A SPECTRAL GLOTTAL FLOW MODEL FOR SOURCE-FILTER SEPARATION OF SPEECH

Olivier Perrotin

Univ. Grenoble Alpes, CNRS, Grenoble INP
GIPSA-lab, F-38000 Grenoble, France

Ian McLoughlin

School of Computing
University of Kent, Medway, UK

ABSTRACT

The estimation of glottal flow from a speech waveform is an essential technique used in speech analysis and parameterisation. Significant research effort has been addressed at separating the first vocal tract resonance from the glottal formant (the low-frequency resonance that describes the open-phase of the vocal fold vibration), but few methods are capable of estimating the high-frequency spectral tilt, characteristic of the closing phase of the vocal fold vibration (which is crucial to the perception of vocal effort). This paper proposes an improved Iterative Adaptive Inverse Filtering (IAIF) method based on a Glottal Flow Model, which we call GFM-IAIF. The proposed method models the wide-band glottis response, incorporating both glottal formant and spectral tilt characteristics. Evaluation against IAIF and recently proposed IOP-IAIF shows that, while GFM-IAIF maintains good performance on vocal tract modelling, it significantly improves the glottis model. This ensures that timbral variations associated to voice quality can be correctly attributed and described.

Index Terms— Glottal inverse filtering, Glottal flow, Spectral model, Spectral tilt, Voice quality

1. INTRODUCTION

Speech communication combines linguistic attributes to convey phonetic information through articulation, and prosodic attributes that encode speech expression through variation of pitch, intensity, rhythm and timbre. The widely used linear source filter model [1] combines those components in four parts; an excitation E, vocal tract (VT) filter V, lip radiation filter L and glottis component G to yield speech $S(f) = E(f)G(f)V(f)L(f)$. The role of G is to model the vibration shape of the vocal folds to convey voice quality. It is often combined with L into a glottal flow derivative (GFD).

Glottal inverse filtering (GIF) [2] is used to separate these components from recorded speech. It first estimates the glottis and VT filters, then deconvolves the VT filter from the speech signal. While glottal spectra are broadband, most current GIF methods simply assign the lower part of the spectrum to the glottis G and the higher part to the VT V. In [3] the authors proposed a new GIF method that extended the well-known Iterative Adaptive Inverse Filtering (IAIF) process [4], to wideband characteristics to more accurately apportion spectral influences between G and V and hence better model vocal effort. More recently, Mokhtari et al. [5] evaluated this glottal flow model (GFM) against IAIF and the IOP-IAIF method [6] using computational physical modelling. In the current paper, we perform an in-depth evaluation of each method using optimal parameters, against well established evaluation criteria [7] on both synthetic speech and natural speech. Results demonstrate the ability of GFM to (a) properly respond to changes in vocal effort, (b) accurately model the wide-band glottis frequency response, yielding better glottal estimation in the high frequency part of the spectrum.

2. GLOTTAL INVERSE FILTERING METHODS

2.1. Spectral glottal flow model

Vocal folds vibrations are asymmetric, with a slow opening phase responsible for a spectral resonance called 'glottal formant', and a more abrupt closing phase contributing to the higher spectral frequencies. These observations motivate a 3$^{\text{rd}}$ order spectral model of glottal flow that combines a complex conjugate pole pair $\{a, a^*\}$ accounting for the glottal formant, with one real pole b modelling the high frequency attenuation called 'spectral tilt', giving: $G(z) = \left\{(1 - az^{-1})(1 - a^*z^{-1})(1 - bz^{-1})\right\}^{-1}$ [8, 9]. Applications to voice quality modification [10] and expressive singing or speech synthesis [11, 12] has demonstrated the close relationship between this glottal flow model (GFM) and perception of voice quality (e.g. tenseness, effort).

2.2. Glottal inverse filtering methods

GIF has been a topic of research for more than 60 years [13, 14] with most current methods based on linear prediction [15] to extract the VT after pre-emphasis. The glottis filter is often reduced to the glottal formant contribution ($b = 0$), and assuming that the a coefficient in G is close to the coefficient of the lip radiation filter $L(z) = 1 - d$, pre-emphasis removes the contribution of GL by first-order high-pass filtering [16]. This is the case in IAIF which uses 1^{st} order LPC analysis to define the pre-emphasis filter [4]. While this method uses straightforward computation without *a priori* knowledge of the signal (no estimation) and is noise robust [7], it fails to encompass the spectral tilt of the glottis filter, which is an important attribute related to the perception of vocal effort. IOP-IAIF [6] is a recent attempt to include spectral tilt through unconstrained high-order filtering for signal pre-emphasis. We believe that while improvements to IAIF are merited, an unconstrained filter order risks endowing the glottal model with too much complexity. It also complicates the extraction of perceptual parameters for applications such as analysis, voice modification or coding. In [3] we therefore proposed replacing the 1st order IAIF glottal model with a 3rd order filter matching the GFM, based on strong evidence that this degree of complexity is sufficient [8, 10, 11]. We will demonstrate that this is significantly better than IAIF at conveying vocal effort information. It also allows simple spectral parameters to describe the model, related to the three poles, which is beneficial for voice transformation, coding, or synthesis.

