ARTIFICIAL BANDWIDTH EXTENSION USING A GLOTTAL EXCITATION MODEL

Sebastian Barth, Simon Stone, Peter Birkholz*

Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany
simon.stone@tu-dresden.de

Abstract: The historical bandwidth of telephone speech (0.3 kHz to 3.4 kHz), which is still used today for speech transmission (e.g. in the AMR-codec [1]) leads to reduced intelligibility and naturalness of the transmitted speech. New mobile devices may use artificial bandwidth extension (ABE) to improve the received narrow-band (NB) speech quality. Aiming to reconstruct missing frequency components of NB speech on the receiving end, ABE often adopts the source-filter-model of human speech to reconstruct excitation and spectral envelope of the speech signal separately. In the extension of the excitation, no existing method exploits the fact that the wide-band (WB) excitation for vowel sounds can be modeled by parametric functions with nearly no perceptible differences [2]. This work investigated the possibility to extract optimal model parameters from the NB speech to use them for high quality ABE of the excitation for vowels. The proposed algorithm objectively meets or exceeds a state-of-the-art reference algorithm, but is currently subjectively slightly inferior.

1 Introduction

Human speech frequency components cover the entire range of human-perceptible frequencies. Therefore, the historical telephone speech bandwidth of 0.3 kHz to 3.4 kHz, which is still used for speech transmission (for example in the AMR codec [1]), leads to reduced intelligibility and naturalness of the transmitted speech. Newer codecs like EVS [3] address this by using sampling rates of up to 48 kHz. However, due to legacy devices transmitting narrow-band (NB) speech signals only, newer mobile devices will also suffer from reduced speech quality and naturalness. One solution to improve speech quality on mobile devices receiving NB speech is artificial bandwidth extension (ABE): the reconstruction of the missing frequency content on the receiving end. ABE methods usually adopt the source-filter-model of human speech (see e.g. [4]) to reconstruct the frequency components of the excitation signal and the spectral envelope separately. Most often, excitation signal and spectral envelope are separated using linear predictive coding (LPC)-analysis. Following this approach, [5, 6, 7, 8] copy or shift the frequency components of the NB excitation signal to frequency ranges where excitation signal components are missing. [5, 9, 8, 10] apply non-linearities to the NB excitation signal or the NB speech signal to generate missing excitation signal frequency components, while [11, 12, 13] try to reconstruct missing frequency components by sinusoidal synthesis. [14] tries to estimate missing frequency components from learned basis functions. Apart from [14], all of the reviewed studies extended the bandwidth of the flat-envelope LPC-analysis residual signal, thereby ignoring the excitation model information captured by glottal models like the ones proposed by Rosenberg [2], Liljencrants and Fant [15] or others [16, 17, 18, 19, 20].

*Corresponding author

The excitation model information is therefore divided among the flat-envelope LPC residual (containing the fundamental frequency and the harmonic structure) and the LPC synthesis filter (containing the spectral envelope). This work instead uses a glottal flow signal or the glottal flow derivative (GFD) signal as the excitation signal rather than the LPC residual signal and thus attempts to model full-bandwidth excitation signal more accurately.

2 Methods and algorithms

2.1 Source-filter model and inverse filtering

According to the source-filter model, a speech signal $s(n)$ may be written as

$$S(z) = U(z)V(z)R(z) = \dot{U}(z)V(z), \tag{1}$$

in the z-domain, where $S(z)$ is the z-transform of the speech signal $s(n)$, $U(z)$ is the z-transform of the glottal flow signal $u(n)$, $V(z)$ the vocal tract filter and $R(z)$ the radiation characteristic. A simple approximation of the radiation characteristic is $R(z) = 1 - z^{-1}$, which is a first-order high-pass filter approximating differentiation at low frequencies. Therefore, using the commutative property of equation 1, the product of $U(z)$ and $R(z)$ approximates the GFD signal $\dot{u}(n)$ with z-transform $\dot{U}(z)$ as given on the right-hand-side in equation 1. The GFD $\dot{u}(n)$ can then be extracted by a simple implementation of inverse filtering. Assuming $V(z)$ can be determined (e.g., using LPC analysis), it is possible to extract $\dot{u}(n)$ using

$$\dot{U}(z) = \frac{S(z)}{V(z)}. \tag{2}$$

In this work, $V(z)$ was extracted using LPC analysis of an order of 20 of the pre-emphasis filtered $s(n)$, where the pre-emphasis is done using the filter $P(z) = 1 - z^{-1}$.

