AN END-TO-END NETWORK TO SYNTHESIZE INTONATION USING A GENERALIZED COMMAND RESPONSE MODEL

François Marelli[1,2,3], *Bastian Schnell*[1,2], *Hervé Bourlard*[1,2], *Thierry Dutoit*[3], *Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[3]Université de Mons (UMONS), Institut NUMEDIART, Belgium
{francois.marelli, bastian.schnell, phil.garner}@idiap.ch

ABSTRACT

The generalized command response (GCR) model represents intonation as a superposition of muscle responses to spike command signals. We have previously shown that the spikes can be predicted by a two-stage system, consisting of a recurrent neural network and a post-processing procedure, but the responses themselves were fixed dictionary atoms. We propose an end-to-end neural architecture that replaces the dictionary atoms with trainable second-order recurrent elements analogous to recursive filters. We demonstrate gradient stability under modest conditions, and show that the system can be trained by imposing temporal sparsity constraints. Subjective listening tests demonstrate that the system can synthesize intonation with high naturalness, comparable to state-of-the-art acoustic models, and retains the physiological plausibility of the GCR model.

***Index Terms*—** Neural Networks, Digital IIR Filters, Speech Synthesis, Prosody Modelling, Fujisaki Model

1. INTRODUCTION

Intonation is a prosodic feature of speech that carries non-linguistic information such as emotion and emphasis. It is crucial to correctly model it in speech-to-speech translation systems that intend to transfer paralinguistics between languages, as a distorted pitch may change the meaning of a sentence. A good model of intonation is also crucial for the transfer and the synthesis of emotion. In previous work with colleagues [1], we investigated a physiologically plausible intonation (F_0) model based on the Command-Response (CR) model of Fujisaki [2]. This Generalized CR (GCR) model represents the intonation contour as the response of muscle models to spike command signals.

We studied how a Recurrent Neural Network (RNN) can generate the command signals of the GCR model to generate intonation by emulating a spiking neural network [3]. The RNN predicts the position and amplitude of the command spikes for a given text, which are filtered by the GCR muscle models to generate the pitch contour.

However, this model has three limitations. First, the muscles models that filter the spikes are not trainable. Their parameters are imposed before training the system and may not be optimal. Second, the trained RNN is not able to generate true spikes. The amplitude and position of the command signals are split into two separate channels, requiring post-processing to recover spikes. This post-processing operation prevents gradient back-propagation from the pitch curve. Third, the system cannot be used to predict the phrase component used to reconstruct the log of F_0 (LF_0).

In this paper we propose to overcome the aforementioned limitations by training an End-to-End (E2E) neural network to generate LF_0. This system includes trainable muscle models and the generation of phrase components, both without post-processing. To build this system, we take advantage of the existing model and replace the post-processing steps by trainable muscle models, as shown in Figure 1. The model is a source-filter model which differs from the commonly used speech production models by generating only LF_0 and using temporal static but trained filters. Additionally, the source and filter prediction models are trained together in an E2E fashion. The filter itself is similar to a second-order Infinite Impulse Response (IIR) synapse, which has been analysed before [4, 5]. We extend that analysis by investigating and solving the gradient explosion issues that can prevent the convergence of recurrent units [6], when applying back-propagation through time [7].

Of course, any modern TTS systems can predict LF_0 (e.g. [8]). The proposed system differs in the sense that it retains the physiologically inspired behaviour of the GCR model, by enforcing spiky command signals and muscle model filtering. This will allow us to conserve its transfer capability and physiological interpretation.

