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

Articulatory Text-to-Speech Synthesis using the Digital Waveguide Mesh driven by a Deep Neural Network

Amelia J. Gully[1]*, Takenori Yoshimura*[2]*, Damian T. Murphy*[1]*, Kei Hashimoto*[2]*, Yoshihiko Nankaku*[2]*, and Keiichi Tokuda*[2]

[1]Audio Lab, Department of Electronics, University of York, UK
[2]Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Japan
amelia.gully@york.ac.uk

Abstract

Following recent advances in direct modeling of the speech waveform using a deep neural network, we propose a novel method that directly estimates a physical model of the vocal tract from the speech waveform, rather than magnetic resonance imaging data. This provides a clear relationship between the model and the size and shape of the vocal tract, offering considerable flexibility in terms of speech characteristics such as age and gender. Initial tests indicate that despite a highly simplified physical model, intelligible synthesized speech is obtained. This illustrates the potential of the combined technique for the control of physical models in general, and hence the generation of more natural-sounding synthetic speech.

Index Terms: speech synthesis, digital waveguide mesh, deep neural network

1. Introduction

Recent studies on speech synthesis have been heavily focused upon statistical parametric techniques [1, 2]. This is a powerful text-to-speech method, which produces increasingly natural synthetic speech. One of the advantages of the method is that characteristics of speech such as speaker individuality, speaking style, and emotion, can be controlled by transforming the model parameters [3, 4]. However, the relationship between these parameters and actual vocal tract shapes is unclear. This makes it difficult to synthesize speech with different speaker characteristics such as age and gender. Although the transformation can be estimated using data-driven techniques, reliable estimation requires a large amount of speech data containing the desired characteristics. A more flexible approach is therefore desirable.

Another speech synthesis technique, known as physical modeling, aims to reproduce the physics of the vocal system. If sufficiently detailed, such a model will inherently produce natural-sounding synthetic speech. Physical modeling has begun to receive significant attention (see, e.g. [5, 6]) as increasing computational power makes detailed three-dimensional (3D) vocal tract modeling more feasible. True physical models offer great potential for natural output and intuitive gestural control. Furthermore, changes to speaker characteristics may be implemented simply, by changing the size, shape, and other parameters of the modeled vocal tract. However, physical models require detailed vocal tract shape information, which is typically obtained from magnetic resonance imaging (MRI) data. MRI data is costly and inconvenient to obtain, so large datasets representative of multiple speakers are not available. In addition, the long data acquisition times, equipment noise, and supine posture required, lead to unnatural articulations [7] and an inability to capture simultaneous high quality audio recordings. For these reasons, a method of obtaining reliable vocal tract shape information by other means is required.

In this paper, we propose a novel method to directly estimate a physical model from speech waveforms rather than MRI data. Although a previous attempt has been made to obtain vocal tract shape data from speech recordings using a genetic algorithm [8], this was not linked to a text-to-speech front end system. The proposed method optimizes a physical model based on a framework of neural networks representing the relation between speech waveforms and linguistic features derived from text. Deep neural networks (DNNs) have been shown to be effective for modeling speech waveforms in [9, 10]. Inspired by them, we utilize a DNN with a specially designed output layer as an estimator of a physical model. Note that as far as we know, this is the first attempt to combine statistical parametric and physical modeling approaches. Detailed 3D vocal tract models such as [5, 6] require hundreds of thousands of variables. This results in computations that are thousands of times slower than real-time, and presents too many degrees of freedom for use with DNN methods at present. Therefore, the proposed method makes use of a simpler, two-dimensional (2D) digital waveguide mesh (DWM) model based on [11]. The DWM is a technique used to simulate acoustic propagation that has previously been successfully used to generate synthetic speech [11, 12]. The proposed technique allows model parameters to be inferred directly, with no reduction in dimensionality required. Although this is a highly simplified model, intelligible speech is obtained, illustrating the potential of the proposed method for the control of physical models.

The remainder of this paper is laid out as follows: Section 2 introduces the digital waveguide mesh, and Section 3 describes the method of controlling the DWM model using a DNN. Section 4 describes the experimental procedure and discusses the results, and conclusions and avenues of further investigation are presented in Section 5.

