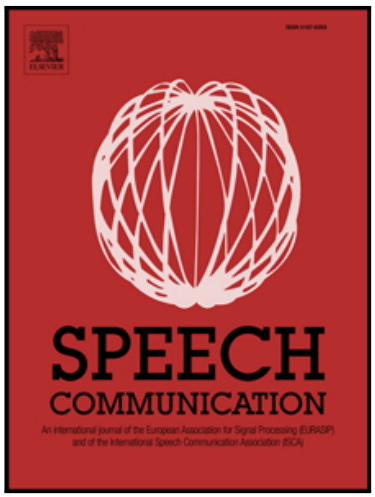

Investigating Very Deep Highway Networks for Parametric Speech Synthesis

Xin Wang, Shinji Takaki, Junichi Yamagishi

PII: S0167-6393(16)30370-3
DOI: 10.1016/j.specom.2017.11.002
Reference: SPECOM 2498

To appear in: *Speech Communication*

Received date: 18 December 2016
Revised date: 24 September 2017
Accepted date: 7 November 2017

Please cite this article as: Xin Wang, Shinji Takaki, Junichi Yamagishi, Investigating Very Deep Highway Networks for Parametric Speech Synthesis, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.11.002

Investigating Very Deep Highway Networks for Parametric Speech Synthesis

Xin Wang[a,b,*], Shinji Takaki[a], Junichi Yamagishi[a,b,c]

[a]National Institute of Informatics, 2-1-2, Hitotsubashi-cho, Chiyoda-ku, Tokyo, 101-8430, Japan.
[b]SOKENDAI, 2-1-2, Hitotsubashi, Chiyoda, Tokyo, 101-8430, Japan.
[c]The Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9LW, United Kingdom.

Abstract

Deep neural networks are powerful tools for classification and regression tasks. While a network with more than 100 hidden layers has been reported for image classification, how such a non-recurrent neural network with more than 10 hidden layers will perform for speech synthesis is as yet unknown. This work investigates the performance of deep networks on statistical parametric speech synthesis, particularly the question of whether different acoustic features can be better generated by a deeper network. To answer this question, this work examines a multi-stream highway network that separately generates spectral and F0 acoustic features based on the *highway* architecture. Experiments on the Blizzard Challenge 2011 corpus show that the accuracy of the generated spectral features consistently improves as the depth of the network increases from 2 to 40, but the F0 trajectory can be generated equally well by either a deep or a shallow network. Additional experiments on a single-stream highway and normal feedforward network, both of which generate spectral and F0 features from a single network, show that these networks must be deep enough to generate both kinds of acoustic features well. The difference in the performance of multi- and single-stream highway networks is further analyzed on the basis of the networks' activation and sensitivity to input features. In general, the highway network with more than 10 hidden layers, either multi- or single-stream, performs better on the experimental corpus than does a shallow network.

Keywords: Text-to-Speech, Statistical parametric speech synthesis, Deep neural network, Highway neural network

1. Introduction

Speech synthesis aims at creating natural-sounding speech waveforms and is used in various types of application with speech waveforms as output. A widely used application is Text-to-Speech (TTS) synthesis [1], where the speech is synthesized to read aloud the input text. Its social value is obvious in human-machine and human-human communication, e.g., when a disabled human speaker cannot articulate sounds.

TTS is difficult because of the ambiguous association between text and speech. Despite the recent trend towards end-to-end TTS [2], most existing TTS systems consist of front- and back-ends. The front-end infers the phonemic and prosodic information from the text, and then the back-end synthesizes the speech waveform given the output of the front-end. The TTS back-end, or the acoustic model, can be implemented on the basis of statistical parametric speech synthesis (SPSS), a framework that uses statistical models such as a hidden Markov model (HMM) to generate speech acoustic features and then construct the waveform [3, 4]. Recently, various acoustic models based on neural networks (NNs) have been proposed to augment the HMM-based SPSS framework [5, 6, 7]. One reason is that well designed NNs can model various aspects of speech that have been ignored by HMM, including the complex correlation among linguistic features [6], cross-time [8, 9], and cross-dimension dependency [10] of speech acoustic features. Additionally, NNs can better describe the distribution of high-dimensional acoustic features [2, 11, 12] or waveform [13, 14], which may alleviate the artefact due to vocoding.

Despite the progress of NN-based SPSS, some questions remain unaddressed. One such question is how a NN's depth affects its performance on acoustic feature modeling, particularly on the commonly used low-dimensional spectral and F0 features. We try to answer this question in the context of non-recurrent NNs. Although theories [15] and research works in image classification [16] have shown that very deep NNs can perform better than shallow networks, SPSS may need specific experiments because it is a regression task with heterogeneous target features. It would be useful to know whether both F0 and spectral features can be better modeled by a deeper NN and whether the classical feedforward NN should be used as it is.

To explore the influence of network depth, this work conducts experiments on the multi-stream highway network because it facilitates the training of very deep networks [17] and reduces the interaction between sub-networks for spectral and F0 modeling [18]. Experiments comparing the network with different depths show that spectral features can be better generated by a network with up to 40 hidden layers, while F0 can be well generated by a network with just 4 hidden layers. Then, this work conduct experiments on single-stream highway and conventional feedforward networks, where all acoustic features are generated from a single network. It was

*Corresponding author
Email addresses: wangxin@nii.ac.jp (Xin Wang), takaki@nii.ac.jp (Shinji Takaki), jyamagishi@nii.ac.jp (Junichi Yamagishi)