INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Unit selection with Hierarchical Cascaded Long Short Term Memory Bidirectional Recurrent Neural Nets

Vincent Pollet, Enrico Zovato, Sufian Irhimeh, Pier Batzu

TTS Research dept., Nuance Communications
{vincent.pollet;enrico.zovato;sufian.irhimeh;pier.batzu}@nuance.com

Abstract

Bidirectional recurrent neural nets have demonstrated state-of-the-art performance for parametric speech synthesis. In this paper, we introduce a top-down application of recurrent neural net models to unit-selection synthesis. A hierarchical cascaded network graph predicts context phone duration, speech unit encoding and frame-level logF0 information that serves as targets for the search of units. The new approach is compared with an existing state-of-art hybrid system that uses Hidden Markov Models as basis for the statistical unit search.

Index Terms: speech synthesis, recurrent neural networks, prosodic prediction, unit selection, embedding, deep learning

1. Introduction

Recurrent Neural Networks (RNN) can model the complex dependencies in which information interacts over varying time lags and have been successfully applied across various dynamic sequence modeling tasks. In a joint research activity with IBM [1], bidirectional Long Short-Term Memory networks (LSTM) [2,3,4] were introduced to model prosody for parametric speech synthesis. An LSTM network is similar to an RNN except that the nonlinear hidden units are replaced by a special kind of memory blocks. Each memory block has one or more self-connected memory cells and three multiplicative units – input, output and forget gates – which provide the cells with analogues of write, read and reset operations. With the bidirectional structure, multiplicative gates allow LSTM memory cells to store and access information over long sequences of both past and future events. State-of-the-art performance was demonstrated for predicting pitch and duration targets: a gap reduction up to 47% in subjective rating was measured between a baseline Deep belief Neural Network (DNN) predictor and a ceiling system using copy prosody [1]. Besides prosody prediction, LSTM-RNNs have shown to improve the prediction of all acoustic parameters that drive a parametric speech synthesizer. Zen and Sak [5] demonstrated a single hidden-layer unidirectional LSTM-RNN to predict frame sized mel-cepstrum, logarithmic fundamental frequency and band aperiodicity parameters, following an LSTM-RNN predicting phoneme unit durations. This system equipped with a recurrent output layer achieved an improvement of 0.35 MOS compared to a baseline DNN system. Fan et al. [6] proposed a hybrid neural network of two hidden DNN layers and two hidden bidirectional LSTM-RNN layers to predict frame unit voiced/unvoiced, logF0, line spectral pairs and gain parameters resulting into better quality compared to a baseline DNN model. These recent advances in prediction of acoustic parameters from text derived input features combined with the advances of parametric speech synthesizers [4,5] and adaptive enhancement [4] have narrowed the gap between parametric and unit selection speech synthesis. For some tonal languages this gap was closed, as claimed by Zen et al. [17].

More recently, a revolutionary approach named *WaveNet* was introduced by van den Oord et al. [15] that can generate speech sample outputs from wide convolutional neural networks without the requirement (and limitation) of an intermediary speech representation or of a speech sample database. This type of neural net covers and generalizes effectively all samples, their interdependencies and relations to linguistic and pitch input features. High quality speech can be obtained, however *WaveNet's* extreme consumption of computational resources makes it impractical for commercial applications. Furthermore synthesis tuning, an effective means to optimize and overcome unit selection problems, is yet to be explored for such generative techniques.

Despite these recent advances in parametric and generative synthesis, unit selection remains today's leading commercial solution because:

- The upper bound of a parametric TTS system is determined by its synthesizer model and is currently below the fidelity of wide-band recordings.
- TTS applications often include inaudible mixing of selected units and prompts. A prompt can either be an indexed recording of an utterance or a 'tuned and stored' sequence of unit identifiers (in its simplest form). With the aid of tuning, a TTS system can generate a better result.
- A TTS engine is often a sub-component of a larger system, deployed on a device or on the cloud. Low usage of computational resources is required for a viable and cost-effective deployment.