2. The Digital Waveguide Mesh

The digital waveguide mesh (DWM) is a time-domain algorithm for simulating acoustic wave propagation within a domain. It is equivalent to the finite difference time domain (FDTD) method under certain conditions [13]. The domain to be simulated is approximated by a regular grid of scattering junctions connected by unit waveguides. Although a domain may be simulated in any number of dimensions using the DWM technique, we use a 2D rectilinear mesh for this proof-of-concept study. The construction of such a mesh is illustrated in Figure 1. For every temporal sample, the sound pressure (or other acoustic variable) must be updated for every scattering

http://dx.doi.org/10.21437/Interspeech.2017-900

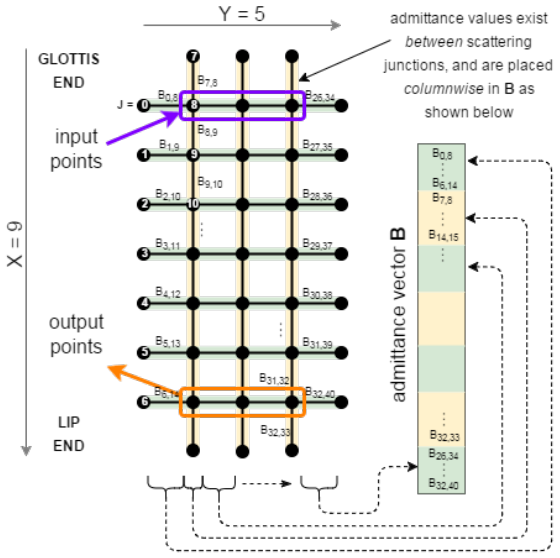
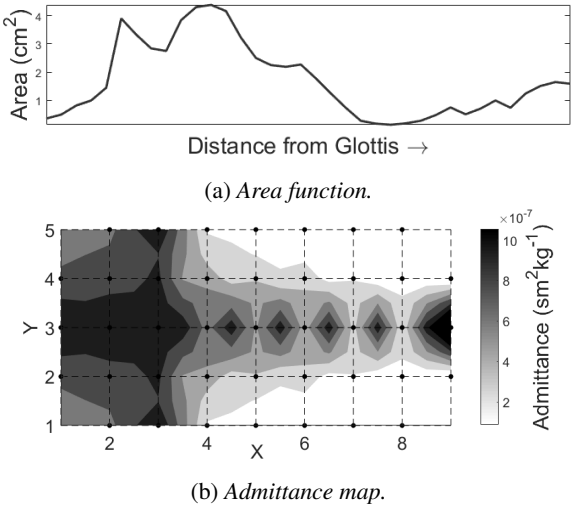


Figure 1: *A two-dimensional digital waveguide mesh, illustrating connections between scattering junctions, input and output locations, and the construction of admittance vector* $\boldsymbol{B}$.



(a) *Area function.*

(b) *Admittance map.*

Figure 2: *Construction of a 2D raised-cosine admittance map from a vocal tract area function, for an English vowel /i/.*

junction in the simulation domain [14]:

$$p_J(n) = \frac{2\sum_{I\in J_{nei}} B_{J,I}\, p_I(n-1)}{\sum_{I\in J_{nei}} B_{J,I}} - p_J(n-2) \quad (1)$$

where $p_J(n)$ is the acoustic pressure at scattering junction J at time step n, $B_{J,I}$ is the admittance of the waveguide connecting junction J to neighbouring junction I, and J_{nei} is the set of scattering junctions immediately adjacent to junction J. For a 2D rectilinear mesh, $|J_{nei}| = 4$. At the edge of the simulation domain, energy is reflected according to boundary conditions. For the simplest DWM boundaries, as described in [14], with a boundary junction connected to the mesh by a single waveguide, boundary pressures are updated as:

$$p_{J_b}(n) = (1-r)p_{I_b}(n-1) - r p_{J_b}(n-2) \quad (2)$$

where $p_{J_b}(n)$ is the acoustic pressure at boundary junction J_b at time step n, I_b denotes the single neighbouring scattering junction connected to the boundary junction (see Figure 1), and r is the reflection coefficient at the boundary. The 2D DWM vocal tract model uses reflection coefficient values of 0.92, -0.90, and 0.97 at glottis, lips, and walls, respectively [11].

