ON THE ANALYSIS OF TRAINING DATA FOR WAVENET-BASED SPEECH SYNTHESIS

Jakub Vít, Zdeněk Hanzlíček, and Jindřich Matoušek

Department of Cybernetics, New Technology for the Information Society (NTIS),
Faculty of Applied Sciences, University of West Bohemia, Czech Republic
{jvit, zhanzlic, jmatouse}@kky.zcu.cz

ABSTRACT

In this paper, we analyze how much, how consistent and how accurate data WaveNet-based speech synthesis method needs to be able to generate speech of good quality. We do this by adding artificial noise to the description of our training data and observing how well WaveNet trains and produces speech. More specifically, we add noise to both phonetic segmentation and annotation accuracy, and we also reduce the size of training data by using a fewer number of sentences during training of a WaveNet model. We conducted MUSHRA listening tests and used objective measures to track speech quality within the conducted experiments. We show that WaveNet retains high quality even after adding a small amount of noise (up to 10%) to phonetic segmentation and annotation. A small degradation of speech quality was observed for our WaveNet configuration when only 3 hours of training data were used.

Index Terms— speech synthesis, WaveNet, deep neural network

1. INTRODUCTION

WaveNet is a neural network for generating high-quality synthetic speech. It was introduced by Oord et al. [1] in 2016 as a new generation speech synthesis algorithm outperforming both statistical parametric and concatenative methods (two most popular and widely used methods for speech synthesis).

WaveNet is a very powerful deep neural-network based architecture capable of generating speech signal directly sample-by-sample without a need of a vocoder parametrization (a necessary step in statistical parametric speech synthesis, considered as one of the factors causing a degradation of speech quality) or a speech segment concatenation (a fundamental principle of concatenative speech synthesis that is prone to introduce local artifacts to synthetic speech). Simply said, WaveNet models the conditional probability of a next sample, given previous samples and linguistic and prosodic conditions derived from to-be-synthesized information (usually in the form of a textual or phonetic representation). The generated speech is very natural and of a high quality, opening new possibilities for speech generation [1, 2, 3]. Recently, WaveNet was also used for statistical voice conversion [4]. While the WaveNet architecture is based on convolutional neural network (CNN), a similar method for speech synthesis based on recurrent neural-network (RNN) architecture was also proposed [5].

This work was supported by the Czech Science Foundation (GA CR), project No. GA16-04420S, and by the grant of the University of West Bohemia, project No. SGS-2016-039. Access to computing facilities provided by LINDAT/CLARIN, project of the Ministry of Education of the Czech Republic No. CZ.02.1.01/0.0/0.0/16_013/0001781, is greatly appreciated.

Traditional methods of speech synthesis have a known behavior and data requirements. Concatenative methods like unit selection [6] are susceptible to the way and accuracy of annotating the source speech data and its segmentation to phone-like units because errors in these processes could cause a concatenation of incompatible speech segments. As a result, artifacts can occur in synthetic speech. On the other hand, parametric speech synthesis [7, 8] is more forgiving of individual errors which are smoothed out by a statistical model. But still, to train a good parametric model, a consistent speech data is advantageous [9]. It is also known that concatenative methods require much more data. The problems mentioned above also apply to so-called hybrid approaches in which the concatenation is driven by a parametric model [10, 11, 12].

WaveNet speech synthesis is a novel approach and thus very little is known about its data requirements. In this paper, we experiment with different quality of the source speech data and see how much WaveNet is sensitive to various kinds of errors and data imperfections. We intend to reveal some insight and intuition into the network. Knowing what is important can help us to focus on right properties when recording a new voice or creating a new voice building pipeline.

Adding noise to annotation (phonetic transcription of source speech data) and segmentation (phone boundary placements) can show us how precisely the data must be prepared for WaveNet-based speech synthesis. These tasks are usually done automatically with optional human inspection or correction. There is always a possibility of errors occurring during these processes.

Recording a new voice is a very laborious task. Higher the number of recorded hours higher the cost and effort. Knowing the sufficient amount of speech data is also very valuable information. In this work, we also experiment with the amount of data WaveNet is trained on. We present several experiments in which we train the WaveNet model with various variants of noise in annotation and segmentation together with a various amount of training data. We conducted MUSHRA listening tests and used objective measures to track speech quality within the conducted experiments.

2. WAVENET ARCHITECTURE

Our WaveNet implementation is based on [1] and [2]. We used a stack of 20 dilated convolution layers with gated activation functions

$$z = tanh(W_f * x) \odot \sigma(W_g * x),$$

where σ is a sigmoid function, f and g denote filter and gate and W is a learnable convolution filter.

