OVERVIEW OF SPEECH SYNTHESIS USING LSTM NEURAL NETWORKS

G. Navickas, G. Korvel, J. Bernatavičienė
Vilnius University, Institute of Data Science and Digital Technologies
Vilnius, LITHUANIA
e-mail: gediminas.navickas@mii.vu.lt

Abstract

Currently, the most popular speech recognition systems are based on unit selection – decision tree algorithms. In literature, new speech synthesis methods based on Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) are proposed. In this paper, an overview of speech synthesis and their realization called LSTM is given. Directions for further investigations are highlighted.

Keywords: data science, speech synthesis, neural network

1 Introduction

Much effort is given by scientists and engineers for human speech modelling at least for half of a century. The state-of-the-art methods applied to speech modelling are a concatenative synthesis, formant synthesis, articulatory synthesis, HMM-based synthesis, sine-wave synthesis. The literature review reveals that speech synthesis beyond the typically used techniques is extended towards exploring Deep Neural Networks (DNNs). Deep learning-based synthesizers use Artificial Neural Networks (ANNs), which are trained on recorded human speech data. One of the types of DNN is Recurrent Neural Network (RNN) architecture called Long Short Term Memory (LSTM) network [1]. These days RNN and LSTM networks are used by researchers for speech synthesis. They are used for English and several other languages and show good results. The goal of this research is the overview of these results. A comparison of the effectiveness of using LSTM networks for speech synthesis collected from the literature for different languages is given in this paper. Neural networks are not used for Lithuanian speech synthesis. Lithuanian speech synthesis systems are implemented using unit selection - decision tree algorithms that are one of the concatenative synthesis methods. In this paper, the possibilities of using LSTM networks for Lithuanian speech synthesis are discussed.

2 Recurrent Neural Networks and LSTM

Recently ANNs and especially DNNs are successfully used for solving machine speech problems [2], [3], [4]. RNN is a type of neural network designed for capturing information from sequential or time series data [2]. RNN is the repetition of simple units, which takes as an input the past, new input and produces a new prediction and connects to the future. Due to the fact that this network has short term memory, it does not work