The admittance term $B_{J,I}$ in (1) indicates scattering within the mesh, and hence the effective mesh shape, may be altered by changing the acoustic admittance $B_{J,I}$ - or its reciprocal, acoustic impedance $Z_{J,I}$ - in the waveguides connecting any junctions J and I [14]. Therefore, the scattering behaviour of a fixed-size mesh is completely described by its admittance values and boundary reflection coefficients. In [11], vocal tract shape is specified by mapping a vocal tract area function to raised-cosine impedance contours across a fixed-size rectangular mesh, with minimum impedance at the mesh centre and area-function dependent impedance at the mesh edges. Impedance is inversely proportional to the cross-sectional area of the vocal tract, so the maximum impedance values (and conversely, the minimum admittance values) represent the narrowest constrictions. An example admittance map generated using this procedure is presented as a filled contour plot in Figure 2. The area function in Figure 2a illustrates the varying cross-sectional area of the vocal tract along its length, which is then interpolated to a suitable length and converted to admittance values as described above to produce Figure 2b. This technique creates a central channel of high admittance, with the effective vocal tract width governed by the lower admittance values at the tract edges.

The resolution of the DWM depends on the temporal sampling frequency of the output signal. The relationship between mesh spacing, d, and the temporal sampling frequency of the output signal, f_s, is given by

$$f_s = \frac{c\sqrt{2}}{d} \quad (3)$$

for a 2D DWM mesh [14], where c is the speed of sound. The mesh spacing d therefore must be selected under the constraint of the size of the human vocal tract. Values given in [15] for an adult male vocal tract place the length at between 15.88cm and 18.25cm. The mesh must also be an odd number of junctions wide to allow for a central channel of maximum admittance. In this study, to obtain an acceptable trade-off between the resolution of the mesh and the computational cost, $f_s = 24$kHz is selected with $c = 350$m/s, giving $d \approx 2$cm. We therefore use a mesh size of 9×5 junctions, corresponding to a physical size of approximately 16cm $\times$ 8cm. While this resolution is sufficient as proof-of-concept, more accurate and natural-sounding 2D DWM vocal tract models will necessarily require higher grid resolution and therefore more parameters for estimation. In addition, the DWM method has a usable bandwidth of $f_s/4$ [14], and in practice even lower due to dispersion error [14], so the synthetic speech output in this study is valid to less than 6kHz: this is sufficient for intelligibility, but a higher f_s, and hence additional parameters, are necessary for natural-sounding output.

3. DNN-driven DWM

Inspired by recent work [9, 10] that directly models speech waveforms using cepstral coefficients based on a DNN framework, we propose a method to derive a physical model based on a 2D DWM from speech waveforms using a DNN with a specially designed output layer. In the proposed method, the

cosine-mapping constraint used in [11] is removed, taking advantage of the 2D structure by permitting the generation of asymmetrical admittance maps such as those required for nasal sounds.

Let us define a vector, $\boldsymbol{B}$, that concatenates the admittances throughout the mesh as illustrated in Figure 1. A 2D DWM vocal tract model is completely described by the admittance vector $\boldsymbol{B}$ and a set of reflection coefficients. In the proposed method, the admittance vector at frame t, $\boldsymbol{B}_t$, is represented as an intermediate output of a DNN whose input is a linguistic feature vector, $\boldsymbol{l}_t$, introducing a physically-informed relationship between the text and the output speech waveform[1]:

$$\boldsymbol{B}_t = \mathcal{H}(\boldsymbol{l}_t) \tag{4}$$

where $\mathcal{H}$ denotes the nonlinear function represented by a DNN. The objective function to be maximized with respect to $\boldsymbol{B}_t$ is the log likelihood computed using the following Gaussian distribution:

$$P(\boldsymbol{s}_t \mid \boldsymbol{B}_t) = \mathcal{N}(\boldsymbol{s}_t; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{B}_t}) \tag{5}$$

where $\boldsymbol{s}_t \in \mathbb{R}^M$ is a discrete-time widowed speech signal based on a zero-mean stationary Gaussian process [16], $\boldsymbol{0} \in \mathbb{R}^M$ is the zero vector, and $\boldsymbol{\Sigma}_{\boldsymbol{B}_t} \in \mathbb{R}^{M \times M}$ is the covariance matrix that can be decomposed as follows:

$$\boldsymbol{\Sigma}_{\boldsymbol{B}_t} = \boldsymbol{H}_{\boldsymbol{B}_t}^{\mathsf{T}} \boldsymbol{H}_{\boldsymbol{B}_t} \tag{6}$$

where

$$\boldsymbol{H}_{\boldsymbol{B}_t} = \begin{bmatrix} h_t(0) & & & 0 \\ \vdots & h_t(0) & & \\ h_t(N-1) & \vdots & \ddots & \\ & h_t(N-1) & \vdots & h_t(0) \\ & & \ddots & \vdots \\ 0 & & & h_t(N-1) \end{bmatrix} \tag{7}$$

and $h_t(n)$ is the impulse response of the 2D DWM, and N denotes the impulse response length. The impulse response $h_t(n)$ is calculated in a recursive manner based upon (1):

$$h_t(n) = \frac{1}{|J_{out}|} \sum_{J \in J_{out}} p_J(n) \tag{8}$$

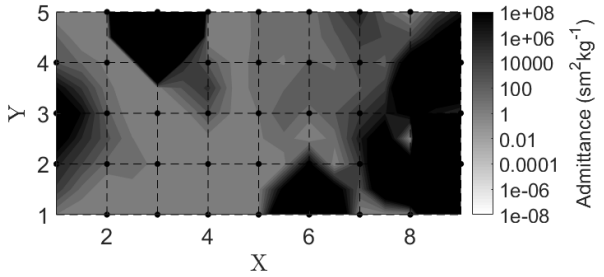
where J_{out} is a set of scattering junctions near the lips (see Figure 1). The initial conditions of the pressures at $n = 0$ are

$$p_J(0) = \begin{cases} 1 \,/\, |J_{in}|, & \text{if } J \in J_{in} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

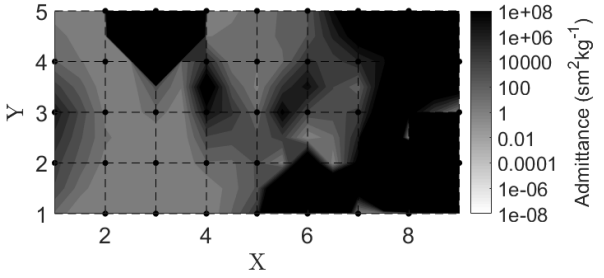
where J_{in} is a set of scattering junctions near the glottis. In this framework, speaker characteristics can be controlled by using adaptation techniques, e.g., feeding speaker codes [17] to the DNN $\mathcal{H}$, used in the standard DNN-based speech synthesis.

In order to generate a smooth speech trajectory, the dynamic behaviour of the mesh must be captured by a model. However, since the present study focuses only on a proof-of-concept of the proposed system, the DNN is optimized in a frame-by-frame manner rather than sequence-to-sequence one. In the synthesis stage, a speech waveform is produced using the DWM method with the values of $\boldsymbol{B}$ predicted by the DNN, or the impulse response of the estimated DWM system, and a suitable excitation signal.

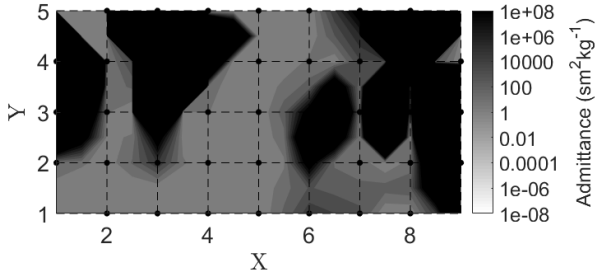
[1] Although reflection coefficients can also be modeled by a DNN, they are held constant in this study for simplicity.



(a) *DNN-generated admittance map for /i/.*



(b) *DNN-generated admittance map for /n/.*



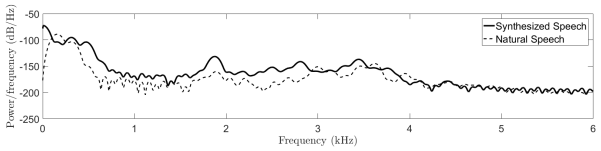
(c) *DNN-generated admittance map for /k/.*