2.2 Spectral envelope extension

The envelope extension, which is part of a complete ABE system, was beyond the scope of this work. Therefore, the proposed excitation extension as well as the excitation extension reference algorithm were combined with the known wide-band (WB) spectral envelopes for evaluation. In practise, an additional algorithm for envelope extension is necessary and will affect the overall system performance. As mentioned in [14], for systems using the LPC residual signal excitation, the envelope extension seems to be critical for high frequencies. High quality excitation signal extension seems to be mainly important for the recovery of low frequency components. It is reasonable to expect that the findings of [14] also hold for the proposed system to a certain extend. But since the glottal excitation contains more modeling information in contrast to the widely used LPC residual excitation, using the proposed method, the excitation extension gains importance over envelope extension in the overall system.

2.3 Frame-based signal processing

The signal processing was done by segmenting the WB and NB speech signals s_{WB} and s_{NB} into frames $s_{\mathrm{WB}}^{(i)}$ an $s_{\mathrm{NB}}^{(i)}$ of 20 ms length and 75 % overlap of successive windows, where i is the frame index. Using the autocorrelation method, the LPC-coefficients $\underline{a}_{\mathrm{WB}}^{(i)}$ were extracted from the pre-emphasized WB speech signal frame $s_{\mathrm{WB,pre}}^{(i)}$ for LPC-analysis filtering of the NB speech signal frames $s_{\mathrm{NB}}^{(i)}$ yielding the GFD $\dot{u}_{\mathrm{NB}}^{(i)}$. The model-based excitation extension (MBEE) then

