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

This paper presents a technique to synthetise speech that is pathological on the articulation level. The technique presented considers a vocoder combined with a speaker-independent articulatory to acoustic neural network using electromagnetic articulography recordings using the three largest articulatory datasets. A speech language pathologist indicated that the synthetised samples reflect some qualities of disordered speech. A visualisation technique is also described to shed light on what these neural networks learn.

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] [2]. Recently, deep learning methods became popular to understand articulation. These methods use a measurement tool, called electromagnetic articulography (EMA) to obtain articulation data [3] [4] [5] along with recurrent neural networks, which are neural networks that are able to deal with the sequential nature of data [6]. 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 [7], which gave intelligeble 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.

Pathological speech synthesis could have many potential applications, for example, improving pathological speech detection using synthetic data. It could be also used as a clinical tool, in patient counselling to prepare them for post-operative speech quality. A deeper understanding of the articulatory to speech relationship could also offer improvements in speech therapy tools.

In this paper, a technique is described which combines healthy speech from the three largest articulatory datasets, MNGU0 [8], MOCHA-TIMIT and TORGO [9] 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

The research code is also available as a Github repository online. [10]

2. Method

2.1. Dataset preprocessing

Preprocessing technique of previous publications are summarised in Table 2.

2.1.1. Electrode preprocessing

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. Seven electrodes were 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. These are then standardised on a per speaker basis. These alleviate some of the speaker-wise 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 artifacts, these have been excluded. The signal to noise ratio in these spiky regions was low enough to affect training.

Previously [11], 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 model will be introduced which is not causal, meaning there is no need for delays.

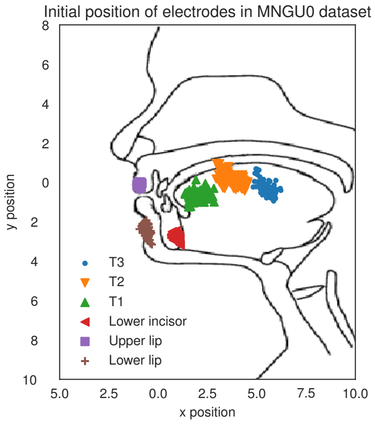
2.1.2. Speech data processing

The total dataset contains speech from six British male and three British female speakers, with a total of 6117 utterances, approximately 10 hours of recorded speech. The recordings from the microphones all have a sampling frequency of 16kHz.

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 300 samples from the MNGU0 dataset at their initial position.*



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 [12] and compressed with the PySPTK toolkit [13]. 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.

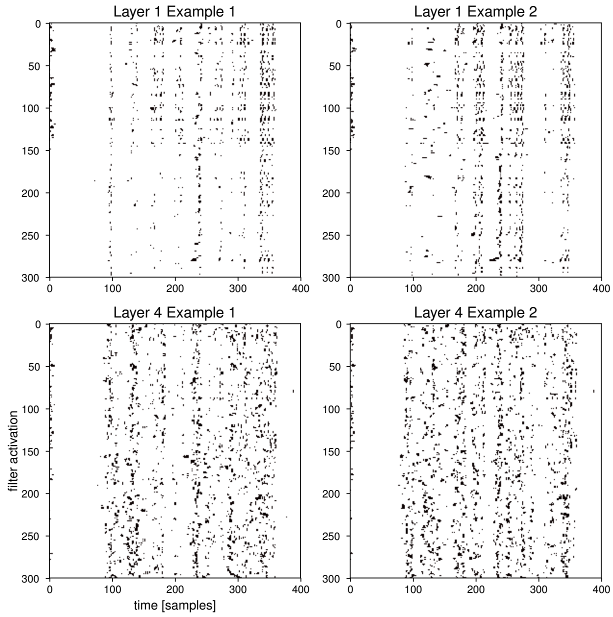
2.1.3. Sampling

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 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.

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



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. [7] An alternative method with exponential interpolation is described in [14].

2.2. Synthesis setup

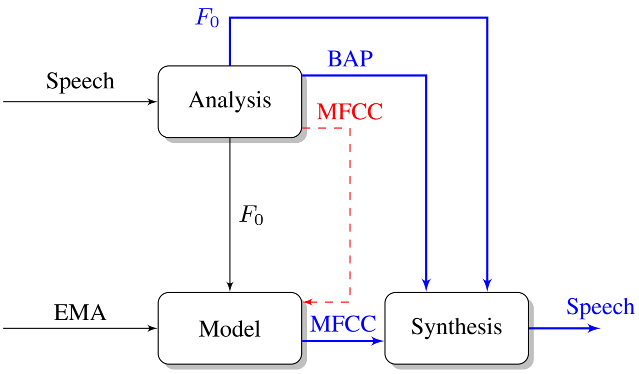
The setup for inference and training can be seen in Figure 3. In the training setup, only the MFCCs are given. The pitch and band aperiodicities are directly fed to the vocoder during synthesis time, as these don't contain information about articulation.