Figure 3: *DNN-generated admittance maps.*

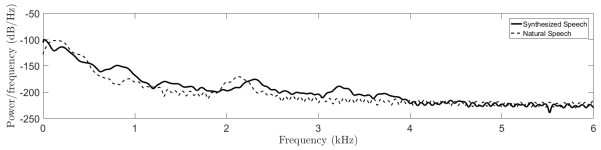
4. Experiments

4.1. Experimental conditions

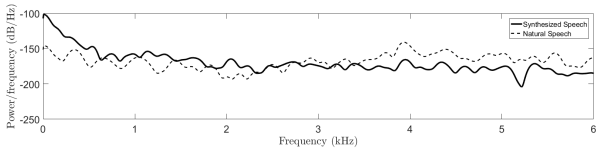
The experiment used 503 Japanese sentences uttered by a male speaker. The contents of the data were the same as the B-set of the ATR phonetically balanced Japanese speech database [18]. A subset of 450 utterances were used for training, with the remaining 53 utterances used for testing. The speech signals were downsampled at 24kHz and split into 25ms frames windowed with a Blackman window, with a 5ms shift between frames. The impulse response length N was set to be 700. The network output was the windowed waveform in 600 (= 24kHz $\times$ 25ms) points, given the 52-dimensional admittance vector. A 411-dimensional linguistic feature vector, consisting of 408 linguistic features including binary features and numerical features for contexts and three duration features, was used as the network input. The architecture of the network was 3-hidden-layer with 256 units per layer. The parameters of the network were randomly initialized, and were optimized using an Adam optimizer [19] with dropout [20]. The network used sigmoid activations. The fundamental frequency and the duration of synthetic speech were derived from natural speech at the synthesis stage.



(a) *Welch PSDs for /i/.*



(b) *Welch PSDs for /n/.*



(c) *Welch PSDs for /k/.*

Figure 4: *Welch power spectral density (PSD) graphs for synthesized and natural speech.*

Table 1: *Errors in formant frequencies for synthesized /i/ and /n/ compared to natural speech*

Phoneme	**Formant Error (%)**		
	F1	F2	F3
/i/	+100%	-0.6%	-6.25%
/n/	+34.8%	-13.1%	+8.56%

4.2. Experimental results

The 53 test utterances were used to assess the performance of the simulation technique. Synthesized versions of these sentences were found to be intelligible and example natural and synthesized speech are included as accompanying multimedia files. Examples of admittance maps, reconstructed from the DNN-generated vector $\boldsymbol{B}$ for individual frames, are provided in Figure 3. Dark areas represent high admittance, while lighter areas represent low admittance, indicating a constriction. Note that the admittance values exist in the waveguides, and are interpolated across the rest of the image for illustration only.

It is immediately apparent from Figure 3 that the admittance maps for different phonemes show several similarities, including a region of high admittance at $x > 8$. As the lip output is taken at $x = 8$, the area beyond this represents lip radiation. Similarly, at $x < 2$ the maps show some indication of a region of higher admittance centered on $y = 3$, indicative of the subglottal tract. Admittance in the vocal tract region $x = 2$ to $x = 8$ is predictably higher than outside this area, and shows some phoneme-dependent variation in both dimensions.

Figure 4 presents power spectral density (PSD) curves for the same frames as Figure 3, providing an estimate of the spectrum. The PSD plots were calculated using a 4096-point FFT with a 300-sample window and 250-sample overlap. Figure 4a compares the PSDs of natural and synthesized /i/ vowels. Despite some error, quantified in Table 1, the formant locations are sufficiently close to those of natural speech that the vowel /i/ is clearly identifiable in the synthesized speech. Figure 4b presents the spectrum for the phoneme /n/. Again, this follows the spectrum of natural speech quite closely, with formant errors given in Table 1. Nevertheless, the phoneme /n/ is occasionally mistaken for /m/ in the output speech. This may be because the synthesized speech features an antiformant at approximately 1.2 kHz, which is characteristic of an /m/ [21]. It is of interest to note that an acceptable nasal consonant is produced, and Figure 3b illustrates a region of higher admittance at $x = 7$, $y = 1.5$ which may be acting as a side branch to facilitate this. Finally, Figure 4c illustrates the spectra of natural and synthesized versions of the phoneme /k/. Although a broadband spectrum is preserved by the synthesized version, there are a large number of deviations from the spectrum of natural speech, and notably reduced energy above 4 kHz compared with natural speech. As a result, /k/ is identifiable as a plosive in the synthesized speech, but not necessarily as /k/. The admittance map for /k/ in Figure 3c contains some indications of a /k/-like constriction at $x = 4$ to $x = 5$ but this is clearly insufficient for accurate reproduction of /k/.

