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

Jakub Vít, Zdeněk Hanzlíček and Jindřich Matoušek
University of West Bohemia, Czechia

Introduction

WaveNet is a recently introduced convolutional deep neural network for generating high-quality synthetic speech.

Novel approach → very little is known about its data requirements.

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.

Experiments

- adding noise to phonetic segmentation accuracy
- adding annotation errors
- reduce the size of training data

Wavenet architecture

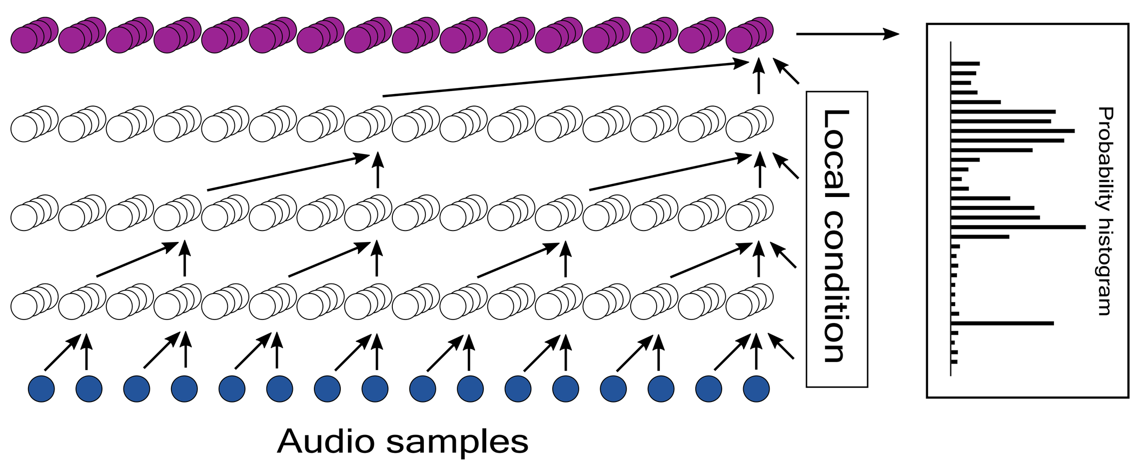
WaveNet models the conditional probability of a sample, given previous samples and linguistic and prosodic conditions derived from to-be-synthesized information.

Implementation is based on the original WaveNet paper.
(Oord et al., Wavenet: A generative model for raw audio, 2016)

Waveform samples were quantized with the μ-law algorithm into 256 discrete values.

Stack of 20 dilated convolution layers:

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



with gated activation functions:

$$z = \tanh(W_f * x + V_f * h) \odot \mathrm{sigmoid}(W_g * x + V_g * h)$$

Local conditioning:

- current and neighboring phone identity
- logarithm of fundamental frequency and voicing
- sample position within the current phone

Listening tests

We conducted MUSHRA listening tests to track speech quality within the conducted experiments. We employed a large Czech speech corpus recorded by a professional male speaker for unit selection speech synthesis.

For each experiment, **20** sentences were used. Original prosodic patterns were imposed. **13** listeners participated in the tests. Each listener evaluated all sentences.

We also used a distance between mel cepstral coefficients as an objective measure to compare speech quality.