Consequently, improving today's widely deployed speech synthesizers is part of a non-zero sum game. In this work, we present a reformulation of the unit selection search with bidirectional LSTM deep recurrent neural nets. The paper is structured as follows. We begin by reviewing previous related work in this area in Section 2. After providing an overview of the proposed system's architecture, and the corpus used for this work in Sections 3 and 4, the new approach is compared and evaluated in 5 before concluding in Section 6.

2. Prior and Related Work

In previous work [12] we introduced a high-quality hybrid speech synthesis system that selects, generates and concatenates speech units using a Hidden Markov Model (HMM) framework. The unit search is formulated as a probabilistic optimization that operates on multi-form segments: either selected from a speech database or generated from predicted parameters. This is the system we use as a

http://dx.doi.org/10.21437/Interspeech.2017-428

baseline for conducting comparisons in this work. However, the hybrid mode is set so that no parametric segments are generated for rendering the final speech output. Inspired by the benefits of LSTMs in modeling prosody over HMMs, further attempts were made by Fernandez et al. [13] to improve an existing unit selection algorithm by inserting LSTM predicted prosody as targets and compute related sub-costs as part of the unit search process. However, the largest benefits reported in this work originate from post-search corrections of prosody by means of pitch synchronous signal manipulation.

Another work that is close to ours is that of Merritt et al. [14], whereby a DNN model was used to generate features that serve as input for a target cost function in an existing unit selection system. It was found that this approach was more effective compared to an HMM system that uses the mel-ceptrum, band-aperiodicity and F0 parameters generated at the output or to using context embeddings as target features resulting from a bottleneck layer.

Compared to Fernandez and Merritt our work is different in the sense that we don't insert outputs from neural networks that work well for parametric synthesis in an existing (statistical) unit selection system. In our work, we top-down design a LSTM neural network graph for the task of unit selection.

3. Proposed Architecture

3.1. Hierarchical Cascaded Mapping

Instead of utilizing a *vanilla* LSTM-RNN mapping from text derived input features to acoustic output features, for the purpose of unit selection we propose a *hierarchical cascaded mapping graph.* The sequence of inputs is mapped by connected cascaded tiers of different resolution and information, namely duration, encoding and prosody tiers. As depicted in Figure1, in the **duration tier** linguistic input vectors ***x*** are mapped to contextual phone-level durations ***d***. In the second **encoding tier**, the durations combined with the linguistic information (***x,d***) are mapped to fixed sized encoding target vectors ***h***, to be used in the unit search. In the third **prosody tier**, these encodings together with the linguistic input features (***h,x***) are mapped to target logF0 frame values ***p***, also used in the unit search.

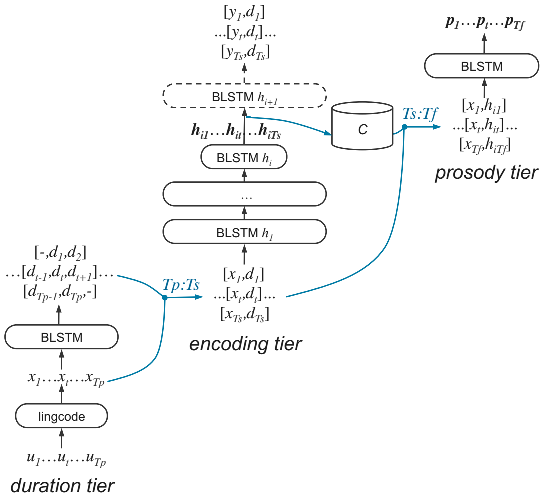


Figure 1: *Hierarchical Cascaded Mapping Bidirectional LSTM graph.*

This mapping by a graph of BLSTM networks *transforms* information derived from text, gradually into a representation that is deep in time and space. The hidden and deep representation of the fixed sized vectors of the encoding tier is similar to the natural language processing task of *word embedding* [6,7], whereby words or phrases from a vocabulary are mapped to vectors of real numbers in a low-dimensional space relative to the vocabulary size. The resulting vectors are assumed to carry a degree of syntactic and semantic information. The size of the basic unit for generating speech by means of unit selection is smaller than a word; it is typically a diphone, half-phone or a state unit. Consequently we adopt the term **speech unit embedding** (SUE) to define the hidden vectors.