Each stack layer has 128 residual connections to the following layer and 128 skip connections. Skip outputs of all layers are concatenated and passed through two ReLU postprocessing layers into

a softmax layer. The model has two ReLU layers for preprocessing local conditions globally and one more in each stack layer.

Waveform samples were quantized with the μ-law algorithm into 256 discrete values. Dilatation pattern was same as in the original paper [1], i.e.,

$$1, 2, 4, ..., 512, 1, 2, 4, ..., 512.$$

We used 24kHz sample rate. The network was therefore conditioned with approximately 85 ms of previous speech. We did not use global conditioning. We trained a new model for each experiment.

The neural network was trained with Adam optimizer set to default parameters. We used TensorFlow framework and GTX1080Ti GPUs to train our models. Training one such model on one graphics card takes approximately two days.

2.1. Local conditioning

Local conditions are used to force WaveNet into generating speech corresponding to target text. The WaveNet model is conditioned with these features:

- phone identity of the current and neighboring phones (represented as one-hot vectors);
- logarithm of fundamental frequency (interpolated in unvoiced parts);
- voicing (binary value);
- sample position within the current phone (coarse coded vector, dimension was experimentally set to 100).

Similarly to [2], we found that including extra prosody features have little effect on generated speech quality.

3. OBJECTIVE MEASURES

Cross-entropy loss value used for backpropagation during the training of a neural network is not a good indicator of generated speech quality (considered as a common problem of generative models). To measure speech quality during training and to compare waveforms generated by various WaveNet configurations we used simple objective measures based on mel cepstral analysis: Compared utterances A and B are represented by sequences of mel cepstral coefficients $\{C_A[k]\}_{k=1}^{N_A}$ and $\{C_B[k]\}_{k=1}^{N_B}$. Since the original phone duration was used for speech generation in our experiments, corresponding utterances are always perfectly aligned (on the phone level) and the distance between them can be simply calculated as

$$D_1(A, B) = \frac{1}{N} \sum_{k=1}^{N} d_{\mathrm{E}} \left(C_A[k], C_B[k] \right)$$

where d_E is the Euclidean distance. In our experiments, this measure is referred to as *fixed*.

Besides, we also used another measure based on the DTW algorithm, which is convenient for general (unaligned) sequences. In our experiments, this measure was expected to improve the sub-phone alignment or to cope with possible inaccuracies in the segmentation of the testing data. This measure is simply referred to as *dtw*.

We also used these measures to ensure convergence because we sometimes experienced (especially when training on very noisy data) that a network diverged from audible speech although loss value was still converging. We used an objective measure threshold as a signal for restarting the training.

4. LISTENING TESTS

We used MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening tests to compare speech quality generated by various WaveNet models. Listening tests followed the ITU-R recommendation BS.1534-2 [13]. For each experiment, the same set of 20 sentences were synthesized. The sentences were excluded from training the WaveNet models. Original prosodic patterns (both duration and pitch contour) were imposed when generating the test sentences. 13 listeners participated in the tests. Each listener evaluated all sentences.

In each MUSHRA test, versions of every single sentence generated by the various WaveNet models that corresponded to the various experiments described further in Section 5 were compared with respect to naturalness. A natural version of each sentence (further referred to as NV) was hidden in each set and used as a reference (upper anchor). Each set also included a version generated by the baseline configuration of our WaveNet speech synthesis in which all available data were used (further referred to as BL). The listener was required to rate the versions between 0 (completely unnatural) and 100 (completely natural). Due to the presence of the reference version in each set, the listener was instructed to give one of the versions a rating of 100. Since it is unclear how to interpret a lower hidden anchor when rating synthetic speech [14, 15], no lower anchor was included in the tests.

5. EXPERIMENTS AND RESULTS

The quality of synthesized speech depends on several aspects: the synthesis method, the quality and amount of training data and the proper description of training data. Besides the word-level annotation, the data description includes primarily the phonetic annotation and segmentation, i.e., the appropriate sequence of phonetic units and the location of boundaries between them. It could also be widely extended with additional levels of description, e.g., with various phonetic, prosodic or higher-level linguistic features. Our initial experiments are focused on the impact of the fundamental level: the accuracy of the phonetic annotation and segmentation.

Disregarding the quality of the speech data, a certain description inaccuracy is related to the natural speech variability. For example, given the continuity of speech, an exact boundary location is not possible in a smooth transition between two similar phones; this is illustrated in our simple experiment described in Section 5.2. Similarly, the real pronunciation of a word can lie somewhere between two alternative pronunciations represented by slightly different phonetic units; so a "correct" pronunciation cannot be selected.

