Blind Bandwidth Extension of Speech based on LPCNet

Konstantin Schmidt
International Audio Laboratories
Friedrich-Alexander-University (FAU)
Erlangen, Germany
konstantin.schmidt@audiolabs-erlangen.de

Bernd Edler
International Audio Laboratories
Friedrich-Alexander-University (FAU)
Erlangen, Germany
bernd.edler@audiolabs-erlangen.de

***Abstract*—A blind bandwidth extension is presented which improves the perceived quality of 4 kHz speech by artificially extending the speech's frequency range to 8 kHz. Based on the source-filter model of the human speech production, the speech signal is decomposed into spectral envelope and excitation signal and each of them is extrapolated separately. With this decomposition, good perceptual quality can be achieved while keeping the computational complexity low. The focus of this work is in the generation of an excitation signal with and autoregressive model that calculates a distribution for each audio sample conditioned on previous samples. This is achieved with a deep neural network following the architecture of LPCNet [1].**

A listening test shows that it significantly improves the perceived quality of bandlimited speech. The system has an algorithmic delay of 30 ms and can be applied in state-of-the-art speech and audio codecs.

***Index Terms*—bandwidth extension, artificial bandwidth expansion, speech enhancement, audio super resolution, speech super resolution**

I. Introduction

Although today there are standardised speech codecs that are able to code almost fullband speech and audio signals [2], [3], todays most used codec for mobile speech communication is still AMR-NB [4] which encodes frequencies from 200 Hz to 3400 Hz (usually named *narrowband*, NB) only. Blind bandwidth extension (BBWE) - also known as artificial bandwidth expansion or audio super resolution - is a simple approach to improve the perceived quality of NB coded speech signals. It extends the frequency range of the speech signal to 7 kHz (*wideband*, WB) or beyond without transmitting information from the encoder. A BBWE can be added to the decoder toolchain and no adaption of the transmission network needs to be done. Thus it can serve as an intermediate solution to improve perceived audio quality and intelligibility [5]–[7] until better codecs will be implemented in the network. BBWE has a long tradition in the audio signal processing community [8] but recently there has been increased interest in BBWEs based on DNNs. These systems not only increase the perceived quality of speech but also can improve word error rates of automated speech recognition systems [9]. End-to-end training of convolutional or recurrent deep neural networks led to a significant improvement compared to approaches based e.g. on HMMs [10].

State-of-the-art DNN based BBWEs can be divided into three categories, two of them are DNNs that output the whole waveform either by predicting the probability density function of speech samples as in [11], [12] or by implicitly modelling the probability density function by a generative adversarial network (GAN) as in [9]. The system in [11] is special because it uses bitstream parameters of NB coded speech to condition the network. Here the network acts as a decoder that implicitly does bandwidth extension.

The third category are BBWEs that - motivated by the source-filter model of the human speech production - model only the spectral envelope of the missing frequency range, see e.g. [7], [13]–[15]. The excitation signal of such systems is generated by nonlinearities or by a simple copy-operation in frequency domain. The advantage of such systems is that spectral magnitudes are much easier to model and the DNNs for such tasks can be much smaller with lower computational complexity.

All networks used in the above systems usually contain convolutional layers as in WaveNet [16], recurrent layers - usually gated recurrent units (GRUs) - as in [17] or a mixture of both [7].

In case the speech signal is modelled directly in time domain, $\mu-$law shaping is often used [16]. This makes the probability density function more tractable and introduces some basic psychoacoustics. If spectral magnitudes are modelled, they are shaped by logarithmic- or other loudness-derived functions [7] for the same reason.

The presented BBWE belongs to the first category since it outputs the whole WB speech signal in time-domain. It benefits from an internal decomposition into vocal tract envelope and excitation signal as explained in detail in the next section. The main motivation behind this is to keep the computational complexity low compared to e.g. [12]. The next section describes the system and the following Sec. III describes the experimental setup and parameterisation. Finally the system will be evaluated objectively and by a listening test in Sec. IV followed by a conclusion in Sec. V.

