Towards pathological speech synthesis from articulation

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

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

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 discovering that speech production could be described by the geometry of the vocal tract and the glottal wave. Mathematically, the source-filter model synthetises 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 technique called permanent magnetic articulography by [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. However, it is possible to obtain satisfactory values for the cepstrum.

This indicates, that this technique could be a good candidate for synthetising pathological speech where the pitch of the voice is natural. For example, in the case of tongue cancer speech, the laryngeal function remains intact, meaning the pitch remains natural. Thus, it is proposed that the F_0 could be simply obtained through vocoder analysis and only predict the cepstral values through articulation to model pathologies.

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

The main contributions of this paper are,

- a description of a method for speaker-independent MFCC prediction in Section 2
- a technique to incorporate articulation domain-knowledge into pathological speech synthesis in Section 2.4
- a discussion of the current limitations of this framework in Section 3.5
- an attempt to shed light on what these neural networks might learn in Section 3.4

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 enables us to model articulation. Each dataset contains recordings using this technology, however the articulators recorded are slightly different, meaning particular attention has to be paid to align these datasets. An example of EMA recording locations are shown on Figure 1.

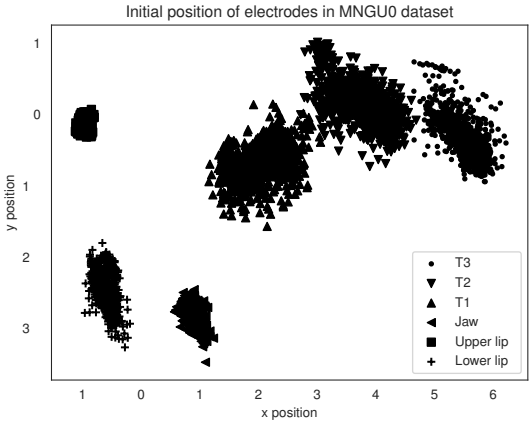
Firstly, it has been decided that seven electrodes will be used for this experiment out of the total eight, Table 1 includes the alignment of the channels that were used. This ensures that each input channel records reasonably similar information, meaning that the channels should have similar variance.

However, this alone does not deal with the speaker-wise variations. In order to ensure that the input represents a general articulation distribution, the articulatory trajectories of the datasets were standardised to have zero mean and unit variance on a per speaker basis. This alleviates some of the deviations, but does not alleviate problems if an electrode falls off during the experiment or if an electrode needs to be changed.

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 these have been excluded. This decision has been made after implementing some simple signal processing algorithms to remove these spikes. The signal to noise ratio in these spiky regions proved to be sufficiently low to affect training.

Previously [9], the effect of delay on the output signal were investigated. It has been found that delay is beneficial for the case of causal models. In Section 2.3, a bidirectional recurrent

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



model will be introduced which is not causal. This means, that theoretically it is not important to use delays with these architectures, as these can use 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

2.1.2. Speech data processing

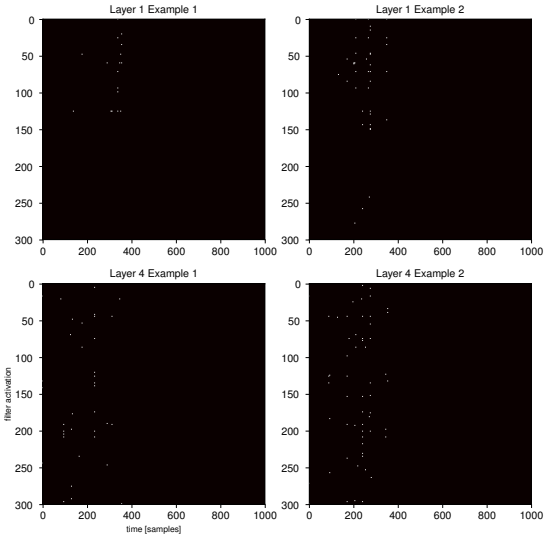
The total combined dataset contains speech from six British male and three British female speakers, with a total of 6117 utterances. The recordings from the microphones all have a sampling frequency of 16kHz. Only the healthy speech has been included from the TORGO dataset. There are 1263 utterances 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 miliseconds. 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

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



pairs contain different amounts of information.