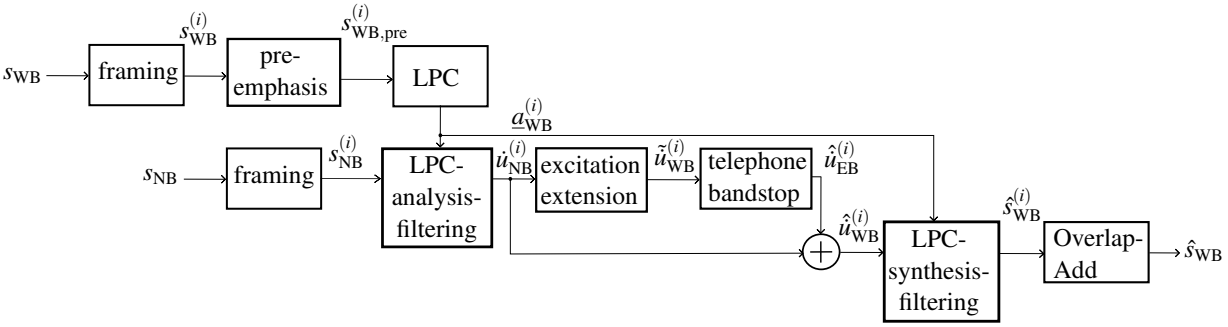

Figure 1 – Signal processing pipeline of the proposed ABE-system

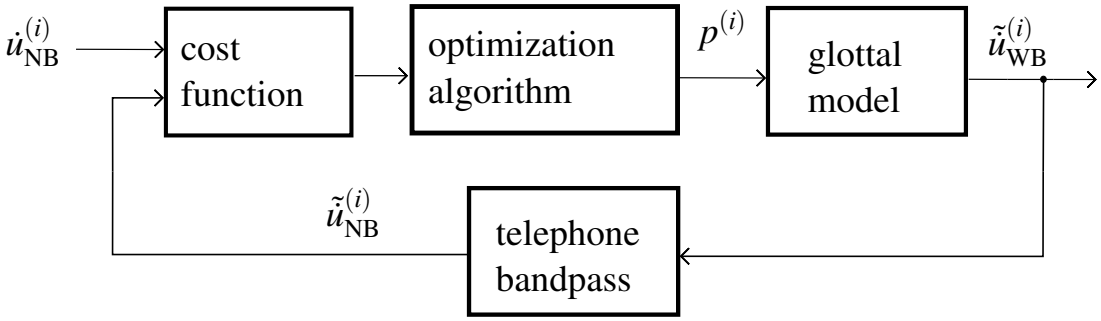

Figure 2 – Flow graph of the model-based excitation extension

determined the GFD WB estimate $\tilde{\dot{u}}^{(i)}_{\mathrm{WB}}$, whose frequency components $\hat{\dot{u}}^{(i)}_{\mathrm{EB}}$ outside of the telephone band are extracted by a telephone bandstop and used to extend $\dot{u}^{(i)}_{\mathrm{NB}}$ to the bandwidth extended GFD $\hat{\dot{u}}^{(i)}_{\mathrm{WB}}$. LPC-synthesis filtering yields bandwidth extended WB speech signal frames $\hat{s}^{(i)}_{\mathrm{WB}}$ which are used to synthesize the bandwidth extended speech signal $\hat{s}_{\mathrm{WB}}$ by overlap-add. Figure 1 shows the signal processing pipeline of the proposed algorithm.

3 Model-based excitation extension

Figure 2 shows the basic idea of the MBEE. For the i-th frame, the parameters $p^{(i)}$ of a glottal model are optimized in order to fit the GFD model function $\tilde{\dot{u}}^{(i)}_{\mathrm{WB}}$ to the observed NB GFD $\dot{u}^{(i)}_{\mathrm{NB}}$ in the telephone band. This is done by repeatedly comparing $\dot{u}^{(i)}_{\mathrm{NB}}$ to the NB version $\tilde{\dot{u}}^{(i)}_{\mathrm{NB}}$ of the model function $\tilde{\dot{u}}^{(i)}_{\mathrm{WB}}$ and corresponding adjustment of the model parameters $p^{(i)}$ by a chosen optimization algorithm.

3.1 Implementation

3.1.1 Glottal model

Different glottal models were compared in terms of modeling capacity. Since it combines a small number of parameters with good modeling accuracy, the Rosenberg model B was chosen [2]. Its GFD is given by

$$\tilde{\dot{u}}_{\mathrm{WB}}(t) = \begin{cases} 6\alpha\left(\frac{t}{T_{\mathrm{P}}^2} - \frac{t^2}{T_{\mathrm{P}}^3}\right) & 0 \le t \le T_{\mathrm{P}} \\ -2\alpha\frac{t-T_{\mathrm{P}}}{T_{\mathrm{N}}^2} & T_{\mathrm{P}} \le t \le T_{\mathrm{P}} + T_{\mathrm{N}} \\ 0 & T_{\mathrm{P}} + T_{\mathrm{N}} \le t \le T_0. \end{cases} \tag{3}$$

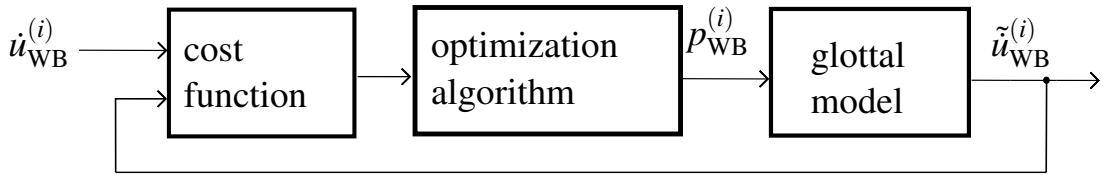

Figure 3 – WB-optimization flow graph

Therefore the model parameters to be determined by optimization were T_P, T_N, α, but also T_0. Furthermore, a parameter T_S was necessary that quantified the shift of the fundamental period given by equation 3 relative to the start of the current frame.

3.1.2 Cost function and performance measure

The mean squared error (MSE) of two length N signals s_1 and s_2

$$d_{\mathrm{MSE}}(s_1, s_2) = \frac{1}{\mathrm{N}} \sum_{n=0}^{\mathrm{N}-1} (s_1(n) - s_2(n))^2. \tag{4}$$

was chosen both as the cost function $d_{\mathrm{MSE}}(\dot{u}^{(i)}_{\mathrm{NB}}, \tilde{\dot{u}}^{(i)}_{\mathrm{NB}})$, or narrow-band mean squared error (NBMSE), and as the performance measure $d_{\mathrm{MSE}}(\dot{u}^{(i)}_{\mathrm{WB}}, \tilde{\dot{u}}^{(i)}_{\mathrm{WB}})$, or wide-band mean squared error (WBMSE), since it is easy to calculate and sensitive to phase differences between the compared signals. The WBMSE reflects the final goal of optimal WB-modeling. For reference, the best possible WBMSE was estimated by averaging the WBMSEs of 20 non-overlapping frames for each of the vowels [aː], [eː], [iː], [oː], [uː] where the WBMSEs were calculated by optimizing model parameters as shown in Figure 3 using three successive runs of MATLABs `GlobalSearch` for optimization where the initial values of the parameters of optimization runs two and three were set to the values found in the respective prior run. The resulting mean WBMSE ($\overline{\mathrm{WBMSE}}$) was calculated as 0.0344.

