

# Pathological speech synthesis from articulation

Bence Halpern<sup>1,2</sup>, Rob J. J. H. van Son<sup>1</sup>, Michiel W. M. van den Brekel<sup>1,2</sup>

<sup>1</sup>NKI-AVL, Amsterdam

<sup>2</sup>ACLC, University of Amsterdam, The Netherlands

b.halpern@nki.nl, r.v.son@nki.nl

## Abstract

Lorem ipsum dolor sit amet, consectetur 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 velut 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, consectetur 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.

**Index Terms:** computational paralinguistics, articulatory-to-acoustic speech synthesis, deep learning, pathological speech

## 1. Introduction

Understanding how articulation affects speech is a central question in speech research. The source-filter model was one of the first models to tackle this problem by noting that speech production could be described by the geometry of the vocal tract and the glottal wave. Mathematically, the source-filter model synthesises speech by exciting an autoregressive (AR) model with a signal, where the AR coefficients capture the geometry of the vocal tract, which is also represented by an area function [1].

Recently, deep learning methods became popular to understand articulation. These methods use a measurement tool, called electromagnetic articulography to obtain articulation data [2] [3] [4] along with recurrent neural networks, which are neural networks that are able to deal with the sequential nature of data [5]. Data-driven methods became of interest also in real-time speech synthesis. The efficacy of real-time speech synthesis has been investigated using a tool called permanent magnetic articulography by Gonzalez et al [6], which gave understandable speech.

The conclusion of these endeavours were that while it is possible to predict some of the pitch from articulation, the quality suffers. It is, however, possible to obtain satisfactory values for the cepstral quality.

In the author's broader research, the aim is to demonstrate how pathologies in articulation are translated to speech. In some types of pathological speech the laryngeal function remains intact, meaning the  $F_0$  remains unchanged. This means, by knowing the original pitch and the pathological deviations in articulation, it should be possible to synthesise pathological speech if the pathology is on the articulation level.

The idea with MFCCs is to separate vocal tract information from the excitation information. Thus based on the electrode signals of the vocal tract configuration it should be possible to

predict part of the spectra only dependent on the vocal tract configuration.

In this paper, a technique is described which combines healthy speech from the three largest articulatory dataset, MNGU0 [7], MOCHA-TIMIT and TORGO [8] to create a general speaker-independent model.

The main contributions of this paper are,

- a set of benchmarks for single-speaker and multi-speaker articulatory to acoustic synthesis
- a discussion of what these neural networks learn
- an example of how to construct pathological speech using this framework

Our code is also available as a Github repository on the link <https://github.com/karkirowle/vocoder-clean>.

## 2. Method

### 2.1. Dataset preprocessing

#### 2.1.1. Electrode preprocessing

Electromagnetic articulography (EMA) is a measurement technology which uses sensor coils which are placed on the articulators of the vocal tract. Using this technology it is possible to record the displacement of the articulators which can be used to approximate the configuration of the vocal tract.

Three public datasets are combined in this study, however note that the combination is not entirely straightforward as the electrodes sometimes don't record the same channel, and there is absolutely no guarantee that even for the same speaker the electrode does not fall off during the experiment.

It has been decided that seven electrodes will be used for this experiment out of the total eight, Table 1 includes the alignment of the channels that were used.

In order to deal with the speaker-wise variations, the articulatory trajectories of the datasets were standardised to have zero mean and unit variance on a per speaker basis. While this doesn't alleviate problems if an electrode falls off during the experiment, given enough per speaker data it will approximate a distribution of articulation movements.

In the case of the TORGO dataset, some of the channels contained spikes, which were attached to the dataset. If the spikes happened in electrode channels, which were used by the neural network, then it has been excluded. This decision has been made after implementing some simple signal processing algorithms to remove these spikes.

Previously [9], the effect of delay on the output signal were investigated. It has been found that delay is beneficial in case of causal models. Our choice of function approximators are restricted to acausal models, so theoretically a speech sample could be synthesised using values from the future.