The observations made above are largely generalizable to the respective phoneme categories across all test sentences: vowels are quite well estimated by the model, nasal and lateral consonants are less well estimated, and turbulent consonants are least well estimated. This may relate to the physical modeling paradigm used, as in general, physical models of vowel reproduction have been highly successful, but consonants have not yet been synthesized reliably. Further investigation into reliable consonant reproduction techniques is therefore expected to improve the model.

All of the $\boldsymbol{B}$ vectors generated with the proposed technique display a much larger range of admittance values than the raised-cosine mapping technique [11], far exceeding the maximum characteristic acoustic admittance for a tube the size of a vocal tract. In addition, many of the DNN-generated admittance maps, including all of those illustrated in Figure 3, have a region of high admittance centered on $x = 3$, $y = 4.5$ which has no apparent physical equivalent. It is evident that, in order to obtain physically-meaningful output, some constraints will be required upon the generated parameters, for example by including the positions and admittances of articulators like the teeth, hard palate and velum. Nevertheless, the synthesized speech has a spectrum similar to natural speech and is sufficient for intelligibility. With further refinement it is anticipated that the quality of the synthesized speech will improve.

5. Conclusions

This study has illustrated the potential of a combined statistical parametric and physical modeling approach for the generation of intelligible synthetic speech. An important priority for future study is a subjective assessment of the intelligibility of the synthesized speech produced using the proposed method. Future work will also involve controlling the speaker characteristics of synthetic speech by feeding metadata to the proposed model, and constraining the parameter generation based on real vocal tract geometries.

6. Acknowledgements

This research was partly funded by Core Research for Evolutionary Science and Technology (CREST) from the Japan Science and Technology Agency (JST). Author A. Gully was supported by the Japan Society for the Promotion of Science (JSPS) Summer Program Fellowship.

7. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *Proceedings of ICASSP 2013*, pp. 7962–7966, 2013.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," *Proceedings of ICASSP 2001*, vol. 2, pp. 805–808, 2001.

[4] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *Proceedings of IEICE Transactions on Information and Systems*, vol. E88–D, no. 11, pp. 2484–2491, 2005.

[5] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall, "Influence of lips on the production of vowels based on finite element simulations and experiments," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2852–2859, May. 2016.

[6] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-different time-domain method," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3724–3738, Dec. 2010.

[7] O. Engwall, "Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA," *Proceedings of Interspeech 2000*, pp. 17–20, 2000.

[8] C. Cooper, D. Murphy, D. Howard, and A. Tyrrell, "Singing synthesis with an evolved physical model," *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 4, pp. 1454–1461, Jul. 2006.

[9] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," *Proceedings of ICASSP*, pp. 4215–4219, 2015.

[10] ——, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," *Proceedings of ICASSP*, pp. 5640–5644, 2016.

[11] J. Mullen, D. M. Howard, and D. T. Murphy, "Real-time dynamic articulations in the 2D waveguide mesh vocal tract model," *IEEE Trans. Audio Speech Language Process.*, vol. 15, no. 2, pp. 577–585, Feb. 2007.

[12] M. Speed, D. Murphy, and D. Howard, "Modeling the vocal tract transfer function using a 3D digital waveguide mesh," *IEEE Trans. Audio Speech Language Process.*, vol. 22, no. 2, pp. 453–464, Feb. 2014.

[13] M. Karjalainen and C. Erkut, "Digital waveguides versus finite difference structures: equivalence and mixed modeling," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 7, pp. 978–989, Jun. 2004.

[14] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley, "Acoustic modeling using the digital waveguide mesh," *IEEE Signal Process. Mag.*, vol. 24, no. 2, pp. 55–66, Mar. 2007.

[15] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 100, no. 1, pp. 537–554, Jul. 1996.

[16] K. Dzhaparidze, "Parameter estimation and hypothesis testing in spectral analysis of stationary time series," *Springer-Verlag*, 1986.

[17] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," *Proceedings of Interspeech 2016*, pp. 2278–2282, 2016.

[18] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357–363, 1990.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1998.