3.2. Generation of Speech Unit Embeddings

Given a sentence with T sub-word unit sequences $u_t, t \in [1,T]$ we first convert the units to fixed sized vectors through a linguistic encoding function $x_t = lingcode(u_t)$ that annotates the sub-word unit with orthographic, syntactic, prosodic and phonetic information. The categorical information such as phone label, lexical stress, phrase type, prosodic boundary-type, part-of-speech, etc. are encoded as one-hot vectors.

Information of higher hierarchy is exploited as well. For instance, word representations are repeated for all the sub-word units. Contextual information on the other hand, i.e. left and right phonetic context labels, is not used.

Numerical information such as syllable counts, phone counts and distances are standardized to zero mean and unit variance.

A bidirectional LSTM-RNN is used to generate the SUE vectors by summarizing information from both directions of units, and therefore incorporates the contextual information we omitted as part of the vectors x. The bidirectional LSTM network contains a forward $\overrightarrow{lstm}$ function, which reads the sentence sequence from u_1 to u_T and a backward $\overleftarrow{lstm}$ function, which reads from u_T to u_1.

Multiple bidirectional LSTM layers are stacked to get deep annotations. We obtain an annotation s_t for a given unit u_t by concatenating the forward hidden layer state $\overrightarrow{h_{t\imath}}$ and backward hidden state $\overleftarrow{h_{t\imath}}$, i.e. $s_t = h_{ti} = [\overrightarrow{h_{t\imath}}, \overleftarrow{h_{t\imath}}]$, which summarizes the deep information at a bidirectional LSTM layer index i of the entire sentence centered around u_t.

$$
\begin{aligned}
x_t &= lingcode(u_t), t \in [1,T] \\
\overrightarrow{h_{t1}} &= \overrightarrow{lstm}(x_t), t \in [1,T] \\
\overleftarrow{h_{t1}} &= \overleftarrow{lstm}(x_t), t \in [T,1] \\
h_{t1} &= [\overrightarrow{h_{t\imath}}, \overleftarrow{h_{t\imath}}] \\
&\dots \\
\overrightarrow{h_{t\imath}} &= \overrightarrow{lstm}(h_{t(i-1)}), t \in [1,T] \\
\overleftarrow{h_{t\imath}} &= \overleftarrow{lstm}(h_{t(i-1)}), t \in [T,1] \\
s_t &= h_{ti} = [\overrightarrow{h_{t\imath}}, \overleftarrow{h_{t\imath}}]
\end{aligned}
\qquad (1)
$$

For space reasons and sufficient coverage elsewhere, the reader is directed to consult one of the references provided for additional details on bidirectional LSTM models including the forward-pass equations [1,2,3,4]. Each speech unit of the inventory is annotated with a SUE vector by means of an offline labeling process. It should be noted that multi-task learning is intrinsic to this cascaded network as the different encoding tiers learn the mapping of text derived linguistic features to unit selection targets of different meaning and size.

The training of the encoding tier is particularly special and includes **auto-encoding**, **regression** to acoustic output targets and **classification** of message types of the sequence. The auto-encoding of duration is realized by presenting the contextual duration d_t, predicted by the first tier simultaneously at the input and the output for each unit u_t. This enables embedding of duration as part of the SUE vector and consequently avoids having to deal with separate unit selection sub-costs for duration. Regression to acoustics of the unit u_t involves an acoustic encoding $y_t = acousticcode(u_t)$, which constructs the sub-unit vector with mel-cepstrum, logF0 and voicing information. Following the encoding tier, the prosody tier stacks the linguistic encodings x_t with the speech unit embeddings h_t as input, and maps this to observable logF0 outputs p_t. Each tier has different unit sizes and therefore different sequence lengths. Propagating through the hierarchical network graph includes transitioning from phone-sized to state-sized sequences, *Tp:Ts*, and from state-sized to frame-sized sequences, *Ts:Tf.*

3.3. Unit Search

As basis of the search for the best sequence of speech units from a set of candidate speech unit vectors $\vec{C}$ given the targets of the cascaded neural net, we concatenate the predictions of the hidden speech unit embeddings and the frame-sized observable logF0 vectors $\tilde{G}_t = [\tilde{s}_t, \tilde{p}_t]$.