Table 1: Articulatory information recorded in datasets

MNGU0	MOCHA-TIMIT	TORGO
Tongue dorsum (T3)	Tongue dorsum (T3)	Tongue back
Tongue blades (T2)	Tongue blades (T2)	Tongue middle
Tongue tip (T1)	Tongue tip (T1)	Tongue tip
Lower incisor (T3)	Jaw	Lower incisor
Upper incisor	Nose	Upper incisor
Upper lip	Upper lip	Upper lip
Lower lip	Lower lip	Lower lip

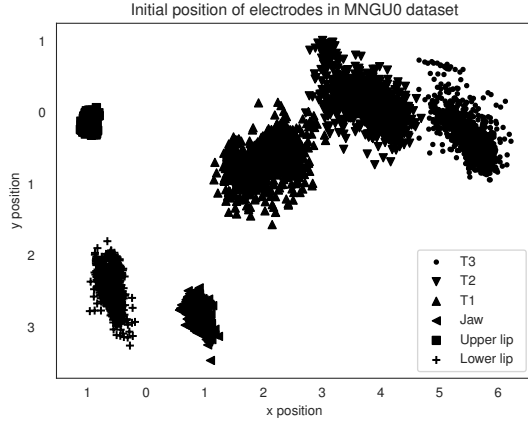


Figure 1: The visualisation of electrode locations for all samples in the MNGU0 dataset at time point  $t = 0$

### 2.1.2. Speech data processing

The total combined dataset contains six British male and three British female speakers, with a total of 6117 utterances. The recordings from the microphones were all 16kHz. Only the healthy speech has been included from the TORGO dataset. There are 1263 utterances from the MNGU0, 920 from the MOCHA-TIMIT and 3934 from the TORGO dataset.

Vocoder features were extracted with the PyWORLD vocoder [10] and compressed with the PySPTK toolkit available at <http://github.com/r9y9/pysptk>. The period between consecutive frames were 5 milliseconds. The resulting 40 MFCC and 1 power parameters were used to generate static and delta parameters, resulting in 82 parameters for the training. As the first step of the MFCC extraction  $\alpha = 0.42$  were used as a pre-emphasis coefficient. The PyWORLD vocoder also provides the  $F_0$  and BAP values, which were not used for training.

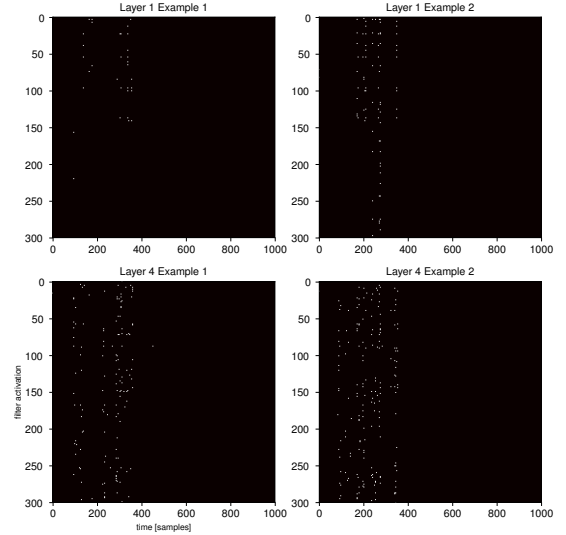
### 2.1.3. Sampling

It is important that the input and output sample rates of the different datasets are matched, because otherwise, the input-output pairs contain different amounts of information.

The sampling frequency of the original EMA signals were 500 Hz, however the MNGU0 was provided to us downsampled to 200 Hz. To match this frequency, the sampling frequency of the other datasets were also downsampled to 200 Hz.

For the MNGU0 dataset, NaN (not a number) values occurred when the measurement precision was low. These values

Figure 2: Thresholded difference mask of activations indicates that a boundary phenomena is learned by the neural network. Best viewed in zoom.



were simply interpolated linearly.

