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

Jakub Vit, 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

Audio samples
with gated activation functions:
\[ x = \text{softmax}(W_1x + V_1y + b) \odot \text{sigmoid}(W_2x + V_2y + b) \]

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.

Proposed solution

Experiment 1
Annotation errors
Two error levels can be distinguished:
- confusion of acoustically similar phones (PS)
- confusion of arbitrary phones (PA)

Results

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

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

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.

Acknowledgement
This work was supported by the Czech Science Foundation (GA CR), project No. GA16-04425S, 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.02.1.01/0.0/00/03.006/002781, is greatly appreciated.