3.1.3 Parameter optimization

The parameter optimization as shown in Figure 2 was done using three successive runs of MATLABs `GlobalSearch` for optimization where the start values for the parameters of optimization runs two and three were set to the values found in the respective prior run. The optimization was configured with parameter boundaries as given in

$$\varepsilon \leq \alpha \leq 1 \tag{5}$$

and

$$0.01 \cdot T_{0,\min} \leq T_\mathrm{P}, T_\mathrm{N} \leq 0.99 \cdot T_{0,\max}, \tag{6}$$

where ε in eq. 5 is the machine epsilon. The upper bound of α was chosen based on experience but much bigger than the values known from the training data being in the order of 0.001. $T_{0,\min}$ and $T_{0,\max}$ are the boundaries of the fundamental period T_0 and were calculated as

$$T_{0,\min} = \frac{1}{0.95 \cdot \hat{f}_0} \text{ and } T_{0,\max} = \frac{1}{1.05 \cdot \hat{f}_0} \tag{7}$$

based on a fundamental frequency estimate $\hat{f}_0$ calculated prior to optimization. An estimate $\hat{T}_\mathrm{S}$ for the parameter T_S was calculated to define its optimization boundaries as

$$\hat{T}_\mathrm{S} - 0.1 \cdot \frac{1}{\hat{f}_0} \leq T_\mathrm{S} \leq \hat{T}_\mathrm{S} + 0.1 \cdot \frac{1}{\hat{f}_0}. \tag{8}$$

$\hat{T}_S$ was calculated as the maximum of the short-time variance detecting sudden changes as the glottal closing instant (GCI) used for definition of the shift of the fundamental period against the start of the current frame. Finally, the inequality constraint

$$T_P + T_N \leq T_0 \tag{9}$$

was used for optimization, to ensure, that the length of a glottal pulse does not exceed the fundamental period.

An optimal solution found in the NB domain did not necessarily correspond to an optimal solution in the WB domain. The $\overline{\text{WBMSE}}$ using a three-run global NB optimization as shown in Figure 2 was calculated as 0.179, which was much more than the reference $\overline{\text{WBMSE}}$ determined in section 3.1.2. So instead of using the optimal NB parameters directly, they were instead mapped to a set of WB parameters using regression.

3.2 Optimization with subsequent regression

While the parameters T_0 and T_S were found reliably by optimization in the NB domain, the parameters α, T_N and T_P found in the NB domain did not correspond to good fits in the WB domain. As a solution, for each of the latter parameters, ensembles of bagged regression trees were used to map the parameters found by optimization to new parameters α, T_N and T_P for better WB modeling. Training was done using a data-set created from recordings of the vowels [aː], [eː], [iː], [oː], [uː], [ɛː], [øː], [yː] uttered with natural, low and high fundamental frequency by a single 24-year old male speaker. 60 non-overlapping frames per recording yielded 1440 frames, for each of which the parameters were determined by NB and WB optimization as shown in Figures 2 and 3. The hyperparameters were trained using Bayesian optimization with a custom *leave-one-vowel-out* cross-validation, where cross-validation was performed holding out all the data of a specific vowel for validation and averaging the results for all vowels. This was done to prevent the validation data set to contain data from vowels occurring in the training data set. The $\overline{\text{WBMSE}}$ of the optimization with subsequent regression was calculated as 0.0674.

4 Evaluation

4.1 Reference algorithm

A reference excitation extension (REFEE) algorithm capable to extend the LPC residual excitation signal in contrast to the glottal excitation signal to low as well as to high frequencies outside of the telephone band was implemented. The REFEE is a simplified version of the algorithm given in [9] which uses a non-linearity to create harmonics outside of the telephone-band for a WB LPC-residual excitation signal. In contrast to [9], the the power ratio for matching the energies of the synthesized signal and the input signal in the telephone-band is not smoothed by a filter since it is not necessary when using the correct WB envelope instead of a bandwidth extended one as done in this work. Also, the phase manipulation described in [9] was omitted due to a lack of precise information of where to extract the phase in the telephone band, which is assigned to the synthesized signal in the extension band.