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

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

In this framework, the F_0 is also used for prediction, thus it needs to be processed to be used by the neural network. Previously, it has been found beneficial to take the logarithm of the pitch to obtain a continous F_0 curve in the prediction setting. When the logarithm is not defined, linear interpolatation 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 fed to the vocoder during synthesis time, as these don't contain information about articulation.

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

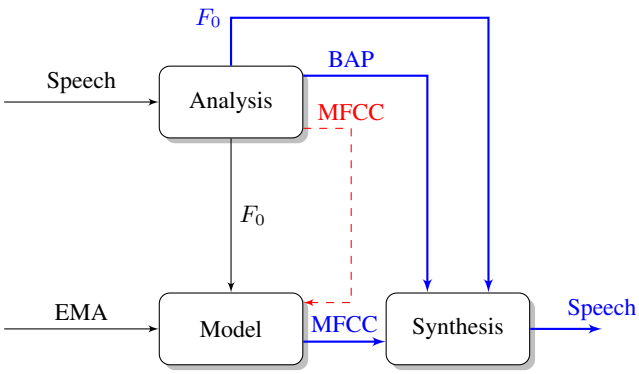


Table 2: *Comparison of preprocessing techniques*

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*
Standardisation	Yes	Yes	Yes
Smoothing	No	Yes	No
Vocoder	STRAIGHT [12]	WORLD	STRAIGHT

*Upsampled to 200 to match analysis rate

2.3. Neural network design

2.3.1. An empirical look at previous architectures

In this paper, a recurrent neural network will be used in order to approximate the articulatory to acoustic mapping. To construct this speaker-independent network, previous speaker-dependent architectures have been studied, to conclude on an appropriate design.

The most pressing problem of recurrent neural networks were the issue of vanishing/exploding gradients. This were somewhat alleviated by the introduction of Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU). This is the reason why [3] used incremental training, and probably the reason why [4] used a learning rate scheduler. However, [6] used Adam optimiser [14] which is known to manage both of these problems with the minor disadvantage of the abscence of good convergence guarantees. The fact that Adam was able to obtain similar results without careful parameter-tuning indicated that it will be an appropriate candidate as an optimiser.

All previous publications on speaker-dependent models reported best performance on bidirectional architectures, however it was unclear whether BLSTM or BGRU architectures are better. Also, [3] resorted to a combination of fully connected and recurrent layers. One fully connected layer was also included in all architectures, which effectively performs linear regression in the end.

In order to determine the best architecture, a pilot study has been performed on all three neural networks which are summarised in Table 5, however all of them were trained with an Adam optimiser, and a learning rate of 0.003, and a batch size of 100 without noise on MNGU0 dataset. The best performing neural network was then trained on the entire dataset.

For training the mean squared error loss function was used,

Table 3: *Performance of speaker-independent articulatory to acoustic neural network for 10-fold cross-validation with 95 % confidence intervals.*

Dataset	Multi-speaker	Single-speaker
Combined result	5.31 $\pm$ 0.09 dB	N/A
MNGU0	5.93 $\pm$ 0.31 dB	4.77 dB
Female MOCHA-TIMIT	5.02 $\pm$ 0.06 dB	5.23 dB
Male MOCHA-TIMIT	**4.06** $\pm$ 0.06 dB	5.83 dB
TORGO Part 1	4.48 $\pm$ 0.03 dB	N/A
TORGO Part 2	4.23 $\pm$ 0.06 dB	N/A
TORGO Part 3	4.81 $\pm$ 0.14 dB	N/A
TORGO Part 4	4.94 $\pm$ 0.09 dB	N/A
TORGO Part 5	4.64 $\pm$ 0.04 dB	N/A
TORGO Part 6	4.70 $\pm$ 0.05 dB	N/A
TORGO Part 7	4.62 $\pm$ 0.04 dB	N/A
TORGO Part 8	15 $\pm$ 0.86 dB	N/A
TORGO Part 9	4.63 $\pm$ 0.11 dB	N/A
TORGO Part 10	4.85 $\pm$ 0.12 dB	N/A

Table 4: *Held out validation for different architectures on MNGU0*

Author	MCD
Gonzalez	4.77 dB
Taguchi	7.28 dB
Liu	4.84 dB

and for evaluation the Mel cepstral distortion (MCD) have been employed. [16]

For the speaker-independent experiments, ten fold cross-validation was performed to estimate the out-ouf-sample generalisation capability of the neural networks.

