature	Output Feature	MCD (dB)	F0 RMSE (Hz)	V/U ER (%)
Phone	Dynamic	5.505	21.70	9.20
State		**5.058**	**20.51**	**6.13**
Phone	Static	5.109	20.41	9.34
State		**4.718**	**19.76**	**6.19**

Table 2: *Objective comparison between the LSTM acoustic models with different position features and output features (XIJUNM corpus).*

Position Feature	Output Feature	MCD (dB)	F0 RMSE (Hz)	V/U ER (%)
Phone	Dynamic	5.973	12.00	4.33
State		**5.515**	**11.69**	**3.23**
Phone	Static	5.694	12.02	4.36
State		**5.072**	**11.43**	**3.33**

tic models for static/dynamic acoustic features are both trained for evaluation. Maximum likelihood parameter generation (MLPG) algorithm with global variance (GV) [16] was conducted to generated the static vocoder parameters for the acoustic models trained with dynamic features. The result is consistent in the experiments that the acoustic models with state-level position features as input have significantly better objective measure than acoustic model without state-level position features, especially that the V/U ER has a large gap for over 3% on ANONYF dataset. However, the gap between V/U ER in XIJUNM corpus is not as high as in ANONYF corpus, which means the divergence between acoustic models with different position features might be eliminated on large corpus.

4.2. Duration Model

Several duration models were trained for objective evaluation. The basic units for modeling were all state durations (in both state sequences and phone sequences). All the phoneme durations were caculated as the sum of state durations in the phoneme. An HSMM system was trained with the union of Train and Dev set as bottom line. DNN duration models [3] and LSTM duration models were both objectively evaluated. The number of nodes and cells in the hidden layers were set to 256. In Table.3 and Table.4, the notations ”S” indicates that state duration was modeled per timestamp, ”5S” indicates that 5-state durations were modeled per timestamp, ”P” indicates that phone duration was treated as a secondary task per timestamp in the model, ”5P” indicates that the weight of secondary task ”P” is set to 5, so that ”5P” in phone sequence model has exact same criteria as the ”P” in state sequence model, ”C” indicates that the explicit constraint was included in the model per phoneme. The objective performance are measured by root mean square error of state duration, phone duration and logarithm of phone duration [9]. Table.3 reports the objective measure of phone sequence modeling on ANONYF corpus. Result shows that both ”P” and ”C” can help improve the performance, and the weights of different cost functions have significant effects, which requires further investigation. Table.5 reports the objective measure of phone sequence modeling on XIJUNM corpus for double checking. The objective improvement on XIJUNM corpus is similar as observed on ANONYF corpus. Meanwhile, the RMSE of phone duration and RMSE of logarithm duration are similar among `"5S+P"`, `"5S+C"` and `"5S+P+C"` on XIJUNM corpus, which means the explicit constraint may not provide extra information since the constraint might be inherently learnt by the network. Also, by comparing `"5S+P"` model and `"5S+C"` model, we can find that the explicit constraint can help reduce the distortion on state-level duration caused by simple add the phone duration as the secondary task.

Table.4 reports the objective measure of state sequence model with different secondary tasks, and the comparison with phone sequence model in ANONYF corpus. The ”P” and ”C” also have improvement for the state sequence model, and can even decrease the distortion of state duration, but the duration RMSE is not as good as the phone sequence model. And it is observed that, `S+C` model and `S+P+C` model can converge to extremely bad values(even much worse than HSMM). We suspect that it is caused by the fact that the gradient coming from the

[3] The DNN architecture is similar to the LSTM-RNN model that the LSTM-RNN layers were replaced by DNN layers.

Table 3: *Duration RMSE of* **phone-level** *LSTM duration models and the baseline HSMM, DNN duration models on ANONYF corpus.*

Duration Model	Output Target	Duration RMSE		
		State	Phone	Log Phone
HSMM	---	3.214	6.239	0.308
DNN	5S	3.225	6.081	0.298
	5S+P	**3.207**	**5.979**	**0.291**
LSTM	5S	3.207	5.862	0.293
	5S+P	3.214	5.816	0.291
	5S+C	3.204	5.792	0.290
	5S+P+C	**3.203**	5.771	0.287
	5S+5P	3.214	5.715	0.286
	5S+5P+C	3.225	**5.662**	**0.282**

Table 4: *Duration MSE comparison between phone and state level LSTM duration models on ANONYF corpus.*

Input Sequence	Output Target	Duration MSE		
		State	Phone	Log Phone
Phone	5S+5P+C	3.225	**5.662**	**0.282**
State	S	3.200	5.923	0.298
	S+P	3.204	5.748	0.285
	S+C	3.200	5.797	0.289
	S+P+C	**3.169**	**5.713**	**0.282**

constraint directly back-propagates to 5 different timestamps, it might get better performance with other optimizers, but we simply give it a lower weight in this experiment.

