The Impact of Deep Learning on Speech Synthesis with Embedded or Mobile Devices

A Systematic Review

Hannes Bohnengel
Technical University of Munich
hannes.bohnengel@tum.de

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

KEYWORDS

Deep Learning, Deep Neural Network, Embedded System, Mobile Device, Speech Synthesis, Text-to-Speech

1 INTRODUCTION

Virtual personal assistants (VPA) like Siri, Cortana or Google Now start having a huge impact on the way of interacting with electronic devices like smartphones or notebooks. Up to now the VPAs help only with rather simple tasks like search queries, starting phone calls or setting a clock, but according to a recent survey from the IT research firm Gartner [1], this will change in the near future. With the Facebook Messenger it is already possible to make purchases or to order an Uber car and new use cases are expected soon. The survey also states, that through the vast increase of devices in the scope of the internet of things (IoT) the way of interacting with machines will go towards minimal or zero touch. Instead of interacting through common touch-displays or buttons, the user simply speaks to the device, like to another person. To enable this, both automatic speech recognition (ASR) and speech synthesis are essential technologies.

In this paper I will only focus on the speech synthesis part. A widely spread technique to synthesize human speech from a given text or from linguistic descriptions is statistical parametric speech synthesis (SPSS); also referred to as statistical parametric speech generation (SPSG) [8]. This technique is based on the usage of hidden Markov models (HMMs). Zen *et al.* [15] show that it has several advantages over its predecessor, the concatenative speech synthesis, for example the flexibility in changing voice characteristics and a smaller memory footprint. However the quality of the generated speech still has potential for improvement. Due to over-smoothing the voice sounds muffled in comparison to natural speech.

This is where recent achievements in deep learning come in. Deep learning is usually referred to as a class of machine learning techniques that achieve tasks like feature extraction or pattern analysis by using many connected layers of non-linear information processing [3, 8]. Since 2006 advances in the training algorithms of deep neural networks (DNNs) have enabled the field of deep learning applications to emerge [2]. Most machine learning models until then had used shallow structures, like for example HMMs, Gausian mixture models (GMMs), conditional random fields (CRFs) or support vector machines (SVMs). In these structures only one layer is responsible for generating features out of the raw input signals. While achieving quite good results with rather simple problems, they reach their limit when it comes to more complex tasks like processing human language or natural images [3]. In the tutorial survey [3] the author also states four different approaches to improve speech synthesis through deep learning models, whereof three are dealing with SPSS. One of those three approaches is described in [14], where the authors implemented a part of the speech synthesis system by using a DNN and observed an improved performance in predicting output features. In [5] a more general approach is conducted by investigating what effects the deployment of a DNN on different parts of the SPSS system has. An improvement of the naturalness of the generated speech was one of the main results.

For implementing speech synthesis on resource-constrained devices like smarthpones or tablets, SPSS is considered the best solution due to the tradeoff between voice quality and acceptable footprint size [11]. Since the computational costs of SPSS are often high, some optimization steps like applying fewer conditional calls are conducted in [11] to make the HMM-based speech synthesis technique SPSS more suitable for mobile devices. Going one step further, in [2] an approach to adapt SPSS for embedded devices by using a deep learning model, an auto encoder (AE), is employed. Four tasks (syllabification, phonetic transcription, part-of-speech tagging and lexical stress prediction) are examined

and tested with the use of this deep learning model. As results the authors highligh hugly reduced model sizes, higher training times, very close performance and a similar run time in comparison to the state-of-the-art models. This shows that the usage of deep learning models for speech synthesis on embedded systems is a reasonable step, not only to improve performance and voice quality, but also towards the independability on online databases for speech synthesis.

The remaining paper is structured as follows: Section 2 first states the motivation, why speech synthesis is a useful technology. Then it describes the conventional approach without deep learning models for speech synthesis and gives an overview of advantages and drawbacks of the used models and techniques. This is followed by a brief explanation of the probably most common used technique SPSS, where the paper [15] has been chosen as commonly cited reference. Thereafter two possibilities how SPSS can be improved by deploying deep learning models are characterized, wherefore the papers [5, 14] are reviewed. In Section 3 the motivation, why speech synthesis is important on embedded or mobile devices is given, followed by two examples on how speech synthesis can be implemented on an embedded system, once without [11] and once with deep learning models [2]. Finally Section 4 summarizes the essential points of this paper and gives some future directions.