978-9-0827-9705-3

II. LPCNet based Blind Bandwidth Extension of Speech

As already observed by Jax et. al. [18] the upper 4 kHz band of speech sampled at 16 kHz (meaning the signal containing frequencies from 4 to 8 kHz) contains elements that can be predicted from the lower 4 kHz band as well as elements that can not be predicted. In their work they calculated mutual information between the upper band spectral envelope and features extracted on lower NB speech signal. From this mutual information they derived a lower bound of log spectral distortion of the upper band spectral envelope extrapolated from the NB features. Their observed lower bound may be the reason why state of the art BBWEs are able to synthesise the missing signal with good perceived quality but are not able to reproduce the precise formant structure - the upper band spectral envelope of BBWE-speech is a rather smooth version of the original envelope. This holds for voiced as well as unvoiced speech. Although this mismatch can be very large, the perceived degradation is often low - particularly when the mismatch happens in very high frequency ranges. This is the motivation of the proposed BBWE that doesn't rely on a very large network that tries to predict all dynamics of the original speech signal.

The network architecture used for this work is based on LPCNet [1], a generative network that achieves similar performance as WaveNet while having low computational complexity. The complexity measured in their paper enables real-time implementation on a AMD A1100 (Cortex-A57) processor while WaveNet is not capable of real-time processing on state of the art GPUs.

LPCNet is designed for either speech synthesis [1] or speech coding [19]. It is based on the decomposition of speech into an excitation signal and an envelope, similar to speech codecs [2], [4]. The envelope in LPCNet is represented by Linear Prediction Coefficients (LPCs). With the LPCs given, a recurrent network models the spectrally flat excitation signal, which is much easier to model than the original speech signal. This network models the distribution of the excitation signal as a product of conditional probabilities:

$$p\left(\boldsymbol{x}\right)=\prod_{t=1}^{T}p\left(x_t|x_1,\ldots,x_{t-1},\hat{x}_1,\ldots,\hat{x}_t\right), \tag{1}$$

where x_t is the output WB speech excitation sample at time t and $\hat{x}_t$ is the NB speech sample at time t. Each audio sample is therefore conditioned on previous samples, and as a result, the network predicts samples that are fed back into the network.

The LPCNet based BBWE is presented in Fig. 1. The main differences to the original paper are:

- the LPCs representing the NB spectral envelope are extrapolated with a simple network to a WB spectral envelope. This is done in the green block in Fig. 1 by a separate GRU
- an additional recurrent path feeding the NB excitation signal - upsampled to 16 kHz - from the input NB speech to the GRU

The presented BBWE has three main building blocks, each containing a DNN. Two of them (shown in orange and green) operate on frames, the third (shown in blue) operates on audio samples. Dashed arrows in the figure represent data flow on frame rate, solid arrows represent data flow on sample rate.

In contrast to the original paper, where the LPCs are derived from spectral magnitudes, they are calculated here by autocorrelation of windowed time-domain frames followed by Levinson recursion [20]. The LPCs are then transformed to Line Spectral Frequency (LSF) coefficients [21]. LSFs are a bijective transformation of LPCs with several advantages: First they are less sensitive to noise disturbances and an ordered set of LSFs with minimum distance between the coefficients will always guarantee a stable LPC filter. Second, the spectral envelope at a particular frequency only depends on one of the LSFs. These properties make them suitable for being extrapolated to a set representing a WB envelope. This is done in the green block by a separate gated recurrent unit (GRU) with two layers. These LSFs are transformed back to the LPC domain for further processing.

