Interspeech 2022
18-22 September 2022, Incheon, Korea

Glottal inverse filtering based on articulatory synthesis and deep learning

Ingo Langheinrich, Simon Stone, Xinyu Zhang, Peter Birkholz

Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany

ingo.langheinrich@mailbox.tu-dresden.de

Abstract

We propose a new method to estimate the glottal vocal tract excitation from speech signals based on deep learning. To that end, a bidirectional recurrent neural network with long short-term memory units was trained to predict the glottal airflow derivative from the speech signal. Since natural reference data for this task is unobtainable at the required scale, we used the articulatory speech synthesizer VocalTractLab to generate a large dataset containing synchronous connected speech and glottal airflow signals for training. The trained model's performance was objectively evaluated by means of stationary synthetic signals from the OPENGLOT glottal inverse filtering benchmark dataset and by using our dataset of connected synthetic speech. Compared to the state of the art, the proposed model produced a more accurate estimation using OPENGLOT's physically synthesized signals but was less accurate for its computationally simulated signals. However, our model was much more accurate and plausible on the connected speech signals, especially for sounds with mixed excitation (e.g. fricatives) or sounds with pronounced zeros in their transfer function (e.g. nasals). Future work will introduce more variety into the training data (e.g. regarding pitch and phonation) and focus on estimating features of the glottal flow instead of the entire waveform.

Index Terms: Glottal inverse filtering, glottal source estimation, source-filter separation, speech synthesis

1. Introduction

Human speech is produced by the interaction of respiration, phonation and articulation [1]. The lungs provide an airflow that can cause flow-induced, quasi-periodic vibration of the vocal folds (called voiced phonation). The resulting pulsed glottal airflow serves as the primary acoustic source of the human voice. This source signal is then filtered by the vocal tract during articulation and radiated at the mouth as the speech signal. The glottal airflow carries diverse information about the speaker's anatomy, the phonation type and the movement of the vocal folds [2]. Unfortunately, the glottis is inaccessible without invasive and obstructive equipment. The glottal flow is therefore very difficult to measure directly. However, there are numerous techniques to estimate the glottal flow from the radiated, vocal-tract-filtered speech signal. Despite an inevitable estimation error, these techniques can still be helpful, e.g. for the detection of voice disorders [3], diseases [4, 5], speaker identification and verification [6, 7], and emotion recognition [8]. The process of eliminating the effects of the vocal tract filter from the speech signal is known as glottal inverse filtering (GIF) (see [9] for a recent review). Established algorithms for GIF are generally based on linear prediction (LP) [10], specifically iterative adaptive inverse filtering (IAIF) [11] and the quasi-closed phase (QCP) analysis method [12]. While these methods are quite popular and achieved considerable successes, they require the identification of the glottal closure and opening instants, because the estimation of the vocal tract model takes place during the closed phase of the glottis, i. e. when the glottal flow signal $U_G(z) \approx 0$ [13]. These methods therefore depend on the precision of algorithms which estimate these instants. Moreover, the order of the LP filter must be adapted to the signal, which often requires manual intervention. It was also reported that an increase of the fundamental frequency will decrease the performance [14]. To avoid these limitations, some GIF algorithms are based on spectral decomposition (e.g. zeros of the $\mathcal{Z}$-transform or complex cepstrum decomposition) [15]. The idea here is to decompose the speech signal into minimum-phase (causal) and maximum-phase (anticausal) signal components [16]. The maximum-phase contribution to the speech signal corresponds to the glottal flow during the glottal open phase, whereas the minimum-phase is (among others) associated with the vocal tract impulse response. The decomposition requires the exact identification of the glottal phases and also depends on the (often manual) choice of window length and function [17]. Neural networks have been successfully applied to inverse filtering or deconvolution problems in non-speech-related contexts for several decades (e.g., [18, 19]). In speech research, the lack of available reference glottal flow data for training all but rules out supervised training approaches. A possible workaround is to use one of the non-neural methods to generate reference glottal flow data for training and validation purposes. This limits such a system to the performance of the employed reference algorithm, however, and constrains the approach conceptually to rather specific scenarios (e.g. [20]).