Remarks and notes starting here:

Open points:

- Improvements of deep learning in speech synthesis also about voice quality. Does that make sense in this context?
- Where the focus of this paper should be (deep learning / speech synthesis / embedded devices)? Most papers are available for:
  - deep learning <—> speech synthesis
  - deep learning <—> embedded devices
- Where to go into technical detail? Suggestion: One way of improving speech synthesis with deep learning.
- Core paper on SPSS [15] available in two versions (4 and 23 pages). Which one to use as core paper? Suggestion: 4 page version as core paper and 23 page version as reference.
- Connection between the improvements of deep learning on speech synthesis (Section 2.4) and the implementation of speech synthesis with/without deep learning (Section 3)?
- Strictly spoken a smartphone/tablet is not an "embedded" device. Is some form of further declaration necessary in this context?

These are the core papers (the red colored):

(1) **Conventional SPSS**

- Statistical parametric speech synthesis [15]
  The most cited paper for SPSS. Two versions available, one with 4 pages (IEEE Xplore, 2007) and one with 23 pages ()[15], 2009 , No access over ScienceDirect –> PDF from http://mlsp.cs.cmu.edu/courses/fall2012/lectures/spss_specom.pdf).

(2) **SPSS with deep learning in general**

- Statistical parametric speech synthesis using deep neural networks [14]
  "The relationship between input texts and their acoustic realizations is modeled by a DNN. The use of the DNN can address some limitations of the conventional approach. (...) The objective evaluation showed that the use of a deep architecture improved the performance of the neural network-based system for predicting spectral and excitation parameters. (..) One of the advantages of the HMM-based system over the DNN-based one is the reduced computational cost."
- The effect of neural networks in statistical parametric speech synthesis [5]
  "This paper investigates how to use neural networks in statistical parametric speech synthesis. (...) In this paper, the effect of DNNs for each component is investigated by comparing DNNs with generative models. Experimental results show that the use of a DNN as acoustic models is effective and the parameter generation combined with a DNN improves the naturalness of synthesized speech."
- Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis [12]
- Efficient deep neural networks for speech synthesis using bottleneck features [6]
- On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis [9]
- TTS synthesis with bidirectional LSTM based recurrent neural networks [4]

(3) **Speech Synthesis WITHOUT DL on Emb. Systems**

- Efficient memory compression in deep neural networks using coarse-grain sparsification for speech applications [7]
- Speeding up deep neural networks for speech recognition on ARM Cortex-A series processors [13]
- Optimizing HMM Speech Synthesis for Low-Resource Devices [11]
  Chosen, because HMM-based synthesis is part of SPSS.
  "Several optimization steps, e.g., changing HMM parameters, applying performance-specific programming methods, are analyzed on three different smartphones in terms of speed, footprint size, and subjective speech quality. The goal is to approach real-time functionality while keeping the speech quality as high as possible"
- Some Aspects of HMM Speech Synthesis Optimization on Mobile Devices [10]

(4) **Speech Synthesis WITH DL on Emb. Systems**

- Robust Deep-learning Models for Text-to-speech Synthesis Support on Embedded Devices [2]
  "This paper focuses on the development of small robust deep-learning models that are designed to provide high quality text-to-speech (TTS) functionality (one of the three main components of HCI) on smart devices, without requiring network access. We obtain very good results in TTS text sub-tasks using models significantly smaller than those used in state-of-the-art approaches"

These are interesting references:

- Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends [8]
- A tutorial survey of architectures, algorithms, and applications for deep learning [3]

Page count estimation:

• Section 1	1.25	(incl. Abstract)
• Section 2.1	0.25	
• Section 2.2	0.5	
• Section 2.3	0.5	
• Section 2.4	1	
• Section 3.1	0.25	
• Section 3.2	0.75	
• Section 3.3	0.75	
• Section 4	0.25	
• References	0.5	
Total:	**6 pages**	

2 SPEECH SYNTHESIS

2.1 Motivation

Why speech synthesis is useful/important?
What are use cases in daily life?
Why there is need to further improve this technology?