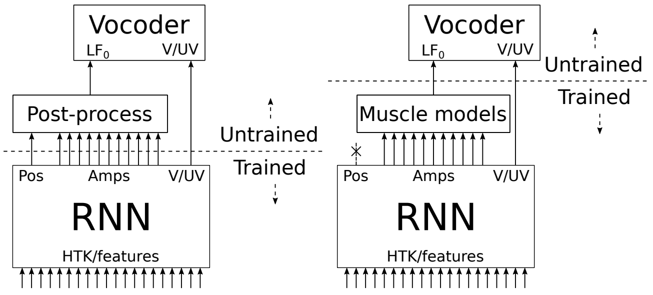

Fig. 1: *GCR intonation synthesis systems. Left: previous model. Right: proposed E2E model. The post-processing step is replaced by trainable muscle models to enable training directly on the LF_0 curve in an E2E fashion.*

In the following sections, we explain how muscle models can be embedded in an E2E neural network, and we derive the associated stability conditions that prevent gradient explosion during training. We then describe the architecture of the proposed system and how it improves on the previous implementation by adding phrase component modelling ability while retaining the behaviour of a GCR model. Finally, the obtained system is compared to a strong baseline through objective and subjective scores.

2. E2E INTONATION SYNTHESIS

2.1. Muscle models

In the GCR model, the output filters approximate muscle activation. Different models for muscle response are investigated in [9]. Even though previous research [10, 11] has shown that higher order systems can improve intonation modelling performance, we use a second-order Spring-Damper-Mass (SDM) muscle model in this work. This choice is consistent with Fujisaki's assumption that intonation is generated by second order systems [12]. Moreover, it is possible to obtain more complex responses by combining multiple second order models, so that this choice is not restrictive.

The generic discrete transfer function of an SDM system is

$$y_{(k)} = Gx_{(k)} + \alpha y_{(k-1)} + \beta y_{(k-2)} \tag{1}$$

with x the command signal, y the response, G the gain, α and β the recurrence coefficients of the model, and k the discrete time step. It is equivalent to the equation of a second-order linear all-pole digital filter (Figure 2).

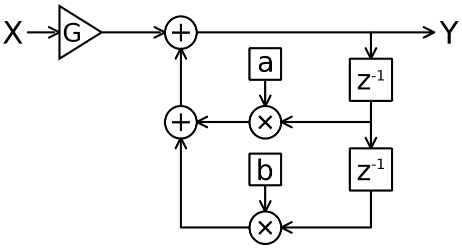

Fig. 2: *Second-order linear all-pole digital filter.*

Given that the SDM system is second-order time recurrent, it would make sense to model it by an RNN. Implementations such as LSTMs or GRUs can indeed learn second order recurrence. Nevertheless their behaviour is strongly non-linear and they are overparametrized for an SDM model. Therefore we propose a simpler implementation of linear second-order recurrent units based on (1).

We expect that gradient descent optimization can be used for the extraction of the filter parameters, as iterative methods have proven to be efficient for digital filter design problems [13]. However, gradient explosion issues are inherent to recurrent units training [6]. To study the convergence properties of the model, we derive the gradient equations of (1) using back-propagation through time [7]. The obtained expressions are given in (2 – 4), with K_n defined in (5).

$$\frac{\partial y_{(k)}}{\partial \alpha} = \sum_{n=0}^{k-1} \left[y_{(k-1-n)} \cdot K_n \right] \tag{2}$$

$$\frac{\partial y_{(k)}}{\partial \beta} = \sum_{n=0}^{k-2} \left[y_{(k-2-n)} \cdot K_n \right] \tag{3}$$

$$\frac{\partial y_{(k)}}{\partial x_{(k-n)}} = GK_n \tag{4}$$

$$K_n = \begin{cases} \alpha K_{n-1} + \beta K_{n-2} & \text{if } n > 0 \\ 1 & \text{if } n = 0 \\ 0 & \text{if } n < 0 \end{cases} \tag{5}$$

The gradient explosion is caused by the recurrence in K_n. The analysis of (5) reveals that a sufficient condition to prevent gradient explosion is that all the poles of the model have a modulus lower than one (i.e. the modelled system stability implies the stability of the gradient). We can express the model in polar notation if we assume that we are only targeting complex conjugate poles

$$y_{(k)} = Gx_{(k)} + 2\rho\cos(\phi)\; y_{(k-1)} - \rho^2\; y_{(k-2)} \tag{6}$$

with ρ the modulus and ϕ the phase of the poles. This assumption is not a limitation for muscle modelling, as muscle responses tend to behave as under-damped or critically damped systems [14]. This, in turn, allows to express the stability constraint as (7), which can be imposed by using a compressing transformation [5] as sigmoid (8). The cosine of the pole angle can also be transformed to use the whole parameter space by defining c in (9).