The present study adopted a novel approach to glottal flow estimation to overcome the above problems: We trained a recurrent neural network to directly map the time-domain representations of speech signals to the corresponding time-domain signals of the glottal flow derivative. The required training data was obtained by synthesizing paired speech and glottal flow signals using the articulatory speech synthesizer `VocalTractLab`[1] [21]. The model was therefore trained entirely on synthetic signals. However, it could still be used to infer glottal flow signals from natural speech signals due to the similarities between the natural and synthetic domains. The model performance was evaluated on the OPENGLOT dataset [22] and compared to the reference algorithm IAIF. While IAIF achieved slightly higher marks (in terms of correlation and open quotient error) on the computationally and physically synthesized isolated vowel sounds contained in OPENGLOT, our model produced much more plausible glottal flow signals on continuous utterances than IAIF without manual intervention, especially for nasal sounds, voiced fricatives, and unvoiced/voiced transitions.

2. Datasets for training and validation

As described in section 1, the reference glottal flow data required for a supervised learning approach is prohibitively dif-

[1] `https://www.vocaltractlab.de`

10.21437/Interspeech.2022-10119

ficult to obtain with human speakers. We therefore generated synthetic speech signals and the corresponding glottal flow data using articulatory synthesis. The idea was to exploit the similarity between the synthetic and the natural domain, train entirely on *synthetic data* but still use the trained model to infer the glottal flow corresponding to *natural speech*. However, an objective evaluation of the model performance on natural speech is also difficult because of the same lack of reference glottal flow data. We therefore used three different datasets: an articulatorily synthesized dataset for training, the physically and computationally synthesized OPENGLOT dataset [22] for objective evaluation, and the BITS Unit Selection corpus [23] for a qualitative sample and plausibility check of the performance on real human speech.

2.1. Articulatorily synthesized dataset

The training data was generated using `VocalTractLab` (VTL). VTL is an articulatory speech synthesis system that simulates the entire speech production process. It combines aerodynamic, articulatory, and acoustic models to produce speech of a quality comparable to other academic, natural-speech-based synthesis systems [24] (albeit still much less natural than the commercial state of the art). In VTL, the trachea, the glottis and the vocal tract are approximated geometrically by a series of cylindrical tube sections with variable length and diameter whose dimensions are obtained from a 3D model of the vocal tract [21] and a geometric model of the vocal folds [25]. For the aerodynamic-acoustic simulation, a transmission-line model is used, where each tube section is represented as an electrical two-port network [26]. The models of the vocal tract and the vocal folds are controlled by a set of parameters, which are varied by means of a so-called gestural score. A gestural score contains eight articulatory trajectories, whereby the first five control the movements of the active articulators, the sixth the phonation type, the seventh the fundamental frequency contour and the eighth the subglottal air pressure from the lungs. The glottis model, which was based on an original version by Titze [27], describes the glottal area between the lower and upper vocal fold edges as a function of time and is able to produce skewed asymmetric glottal area waveforms as well as diplophonic double pulsing. The vocal tract is controlled by concatenating a series of target vocal tract shapes with a specified duration, and then interpolating these targets using the Target Approximation Model (TAM) [21]. While gestural scores are usually specified by hand, version 2.3 of VTL introduced a new function that can initialize a gestural score based solely on the phone labels and their durations. To obtain the phone labels and durations from normalized orthographic text, a recently-presented (closed-source) preprocessing frontend [28] was used. The text material used for the synthesis was selected based on three sentence comprehension tests from the field of audiometry: The *Berliner* [29], *Marburger* [30], and *Oldenburger Satztest* [31], which were all designed to be phonemically balanced. As the text preprocessing module could only process declarative sentences, all imperative sentences were treated as declarative and all interrogative and exclamatory sentences were excluded. The 163 remaining sentences from the Berlin and Marburg sets were selected and supplemented by 30 sentences generated according to the procedural Oldenburg test to restore phonemic balance for a total of 193 sentences. Using the pipeline described above, 193 gestural scores were generated based on these sentences with the respective default fundamental frequency contour and speech rate (as determined by the preprocessing module), and with modal phonation of the voiced sounds. These 193 initial scores only corresponded to roughly seven minutes of audio. Since articulatory synthesis is entirely parametric, however, we created a number of variants of each sentence by changing the phonation type of the voiced segments, the mean fundamental frequency, the speech rate (in terms of a linear stretch factor of the phone durations), and by adding noise. In order to do that, we randomly sampled values from the discrete ranges and continuous distributions given in Table 1 and manipulated the parameters of the baseline gestural scores accordingly. In total, we generated scores corresponding to an additional 113 min of speech for a total of about 120 min.