4.2 Objective evaluation

The MBEE as well as the REFEE results were compared to original signals for different vowels in terms of the spectral distortion

$$d_{\mathrm{LSD}}(\lambda) = \sqrt{\frac{1}{M} \cdot \sum_{\mu=1}^{M} \left(20 \log_{10} \frac{H(\mu,\lambda)}{\hat{H}(\mu,\lambda)} \right)^2} \tag{10}$$

of frame λ as given in [21] with M being the number of frequency bins and $H(\mu,\lambda)$ and $\hat{H}(\mu,\lambda)$ the DFT spectra at bin μ. The mean spectral distortion for each vowel was calculated from non-overlapping 20 ms frames taken from 1 s speech signals as shown in Table 1. Although the performance of each algorithm differs depending on the vowel, they seem to perform nearly equally well averaged over all vowels, with the MBEE being slightly better in terms of averaged performance.

d_{LSD}	LSD REFEE	LSD MBEE
[aː]	7.0019 dB	6.4650 dB
[eː]	6.0879 dB	9.4833 dB
[iː]	6.7672 dB	7.1764 dB
[uː]	9.2943 dB	7.5153 dB
[oː]	10.0271 dB	8.1638 dB
[ɛː]	6.5775 dB	6.9834 dB
[øː]	7.6535 dB	6.6201 dB
[yː]	8.1960 dB	7.9715 dB
mean	7.7007 dB	7.5473 dB

Table 1 – LSD for vowels bandwidth extended using REFEE- and the MBEE algorithm respectively

4.3 Subjective evaluation

To evaluate if the proposed ABE method may improve NB speech, a listening experiment with 10 participants (5 female, 5 male, age 20 to 36, median age 24) was conducted using Praat [22]. Unfortunately, due to the CoViD-19 pandemic, the test had to be conducted remotely leading to differing test set-ups. All participants used headphones (of varying make and model) and were asked to conduct the test in a quiet environment. The stimuli comprised excerpts of length of 1 s from natural speech recordings of the vowels [aː], [eː], [iː], [oː], [uː], [ɛː], [øː], [yː] uttered by a 24 year-old male speaker. Each recording was weighted with a Tukey-window for 0.25 s fading in and 0.25 s fading out. In each trial in the experiment, the participants were presented with three versions of a vowel sound. The first version was always the original WB sound and the participants were asked to identify, which of the following two items they considered more similar to the original. The two items following the original were either NB versions, again the original WB sound or bandwidth extended speech signals using either the reference algorithm or the proposed algorithm for ABE. While the original recordings were done at 44.1 kHz with 16 bit quantization, all excerpts presented during the listening test were processed versions at a sampling frequency of 16 kHz. Each possible combination was presented once during the test. To avoid bias towards the stimulus played first after the WB original, each combination of two stimuli to choose occurred in both possible orders after the original at some point during the test. The participants were able to replay the sequence of the three signals one time before

making their decision. Table 2 summarizes the results. They indicate that despite the changing test conditions, the participants were on average able to differentiate very well between the WB and NB signals as well as between the WB and extended signals. The table also shows, that the extended signals were preferred to the NB signals for both ABE-algorithms. However, it seems that the reference algorithm is slightly preferred against the proposed one in direct comparison.

competing recordings	preferred recording	percentage
WB-NB	WB	98.75
WB-REFEE	WB	91.25
WB-MOD	WB	96.25
REFEE-NB	REFEE	95.00
MBEE-NB	MBEE	94.37
REFEE-MBEE	REFEE	63.75

Table 2 – Recordings preferred in subjective evaluation (all preferences are significant at $p < 0.01$ according to Fisher's Exact Test)

5 Conclusion

This work investigated the use of glottal model information for ABE. It was observed that the parameters for optimal WB modeling of the glottal flow signal could not be found using the investigated optimization techniques and the NB GFD signal. It was possible to use regression to map the parameters found by optimization to parameters for optimal WB modeling using a small vowel corpus for training. The proposed algorithm objectively meets or exceeds a state-of-the-art reference algorithm, but is currently subjectively slightly inferior in direct comparison. Future work should comprise more detailed investigation of the inconsistencies between the optimal NB- and WB parameters. Furthermore, future work might investigate machine learning techniques instead of optimization followed by regression to map the NB signal or extracted features directly to model parameters for optimal WB modeling. Finally, it would be interesting if the results in terms of modeling inconsistency and performance observed during this work also hold for a larger corpus consisting of a more diverse set of utterances.