Envelope representations together with fundamental frequency (pitch) are the most important elements in the human speech production process. With these two parameters given, it is already possible to build a rudimentary speech synthesiser [22]. In LPCNet, the extrapolated LPCs and pitch values are input to the second network operating on frames. This network comprises two convolutional layers with residual connection followed by two fully connected layers to generate a compact representation of the input values. The 1X3 filter kernels of this network (orange) are calculated on past and future frames and are the main source of algorithmic delay. The output of this network is used as conditioning parameter to the sample-rate network generating the WB speech excitation signal (shown in blue in Fig. 1). Other inputs to the blue network are the LPC excitation signal calculated in the *LPC residual* block by the well know LPC recursion [20], the delayed output of the network before and after sampling, and the prediction (explained below). This network consists of two layers of GRUs, followed by a dual fully connected layer, ending in a softmax activation.

The *compute prediction* block in Fig. 1 computes the WB speech prediction s_t on previous output speech samples $y_{t-(1\ldots M)}$:

$$s_t=\sum_{i=0}^{M}a_i y_{t-i}, \tag{2}$$

which is added to the excitations signal x_t to form the output signal:

$$y_t=x_t+s_t, \tag{3}$$

where a_i are the LPCs (without the first coefficient always being 1). This is just the same IIR filter operation to shape the flat excitation signal x as in well known LPC applications [20].

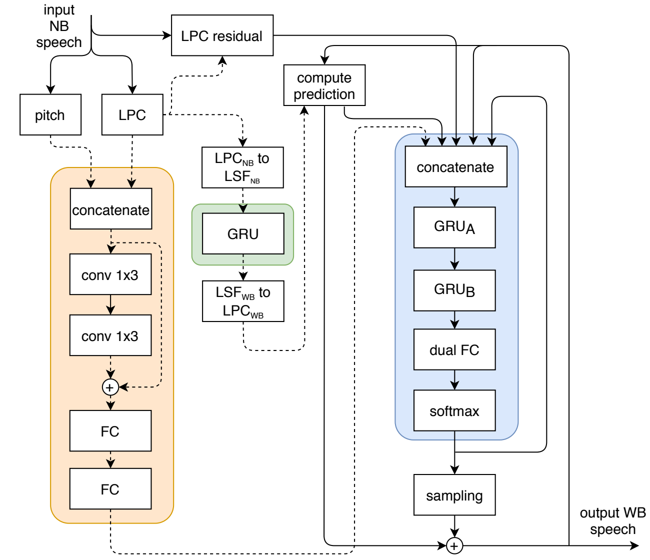


Fig. 1: Proposed system containing three main building blocks, each containing a DNN. Two of them (shown in orange and green) operate on frames, the third (shown in blue) operates on audio samples. Input is time-domain NB speech, output is time-domain WB speech.

A. Speech Data Preprocessing

Initial experiments feeding NB speech upsampled to 16 kHz into the network caused an annoying emphasis in the output signal close to 4 kHz. This is caused by relative sharp stop of spectral content at 4 kHz of the NB speech. To circumvent this artefact, the upper band containing zero energy is filled with either high-pass filtered noise or with high-pass filtered NB signal distorted by some nonlinear function. We used the latter approach with simply squaring the NB signal while retaining the sign followed by a linear-phase high-pass filter with 4 kHz transition frequency [15].

B. Data Representations and Loss

As already observed in [16] training a generative network with mean square error loss on floating-point speech data will not produce good results. Instead it is common to predict samples as classes and use sparse categorical cross-entropy as loss. When using this loss, the speech data needs to be quantised and the number of quantisation levels has influence in the architecture and complexity of the generating network. Pre-shaping the time-domain data going into the network as well as the targets with the well known $\mu-$law function will make the probability density function of the quantised data more Gaussian and thus easier to model. The $\mu-$law function is already used in the first ever standardised audio codec [23] for this reason. Furthermore the quantisation noise will be shaped to be less audible. To keep the quantisation noise in high frequencies less audible a first-order pre-emphasis filter is applied as in the original paper.