To ease training, the input signals were either truncated or padded so there were a total of  $T = 1000$  samples for each training example. For input signals which are shorter, it is assumed that the last part is silence, so it is padded with the last element. These are not propagated back during training, to avoid the neural network making inference based on the length of the last element.

### 2.1.4. Fundamental frequency interpolation

Previously, it has been found beneficial to take the logarithm of the pitch to obtain a continuous  $F_0$  curve. When the logarithm is not defined, linear interpolation has been done. [6] An alternative method also exists with exponential interpolation which is described in [11]. It has been decided that the linear interpolation technique will be used.

## 2.2. Synthesis setup

The setup for inference and training can be seen on Figure 3. In the training setup, the MFCCs are given to be predicted only. The pitch and band aperiodicities are directly feed to the vocoder during synthesis time, as these don't contain information about articulation.

# 3. Experiment

## 3.1. Neural network experiments

### 3.1.1. Previous architectures

Three recurrent neural networks architectures were trained based on previous papers tackling similar problems. Slight modifications have been made to each architecture.

Figure 3: Red dashed line indicates training-only setup, and blue thick lines indicate inference for speech synthesis. Best viewed in colour.

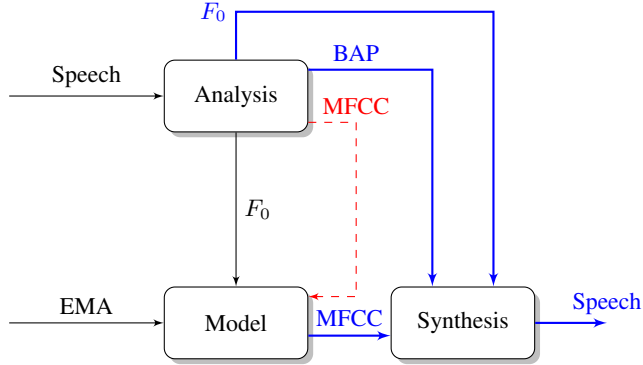


Table 2: Comparison of preprocessing techniques

Author	Liu	Taguchi	Gonzalez
EMA/PMA	EMA	EMA	PMA
MFCC	40 + 1	40 + 1	24 + 1
Delta	No	Yes	Yes
EMA sampling	200 Hz	200 Hz	100 Hz*
Input standardisation	Yes	N/A	Yes
Trajectory smoothing	No	Yes	No
Output standardisation	Yes	Yes	Yes
Vocoder	STRAIGHT [12]	WORLD	STRAIGHT [12]

\*Upsampled to 200 to match analysis rate

Based on [4], a bidirectional LSTM was trained with four layers. The final layer was a fully connected layer matching the output dimensions. This neural network was trained using stochastic gradient descent and a learning rate of  $\alpha = 0.01$ . ii Another publication [3] used a neural networks with three fully connected layers with 128 hidden units, each layer has a linear activation function, which is followed by Layer Normalisation and a sigmoid activation function. This is followed by two BLSTM layers with 256 hidden units. Finally, a fully connected layer is used. This neural network was trained using Grave’s RMSProp [13], with a learning rate of  $\alpha = 0.01$ .

Finally, a BGRU was trained based on the architecture of [6].

### 3.1.2. Justification for layer choices

Determining a good set of architecture and parameter settings is the most difficult part of the experiment design. Bidirectional LSTMs have performed the best on predicting the MFCC spectra in all related papers. Taguchi uses fully-connected layers and while it not justified in the paper, it is reasonable to assume that this does pre-processing of the time-series.

It can be easily seen mathematically that after after the last LSTM layer, a fully connected layer is also needed as this acts as a linear regression based on the LSTM parameters. This is because the range of the activation function is not sufficient to capture the range of the normalised MFCCs. It is possible to introduce some structural bias by using the same linear regression from each frame, somewhat limiting complexity.