$$|\rho| \leq 1 \tag{7}$$

$$\rho = \sigma(p) \tag{8}$$

$$\cos(\phi) = \tanh(c) \tag{9}$$

The reformulation of the system

$$y_{(k)} = Gx_{(k)} + 2\;\sigma(p)\;\tanh(c)\; y_{(k-1)} - \sigma^2(p)\; y_{(k-2)} \tag{10}$$

is used to implement trainable muscle models integrated in neural networks, and prevents gradient explosion issues.

2.2. Network architecture

We integrate the trainable muscle models into the existing system, which we will refer to as *atom model*, by replacing the post-processing step with a new muscle models layer (Figure 1). In contrast to the existing system, the command signals are not split into separate position and amplitude channels. The former position output signal is therefore removed from the system. The amplitude outputs become the command inputs for the new muscle models layer. The Voiced/Unvoiced (V/UV) prediction output remains unchanged as it is independent from the post-processing.

The new muscle models layer has one recurrent unit φ_i per GCR muscle (Figure 3). Each output is multiplied by a gain that normalizes the L2 norm of the impulse response of the filters (the linearity of the models implies that the gains can be applied on the input or the output signals equally). The normalization allows an easier interpretation of the command signals, and is consistent with the atom model that also uses normalized muscle responses. Moreover, normalization gains help balance the impact of the gradients on the different muscle models. Trainable gains could create dominant filters that would receive higher gradients, resulting in uneven convergence speed across the muscle models.

The gain that normalizes the impulse response of a filter depends only on its poles. This relationship can be computed analytically. However, as the exact expressions are difficult and computationally heavy, we use a numerical approximation.

The final LF_0 contour is given by summing up the normalized filter responses and adding a trainable bias, which compensates for the non-zero mean of LF_0. We believe that a trainable bias, which can depend on a global speaker ID input, is more flexible than training on a normalized LF_0 target. Compared to the atom model, the bias compensates the main shift of the phrase component.

Instead of starting from a random initialization point we can start with a pre-trained atom model. Nevertheless, the muscle models

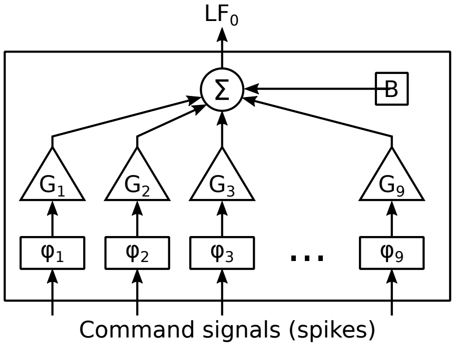

Fig. 3: *Muscle models layer. Each muscle model φ_i is associated to a normalization gain G_i. All the outputs are summed and the phrase bias B is added to reconstruct LF_0.*

used in the former architecture are gamma-shaped atoms whose impulse response is

$$f_{K,\theta}(t) = \frac{1}{\theta^K\,\Gamma(K)}\,t^{K-1}\,e^{-t/\theta} \qquad \text{for } t \geq 0 \qquad (11)$$

with K and θ the shape and scale of the atom and t the continuous time variable. To use the previous muscle parameters a relationship between gamma atoms and discrete linear filters is required. Setting $K = 2$ and using an impulse-invariant transformation [15] to discretize (11) relates the pole modulus of discrete linear filters to the scale of gamma atoms

$$\rho = \exp\left(\frac{-T_s}{\theta}\right) \qquad (12)$$

with T_s the sampling period. Since the filters are normalized, the gain relationship can be ignored.

3. EXPERIMENTS

In running experiments, we want to validate the assumption that the proposed E2E system can reproduce the behaviour of the GCR model, and generate the phrase contribution. We are also testing the hypothesis that the fixed muscle parameters used in the atom model are not optimal for intonation generation by a neural network, and that trainable models will converge to values giving better performance matching the quality of a strong baseline.