2.2 Conventional Approaches

Brief overview of conventional approaches how to implement speech synthesis, with highlighting advantages and drawbacks.

- concatenative & unit-selection
- formant-based
- diphone-based
- SPSS
- etc. ??

2.3 HMM based Speech Synthesis: SPSS

Description of one approach (SPSS) more in detail [15].

2.4 SPSS with Deep Learning Models

Description of the improvements of the approach in previous subsection by using deep learning models.

2.4.1 General ways for improvement

The effect of neural networks in statistical parametric speech synthesis [5]

2.4.2 One specific approach for improvement

Statistical parametric speech synthesis using deep neural networks [14]

3 SPEECH SYNTHESIS ON EMBEDDED DEVICES

3.1 Motivation

Why is it important to implement speech synthesis on embedded platform?
What needs to be thought about when dealing with embedded or mobile devices?

3.2 HMM-based Approach

An example of how speech synthesis can be implemented on embedded platform without deep learning (core paper 3).

3.3 Deep Learning-based Approach

An example of how speech synthesis can be implemented on embedded platform WITH deep learning (core paper 4).

4 CONCLUSIONS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

To do:

- check references
- check correct citing
- check for typos
- check page count (6 pages)
- disable sreen mode on last draft

REFERENCES

[1] 2017. Survey on usage of virtual personal assistants by Gartner. (22 May 2017). http://www.gartner.com/newsroom/id/3551217

[2] Tiberiu Boroş and Stefan Daniel Dumitrescu. 2015. Robust Deep-learning Models for Text-to-speech Synthesis Support on Embedded Devices. In *Proceedings of the 7th International Conference on Management of Computational and Collective intElligence in Digital EcoSystems (MEDES '15)*. ACM, New York, NY, USA, 98–102. https://doi.org/10.1145/2857218.2857234

[3] Li Deng. 2014. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing* 3 (January 2014). https://doi.org/10.1017/atsip.2013.9

[4] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*. 1964–1968. http://www.isca-speech.org/archive/interspeech_2014/i14_1964.html

[5] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. 2015. The effect of neural networks in statistical parametric speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4455–4459. https://doi.org/10.1109/ICASSP.2015.7178813

[6] Y. S. Joo, W. S. Jun, and H. G. Kang. 2016. Efficient deep neural networks for speech synthesis using bottleneck features. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. 1–4. https://doi.org/10.1109/APSIPA.2016.7820721

[7] D. Kadetotad, S. Arunachalam, C. Chakrabarti, and Jae sun Seo. 2016. Efficient memory compression in deep neural networks using coarse-grain sparsification for speech applications. In *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. 1–8. https://doi.org/10.1145/2966986.2967028

[8] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng. 2015. Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine* 32, 3 (May 2015), 35–52. https://doi.org/10.1109/MSP.2014.2359987

[9] Y. Qian, Y. Fan, W. Hu, and F. K. Soong. 2014. On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3829–3833. https://doi.org/10.1109/ICASSP.2014.6854318

[10] Bálint Tóth and Géza Németh. 2011. Some aspects of HMM speech synthesis optimization on mobile devices. In *2011 2nd International Conference on Cognitive Infocommunications (CogInfoCom)*. 1–5.

[11] Bálint Tóth and Géza Németh. 2012. Optimizing HMM Speech Synthesis for Low-Resource Devices. In *Journal of Advanced Computational Intelligence and Intelligent Informatics*. 327–334.

[12] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King. 2015. Deep neural networks employing Multi-Task Learning and stacked bottleneck features for speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4460–4464. https://doi.org/10.1109/ICASSP.2015.7178814

[13] A. Xing, X. Jin, T. Li, X. Wang, J. Pan, and Y. Yan. 2014. Speeding up deep neural networks for speech recognition on ARM Cortex-A series processors. In *2014 10th International Conference on Natural Computation (ICNC)*. 123–127. https://doi.org/10.1109/ICNC.2014.6975821

[14] Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7962–7966. https://doi.org/10.1109/ICASSP.2013.6639215

[15] Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. (April 2009), 1039 - 1064 pages. https://doi.org/10.1016/j.specom.2009.04.004