The search is implemented as a dynamic programming optimization whereby the best units C^* are obtained by computing the arguments that minimize the loss function *f*:

$$C^* = \underset{C}{argmin}\, f(\tilde{G}, C) \qquad (2)$$

The loss function *f* consists of; a *target* loss function f^t and a *connection* loss function f^c. The regularization term Q balances the contribution of the connection loss.

$$f(\tilde{G}, c \in C) = \sum_{t=1}^{T} f^t(\widetilde{G_t}, c_t) + Q \sum_{t=2}^{T} f^c(c_{t-1}, c_t) \qquad (3)$$

The *target loss function* can be formulated as:

$$f^t(\widetilde{G_t}, c_t) = -logP(c_t|\widetilde{G_t}) \qquad (4)$$

Applying the *softmax* function and hereby using a similarity between vectors $\widetilde{G_t}$, c_t and a set of k_t sample candidates including c_t:

$$P(c_t|\widetilde{G_t}) = \frac{e^{sim(\tilde{G}_t, c_t)}}{\sum_{k_t \in K_t} e^{sim(\tilde{G}_t, k_t)}} \qquad (5)$$

Using the cosine similarity, substituting and simplifying the denominator results into a target distance metric that satisfies the triangular inequality theorem.

$$f^t(\tilde{G}_t, c_t) \cong \frac{1}{\pi} cos^{-1}\left(\frac{\tilde{G}_t^{\top} c_t}{|\tilde{G}_t||c_t|}\right) \qquad (6)$$

The connection loss is a Markov process after applying a class of boundary functions *r* to edges of speech units.

$$f^c(c_{t-1}, c_t) = -logP(r_L(c_t)|\, r_R(c_{t-1})) \qquad (7)$$

The boundary functions resulting into co-occurrence observations are obtained by another neural network that learns to classify join types from a set of acoustic features. The conditional probabilities of the resulting co-occurrence observations can be modeled by a probabilistic model as in [12] or a connection neural work.

4. Experiment

4.1. Data

A large American English female speech database of formal and conversational speaking styles and including various sub-domains such as navigation, weather forecast, greetings, numbers etc. was used for this experiment. The speech signals, sampled at 22.05 kHz, were analyzed with a window of 24 ms and a frame-shift of 8 ms. Acoustic analysis resulted into logF0, voicing, 34th order mel-cepstrum and speech excitation parameters. The data was split into a training set (84% of the data), a validation set (15%) and 1% was held out for testing. Both the proposed and reference [12] systems share the same data set, speech analysis, unit alignment, unit size (one-third of a phone) and linguistic processing function.

4.2. Model Setup and Training

The tiers of the cascaded network graph were trained by means of stochastic gradient descent. The network weights were initialized randomly according to a normal distribution with 0 mean and 0.1 standard deviation. The training was stopped after 20 epochs of no improvements on the validation set or in case a maximum number of epochs was reached. The *lingcode* function extracts and normalizes a vector x_t of size 367 for each unit. The duration tier has three bidirectional LSTM layers of sizes; 256, 160, 40 and outputs current, left and right phoneme duration expressed in log seconds. The encoding tier has four bidirectional LSTM layers of sizes; 256, 256, 160, 256, whereby the smallest layer acts as the SUE layer.

The input is the x_t vector aggregated with the duration output of the previous tier and consequently of size 370. The output targets are the mean mel-cepstral coefficients, the mean continuous logF0, variance of continuous logF0, phone duration, voicing type, message style and message type vectors expressed as one-hot vectors for each style and type category.

The up-sampling ratio from the phone-level duration tier to the unit-level encoding tier is 1:3. Also the prosody tier has three bidirectional LSTM layers of the sizes: 384, 192 and 64. The input is the x_t vector is aggregated with the SUE vectors. The outputs are logF0 values. The up-sampling ratio from the unit-level encoding tier to the frame-level prosody tier is 1:2.

5. Results

5.1. Objective Analysis

We express the distance between real and predicted observable output features that are used either explicitly or implicitly (i.e. encoded as part of the SUE vector) in terms of cross-correlation (XCOR) and Root Mean Squared Error (RMSE).