3.1. Experimental setup

We use a subset of the 2008 Blizzard Challenge speech database [16] (about 5.7 hours) to test our model. 5% (17 minutes) are set aside for test and evaluation set respectively. The phone sequences are extracted from text and force-aligned by context-independent HMMs using Festival [17]. The inputs are 425 text-derived binary and numerical features normalized to [0.01, 0.99].

The WORLD vocoder [18] (D4C edition [19]) is used for the extraction of LF_0, 60-dimensional MGC, and one Band Aperiodicity (BAP) at 5 ms frame step. LF_0 is interpolated before training and a binary V/UV flag is used to capture voicing information. For the baseline system dynamic features are computed as well. We use a set of nine muscles for the GCR, initialized with gamma scales $\theta \in \{0.03, 0.045, ..., 0.15\}$, approximating the ones used in previous research [11].

3.2. Network Topologies and Training

We use a state-of-the-art acoustic baseline system consisting of two feed-forward RELU layers of 1024 nodes, three bi-directional LSTMs with 512 nodes each, and a linear output layer with 187 nodes (features + Δ + $\Delta\Delta$). The model is trained for 35 epochs with a learning rate (LR) of 0.002.

The E2E system is initialized with a pre-trained atom model, which uses the same topology and training as described in our previous work [3]. At first the E2E model is trained for 50 epochs (LR of 0.001), without the phrase bias, on LF_0 from which the phrase contribution is removed. It is then trained with phrase bias on non-normalized LF_0 for another 50 epochs (LR of 0.0006). Note that training the system without initialization converges but deteriorates the reconstruction performance.

The loss is computed by summing the Mean-Squared-Error (MSE) of LF_0 on the voiced frames and the MSE of the V/UV output weighted by 0.3. In order to generate spiky command signals that fit the behaviour of an GCR model, we apply a temporal L1 constraint [20] on the outputs of the atom model weighted by 0.3. Without this penalization the generated command signals are not sparse and cannot be assimilated to GCR spikes.

The LR is reduced using a plateau scheduler with a patience of five, a relative threshold of 0.001 and a factor of 0.3. All the networks are trained using Adam [21] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

3.3. Objective Scores

The performance of the models is objectively measured by the Root-Mean-Squared-Error (RMSE) of F_0 on all the target voiced frames and the V/UV error rate. Table 1 shows that the proposed E2E system significantly improves the performance on the atom model, and the obtained objective performance closely matches the score of the strong baseline system. The signals generated by the trained system are plotted in Figure 4, which shows the ability of the model to generate spiky signals and to synthesize the phrase component. For the tested E2E system the muscle parameters converge to the values $\theta \in \{0.01, 0.019, 0.02, 0.021, 0.037, 0.042, 0.145\}$ which are different from the initial ones. This validates our hypothesis that the fixed values used in the atom model are not optimal for this task. Whether the parameters always converge to the same values remains future research.

Table 1: *Objective scores*

Model	F$_0$ RMSE	V/UV error
Baseline	21.3 Hz	10.4 %
Atom	28.8 Hz	14.9 %
E2E	22.3 Hz	10.7 %

3.4. Subjective Measurements

We synthesize the samples with the WORLD vocoder using the original durations, MGCs, and BAPs; only the impact of LF_0 is measured. For the baseline LF_0 is improved by maximum likelihood parameter generation [22]. The naturalness of the synthesized speech is evaluated through a MUSHRA test conducted using the BeaqleJS toolkit [23]. It compares our model (E2E) to the previous system