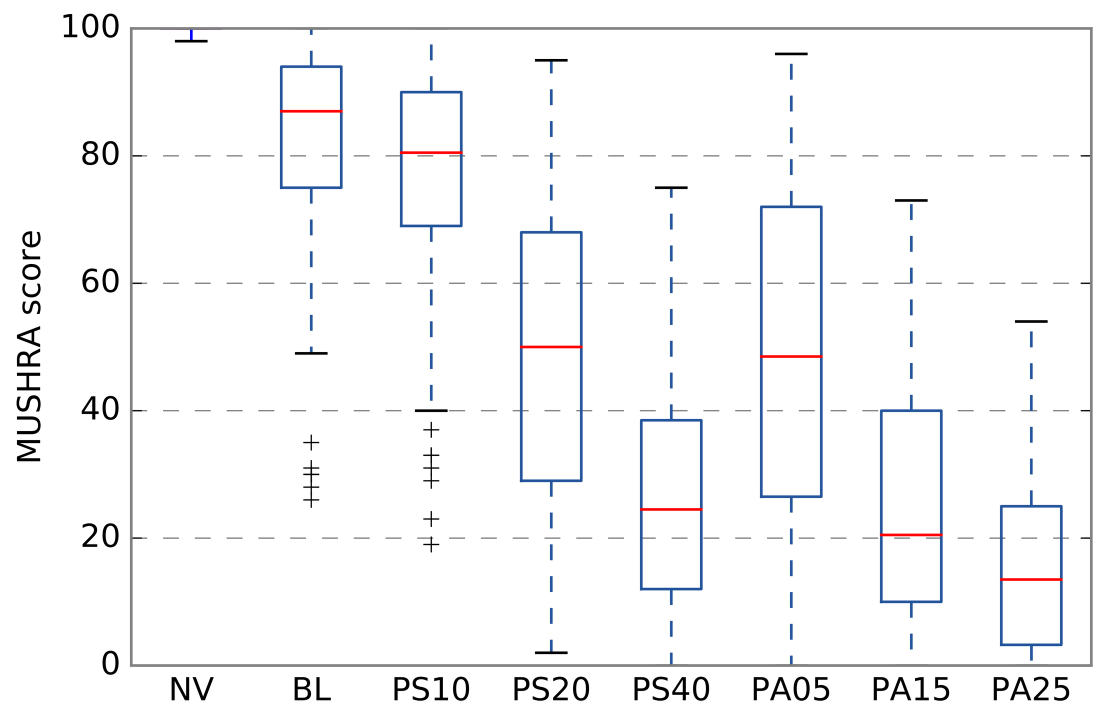
Experiment 1

Annotation errors

Two error levels can be distinguished:

- confusion of acoustically similar phones **(PS)**
- confusion of arbitrary phones **(PA)**

Results



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

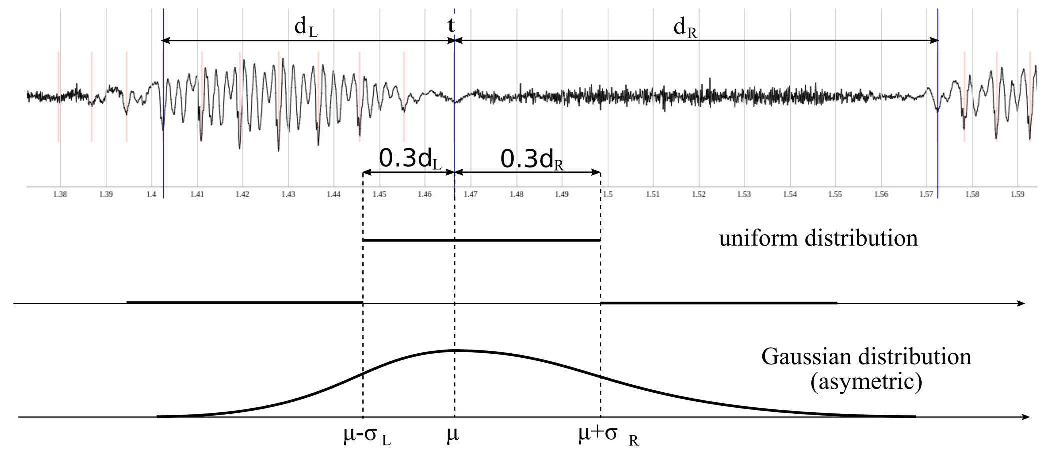
Experiment 2

Segmentation errors

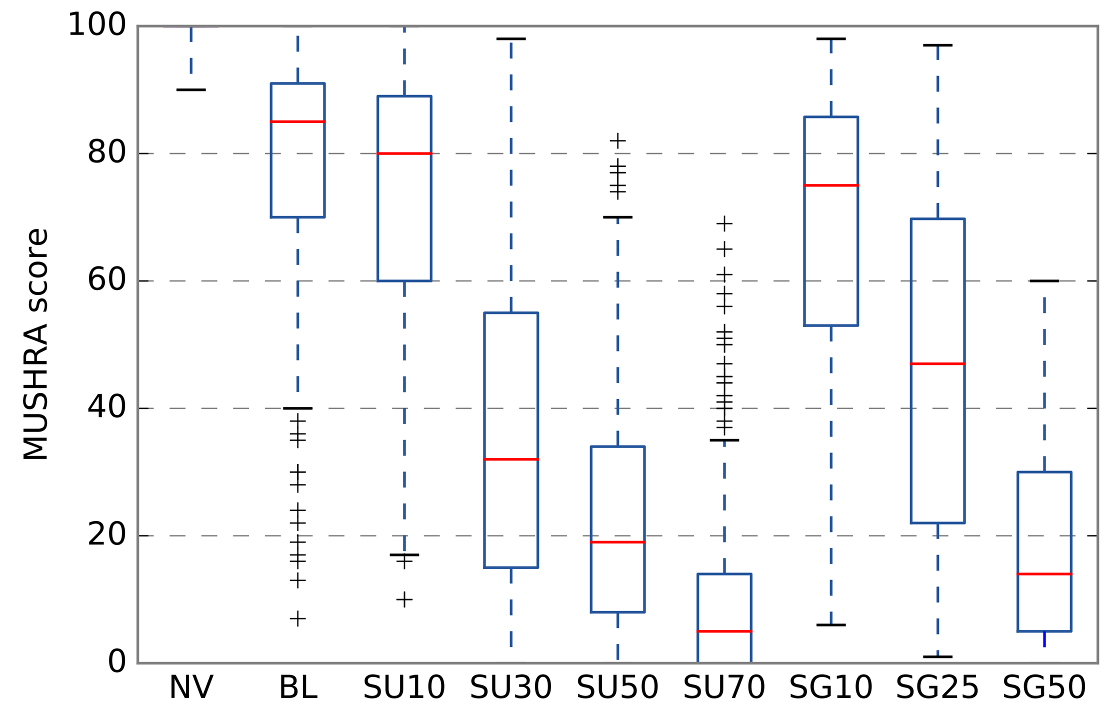
To analyze the robustness of WaveNet to segmentation errors, artificial noise was added to the default segmentation.