Table 1: *Distributions and value ranges for the training data augmentation*

Voiced phonation type:	[breathy, modal, pressed]
Fundamental frequency f_0 [32]:	$\mathcal{N}(\mu_{f_0} = 120\,\text{Hz}, \sigma_{f_0} = 19\,\text{Hz})$
Speech rate factor sr [33]:	$\mathcal{N}(\mu_{\text{sr}} = 1, \sigma_{\text{sr}} = 0.15)$
Noise models:	[white, babble]
Noise levels:	[clean, 20 dB, 30 dB]

Finally, the gestural scores were used to synthesize the speech signals and the corresponding glottal flow signals according to VTL's geometric glottis model. All signals were downsampled to 8 kHz to reduce the computational load of the downstream processing. The entire dataset is available in the supplemental material accompanying this paper[2].

2.2. OPENGLOT

OPENGLOT is a free and open collection of data designed to evaluate GIF algorithms [22]. It consists of four subsets called repositories. Three of these repositories contain glottal flow data (see Table 2) and were therefore chosen for this study. Since VTL is currently only based on a male speaker, we excluded the female data from OPENGLOT. The phoneme set included in OPENGLOT is fairly limited and only includes a few vowels. The duration of each utterance is also very short (less than 1 s) and there are no transitions included. These limitations along with the employed synthesis methods bias the corpus towards the (unrealistic) linearly separable source-filter assumption and the all-pole filter model underlying IAIF. For lack of a more realistic and diverse alternative, it was still used for the objective evaluation.

2.3. BITS Unit Selection corpus

The BITS Unit Selection corpus [23] contains recordings of 1683 German sentences each spoken by four speakers (two male, two female) originally intendend for unit selection speech synthesis. In addition to the audio signal, the corpus also includes synchronous electroglottography (EGG) signals made with a LaryngoGraph PCLX. The EGG signals are quite similar to glottal flow signals but have some key differences that make it unwise to use them as a quantitative reference. The main reason is that the EGG ultimately measures the contact area of the vocal folds, which is correlated with but not identical to the glottal flow. However, the continuous natural speech in this corpus is clearly a more useful and practically relevant domain to evaluate a GIF algorithm than synthetic isolated sounds. In this study, we therefore used this dataset only for qualitative comparative

[2] `https://vocaltractlab.de/index.php?page=birkholz-supplements`

Table 2: *Details of the three repositories from OPENGLOT used in this study*

	Data generation method	**Content**
Repository I (144 samples)	linear source-filter model (Liljencrants-Fant excitation, all-pole vocal tract filter, $f_s = 8\,\mathrm{kHz}$)	f_0: from 100 Hz to 200 Hz in 20 Hz steps phonation type: normal, breathy, whispery, creaky vowels: /ɑ, æ, i, u, o, e/
Repository II (48 samples)	physical computer simulation of speech production ($f_s = 44.1\,\mathrm{kHz}$)	f_0: 82 Hz, 110 Hz, 156 Hz, 220 Hz vocal fold adduction ξ_{02}: 0.09 cm, 0.06 cm, 0.03 cm vowels: /ɑ, æ, i, u/
Repository III (44 samples)	physical system with an acoustic source and 3D printed vocal tract ($f_s = 44.1\,\mathrm{kHz}$)	f_0: from 100 Hz to 200 Hz in 10 Hz steps vowels: /ɑ, æ, i, u/

analysis of the glottal flow signals produced by our proposed model.