The output of the network is converted to a time-domain signal by sampling from a multinomial probability distribution parameterised by the network prediction instead of simply taking the argument with maximum likelihood. This multinomial distribution is parameterised by the output of the dual fully connected layer (dual FC) which is a variant of a fully connected layer that helps determining whether an output value falls within a certain range ($\mu-$law quantised intervals). Since it is beneficial to have a more noisy output for unvoiced speech and a cleaner output for voiced speech the distribution temperature is parameterised by the pitch correlation estimated on the NB speech. This is done as in the original paper [1].

C. Sparse Matrices

Initial large matrix sizes in the GRUs causing high computational complexity are successively decreased by sparsification. Sparse weight matrices have a large number of entries with a value of exactly zero and thus do not need to be calculated. The network is trained with initially large matrices and during training sub-matrices are identified with a norm of their elements falling below a threshold. These sub-matrices are set to zero and the loss is no more backpropagated through them.

D. Reference Systems

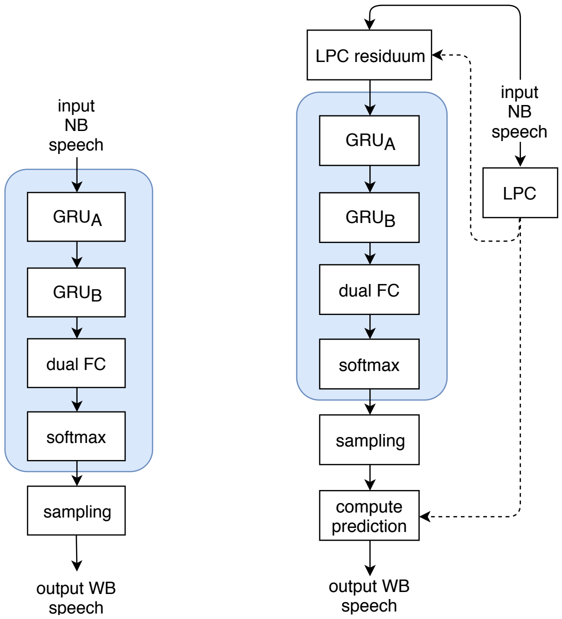
To justify the proposed decomposition of speech into spectral envelope and excitation signal for bandwidth extension together with it's complex structure, we compare it to two simpler systems shown in Fig. 2 (a) and (b) as well as a previously published system [7]. System (a) is basically the sample-rate network of the proposed system (Fig. 1) without the recurrent paths. The input is NB speech is fed to the same GRU as in the proposed system and trained to predict WB speech. No LPC residual is calculated here. The GRU needs to model the distribution of WB speech with much higher dynamics than the LPC residual. The evaluation Sec. IV will show that the GRU is not capable of improving NB speech significantly.

Alternatively the GRU can be used to extrapolate the NB speech excitation signal as shown in (b). Here the GRU needs to model a similar distribution as in the proposed system with the difference that the GRU has no information about the prediction from the LPC-shaping. This results in a mismatch between the excitation signal and the LPCs causing noisy artefacts as shown in the evaluation Sec. IV.

Furthermore the proposed system will be compared to the previously published system [7]. This system uses a mixture of convolutional layers and recurrent layers to predict the energy of frequency bands of about 1 Bark width. The excitation signal is generated - similar to spectral folding - by copying the NB spectrum to the missing upper band (a common practice in speech and audio coding [3]). These frequency band energies contain similar information as LPC envelopes so the DNN in this system performs similar to the GRU extrapolating the LPC envelope see Fig. 1.

III. Experimental Setup

All network parameters are as in the original paper. The additional frame-rate network extrapolating LPCs in LSF



(a) (b)

Fig. 2: Reference systems for comparison. Input is time-domain NB speech, output is time-domain WB speech.

domain has two GRUs with kernel sizes 8x32 and 32x8. A global mean is removed from the LSF before feeding them to the GRU and is added back after extrapolation.