Besides this ambiguity (which corresponds rather to the nature of speech), real speech data can also contain pronunciation inaccuracies or even failures. Various errors can also arise during the annotation process, either made by a human or caused by a partly/fully automatic process of converting text to its phonetic representation (pronunciation). Additionally, phone boundaries estimated by automatic phonetic segmentation methods (typically based on hidden Markov models [16, 17]) can be appreciably misplaced.

As a result, real speech data naturally contains a combination of both annotation and segmentation errors. Since the examination of both problems together would be challenging, we performed independent experiments on the influence of the annotation and segmentation accuracy.

Besides the experiments on errors in speech data description, the amount of training data was analyzed as well. Naturally, the more training data, the better quality of synthesized speech can be expected.

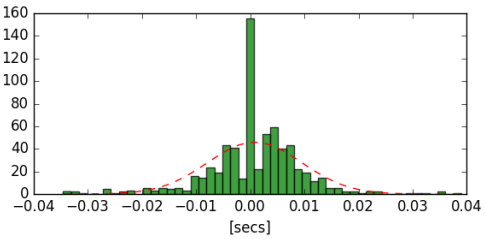

Fig. 1. Difference between the original (automatic) and average manual segmentation. As a result of the alignment to pitch marks, the difference is discrete in voiced speech segments, and histogram has secondary peaks.

However, the process of recording high-quality speech data is costly and time demanding [18], or the amount of available data can be limited.

Similarly, as the annotation and segmentation errors are closely related, the necessary amount of data probably depends on its quality, i.e., the required amount may be lower for data of high quality and consistency.

5.1. Experimental data

For our experiments, we employed a large speech corpus recorded by a professional male speaker for unit selection speech synthesis [18]. Although the corpus language is Czech, we believe our results would be valid also for other languages.

The selected training data set contained 10,000 utterances (about 14 hours of speech). Since this speech corpus has been used in many speech synthesis experiments and is still being used in a commercial unit-selection based text-to-speech system, it is considered to be a suitable experimental data, since all the revealed bugs have been fixed.

5.2. Segmentation accuracy

To evaluate the accuracy of phonetic segmentation of our experimental data, ten utterances were taken and their (automatic) phonetic segmentation was blurred so that each boundary was randomly shifted in the range of neighboring phones. Then, two speech processing experts tried to fix the segmentation. They utilized a specialized editor that displayed the waveform, spectrogram, pitch, and the segmentation to fix. The segmentation (in voiced segments) was automatically aligned to the moments of glottal closures (pitch marks) since it was also done in the original segmentation.

Results of this simple experiment are presented in Fig. 1. Mean segmentation difference ($\pm$ standard deviation) between the original automatic and average manual segmentation was 0.9 ± 9.5 ms; in the percentage of the duration of the neighboring phones it corresponds to 0.7 ± 7.2 %.

However, these values should not be understood only as a segmentation error in the data; they also include a natural variability/ambiguity of the phone boundaries is speech waveform. For comparison, the difference between boundaries placed by the two annotators was 2.1 ± 8.8 ms (1.6 ± 6.8 %).

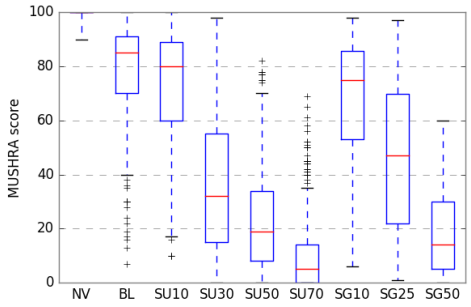

Fig. 2. Results of listening test: segmentation accuracy.

System	Objective metric		MUSHRA score	
	fixed	dtw	mean	median
NV	n/a	n/a	99.90	100
BL	0.0560	0.0422	77.40	85
SU10	0.0621	0.0458	71.70	80
SU30	0.0741	0.0477	36.41	32
SU50	0.0811	0.0472	23.50	19
SU70	0.0962	0.0534	11.37	5
SG10	0.0617	0.0433	68.24	75
SG25	0.0748	0.0468	46.17	47
SG50	0.1068	0.0553	17.43	14

Table 1. Results of experiment on segmentation accuracy.

5.3. Segmentation errors

To analyze the robustness of WaveNet to segmentation errors, artificial noise was added to the default segmentation. Two different probability distributions of noise were used: uniform and Gaussian distribution. The segmentation error in real speech corresponds rather to the Gaussian distribution. On the other hand, the uniform distribution controls the extent of the error directly.