3. Model training

Estimating the glottal flow signal from the speech signal is a sequence-to-sequence transformation task. To avoid having to deal with constant offsets, scaling issues, and to limit the influence of the radiation characteristic at the mouth on the glottal flow $u_\mathrm{g}(k)$, we instead used its first derivative $u'_\mathrm{g}(k)$ throughout this study. All presented glottal flow signals were obtained by trapezoidal numerical integration of the derivative. For the transformation, we used a comparatively simple yet powerful bidirectional recurrent neuronal network with long short-term memory units (BiLSTM), which mapped the speech pressure signal $s(k)$ as input to the first derivative of the glottal flow as the output $\hat{u}'_\mathrm{g}(k)$. The model was implemented in PyTorch [34]. Only the VTL dataset (speech and corresponding reference glottal flow) was used for the training. It was split into a 100 min training subset and a 20 min validation subset, partitioned along the original 193 sentences. Thus, no particular phone sequence appeared both in the training and the validation set. During the training process, the speech and the corresponding glottal flow derivative signals were subdivided into sequences with K time steps. To ensure at least one glottal cycle in the training sequences given the expected fundamental frequency ranges of German males [32], we chose a length of $K = 120$ samples ($\approx 15\,\mathrm{ms}$ corresponding to an f_0 floor of 67 Hz). The L2 loss between the estimated glottal flow derivative $\hat{u}'_\mathrm{g}(k)$ and the reference glottal flow derivative $u'_\mathrm{g}(k)$, backpropagation through time, gradient clipping, and stochastic gradient descent were used for training. The BiLSTM consisted of a single LSTM layer with a hidden size N (processing the input in forward and backward direction) and a dense layer of size $2N$. According to [35], the learning rate α and hidden size N have the strongest influence on the performance of a BiLSTM model. We therefore trained models for all permutations of $\alpha \in \{0.01, 0.1\}$ and $N \in \{5, 10, 20, 30, 40, 50\}$. The method to choose the hyperparameters of the BiLSTM model deviated from the best practice in machine learning to account for the different domains in the training (synthetic speech) and the production (natural speech) environment: The final model was not chosen based on the performance on the test subset of the VTL dataset, but on the performance on the OPENGLOT dataset. While the OPENGLOT data was still not truly natural data, this still better represents the domain shift that the model endures between training and production. Another aspect of the domain shift is that a certain level of underfitting the model was actually desirable, in order to avoid learning details that are only present in synthetic but not in natural speech data. We therefore trained models on random, overlapping subsets of the training data of the lengths $T_\mathrm{train} \in \{5, 10, 20, 40, 60, 80, 100\}$ (in minutes). In total, 84 models were trained (2 learning rates $\times$ 6 hidden sizes $\times$ 7 training subsets).

4. Results and discussion

All 84 trained models were evaluated on the OPENGLOT subset described in subsection 2.2. The performance was determined both in terms of the difference of the open quotient ΔOQ [36] and the cross-correlation r between $u_\mathrm{g}(k)$ and $\hat{u}_\mathrm{g}(k)$. The results of the best-performing model ($N = 50$, $\alpha = 0.01$, $T_\mathrm{train} = 10$) are shown in Table 3 using IAIF as a reference (implemented in Aalto Aparat using default parameters and the recommended value of 4 formants [37]). The performance of both IAIF and the above BiLSTM on the VTL test subset (limited to the most similar German vowels to the ones contained in OPENGLOT) is also listed. A representative sample of the estimated glottal flow signals is shown in Figure 1.

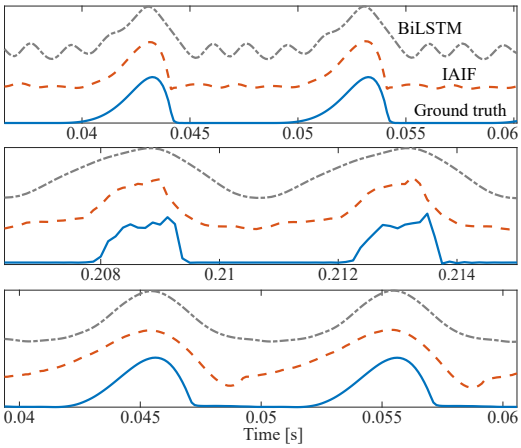

Figure 1: *Examples of glottal flow estimations (top to bottom): average error (repository I, [ɔ], creaky, $f_0 = 100\,\mathrm{Hz}$), highest error (repository II, [u], $\xi_{02} = 0.03\,\mathrm{cm}$, $f_0 = 220\,\mathrm{Hz}$), and lowest error (repository III, [a], $f_0 = 100\,\mathrm{Hz}$).*