The training material is from the VCTK database [24] as well as other speech items of different languages. In total 6 hours of training material were used, all of it resampled to 16 kHz sampling frequency. Silent passages in the training data were removed with a voice-activation-detection. The features for the conditioning were generated with the provided code from the original paper, but with the data augmentation switched off. The signal is split into 10 ms audio frames resulting in an algorithmic delay of 30 ms due to lookahead (see Sec. II). The network was trained with a variant of the mini-batch stochastic gradient descent algorithm (SGD) called *Adam* [25]. The DNN was trained with the deep learning library Keras [26].

IV. Evaluation

Since the main intention of this work is to show the superiority of the proposed system over the reference systems, a MUSHRA listening test [27] was conducted. According to the MUSHRA methodology, the test items contain the reference marked as such, a hidden reference and the NB signal serving as anchor. 8 experienced listeners participated in the test.

The results are presented in Figure 3 per item and averaged over all items. The proposed BBWE is shown in brown. The box plots show mean values and 95% confidence intervals. The speech items are about 10 seconds long and neither part of the training nor the test set. The items are from native speakers with different nationality: Chinese, English, French, German and Spanish.

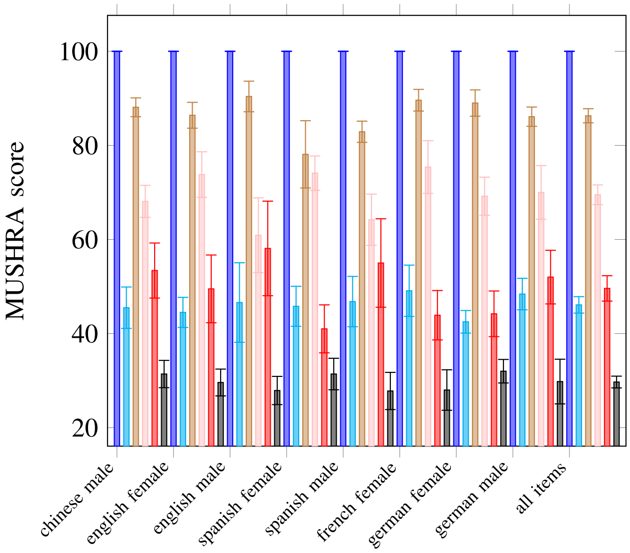


Fig. 3: Results form a MUSHRA listening test [27] with 8 experienced listeners. The box plots show mean values and 95% confidence intervals for each speech item as well as all averaged. The MUSHRA scale estimates the perceived quality from 0 or ”bad” to 100 or ”excellent”. Results are: blue: hidden reference cyan: NB anchor brown: proposed bandwidth extension pink: previously published system [7] red: reference system (a) black: reference system (b)

The results show that the proposed BBWE enhances NB speech by more than 35 MUSHRA points for almost all items except the Spanish female and male items. Both items have a very pronounced pitch with harmonics ranging till nyquist-frequency. At least for the male item this is untypical. The proposed BBWE fails here retain the tonal structure resulting in buzzing (female item) or noisy (male item) artefact.

The reference system (a) only generates little extension signal and adds noise to the speech signal. As result it does not perform better than the NB anchor on most items. The reference system (b) suffers from the mismatch between the excitation signal and the LPC envelope resulting in strong broadband noise.

V. Conclusion

The presented BBWE improves NB speech with low computational complexity and an algorithmic delay of 30 ms - suitable for real-time communication. This is achieved by decomposing the signal into en envelope representation and an excitation signal. Although the network used for extrapolating the envelope is a very simple network, the presented system is able to generate good quality speech. Furthermore the blind extrapolated envelope can be replaced by an envelope quantised at the encoder and transmitted with few bits as e.g. in [15].

REFERENCES

[1] J. Valin and J. Skoglund, “Lpcnet: Improving neural speech synthesis through linear prediction,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5891–5895.

[2] Stefan Bruhn, Harald Pobloth, Markus Schnell, Bernhard Grill, Jon Gibbs, Lei Miao, Kari Järvinen, Lasse Laaksonen, Noboru Harada, N. Naka, Stéphane Ragot, Stéphane Proust, T. Sanda, Imre Varga, C. Greer, Milan Jelinek, M. Xie, and Paolo Usai, “Standardization of the new 3GPP EVS codec,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, 2015, pp. 5703–5707.