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 [15]	WORLD	STRAIGHT

*Upsampled to 200 Hz to match analysis rate

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



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 [4] used incremental training, and probably the reason why [5] used a learning rate scheduler. However, [7] used Adam optimiser [16] 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.

Previous publications on speaker-dependent models reported best performance on bidirectional architectures, however it was unclear whether BLSTM or BGRU architectures are better. Also, [4] 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 4, 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, and for evaluation the Mel cepstral distortion (MCD) have been employed. [18]

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. Articulatory space modification

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.

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

Dataset	Multi-speaker MCD	Single-speaker MCD
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: *Comparison of different training methods used in previous publications with the results of the pilot study using held-out validation. The method described in the paper of Gonzalez performed best.*

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 [17]	No	Yes	Yes
Maximum epochs	32	N/A	100
Batch size	N/A	8	100
Incremental training	No	Yes	Yes
MCD* **	4.84 dB	7.28 dB	4.77 dB

* with decay after Epoch 11
** from author communication
*** results of our training with Adam optimiser

In order to make the articulation pathological, two methods were employed. The first method fixes the position of the tongue tip in the initial position. The second method decreases the velocity of the tongue via thresholding discrete time differences, and reconstructing the position with performing a cumulative sum.

3. Results and discussion

3.1. Pilot study

The pilot study results are summarised in Table 2.

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. [19]

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

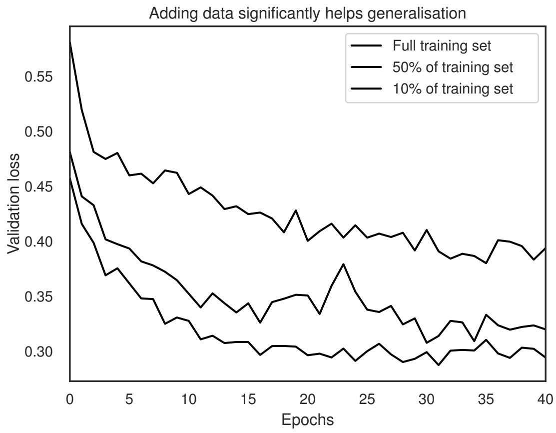


Figure 4: *Partial data retraining shows that adding more data would decrease loss*

the speaker-independent architectures clearly performed better than the speaker-dependent architectures on the MOCHA-TIMIT datasets.

3.3. Learning curves

The neural network was retrained using incrementally more training data in twenty percent batches. The mean squared error (MSE) 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, and it has been found that there is a statistically significant improvement with each addition of the training data, meaning more data improves the model.

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 mean square error (MSE), it has been found that difference between the vocoder resynthesised speech and the predicted speech has an MSE of 11. The MSE between analysis-resynthesis and the vocoder is 80. That means that 88% of the performance loss is due to the vocoder analysis-resynthesis. This indicates 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 [20]. Informal discussions with speech language pathologist indeed confirmed that some of these synthetised samples resemble dysarthic or disordered speech, but these simple heuristics usually don't incorporate enough knowledge about a particulator pathology to show it consistently.

This confirms that that the framework can be used as platform to implement a pathological articulation by reflecting the physiological changes in the articulatory space. 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

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 from impaired articulation. Further work needs to be done on improving speech quality and creating models which consistently show a certain pathology.

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. The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Malm, Sweden), which contributes to the existing infrastructure for quality of life research.

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] G. Fant, "The source filter concept in voice production," *Quarterly Progress and Status Report*, 1981.

[3] 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

[4] 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

[5] 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

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

[7] 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.

[8] 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.

[9] 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.

[10] B. Halpern, "Articulatory vocoder." [Online]. Available: http://github.com/karkirowle/vocoder-clean

[11] 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.

[12] 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.

[13] "PySPTK toolkit." [Online]. Available: http://github.com/r9y9/pysptk

[14] 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.

[15] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, 2006.

[16] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic gradient descent," *ICLR: International Conference on Learning Representations*, 2015.

[17] 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.

[18] 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.

[19] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An Empirical Exploration of Recurrent Network Architectures," *JMLR*, 2015.

[20] B. Halpern, "Speech Examples." [Online]. Available: http://karkirowle.github.io/paper1