2.4. How to construct pathological speech?

Using this framework, the problem of making pathological speech can be traded for the problem of making pathological articulation and feeding pathological articulation through the neural network.

1. This is most easily done in a data-driven fashion, by recording healthy and pathological articulation and creating a pathology-dependent mapping in the articulatory space. Because it is a low dimensional space, this is a much easier problem than learning the same mapping in the cepstral space.
2. Another approach is to perform a signal processing in the articulatory space which reflects some kind of physiological change.
   (a) For example, in tongue cancer, articulation of the tongue is impeded. This could be modelled by restricting the values of the tongue articulatory recordings by zeroing, as due to the normalisation, zeroing will mean a resting tongue state.
   (b) One can also model decrease in velocity via discrete time difference equations, to demonstrate impeded control.

3. Results and discussion

3.1. Pilot study

The pilot study results are summarised in Table 4.

Table 5: *Comparison of different trainng methods*

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

* with decay after Epoch 11
** from author communication

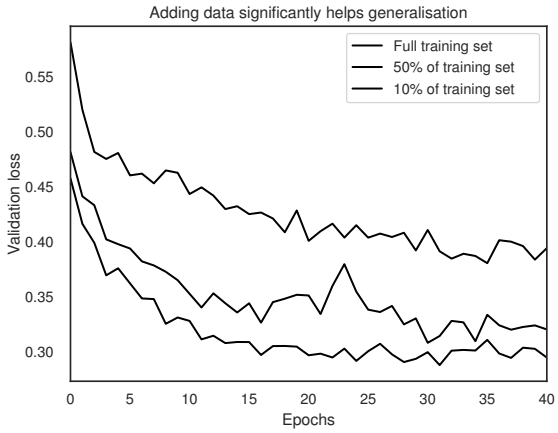


Figure 4: *Partial data retraining shows that adding more data would decrease loss*

Based on our training, it seems clear that the GRU architecture was superior to an LSTM architecture in our case, when used with an Adam optimiser. There is no general consesus whether GRU or LSTM is better for particular datasets. [17]

3.2. Prediction of MFCC values

The prediction results for the MFCC values are summarised in Table 3. Results were all in similar range as previously reported values for speaker-dependent datasets, and in our framework the speaker-independent architectures clearly performed better than the speaker-dependent architectures on the MOCHA-TIMIT datasets.

3.3. Learning curves (TODO)

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 againts the amount of training data included in number of samples. This fit can be seen on Figure **??**. 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.

3.4. What do these neural networks learn?

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 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.

3.5. The current limitations of the synthesis

According to our observations, the quality is bounded more by the quality of the vocoder, than the synthesis itself. In terms of root mean square error, it has been found that difference between the vocoder resynthesised speech and the predicted speech has a mean squared error of 11.15. The mean squared error between analysis-resynthesis and the vocoder is 80.32, meaning that future improvements should focus on better vocoding rather than better acoustic mapping.

3.6. Pathological speech examples

Some synthetised pathological speech examples can be found on the webpage of the author, see `http://karkirowle.github.io/paper1`.

Informal discussions with speech language pathologist indeed confirmed that some of these synthetised samples resemble dysarthic or lisping speech, but these simple heuristics usually don't incorporate enough knowledge about a particulator pathology is to show it consistently.

4. 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.

It can be concluded that it is possible to make speech that resembles pathological speech, however further work need to be done on improving speech quality and creating models which consistently show a certain pathology.

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.

5. 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. This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

6. 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] 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, INTERSPEECH*, 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-cepstral 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.