Considering the within-domain case first (evaluation on the VTL test subset), it is evident that the proposed model was able to generalize quite well from relatively little training data. Since the model was selected based on the cross-domain performance, the ideal amount of training data was much less than what was expected (just 10 min) but the model still achieved

Table 3: *Performance of the proposed BiLSTM model and the reference algorithm IAIF. Results are shown as mean ± std (median).*

	Proposed BiLSTM		**IAIF**	
	$\lvert\Delta\mathrm{OQ}\rvert$	$r\,(u_g(k), \hat{u}_g(k))$	$\lvert\Delta\mathrm{OQ}\rvert$	$r\,(u_g(k), \hat{u}_g(k))$
VTL test subset /a, e, i, o, u, ɛ/	$\mathbf{0.0291 \pm 0.0177 (0.0269)}$	$\mathbf{0.9501 \pm 0.037 (0.9612)}$	$0.0879 \pm 0.0489 (0.0765)$	$0.8451 \pm 0.0516 (0.8475)$
OPENGLOT				
Repository I	$0.0461 \pm 0.0461 (0.0361)$	$0.9128 \pm 0.0785 (0.9413)$	$\mathbf{0.0227 \pm 0.0627 (0.0105)}$	$\mathbf{0.9756 \pm 0.0174 (0.9801)}$
Repository II	$0.0671 \pm 0.0684 (0.0359)$	$0.9167 \pm 0.0608 (0.9347)$	$\mathbf{0.0206 \pm 0.0164 (0.0171)}$	$\mathbf{0.9782 \pm 0.0204 (0.9878)}$
Repository III	$\mathbf{0.1937 \pm 0.0953 (0.1758)}$	$\mathbf{0.8578 \pm 0.1010 (0.8854)}$	$0.2166 \pm 0.0421 (0.2029)$	$0.8657 \pm 0.0304 (0.8683)$

this rather high within-domain accuracy. This appears to validate the assumption that cross-domain training is a suitable approach to compensate the lack of reference in the desired domain if proper care is take to underfit the training domain data. In the cross-domain case, the performance of the BiLSTM was slightly worse for repository I and II of OPENGLOT compared to IAIF and slightly better in repository III. However, as discussed in subsection 2.2, the OPENGLOT corpus makes some of the same assumptions as IAIF regarding source-filter separability and the all-pole structure of the vocal tract filter. In natural speech, both of these assumptions are routinely violated. Voiced consonants, for example, have a glottal and a supraglottal excitation and the transfer functions of nasal sounds contain zeros. As shown in Figure 2, the signal estimated by IAIF was much less accurate compared to the proposed model in such cases, even in synthetic speech.

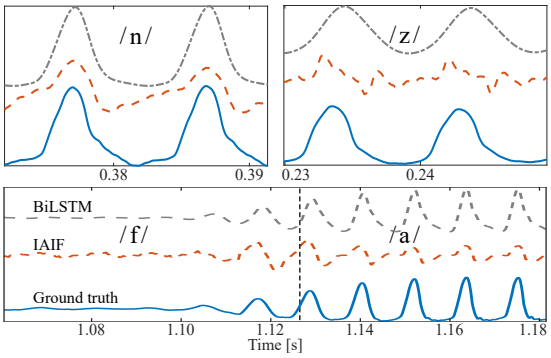

Figure 2: *Example glottal flow segments for the phonemes /n, z/ and an unvoiced/voiced transition from the VTL test set of synthetic speech*

Furthermore, the degree of realism in the signal generation technique was higher in repository III, where the proposed model outperformed IAIF. A qualitative analysis of the glottal flow estimations based on continuous speech signals from the BITS corpus by both the proposed BiLSTM and IAIF (this time as implemented in COVAREP [38] using default settings) further support this observation: Since no objective reference is available for the natural data, we instead inspected only a few representative examples. Figure 3 shows results for the same sounds examined in Figure 2. As before on the synthetic data, the estimations by IAIF seem rather unrealistic while the proposed model's output is more in line with what the EGG data suggests.