References

[1] *3GPP TS 26.090 V15.0.0, Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding Functions*, 2018.

[2] ROSENBERG, A. E.: *Effect of glottal pulse shape on the quality of natural vowels*. *The Journal of the Acoustical Society of America*, 49(2B), pp. 583–590, 1971.

[3] *3GPP TS 26.445 V15.2.0, Codec for Enhanced Voice Services (EVS); Detailed Algorithmic Description*, 2019.

[4] RABINER, L. R. and R. W. SCHAFER: *Digital processing of speech signals*, vol. 100. Prentice-hall Englewood Cliffs, NJ, 1978.

[5] KORNAGEL, U.: *Spectral widening of the excitation signal for telephone-band speech enhancement*. In *Proc. International Workshop on Acoustic Echo and Noise Control*, pp. 215–218. 2001.

[6] JAX, P. and P. VARY: *On artificial bandwidth extension of telephone speech. Signal Processing*, 83(8), pp. 1707–1719, 2003.

[7] FUEMMELER, J. A., R. C. HARDIE, and W. R. GARDNER: *Techniques for the regeneration of wideband speech from narrowband speech. EURASIP Journal on Applied Signal Processing*, 2001(1), pp. 266–274, 2001.

[8] MAKHOUL, J. and M. BEROUTI: *High-frequency regeneration in speech coding systems*. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1979)*, vol. 4, pp. 428–431. IEEE, 1979.

[9] ISER, B. and G. SCHMIDT: *Neural networks versus codebooks in an application for bandwidth extension of speech signals*. In *Eighth European Conference on Speech Communication and Technology*. 2003.

[10] ISER, B. and G. SCHMIDT: *Bandwidth extension of telephony speech*. In *Speech and Audio Processing in Adverse Environments*, pp. 135–184. Springer, 2008.

[11] MIET, G., A. GERRITS, and J.-C. VALIERE: *Low-band extension of telephone-band speech*. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, pp. 1851–1854. IEEE, 2000.

[12] PARK, J. S., M. Y. CHOI, and H. S. KIM: *Low-band extension of CELP speech coder by harmonics recovery*. In *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2004)*, pp. 147–150. IEEE, 2004.

[13] EPPS, J. and W. H. HOLMES: *Speech enhancement using STC-based bandwidth extension*. In *Fifth International Conference on Spoken Language Processing*. 1998.

[14] THOMAS, M. R., J. GUDNASON, P. A. NAYLOR, B. GEISER, and P. VARY: *Voice source estimation for artificial bandwidth extension of telephone speech*. In *International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, pp. 4794–4797. IEEE, 2010.

[15] FANT, G., J. LILJENCRANTS, and Q.-G. LIN: *A four-parameter model of glottal flow. STL-QPSR*, 4(1985), pp. 1–13, 1985.

[16] HEDELIN, P.: *A glottal LPC-vocoder*. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1984)*, vol. 9, pp. 21–24. IEEE, 1984.

[17] ANANTHAPADMANABHA, T.: *Acoustic analysis of voice source dynamics. STL-QPSR*, 2(3), pp. 1–24, 1984.

[18] VELDHUIS, R.: *A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation. The Journal of the Acoustical Society of America*, 103(1), pp. 566–571, 1998.

[19] CUMMINGS, K. E. and M. A. CLEMENTS: *Glottal models for digital speech processing: A historical survey and new results. Digital signal processing*, 5(1), pp. 21–42, 1995.

[20] FUJISAKI, H. and M. LJUNGQVIST: *Proposal and evaluation of models for the glottal source waveform*. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1986)*, vol. 11, pp. 1605–1608. IEEE, 1986.

[21] SCHLIEN, T., P. JAX, and P. VARY: *Acoustic tube interpolation for spectral envelope estimation in artificial bandwidth extension.* In *ITG-Fachbericht 282: Speech Communication*. 2018.

[22] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer. Version 6.1.27*. 2020. URL `http://www.praat.org/`.