Two different probability distributions of noise were used:

- uniform distribution **(SU)**
- gaussian distribution **(SG)**



Results



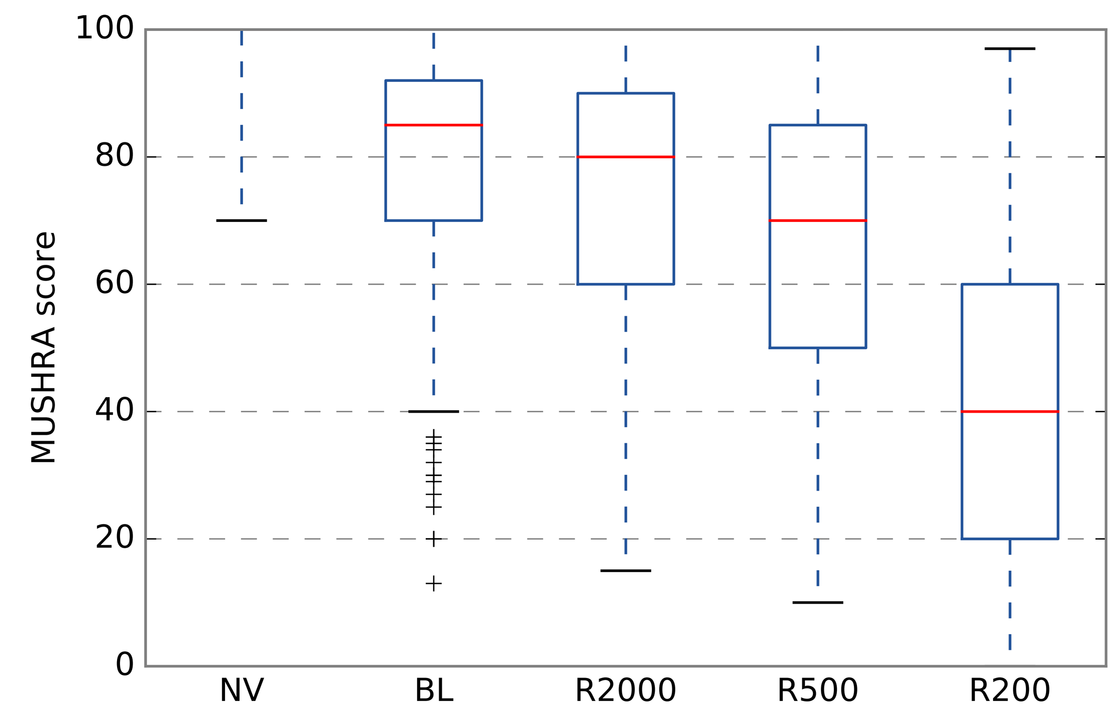
System	Objective metric	MUSHRA score	
		mean	median
NV	n/a	99.90	100
BL	0.0560	77.40	85
SU10	0.0621	71.70	80
SU30	0.0741	36.41	32
SU50	0.0811	23.50	19
SU70	0.0962	11.37	5
SG10	0.0617	68.24	75
SG25	0.0748	46.17	47
SG50	0.1068	17.43	14

Experiment 3

Data reduction

From the original speech data set with approx. 14 hours and 10,000 utterances **(BL)**, several smaller inclusive subsets containing **2000**, **500**, and **200** utterances were gradually selected.

Results



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

Conclusions

WaveNet retains high speech quality even after adding a small amount of noise **(up to 10%)** to phonetic segmentation and annotation of training data.

A small degradation of speech quality was observed for our WaveNet configuration when only **3 hours (2000 sentences)** of training data were used.

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.

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

Contact

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

Acknowledgment

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.