The targets predicted by the proposed and baseline systems were compared with the ground-truth of the held-out data that was unit-aligned and labeled with frame-level logF0 values. We calculated the distances for phone durations and unit level average logF0 of the selected unit sequences.

Table 1 shows an overview of the results. A 21% and 19% increase in XCOR and a 12% and 10% decrease in RMSE for logF0 and phone duration (pDur) is obtained with the proposed system compared to the baseline.

System	Target	RMSE	XCOR
Baseline	logF0	0.1622±0.0365	0.5142±0.2047
Proposed		0.1464±0.0262 **(-12%)**	0.6226±0.1309 **(+21%)**
Baseline	pDur	0.0487±0.0153	0.5861±0.1744
Proposed		0.0429±0.0104 **(-10%)**	0.6993±0.0948 **(+19%)**

Table 1: *Objective analysis results, Root Mean Squared Error and Cross Correlation vs. ground-truth.*

5.2. Subjective Analysis

In case of non-hybrid unit selection, the targets are hit or missed depending on the coverage of the unit inventory. In general, the training of the search models typically occurs on the full unit inventory, and consequently the unit coverage is well aligned with the prediction.

To study how the prediction improvements translate into perceptual gains, a Mean Opinion Score (MOS) test was conducted. The test stimuli consisted of 59 American English synthesized sentences, including declarative and interrogative phrases of various lengths.

The test was designed so that each sample had to be evaluated by at least 40 American English mother tongue listeners that were recruited via Amazon Mechanical Turk. The subjects were asked to judge the overall quality of the stimuli on the basis of a 5 points MOS scale, where 1 means “very poor” and 5 “excellent”. An analysis of their responses, based on correlation was conducted to reject scores from unreliable listeners [16]. Final results depicted in figure 2 show that the proposed system is better than the baseline with an average score of **3.94** MOS.

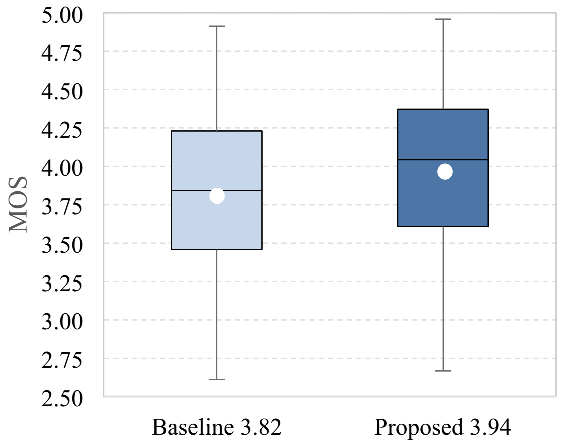


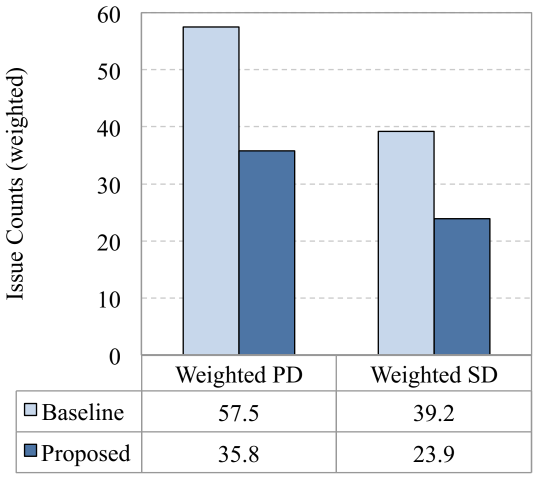
Figure 2: *Boxplot of Mean Opinion Score results. The white dots depict the averages.*

5.3. Stability Analysis

Stability is an important qualitative metric for unit selection systems. Unit scarcity or an ill-conditioned unit search leads often to salient unnatural glitches in synthesis. With MOS tests, assessing high quality systems that rarely exhibit subtle glitches is difficult. Listening conditions of subjects may vary, subjective judgment involves many aspects and enforces a single vote for the entire sentence. The averaging operation on the votes filters the degree and type of glitches in speech synthesis. Care must be taken when dealing with different systems, i.e., parametric systems fit better the overall averaging character of the test compared to unit selection systems.