[3] Sascha Disch, Andreas Niedermeier, Christian R. Helmrich, Christian Neukam, Konstantin Schmidt, Ralf Geiger, Jérémie Lecomte, Florin Ghido, Frederik Nagel, and Bernd Edler, “Intelligent gap filling in perceptual transform coding of audio,” in *Audio Engineering Society Convention 141, Los Angeles*, Sep 2016.

[4] 3GPP, “TS 26.090, Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions,” 1999.

[5] Patrick Bauer, Rosa-Linde Fischer, Martina Bellanova, Henning Puder, and Tim Fingscheidt, “On improving telephone speech intelligibility for hearing impaired persons,” in *Proceedings of the 10. ITG Conference on Speech Communication, Braunschweig, Germany, September 26-28, 2012*, 2012, pp. 1–4.

[6] Patrick Bauer, Jennifer Jones, and Tim Fingscheidt, “Impact of hearing impairment on fricative intelligibility for artificially bandwidth-extended telephone speech in noise,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 7039–7043.

[7] K. Schmidt and B. Edler, “Blind bandwidth extension based on convolutional and recurrent deep neural networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5444–5448.

[8] J. Makhoul and M. Berouti, “High-frequency regeneration in speech coding systems,” in *ICASSP ’79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, April 1979, vol. 4, pp. 428–431.

[9] Xinyu Li, Venkata Chebiyyam, and Katrin Kirchhoff, “Speech audio super-resolution for speech recognition,” in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, September 15-19, 2019*, 09 2019.

[10] Peter Jax and Peter Vary, “Wideband extension of telephone speech using a hidden markov model,” in *2000 IEEE Workshop on Speech Coding. Proceedings.*, 2000, pp. 133–135.

[11] W. B. Kleijn, F. S. C. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, “Wavenet based low rate speech coding,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 676–680.

[12] Archit Gupta, Brendan Shillingford, Yannis M. Assael, and Thomas C. Walters, “Speech bandwidth extension with wavenet,” *ArXiv*, vol. abs/1907.04927, 2019.

[13] S. Li, S. Villette, P. Ramadas, and D. J. Sinder, “Speech bandwidth extension using generative adversarial networks,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5029–5033.

[14] J. Sautter, F. Faubel, M. Buck, and G. Schmidt, “Artificial bandwidth extension using a conditional generative adversarial network with discriminative training,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 7005–7009.

[15] Konstantin Schmidt and Bernd Edler, “Deep neural network based guided speech bandwidth extension,” in *Audio Engineering Society Convention 147*, Oct 2019.

[16] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, 2016, p. 125.

[17] Z. Ling, Y. Ai, Y. Gu, and L. Dai, “Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 883–894, May 2018.

[18] P. Jax and P. Vary, “An upper bound on the quality of artificial bandwidth extension of narrowband speech signals,” in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 1, pp. I–237–I–240.

[19] Jean-Marc Valin and Jan Skoglund, “A real-time wideband neural vocoder at 1.6 kb/s using lpcnet,” *ArXiv*, vol. abs/1903.12087, 2019.

[20] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[21] Yao Tianren, Xiang Juanjuan, and Lu Wei, “The computation of line spectral frequency using the second chebyshev polynomials,” in *6th International Conference on Signal Processing, 2002.*, Aug 2002, vol. 1, pp. 190–192 vol.1.

[22] Gunnar Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[23] International Telecommunication Union, “Pulse code modulation (pcm) of voice frequencies,” ITU-T Recommendation G.711, November 1988.

[24] Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2017.

[25] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.

[26] François Chollet et al., “Keras,” https://keras.io, 2019.

[27] ITU-R, *Recommendation BS.1534-1 Method for subjective assessment of intermediate sound quality (MUSHRA)*, Geneva, 2003.