The range of uniform distribution is given as

$$\langle t - p \cdot d_L, t + p \cdot d_R \rangle$$

where t is the default time of boundary, d_L and d_R are durations of left and right phones and p defines the relative error magnitude – see Figure 3 for a better understanding. The mean of the Gaussian distribution is set to t and the standard deviation is given in the same manner as the range of the uniform distribution.

We used notation SUx and SGx for experiments/systems with uniform and Gaussian distributions of segmentation noise, respectively, where x is the relative parameter p in percentage.

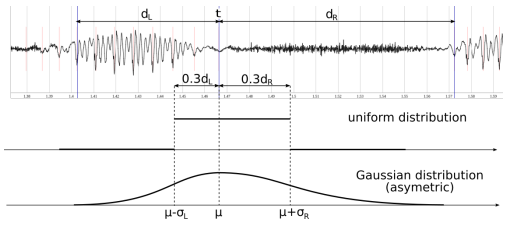

Fig. 3. Segmentation errors – distribution functions for shifting phone boundaries (an example for $p = 0.3$).

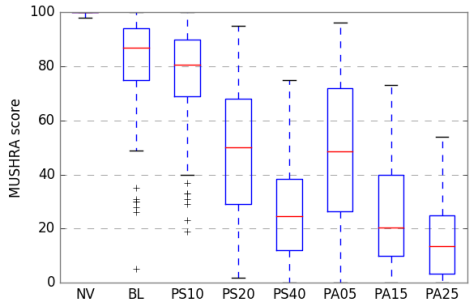

Fig. 4. Results of listening test: annotation accuracy.

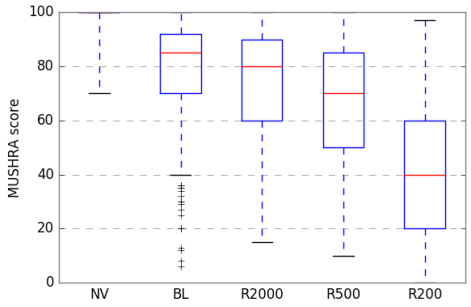

Fig. 5. Results of listening test: training data reduction.

System	Objective metric		MUSHRA score	
	fixed	dtw	mean	median
NV	n/a	n/a	99.98	100
BL	0.0560	0.0422	80.74	87
PS10	0.0573	0.0428	76.66	81
PS20	0.0641	0.0470	48.93	50
PS40	0.0680	0.0496	27.28	24
PA05	0.0643	0.0469	48.32	49
PA15	0.0690	0.0502	25.37	21
PA25	0.0751	0.0537	15.02	14

Table 2. Results of experiment on annotation accuracy.

System	Objective metric		MUSHRA score	
	fixed	dtw	mean	median
NV	n/a	n/a	99.79	100
BL	0.0560	0.0422	77.97	85
R2000	0.0566	0.0398	72.90	80
R500	0.0582	0.0428	65.35	70
R200	0.0599	0.0452	41.10	40

Table 3. Results of experiment on training data reduction.

As shown in Table 1 and Figure 2, adding 10% of noise to segmentation (either uniform or Gaussian) does not cause a considerable drop in the quality of synthetic speech. As shown in Section 5.2, a similar amount of errors (corresponding both to the variability in natural speech data and to errors caused by its automatic processing) is inherently present in the original speech data. This suggests that these errors do not influence the quality of speech generated by WaveNet. On the other hand, any higher values of noise degrade speech quality significantly.

5.4. Annotation errors

Regarding the relevance of annotation errors, two error levels can be distinguished:

- confusion of acoustically similar phones (within the same phonetic category);
- confusion of arbitrary phones (without restriction).

We used notation PSx and PAx for experiments/systems with the confused similar and arbitrary phones, respectively, where x denotes the percentage of errors.

Table 2 and Figure 4 show the results of this experiment, where the natural voice and the baseline system are referred to as NV and BL, respectively. Small errors could be present in the phonetic annotation of speech, but these should be reduced only to a small number of pronunciation ambiguities within the same phonetic categories. Any structural errors in annotation could cause a substantial quality degradation.

5.5. Data reduction

From the original speech data set with 10,000 utterances, several smaller inclusive subsets containing 2000, 500, and 200 utterances were gradually selected (a smaller subset was included in the larger); corresponding systems are denoted as R2000, R500, and R200, respectively.

Results of the experiment with annotation errors are shown in Table 3 and Figure 5. They show that at least few hours of speech (approx. 3 hours in the case of 2000 utterances) is necessary to generate speech of good quality. Some minor quality increase could be observed when using all available data (14 hours).