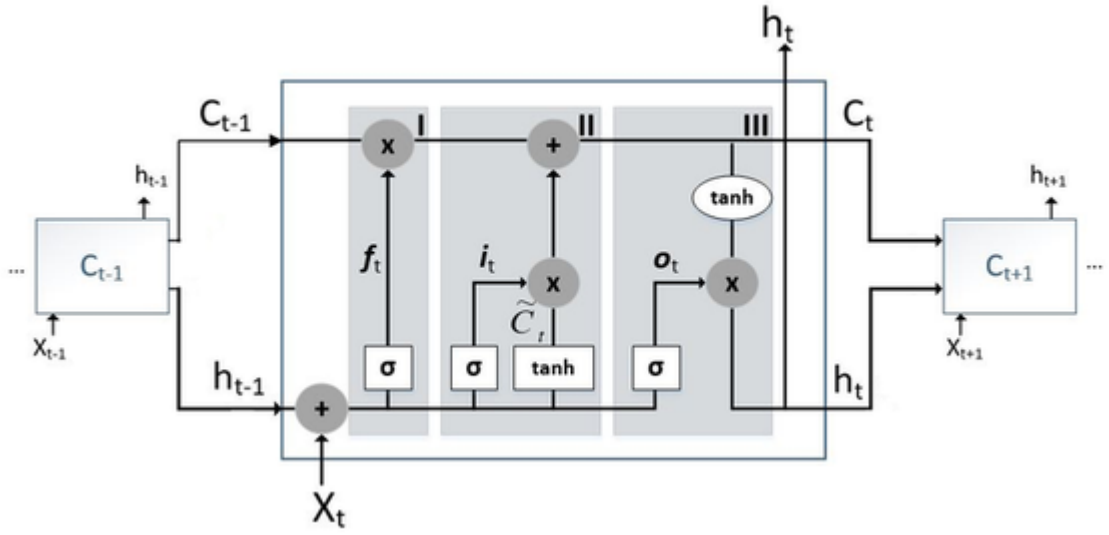

Figure 1: Concept Scheme of Vanilla LSTM network

well for longer sequences. This disadvantage is known as the Vanishing Gradient Problem [5]. One of the most common RNN architectures which solve this problem is the LSTM network. It is the type of RNN designed to work with long data sequences. It uses a mechanism called gates. Gates are used for learning which information to forget or add to the hidden state. LSTM network was presented in 1997 [1]. These days it is the most used RNN architecture. This architecture is also called Vanilla LSTM. The graphical representation of the LSTM network is given in Fig. 1.

The Vanilla LSTM network is RNN network where a repeating module contains three interacting gates. The module is called the LSTM cell. Denote state of LSTM cell of current time by C_t, input of current time by X_t, output of previous time by h_{t-1}. First gate of the LSTM cell is called forget gate f_t, and is given by: $f_t = \sigma(W_f \cdot X_t + R_f \cdot h_{t-1} + b_f)$, where W_f -input weights, R_f - recurrent weights, b_f - bias. It is a *sigmoid* layer with outputs between 0 and 1. It takes h_{t-1} and X_t and for each number in the cell state C_{t-1} returns a number where 0 means completely forget and 1 means completely keep the value.

Second gate is called input gate i_t. It decides what new information we will store in the cell state. At first, this *sigmoid* layer decides which values will be updated according to formula $i_t = \sigma(W_i \cdot X_t + R_i \cdot h_{t-1} + b_i)$, then *tanh* layer creates a vector of new candidate values $\tilde{C}_t = tanh(W_C \cdot X_t + R_C \cdot h_{t-1} + b_C)$. At this step previously calculated values are combined by multiplying old state by f_t (forgetting the values), then $i_t * \tilde{C}_t$ is added (the new candidate values scaled by how much we decided to update each state value) $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$.

Third gate is called output gate o_t. At first, *sigmoid* layer decides which parts of the cell state will go to output $o_t = \sigma(W_o \cdot X_t + R_o \cdot h_{t-1} + b_o)$. Before outputting the cell state is run through *tanh* layer and values are pushed between -1 and 1. Finally,

Table 1: Speech Synthesis using different DNN architectures

Reference	Network architecture	Data Set	Results
Salah Al-Radhi et al. [9], 2017	LSTM: 4 feed-forward hidden lower layers of 1024 hyperbolic tangent units each, followed by a single LSTM hidden top layer with 512 units.	US English female (SLT); speaker from the CMU-ARCTIC database.	Improved naturalness of the speech synthesized significantly over DNN baseline. Object of synthesis: sentences.
Wu and King [8], 2016	Simplified LSTM architecture (uses only the forget gate, has significantly fewer parameters than the vanilla LSTM). 256 units (e.g., LSTM blocks) in the recurrent layer.	A corpus from a British male speaker divided into three subsets: training, development and testing (2400, 70 and 72 utterances).	Forget gate can learn the temporal structure of speech; its activations have a high correspondence with phone boundaries. For this task, the forget gate is the only critical component of the LSTM; other components can be omitted with no reduction in naturalness. Object of synthesis: sentences.
Zen and Sak [10], 2015	Unidirectional LSTM RNNs with a recurrent output layer. The architecture of the LSTM: 1 forward-directed hidden LSTM layer with 256 memory blocks.	US English speech data from a female professional speaker. The training and development data sets consisted of 34 632 and 100 utterances, respectively.	LSTMs produced significantly better speech than DNNs. Object of synthesis: Utterances.
Zen et al. [7], 2016	LSTM-RNNs was 1 x 128-unit ReLU layer followed by 3 x 128-cell LSTMP layers with 64 recurrent projection units with a linear recurrent output layer.	Speech data from a female professional speaker, 26 languages.	The LSTM-RNN-based SPSS systems with proposed optimizations surpassed the HMM-based SPSS systems in speed, latency, disk footprint, and naturalness on modern mobile devices. Experimental results also showed that the LSTM-RNN-based SPSS system with the optimizations could match the HMM-driven unit selection TTS systems in naturalness in 13 of 26 languages. Object of synthesis: Utterances.
Fan et al. [6], 2014	Hybrid system of DNN and BLSTM (Bidirectional LSTM): lower hidden layers with a feed-forward structure which is cascaded with upper hidden layers with a BLSTM.	Female, American English, native speaker, both phonetically and prosodically rich. The corpus consisted of 5,000 training utterances (around 5 hours) and 200 extra utterances were used for testing.	Hybrid system can outperform either the conventional, decision tree-based HMM, or a DNN TTS system, both objectively and subjectively. Object of synthesis: sentences.