### 3.1.3. Optimisation

In terms of the optimisation, Adam [14] and Graves’s RMSPProp is preferred, which is not surprising due to the fact that these known to be better for optimising non-stationary objectives. With the SGD, learning rate scheduling had to be used.

In order to choose the best architecture, all of them have been reimplemented with slight modifications. This is a better practice, because the MCD’s were interpreted differently on the different papers, i.e the error in the silence frames were not taken into consideration. All of them were compared on the mngu0 dataset. The best performing architecture have been re-trained on the full dataset, using the preprocessing techniques mentioned above, and an MCD of 5.48 have been obtained.

Due to the imbalanced dataset this immediately begs the question, how this score is reflected speaker-wise. The results can be seen in the table. According to the table below it seems to be that the true result seems to be even better. Note that the validation set sizes are smaller for the TORGO dataset which could introduce bias.

Table 3: Performance of speaker-independent articulatory to acoustic neural network

Dataset	Multi-speaker	Single-speaker
<b>Combined result</b>	5.48 dB	N/A
<b>MNGU0</b>	6.19 dB	4.77 dB
<b>Female MOCHA-TIMIT</b>	5.73 dB	5.23 dB
<b>Male MOCHA-TIMIT</b>	<b>5.07 dB</b>	5.83 dB
<b>TORGO Part 1</b>	4.06 dB	N/A
<b>TORGO Part 2</b>	4.45 dB	N/A
<b>TORGO Part 3</b>	3.86 dB	N/A
<b>TORGO Part 4</b>	4.80 dB	N/A
<b>TORGO Part 5</b>	4.90 dB	N/A
<b>TORGO Part 6</b>	4.54 dB	N/A
<b>TORGO Part 7</b>	4.62 dB	N/A
<b>TORGO Part 8</b>	14.15 dB	N/A
<b>TORGO Part 9</b>	4.54 dB	N/A
<b>TORGO Part 10</b>	4.83 dB	N/A

Table 4: Transfer learning comparison with single speaker models

Dataset	Speaker only	Transfer preproc
<b>MNGU0</b>	4.77 dB	N/A
<b>Female MOCHA-TIMIT</b>	5.23 dB	11.43 dB
<b>Male MOCHA-TIMIT</b>	5.88 dB	7.86 dB

Table 5: Held out validation for different architectures on MNGU0

Author	MCD
<b>Gonzalez</b>	4.77 dB
<b>Taguchi</b>	7.28 dB
<b>Liu</b>	4.84 dB

For training the mean squared error loss function was used, and for evaluation the Mel cepstral distortion (MCD) have been employed. [16]

Ten fold cross-validation was performed to estimate the out-of-sample generalisation capability of the neural networks.

Table 6: Comparison of different training methods

Author	Liu	Taguchi	Gonzalez
<b>BLSTM layers</b>	4 (128)	2 (256)	4 (150) GRU
<b>Dense layers</b>	1	3+1	1
<b>Regularisation</b>	No	LayerNorm	Noise 0.05
<b>Dropout</b>	No	Yes (50 %)	No
<b>Optimiser</b>	SGD	Grave’s RMSProp	Adam
<b>Learning rate</b>	0.01*	0.01	0.003
<b>Gradient clipping</b>	No	5	No
<b>Early stopping</b>	Yes	Yes	Yes
<b>MLPG [15]</b>	No	Yes	Yes
<b>Maximum epochs</b>	32	N/A	100
<b>Batch size</b>	N/A	8	100
<b>Incremental training</b>	No	Yes	Yes

\* with decay after Epoch 11

\*\* from author communication

### 3.2. Pathological speech synthesis

Pathological speech synthesis is performed by considering the articulatory space and taking knowledge about the change of articulation.

In tongue cancer, articulation of the tongue is impeded. In practice, it is found that teaching patients to speak at a slower rate helps these articulation problems. Thus, it is hypothesised that is the maximum velocity of the tongue that is limited in pathological speech.