6. CONCLUSIONS

This paper has presented experiments which analyzed the robustness of WaveNet-based speech synthesis to training data. Various amounts and kinds of noise were added to a high-quality speech corpus to measure quality degradation of trained WaveNet models. We used objective measures and MUSHRA listening tests for comparison.

We showed that WaveNet retains high quality even after adding a small amount of noise (up to 10%) to phonetic segmentation and annotation. A small degradation of speech quality was observed for our WaveNet configuration when only 3 hours of training data were used. It should be noted that it is however likely that the results are partially dependent on the network configuration (number of layers, number of neurons, etc.).

It seems there is no need to design and record a new speech corpus specifically for WaveNet-based speech synthesis since the speech corpus intentionally built for unit selection could be utilized.

In our future work, we plan to extend the described experiments to other voices and languages as well. Similar experiments could also be carried out for unit selection and statistical parametric speech synthesis methods to get a direct comparison of how the methods are robust to the source speech data. Low-quality speech voices as those recorded by non-professional speakers [19, 20] will be examined as well.

7. REFERENCES

[1] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[2] Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Greg Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi, "Deep voice: Real-time neural text-to-speech," *CoRR*, vol. abs/1702.07825, 2017.

[3] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent WaveNet vocoder," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1118–1122.

[4] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, "Statistical voice conversion with WaveNet-based waveform generation," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1138–1142.

[5] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, "Char2Wav: End-to-End Speech Synthesis," in *International Conference on Learning Representations (ICLR)*, 2017, pp. 44–51.

[6] Andrew Hunt and Alan W. Black, "Unit selection in concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Atlanta, USA, 1996, pp. 373–376.

[7] Heiga Zen, Keiichi Tokuda, and Alan W. Black, "Statistical Parametric Speech Synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[8] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical Parametric Speech Synthesis Using Deep Neural Networks," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Vancouver, Canada, 2013, pp. 7962–7966.

[9] Rasmus Dall, Sandrine Brognaux, Korin Richmond, Cassia Valentini-botinhao, Gustav Eje Henter, Julia Hirschberg, Junichi Yamagishi, and Simon King, "Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Shanghai, China, 2016, pp. 5155–5159.

[10] Yao Qian, Frank K. Soong, and Zhi Jie Yan, "A unified trajectory tiling approach to high quality speech rendering," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 2, pp. 280–290, 2013.

[11] Vincent Wan, Yannis Agiomyrgiannakis, Hanna Silen, and Jakub Vít, "Google's next-generation real-time unit-selection synthesizer using sequence-to-sequence LSTM-based autoencoders," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1143–1147.

[12] Tim Capes, Paul Coles, Alistair Conkie, Ladan Golipour, Abie Hadjitarkhani, Qiong Hu, Nancy Huddleston, Melvyn Hunt, Jiangchuan Li, Matthias Neeracher, Kishore Prahallad, Tuomo Raitio, Ramya Rasipuram, Greg Townsend, Becci Williamson, David Winarsky, Zhizheng Wu, and Hepeng Zhang, "Siri on-device deep learning-guided unit selection text-to-speech system," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 4011–4015.

[13] "Method for the subjective assessment of intermediate quality level of coding systems," *ITU Recommendation ITU-R BS.1534-2*, 2014.

[14] Gustav Eje Henter, Thomas Merritt, Matt Shannon, Catherine Mayo, and Simon King, "Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech," in *INTERSPEECH*, Singapore, 2014, pp. 1504–1508.

[15] Thomas Merritt, Robert A J Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King, "Deep neural network-guided unit selection synthesis," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Shanghai, China, 2016, pp. 5145–5149.

[16] Doroteo Toledano, Luis Gomez, and Luis Grande, "Automatic phonetic segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 617–625, 2003.

[17] Jindřich Matoušek and Jan Romportl, "Automatic pitch-synchronous phonetic segmentation," in *INTERSPEECH*, Brisbane, Australia, 2008, pp. 1626–1629.

[18] Jindřich Matoušek, Daniel Tihelka, and Jan Romportl, "Building of a speech corpus optimised for unit selection TTS synthesis," in *Proceedings of LREC '08*, 2008.

[19] Markéta. Jůzová, Jan Romportl, and Daniel Tihelka, "Speech Corpus Preparation for Voice Banking of Laryngectomised Patients," in *Text, Speech, and Dialogue*, vol. 9302 of *Lecture Notes in Computer Science*, pp. 282–290. Springer, 2015.

[20] Markéta Jůzová, Daniel Tihelka, Jindřich Matoušek, and Zdeněk Hanzlíček, "Voice conservation and TTS system for people facing total laryngectomy," in *INTERSPEECH*, Stockholm, Sweden, 2017.