5. Conclusions and outlook

The proposed BiLSTM model, trained on articulatorily synthesized speech to estimate the glottal flow from a given speech

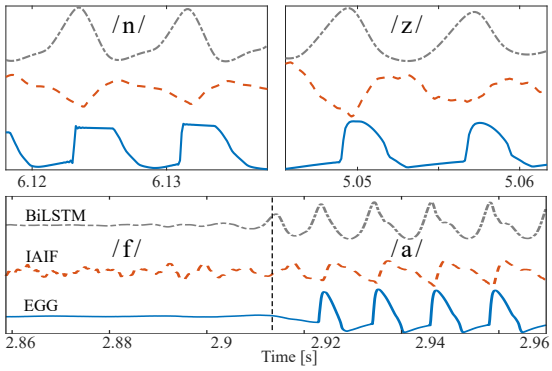

Figure 3: *Example glottal flow and EGG segments for the phonemes /n, z/ and an unvoiced/voiced transition from the BITS corpus of natural speech*

signal, outperformed the state-of-the-art IAIF slightly on physically synthesized speech data of short, isolated vowel sounds. On continuous natural speech, a qualitative analysis showed that the proposed model produces much more plausible glottal flow signals than IAIF. While IAIF requires the manual specification of some parameters based on the speech signal content, the BiLSTM model has no free parameters and can process continuous, arbitrary speech input including voiced/unvoiced transitions without any user intervention. Due to the lack of natural glottal flow reference data, the objective accuracy of the estimations cannot be evaluated. Even though EGG data is available, an EGG waveform is very different from the shape of the true glottal flow. However, future work should analyze metrics that can be derived from an EGG signal, e.g. the glottal closure and glottal opening instants. More applications for the cross-domain training approach, for example the estimation of phone durations or intonation contours, should also be explored.

6. Acknowledgements

The authors are grateful to the Center for Information Services and High Performance Computing at TU Dresden for providing its facilities for the high throughput calculations involved in the synthesis and model training procedures. This project was funded by the Federal Ministry for Economic Affairs and Energy (BMWi) through the AiF Projekt GmbH (German Federation of Industrial Research Associations) under grant no. ZF4443005HB9 - "SEMED".

7. References

[1] G. Fant, *Acoustic Theory of Speech Production*. The Hague, Netherlands: De Gruyter Mouton, 1960.

[2] I. R. Titze, "Comments on the myoelastic-aerodynamic theory of

phonation," *J. Speech Lang. Hear. Res.*, vol. 23, no. 3, pp. 495–510, 1980.

[3] N. Narendra and P. Alku, "Dysarthric speech classification using glottal features computed from non-words, words and sentences." in *Proc. of the Interspeech*, Hyderabad, India, 2018, pp. 3403–3407.

[4] A. Ben Aicha and K. Ezzine, "Cancer larynx detection using glottal flow parameters and statistical tools," in *Proc. of the ISIVC*. Tunis, Tunisia: IEEE, 2016, pp. 65–70.

[5] N. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1925–1936, 2021.

[6] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Speech Audio Process.*, vol. 7, no. 5, pp. 569–586, 1999.

[7] Y. Banaras, A. Javed, and F. Hassan, "Automatic speaker verification and replay attack detection system using novel glottal flow cepstrum coefficients," in *Proc. of FIT*, Islamabad, Pakistan, 2021, pp. 149–153.

[8] Ling He, M. Lech, and N. Allen, "On the importance of glottal flow spectral energy for the recognition of emotions in speech," in *Proc. of the Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 2346–2349.

[9] P. Alku, "Glottal inverse filtering analysis of human voice production — a review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, 2011.

[10] K. D. Vahid Khanagha, "An efficient solution to sparse linear prediction analysis of speech," *Eurasip J. Audio Speech Music Process.*, p. 3, 2013.

[11] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109–118, 1992.

[12] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 3, pp. 596–607, 2014.

[13] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 27, no. 4, pp. 350–355, 1979.

[14] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Commun.*, vol. 38, no. 1-2, pp. 141–160, 2002.

[15] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Commun.*, vol. 53, no. 6, pp. 855–866, 2011.

[16] B. Bozkurt and T. Dutoit, "Mixed-phase speech modeling and formant estimation, using differential phase spectrums," in *Proc. of VOQUAL*, Geneva, Switzerland, 2003, pp. 21–24.

