Deep Elman Recurrent Neural Networks for Statistical Parametric Speech Synthesis

Sivanand Achanta, Suryakanth V Gangashetty

PII: S0167-6393(16)30381-8
DOI: 10.1016/j.specom.2017.08.003
Reference: SPECOM 2480

To appear in: *Speech Communication*

Received date: 28 December 2016
Revised date: 11 May 2017
Accepted date: 2 August 2017

Please cite this article as: Sivanand Achanta, Suryakanth V Gangashetty, Deep Elman Recurrent Neural Networks for Statistical Parametric Speech Synthesis, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.08.003

Deep Elman Recurrent Neural Networks for Statistical Parametric Speech Synthesis

Sivanand Achanta and Suryakanth V Gangashetty

Speech and Vision Laboratory, International Institute of Information Technology Hyderabad, India - 500032.

Abstract

Owing to the success of deep learning techniques in automatic speech recognition, deep neural networks (DNNs) have been used as acoustic models for statistical parametric speech synthesis (SPSS). DNNs do not inherently model the temporal structure in speech and text, and hence are not well suited to be directly applied to the problem of SPSS. Recurrent neural networks (RNN) on the other hand have the capability to model time-series. RNNs with long short-term memory (LSTM) cells have been shown to outperform DNN based SPSS. However, LSTM cells and its variants like gated recurrent units (GRU), simplified LSTMs (SLSTM) have complicated structure and are computationally expensive compared to the simple recurrent architecture like Elman RNN. In this paper, we explore deep Elman RNNs for SPSS and compare their effectiveness against deep gated RNNs. Specifically, we perform experiments to show that (1) Deep Elman RNNs are better suited for acoustic modeling in SPSS when compared to DNNs and perform competitively to deep SLSTMs, GRUs and LSTMs, (2) Context representation learning using Elman RNNs improves neural network acoustic models for SPSS, and (3) Elman RNN based duration model is better than the DNN based counterpart. Experiments were performed on Blizzard Challenge 2015 dataset consisting of 3 Indian languages (Telugu, Hindi and Tamil). Through subjective and objective evaluations, we show that our proposed systems outperform the baseline systems across different speakers and languages.

Keywords: Speech synthesis, Recurrent neural networks, Deep neural networks, Hidden state

1. Introduction

Statistical parametric speech synthesis (SPSS) has emerged as a viable alternative to the unit selection paradigm for text-to-speech systems (TTS) [1]. This can be seen from the results of Blizzard Challenges conducted in recent years [2]. The main reasons for preferring SPSS systems are the flexibility offered as well as the low footprint of these systems. While SPSS can be implemented in