Therefore in addition to a MOS test we performed a stability test. In view of assessing a TTS back-end, we defined two categories of defects: prosody defects (PD) and smoothness defects (SD), each having two degrees of severity. A large set of input text containing various domains and phrase types was used to synthesize samples with both systems. Experts in the field of speech synthesis equipped with high quality headphones listened in randomized order and annotated the degrees and categories of perceived defects by marking up the corresponding areas in the input text. The defects were weighted and summed over listeners and per sample to provide normalized scores. Defects jointly marked by the experts received a higher weight.

Figure 3 presents the results of the stability test and shows that the proposed system outperforms the baseline by reducing the number of weighted smoothness defects by **39%** and weighted prosody defects by **38%**.



	Weighted PD	Weighted SD
□ Baseline	57.5	39.2
■ Proposed	35.8	23.9

Figure 3: *Results Weighted Defects.*

6. Conclusions

We introduced a new approach for unit selection that exploits hierarchical organization of the knowledge modeled by cascaded bidirectional LSTM networks. Compared to an existing state-of-art hybrid system that uses HMMs, clear improvements are shown in terms of: objectively observable predicted quantities, such as target prosody; and subjective quality, measured through MOS and stability tests. In particular, stability tests have shown a reduction in the amount of perceivable prosody and smoothness issues.

7. References

[1] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, “Prosody contour prediction with long short-term memory, bidirectional, deep recurrent neural networks,” in Interspeech, Singapore, 2014, pp. 2268–2272.

[2] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] F. A. Gers, J. Schmidhuber, and F. Cummings, “Learning to forget: Continual prediction with LSTM,” Neural Computaiton, vol. 12, no. 10, pp. 2451–2471, 2000.

[4] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, “Learning precise timing with LSTM Recurrent Networks,” J. of Machine Learning Research, vol. 3, pp. 115–143, 2002.

[5] H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low latency speech synthesis,” in Proc. ICASSP, Brisbane, 2015, pp. 4470–4474.

[6] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in Proc. Interspeech, Singapore, 2014, pp. 1964–1968.

[7] A. Sorin, S. Shechtman, and V. Pollet, “Uniform Speech Parameterization for Multi-form Segment Synthesis,” in Proc. Interspeech, 2011, pp. 337–340.

[8] Y. Agiomyrgiannakis, “Vocaine the vocoder and Applications in speech synthesis”, in Proc. ICASSP 2015

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. ICLR, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proc. NIPS, 2013, pp. 3111–3119.

[11] S. Shechtman and S. Shechtman, "Sinusoidal model parameterization for HMM-based TTS system," in Proc. Interspeech Makuhari, 2010, pp. 26-30

[12] V. Pollet and A. Breen, “Synthesis by generation and concatenation of multiform segments,” in Proc. Interspeech, 2008, pp.1825–1828.

[13] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, “Using Deep Bidirectional Recurrent Neural Networks for Prosodic- Target Prediction in a Unit-Selection Text-to-Speech System,” in Proc. Interspeech, 2015.

[14] T. Merritt, R. Clark, Z. Wu, J. Yamagishi, S. King, “Deep Neural Network-Guided Unit Selection Synthesis”, in Proc. ICASSP, 2016.

[15] A. van den Oord, S. Dieleman, H. Zen, K. Si- monyan, O. Vinyals, A. Graves, N. Kalch- brenner, Nal, A. Senior, and K Kavukcuoglu, “Wavenet: A generative model for raw audio,” CoRR abs/1609.03499, 2016.

[16] M. Legát, and J. Matoušek, “Collection and Analysis of Data for Evaluation of Concatenation Cost Functions”, in Text, Speech and Dialogue, proceedings of the 13th International Conference TSD 2010, Lecture Notes in Artificial Intelligence, p. 345--352, Springer, BerlinHeidelberg, Germany, 2010.

[17] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson and P. Szczepaniak, “Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices”, in Proc. Interspeech, 2016, San Francisco, CA, USA.