A pathological speech transformation could then be constructed for a discretet time signal by first taking the discrete time difference,

$$d[t] = x[t] - x[t - 1], \quad (1)$$

where  $x[t] \in \mathbb{R}^T$  is a signal for one particular electrode channel.

The difference signal then can be thresholded using,

$$d_p[t] = \min(d[t], c) \quad \text{for } d[t] \geq 0, \quad (2)$$

$$d_p[t] = \min(d[t], -c) \quad \text{for } d[t] < 0, \quad (3)$$

where  $c \in \mathbb{R}^+$  is a positive number representing an arbitrary threshold.

After obtaining this signal a cumulative sum could be performed to obtain the pathological EMA signal

$$p[t] = \sum_{i=0}^t x[i]. \quad (4)$$

This signal then could be fed through a feedforward run of a neural network to synthesise pathological speech.

## 4. Results and discussion

### 4.1. Benchmark results

Table 7: Comparison of 10-fold CV performance of neural networks

Architecture	MCD
Liu 2018	5.8 dB
Taguchi 2018	6.2 dB
Modified Liu	6 dB

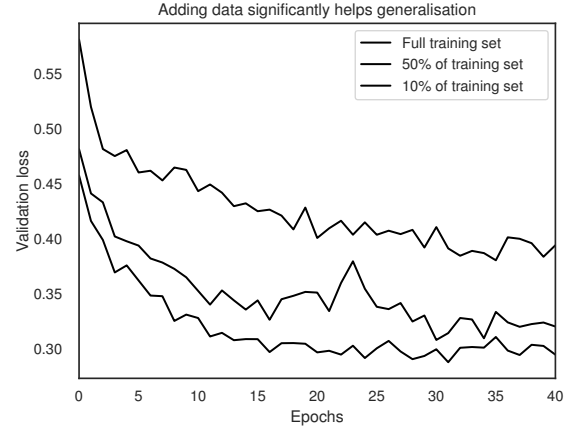


Figure 4: Partial data retraining shows that adding more data would decrease loss

It is important to note that the original networks were trained on single speakers, that is why the MCD values are higher than in the original publications.

### 4.2. Learning curves

The training set was increased from ten percent of it’s total size to it’s total size in increments of ten percents, and the mean squared error was calculated at all epochs of training for the validation set, which can be seen on Figure 4.

After that, a paired t-test was performed to answer whether there is a statistically significant improvement with each addition of the training data. It has been found that for each addition, the paired t-test resulted in statistically significant results, which indicates that it is very likely that addition of more data improves the model.

To estimate how much data would be needed to achieve a “perfect” performance, assuming a linear fit, the mean of the validation loss in last 5 epochs were taken and regressed against the amount of training data included in number of samples. This fit can be seen on Figure 5. Taking the ratio of the slope and the intercept, approximately 8382 training data point would be needed.

Again, note that there are several limitations of this assumption. First, the relationship is most likely not linear. Secondly, given any noise, achieving zero loss is impossible. However, these benchmarks still have merit in future experiment design.

### 4.3. What do these neural networks learn?

Based on our training, it seems clear that a GRU architecture was superior to an LSTM architecture in our case, when used with a more conventional optimiser. General consensus of sequential networks seems to be that sometimes GRU and sometimes LSTM is better [17]

Recently, there have been many advancements in understanding what neural networks learn. Convolutional neural networks can be analysed via conventional methods in filter analysis, classification neural networks can propagate back gradients to find the most important inputs for the prediction. These techniques are not applicable for recurrent neural networks in a regression context, so we resort to exploring the activations

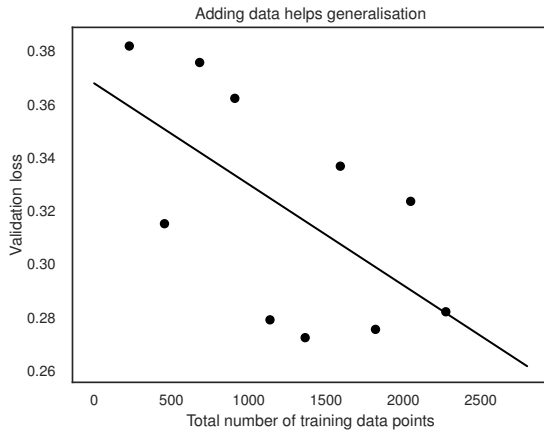


Figure 5: Partial data retraining shows that adding more data would decrease loss.

outputs of the layers.