3. FRAMEWORK OF GFM-IAIF

GFM-IAIF primarily proposes replacing the simple IAIF pre-emphasis filter with a 3rd order glottal model as shown in Fig 1. Like traditional IAIF [4], it comprises four steps, with the main difference being in pre-emphasis (step 1) and fine glottis estimation (step 3). Step 1, also called gross glottis estimation, enables removal of the glottis spectral tilt contribution from the speech signal, in preparation for VT estimation. It is essential in this step not to model any VT formants. Estimation is therefore accomplished by three successive 1st order LPC iterations and the resulting gross glottis filter has three real poles. In VT gross estimation (step 2), the gross glottis and lip radiation filters are deconvolved from the original signal and VT autoregressive coefficients estimated through high order LPC. Next, fine estimation of the glottis (step 3) first removes lip radiation and the estimated VT contributions (hence all VT formants) from the speech signal. From this, the full spectral envelope of the glottis is extracted, including glottal formant and high frequency attenuation. A 3rd order LPC is used, to ensure that the final glottis filter adheres to the glottal flow model. VT fine estimation (step 4) then reverts to the IAIF final step. The glottal flow derivative is finally obtained by deconvolving the fine VT from the speech signal.

4. EVALUATION

In this section, GFM-IAIF, the recently published IOP-IAIF [5], and IAIF [17] are evaluated and compared. We adopt the Drugman et al. [7] GIF evaluation on (a) synthetic speech, to quantify the ability to model glottis characteristics, and (b) natural speech, to evaluate how well voice quality is encoded.

Three frequency-domain features [7] are used to describe the extracted glottal flow: The dB amplitude difference between first and second harmonics *H1H2* [18, 19], relating to the glottal formant position (higher when the latter is closer to the first harmonic); Harmonic richness factor (*HRF*), a measure of the quantity of harmonics in the spectrum, is the ratio between the sum of the 2^{nd} to n^{th} harmonic dB amplitudes over the fundamental frequency amplitude [8]; Spectral tilt (*ST* in dB/decade), computed from a linear regression of the n first harmonic amplitudes on a log-frequency scale. n is chosen to select harmonics below 5 kHz only [17]. All three parameters have been proven to match the perception of voice quality – when the voice becomes louder, *H1H2* reduces while *HRF* and *ST* increase, and vice versa [8, 18, 20].

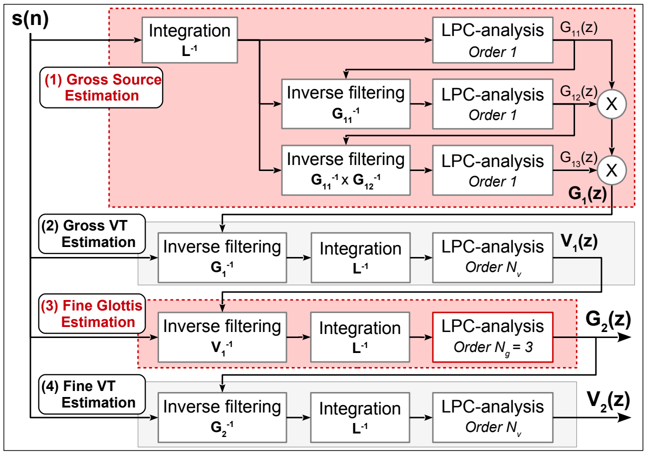

Fig. 1. Architecture of the GFM-IAIF method. The highlighted boxes are steps that differ from the standard IAIF.

4.1. Evaluation on synthetic speech

We synthesised 8700 phone stimuli at $F_s = 16$ kHz by passing LF glottal waveforms [21] through auto-regressive VT filters. The waveforms are defined in the time domain and were parametrised by the R_d coefficient that describes each glottal pulse (width and asymmetry) and is representative of voice quality: small values lead to a tense/loud voice while large values lead to a lax/soft voice [18]. 30 voice qualities (R_d from 0.4 to 2.7, equally spaced on a log-scale) × 29 pitch levels (f_0 from 100 Hz to 240 Hz with a 5 Hz step) × 10 isolated vowels were synthesised, using the formant values provided in [22]. F_1 is the centre frequency of the first vocalic formant.