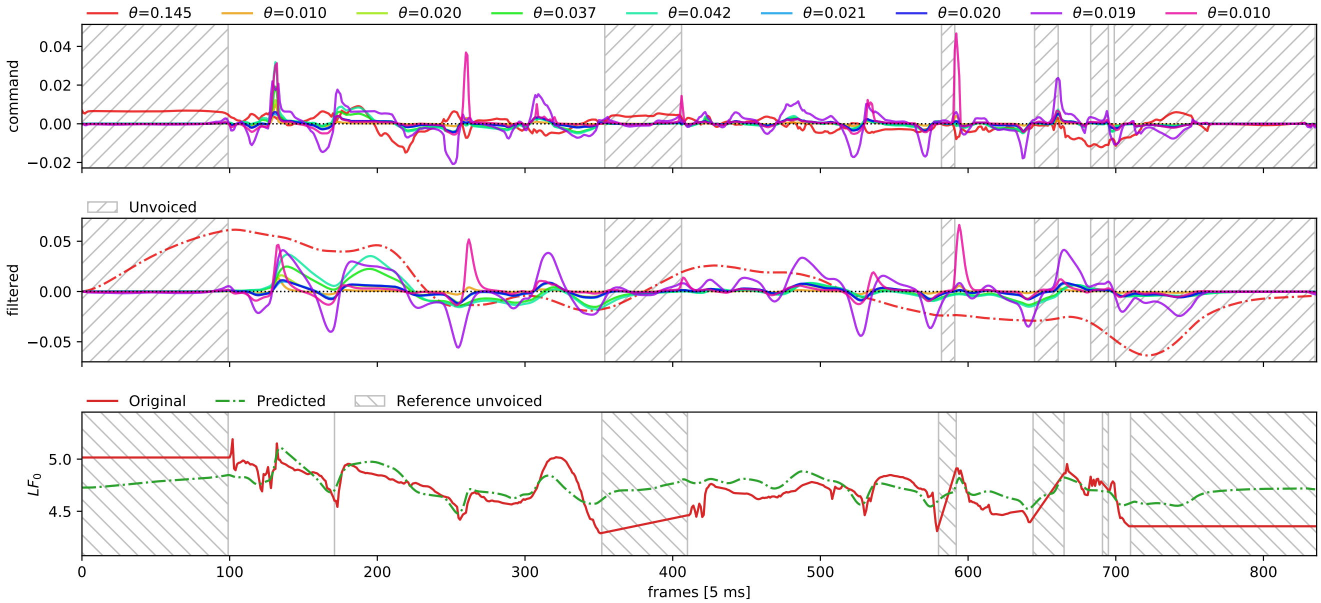

Fig. 4: *Signals generated by the E2E model. From top to bottom:* ***1:*** *Command signals generated by the RNN, which can be assimilated the spikes of a GCR.* ***2:*** *Muscle model responses, where the slow phrase component is clearly visible (dash-dotted line).* ***3:*** LF_0 *reconstruction (dash-dotted) and original (solid). The striped regions represent unvoiced frames.*

(atom), an anchor (only the phrase component), the baseline, and the speech synthesized using the original LF_0 (reference). 20 random samples from the evaluation set have been selected for evaluation by 42 fluent English speakers. Each of them was asked to rate 5 random samples of this subset on a scale from 0 to 100, focusing on prosody only and ignoring minor artefacts.

A two-tailed paired t-test on the individual ratings of the baseline and E2E system gives $p = 0.746 > 0.1$, proving that the proposed model achieves the same perceived quality as a strong baseline system. The quality of the atom model is worse, with the p-value of E2E–Atom being $p = 0.0002 < 0.1$. This is expected because the evaluation set contains multi-phrase samples, while the atom model can correctly model single-phrase samples only. Thus, the proposed model achieves a better performance on a more complex task.

4. CONCLUSIONS

We have shown that an E2E neural network with embedded trainable second-order linear all-pole digital filters can generate natural sounding intonation, provided that suitable stability conditions are imposed. The temporal L1 constraint allows to produce spiky command signals to drive muscle responses, thus reproducing the behaviour of a GCR model. Taking advantage of the flexibility of E2E networks, the system can also generate the phrase component in LF_0. The objective and subjective results of this model closely match those of a strong baseline.

We noticed a clustering effect in the muscle models, which can be further investigated in future work. Furthermore, the obtained muscle parameters and the shape of the command signals allow the psycho-linguistic analysis of the model behaviour. The system capabilities to produce and transfer affect also need to be investigated for exploitation in emotional speech synthesis.