To make these intelligible, a difference mask is thresholded to find oriented peaks in the spectra. On Figure 2, two things can be observed. Firstly, line-like boundaries are learned, and their duration indicates these might approximate phone to word level representations. Secondly, as the activations propagate through the deeper layers of the neural networks, these line like boundaries are better approximated. iii

#### 4.4. Vocoder upper bound

In our experience, one significant bottleneck in this application is the vocoder itself. To quantify this

#### 4.5. Pathological speech examples

The pathological speech examples can be listened on the webpage of the author, see <http://karkirowle.github.io/paper1>.

### 5. Conclusion

This paper is a proof of concept that it is possible to make pathological speech by incorporating changes in an articulatory domain. Benchmarks have been also established and an open source repository is also available in order to reproduce these results. Using cross-validation as the training data increases, bounds have been established on the expected amount of data needed for the model to improve.

It can be concluded that it is possible to make satisfactory quality speech synthesis, and it is possible to present some lisping pathologies.

In future work, changes in articulation during oral cancer will be investigated using oral cancer to use empirical data for speech synthesis instead of physical considerations.

### 6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287. This work was carried out on the Dutch national e-infrastructure with the sup-

port of SURF Cooperative.

### 7. References

- [1] J. Benesty, M. M. Sondhi, Y. Huang, and S. Greenberg, *Springer Handbook of Speech Processing*, 2009, vol. 126, no. 4. [Online]. Available: <http://scitation.aip.org/content/asa/journal/jasa/126/4/10.1121/1.3203918>
- [2] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech and Language*, vol. 36, pp. 260–273, 2016. [Online]. Available: <http://dx.doi.org/10.1016/j.csl.2015.02.003>
- [3] F. Taguchi and T. Kaburagi, "Articulatory-to-speech conversion using bi-directional long short-term memory," [Online]. Available: [https://www.isca-speech.org/archive/Interspeech{\\\_}2018/pdfs/0999.pdf](https://www.isca-speech.org/archive/Interspeech{\_}2018/pdfs/0999.pdf)
- [4] Z. C. Liu, Z. H. Ling, and L. R. Dai, "Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation," *Speech Communication*, vol. 99, no. February, pp. 161–172, 2018. [Online]. Available: <https://doi.org/10.1016/j.specom.2018.02.008>
- [5] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.
- [6] J. A. Gonzalez, L. A. Cheah, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Evaluation of a silent speech interface based on magnetic sensing and deep learning for a phonetically rich vocabulary," *Proceedings of the Annual Conference of the International Speech Communication Association, INTER-SPEECH*, vol. 2017-Augus, pp. 3986–3990, 2017.
- [7] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011.
- [8] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, 2012.
- [9] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech and Language*, 2016.
- [10] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, 2016.
- [11] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," *5th European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997.
- [12] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, 2006.
- [13] A. Graves, "Generating Sequences With Recurrent Neural Networks," aug 2013. [Online]. Available: <http://arxiv.org/abs/1308.0850>
- [14] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," *ICLR: International Conference on Learning Representations*, 2015.
- [15] Z. Wu and S. King, "Improving Trajectory Modelling for DNN-Based Speech Synthesis by Using Stacked Bottleneck Features and Minimum Generation Error Training," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2016.
- [16] R. F. Kubichek, "Mel-cestral Distance Measure for Objective Speech Quality Assessment," *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pp. 125–128, 1993.
- [17] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An Empirical Exploration of Recurrent Network Architectures," *JMLR*, 2015.