Multi-speaker articulatory-to-acoustic speech synthesis with deep neural networks

Bence Halpern[12]*, Rob J. J. H. van Son*[1]*, Michiel W. M. van den Brekel*[12]

[1]NKI-AVL, Amsterdam
[2]ACLC,University of Amsterdam, The Netherlands
b.halpern@nki.nl, r.v.son@nki.nl

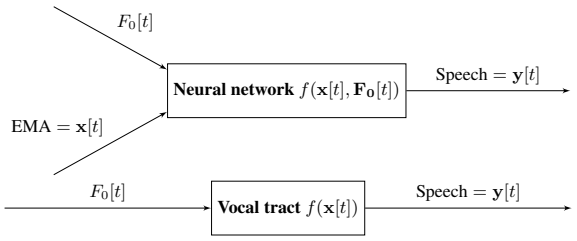

Figure 1: *Black box diagram showing the parallelities of a neural network based model with a source-filter model*

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.

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

1. Introduction

Understading 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 from physical considerations. The model synthetises speech by exciting an autoregressive model with a glottal wave signal, where the model coefficients capture the geometry of the vocal tract via 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. Data-driven methods became of interest also 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 [5], 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 of the speech suffers. It was also possible to obtain satisfactory values for the spectral quality.

In the author's broader research, the aim is to demonstrate how pathologies in articulation are translated to speech. This of interest in many types of pathological speech where the laryngeal functions remains intact, or the deviations in the laryngeal function are known. This way accomodating the articulatory changes to a model is able to aid understanding of the pitch of the voice which has no pathological deviations.

Furthermore, based on the principles of the source-filter model, it is then desirable to construct a data-driven method for multi-speaker articulatory synthesis, which separates the source infomation (excitation and articulation) from the effect (speech).

The main contributions of this paper are,

- a set of benchmarks for multi-speaker articulatory to acoustic synthesis
- estimation on how much data would be needed to construct a better model
- 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. Electromagnetic articulography

Electromagnetic articulography uses sensor coils which are placed on the different articulators in the vocal tract. Using this technology it is possible to record the displacement of the articulators which can then be used to approximate the dynamic area function with a neural network.

The electrodes are placed on the MNGU0 and MOCHA-TIMIT dataset on a total of seven positions.

TODO: Decide whether you want to leave out the lower incisor or jaw ii

2.2. Dataset

The MNGU0 and the MOCHA-TIMIT dataset were used to obtain a total of X samples, which were partitioned to training-validation set. The combined dataset has 2274 speech recordings and electromagnetic articulography data pairs. The dataset contains data from three speakers, two British male and one British female.

Table 1: *Articulatory information recorded in datasets*

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

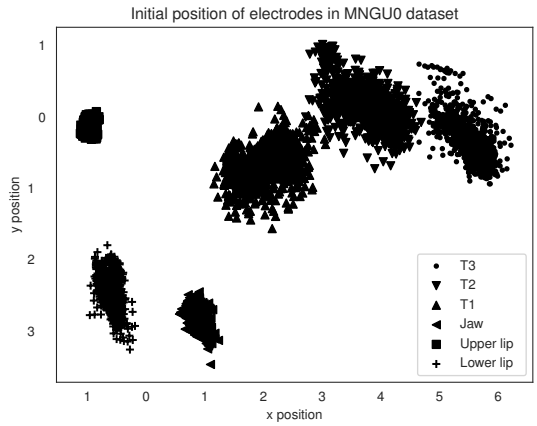

Figure 2: *The visualisation of electrode locations for all samples in the MNGU0 dataset at time point* $t = 0$

2.3. Preprocessing

2.4. 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 amount 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 MOCHA-TIMIT were also downsampled.

For the MNGU0 dataset, NaN values were of EMA occured, this was interpolated linearly.

For the MOCHA-TIMIT dataset, a Savitzky-Golay filter was used to ... (TODO: Savgol filter?)

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.

2.5. Vocoder analysis

The speech was transformed with the PyWORLD vocoder [6] and compressed with the PySPTK toolkit available at `http://github.com/r9y9/pysptk`. The period between consecutive frames was 5 miliseconds. The resulting 40 MFCC and 1 power parameters were used to generatic static and delta parameters, resulting in 82 parameters for the training. As the first step of the MFCC processsing $\alpha = 0.42$ were used a preemphasis coefficient. The PyWORLD vocoder also provides the F_0 and BAP values for synthesis, these were explicitly given for the synthesis.

2.6. Delay and interpolation

Previously [7], the effect of delay on the output signal were investigated. While it has been found that delay is beneficial, the author's choice of function approximators are restricted to acausal models, so it has been decided not to use delay in our final implementation.

2.7. Fundamental frequency

Previously, it has been found beneficial to take the logarithm of the pitch and linearly interpolate. [5] An alternative method also exists with exponential interpolation which is described in [8]. It has been decided that the linear interpolation technique will be used. This makes the F_0 continous, which is important property to have when it is regressed in a function approximation setting.

2.8. Normalisation

Exploratory data analysis indicated that the articulatory trajectories of the datasets were on different scale and bias, so it was normalised on a per speaker basis. The output MFCCs were normalised for each cepstral coefficients.

2.9. Neural network

To approximate the functional relationship, the authors have decided to train a Bidirectional Long-Short Term Memory based neural network [9].

For all BLSTM layers, the CuDNNLSTM implementation have been used to improve efficiency, however this limits the activation function to be stricly $\tanh(\cdot)$

A LSTM is governed by the following equations,

$$\mathbf{i}_t = \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{cf}\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}\mathbf{c}_{t-1} + \mathbf{i}_t \tanh(\mathbf{W}_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c) \quad (3)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{x0}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t + \mathbf{b}_0) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t) \quad (5)$$

2.10. Synthesis

A slighlty different method were undertaken for synthesis. It might seem unconventional, that we don't predidict BAP and F_0, but F_0 is directly related to laryngeal function, so it is at the input side of the actual speech synthesis.

3. Experiment

3.1. Neural network experiments

Three recurrent neural networks architectures were trained based on previous papers tackling similiar problems. aa 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 The ypublication of [3] used a neural networks with... This neural network was trained using RMSProp, with a learning rate of 0.01.

Table 2: *Comparison of different trainng methods*

Architecture	Optimiser	Base learning rate
Liu 2018	SGD	0.01
Taguchi 2018	RMSProp	0.01
Gonzalez 2017	SGD (?)	0.01

For training the mean squared error loss function was used, and for evaluation the Mel cepstral distortion (MCD) have been employed. [10]

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

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,

$$d[t] = x[t] - x[t-1], \tag{6}$$

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

The difference signal then can be thresholded using,

$$d_p[t] = \min(d[t], c) \quad \text{for} \quad d[t] \geq 0, \tag{7}$$

$$d_p[t] = \min(d[t], -c) \quad \text{for} \quad d[t] < 0, \tag{8}$$

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

$$p[t] = \sum_{i=0}^{t} x[i]. \tag{9}$$

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

4. Results and discussion

Table 3: *Comparison of 10-fold CV performance of neural networks*

Architecture	MCD
Liu 2018	5.8 dB
Taguchi 2018	6.2 dB
Gonzalez 2017	6 dB

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.

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

5. Conclusion

6. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287.

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

[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] 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, INTERSPEECH*, vol. 2017-Augus, pp. 3986–3990, 2017.

[6] 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.

[7] 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.

[8] 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.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.

[10] R. F. Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pp. 125–128, 1993.