[17] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Comput. Speech Lang.*, vol. 26, no. 1, pp. 20–34, 2012.

[18] F. Glanz and W. Miller, "Deconvolution and nonlinear inverse filtering using a neural network," in *Proc. of the ICASSP*, vol. 4, Glasgow, UK, 1989, pp. 2349–2352.

[19] P.-R. Chang, C. G. Lin, and B.-F. Yeh, "Inverse filtering of a loudspeaker and room acoustics using time-delay neural networks," *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3400–3408, 1994.

[20] N. Narendra, M. Airaksinen, and P. Alku, "Glottal source estimation from coded telephone speech using a deep neural network." in *Proc. of the Interspeech*, Stockholm, Sweden, 2017, pp. 3931–3935.

[21] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[22] P. Alku, T. Murtola, J. Malinen, J. Kuortti, B. Story, M. Airaksinen, M. Salmi, E. Vilkman, and A. Geneid, "OPENGLOT – an open environment for the evaluation of glottal inverse filtering," *Speech Commun.*, vol. 107, pp. 38–47, 2019.

[23] T. Ellbogen, F. Schiel, and A. Steffen, "The BITS speech synthesis corpus for german," in *Proc. of the LREC*, Lisbon, Portugal, 2004, pp. 2091–2094.

[24] P. K. Krug, S. Stone, and P. Birkholz, "Intelligibility and naturalness of articulatory synthesis with vocaltractlab compared to established speech synthesis technologies," in *Proc. of the SSW 11*, Budapest, Hungary, 2021, pp. 102–107.

[25] P. Birkholz, S. Drechsel, and S. Stone, "Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis," in *Proc. of the Interspeech*, Graz, Austria, 2019, pp. 3765–3769.

[26] P. Birkholz, D. Jackel, and B. J. Kroger, "Simulation of losses due to turbulence in the time-varying vocal system," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1218–1226, 2007.

[27] I. R. Titze, "Parameterization of the glottal area, glottal flow, and vocal fold contact area," *J. Acoust. Soc.*, vol. 75, no. 2, pp. 570–580, 1984.

[28] S. Stone, A. Azgin, S. Mänz, and P. Birkholz, "Prospects of articulatory text-to-speech synthesis," Poster at the ISSP, Providence, RI, USA, 2020, [Online] https://vocaltractlab.de/publications/stone-2020-issp-tts.pdf.

[29] M. Pätzold and A. P. Simpson, "Acoustic analysis of German vowels in the Kiel corpus of read speech," *AIPUK*, vol. 32, pp. 215–247, 1997.

[30] K. Brinkmann, "Die Neuaufnahme des Marburger Satzverständnistestes," *Zeitschrift für Hörgeräte-Akustik*, vol. 13, no. 6, pp. 190–194, 1974, [German, English].

[31] V. Kuehnel, B. Kollmeier, and K. Wagener, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests," *Zeitschrift für Audiologie*, vol. 38, pp. 4–15, 1999, [German].

[32] F. Zimmerer, B. Andreeva, J. Jügler, and B. Möbius, "Comparison of pitch profiles of German and French speakers speaking French and German," in *Proc. of the ICPhS*, Glasgow, UK, 2015, pp. 0183.1–5.

[33] J. Trouvain, J. Koreman, A. Erriquez, and B. Braun, "Articulation rate measures and their relation to phone classification in spontaneous and read German speech," in *ITRW on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, 2001, pp. 155–158.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inf. Process. Syst.* Curran Associates, Inc., 2019, pp. 8024–8035.

[35] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2017.

[36] R. Timcke, H. von Leden, and P. Moore, "Laryngeal vibrations: Measurements of the glottic wave: Part I. the normal vibratory cycle," *AMA Arch. Otolaryngol.*, vol. 68, no. 1, pp. 1–19, 1958.

[37] P. Alku, H. Pohjalainen, and M. Airaksinen, "Aalto Aparat - a freely available tool for glottal inverse filtering and voice source parameterization," in *Proc. of the Subsidia*, Malaga, Spain, 2017, pp. 21–23.

[38] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - a collaborative voice analysis repository for speech technologies," in *Proc. of the ICASSP*. Florence, Italy: IEEE, 2014, pp. 960–964.