sigmoid layer and *tanh* layer values are multiplied and we output only the parts which were decided $h_t = o_t * tanh(C_t)$. More deteiled LSTM description is presented in [11].

3 LSTM for Speech synthesis

LSTM is successfully used for different speech synthesis applications. In this section overview of using LSTM networks for speech synthesis is given. Overview is presented in Table 1.

Based on analysed references we can make the following conclusions: 1) it is difficult to compare the quality of synthesis among different experiments because they use different evaluation systems and criteria (both objective and subjective), 2) it is obvious that most of the experiments and LSTM applications are made for the English language.

4 Further investigations

Speech synthesis using LSTM networks according to recent articles outperforms statistical methods and traditional Neural Network implementations. In addition, it gives more flexibility for signal transformations, and most significant among them are: adding the emotions to synthesized speech and voice conversion. In further investigations, LSTM networks will be used for Lithuanian speech synthesis and later for adding intonation and emotions to synthesized speech.

References

[1] Hochreiter S., Schmidhuber J. (1997). Long short-term memory.*Neural computation*, Vol. **9(8)**, pp. 1735-1780.

[2] Graves A., Mohamed A. R., Hinton G. (2013). Speech recognition with deep recurrent neural networks.*In 2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645-6649.

[3] Korvel G., Treigys P., Tamulevicius G., Bernataviciene J., Kostek B. (2018). Analysis of 2D Feature Spaces for Deep Learning-Based Speech Recognition. *Journal of the Audio Engineering Society*, Vol. **66(12)**, pp. 1072-1081.

[4] Tkachenko M., Yamshinin A., Lyubimov N., Kotov M., Nastasenko M. (2017). Speech Enhancement for Speaker Recognition Using Deep Recurrent Neural Networks.*In International Conference on Speech and Computer*, pp. 690-699.

[5] Hochreiter S., et al.(2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

[6] Fan Y., Qian Y., Xie F. L., Soong F. K. (2014). TTS synthesis with bidirectional LSTM based recurrent neural networks.*In Fifteenth Annual Conference of the International Speech Communication Association.*

[7] Zen H., Agiomyrgiannakis Y., Egberts N., Henderson F., Szczepaniak P. (2016). Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices. arXiv preprint arXiv:1606.06061.

[8] Wu Z., King S. (2016). Investigating gated recurrent networks for speech synthesis. *In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5140-5144).

[9] Al-Radhi M. S., Csapó T. G., Németh G. (2017). Deep Recurrent Neural Networks in Speech Synthesis Using a Continuous Vocoder.*In International Conference on Speech and Computer*,Springer, Cham, pp. 282-291.

[10] Zen H., Sak H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis.*In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4470-4474.

[11] Wu Y., Yuan M., Dong S., Lin L., Liu Y. (2018). Remaining useful life estimation of engineered systems using vanilla LSTM neural networks.*Neurocomputing*, Vol. **275**, pp. 167-179.