4.3. Listening Test

The first preference test was conducted to confirm that the acoustic model with both state-level position features achieved better performance than the acoustic model without state-level position features. 30 sentences from the Test set of ANONYF corpus were used for listening test (with human labeled C-Tobi). Frame-level input sequences were generated from S+P+C duration model. A preference test is first examined between acoustic models with phone-level position features and state-level position features, to show that acoustic model with state-level position features can achieve better performance which means state-level duration modeling still has it advantage. Each sentence was presented to at least 18 listeners at random orders. The listeners were asked to give preference on the naturalness of the speech generated from the acoustic models after postfilering.

Since the objective divergence of V/U ER is very large between the two acoustic model, an extra preference test was conducted with the same continuous F0 sequences to eliminate the preference caused by F0 modeling. The result in Fig.2 shows that the acoustic model with state-level position features outperform the acoustic model without state-level position features (p-value$<$1e-4 and p-value$<$1e-8), which means state-level duration modeling is still valuable (especially in V/U classification). It should be noted that the phone-level durations used in acoustic model training were obtained from state-level forced-alignment with speaker dependent HSMM based speech synthesis system. The performance of acoustic model trained with more accurate phone-level duration has not been investigated.

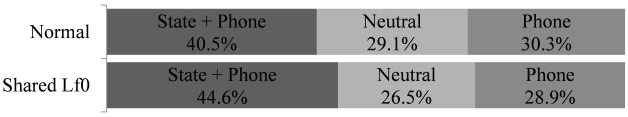


Figure 2: *Preference scores of acoustic models.*

Table 5: *Duration RMSE of* **phone-level** *LSTM duration models on XIJUNM corpus.*

Duration Model	Output Target	Duration RMSE		
		State	Phone	Log Phone
HSMM	—	2.578	4.519	0.279
DNN	5S	2.567	4.279	0.270
	5S+P	**2.551**	**4.214**	**0.264**
LSTM	5S	2.550	4.224	0.267
	5S+P	2.574	4.173	0.263
	5S+C	2.546	4.174	0.263
	5S+P+C	2.544	4.174	0.263
	5S+5P	2.540	**4.078**	0.258
	5S+5P+C	**2.537**	4.080	**0.256**

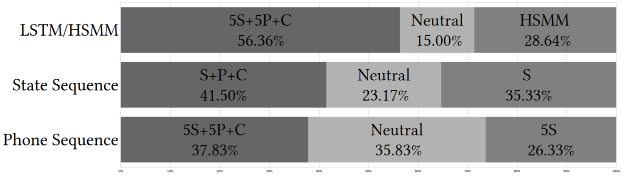


Figure 3: *Preference scores of duration models.*

Three preference tests were conducted to evaluate the performance of different duration models. The speech were generated from acoustic model trained with static acoustic features and state-level position features. 20 sentences of ANONYF corpus were objectively selected for each preference test using the method in [17]. So that each of the speech pairs has at least 1 phoneme with large duration divergence. Each pair was presented to 15 listeners twice in random order. The results are shown in Fig.3. The first test confirmed that the LSTM duration model significantly perferred over the HSMM duration model. The other two tests compares the duration models with the secondary tasks with the best objective measure. We can see that the secondary task can contribute to the voice quality, but the improvement was not so significant.

5. Conclusion

This paper investigated the effectiveness of different methods to model durations in LSTM-RNN SPSS system. Acoustic models with position features at different levels are compared in objective and subjective measure. Result shows that the acoustic model with state-level position features significantly outperform the acoustic model without state-level position features, which indicates that state-level duration model still has its advantage. Methods to model durations in different sequences are compared with multi-tasks learning, the result shows that by state-level duration model can be improved by both modeling the phoneme duration as the secondary task and adopting an explicit constraint. The weight of the secondary tasks has significant effects on objective performance and can contribute to the speech quality.

6. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum , pitch and duration in hmm-based speech synthesis," *Ieice Technical Report Speech*, vol. 99, pp. 33–38, 1999.

[2] 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 *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4460–4464.

[3] 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.

[4] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3829–3833.

[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] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for dnn-based speech synthesis," in *Proceedings interspeech*, 2015.

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

[8] 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.

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

[10] 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.

[11] O. Watts, S. Ronanki, Z. Wu, T. Raitio, and A. Suni, "The nst–glotthmm entry to the blizzard challenge 2015," in *Proc. Blizzard Challenge Workshop*, 2015.

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

[13] 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.

[14] 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.

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

[16] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[17] B. Chen, B. Tianling, and Y. Kai, "Discrete duration model for speech synthesis," in *Interspeech*, 2017.