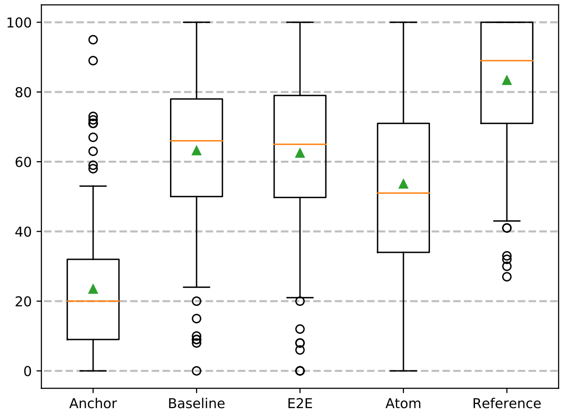

Fig. 5: *Subjective score of MUSHRA intonation test. Medians as lines, sample averages as triangles, outliers as circles.*

Reproducibility: The source code used to train the model and measure its performance will be made available.

5. ACKNOWLEDGEMENTS

This work has been conducted with the support of the Swiss NSF under grant number 165545: Multilingual Affective Speech Synthesis (MASS).
`http://p3.snf.ch/Project-165545`

6. REFERENCES

[1] Pierre-Edouard Honnet, Branislav Gerazov, and Philip N Garner, "Atom decomposition-based intonation modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4744–4748.

[2] Hiroya Fujisaki, Sumio Ohno, and Changfu Wang, "A command-response model for F0 contour generation in multilingual speech synthesis," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[3] Bastian Schnell and Philip N Garner, "A neural model to predict parameters for a generalized command response model of intonation," *Proc. Interspeech 2018*, pp. 3147–3151, 2018.

[4] Andrew D Back and Ah Chung Tsoi, "FIR and IIR synapses, a new neural network architecture for time series modeling," *Neural Computation*, vol. 3, no. 3, pp. 375–385, 1991.

[5] Paolo Campolucci and Francesco Piazza, "Intrinsic stability-control method for recursive filters and neural networks," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 8, pp. 797–802, 2000.

[6] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, "Understanding the exploding gradient problem," *CoRR, abs/1211.5063*, 2012.

[7] Paul J Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[8] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[9] Branislav Gerazov and Philip N. Garner, "An investigation of muscle models for physiologically based intonation modelling," in *Telecommunications Forum Telfor (TELFOR), 2015 23rd*. IEEE, 2015, pp. 468–471.

[10] Santitham Prom-On, Yi Xu, and Bundit Thipakorn, "Modeling tone and intonation in mandarin and english as a process of target approximation," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[11] Branislav Gerazov, Pierre-Edouard Honnet, Aleksandar Gjoreski, and Philip N Garner, "Weighted correlation based atom decomposition intonation modelling," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] Hiroya Fujisaki and Keikichi Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[13] MN Howell and TJ Gordon, "Continuous action reinforcement learning automata and their application to adaptive digital filter design," *Engineering Applications of Artificial Intelligence*, vol. 14, no. 5, pp. 549–561, 2001.

[14] Branislav Gerazov and Philip N. Garner, "An agonist-antagonist pitch production model," in *International Conference on Speech and Computer*. Springer, 2016, pp. 84–91.

[15] F Gardner, "A transformation for digital simulation of analog filters," *IEEE transactions on communications*, vol. 34, no. 7, pp. 676–680, 1986.

[16] Vasilis Karaiskos, Simon King, Robert AJ Clark, and Catherine Mayo, "The blizzard challenge 2008," in *Proc. Blizzard Challenge Workshop, Brisbane, Australia*, 2008.

[17] Alan W Black, Paul Taylor, Richard Caley, and Rob Clark, "The festival speech synthesis system," 1999.

[18] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[19] Masanori Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[20] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 3, pp. 1315–1318.

[23] Sebastian Kraft and Udo Zölzer, "BeaqleJS: HTML5 and Javascript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.