Following [5], the lip radiation coefficient d and the LPC orders of the fine glottis estimation N_g and both vocal tract estimations N_v, were chosen to minimise the root mean square (RMS) difference between the ground truth glottal flow (LF waveform) and the estimated glottal flow. For this sake, d was varied from 0.8 to 0.99 in steps of 0.01, N_v was varied from $\lfloor F_s/1000 \rceil - 2$ to $\lfloor F_s/1000 \rceil + 6$ in steps of 2, and N_g was varied from 3 to 6 in steps of 1. The triplet that provided the smallest RMS distance for each method and each stimuli was kept. Note that the glottis LPC order N_g was fixed to 3 in the case of the GFM-IAIF method.

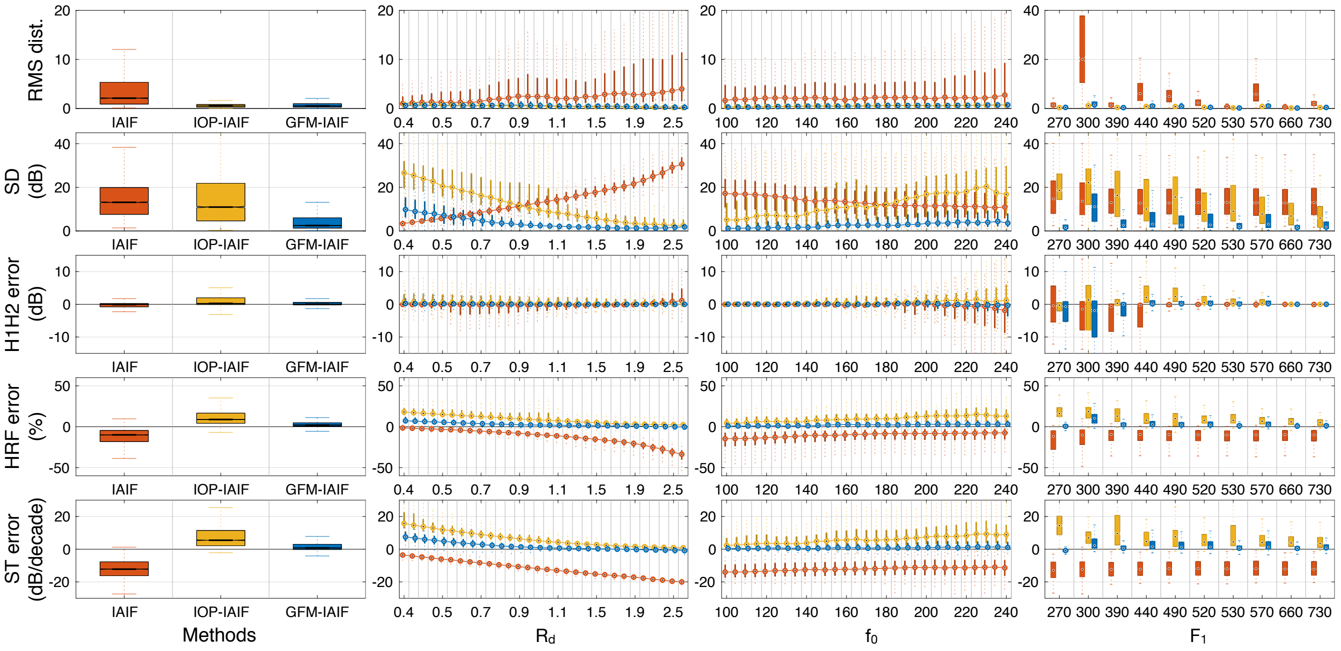

Fig. 2. Distances between extracted and original glottis. From rows 1 to 5: RMS distance; *SD*; *H1H2* error; *HRF* error; *ST* error. Columns 1 to 4: overall scores, R_d, f_0 and F_1 dependencies. Orange: IAIF; Yellow: IOP-IAIF; Blue: GFM-IAIF.

The top row of Fig. 2 shows the RMS distances obtained for each method, and their dependencies on R_d, f_0, and F_1. A Kruskal-Wallis rank-sum (KWRS) test showed a significant effect of the method factor on the RMS distance (left panel), and a Wilcoxon rank-sum test assessed the difference between each pair of distributions relative to the GIF method. All were significantly different ($p < 10^{-4}$). Both IOP-IAIF and GFM-IAIF are shown to outperform IAIF in terms of RMS, indicating the ability of iterative pre-emphasis in general.

However, RMS distance tends to favour the most energetic parts of a spectrum, i.e. the low frequencies in the case of the glottal flow. Therefore, to investigate effects on other parts of the spectrum, we use spectral distortion (*SD*) [7], and mean error E between the three perceptive parameters (defined by Airaksinen et al. [23]):

$$E_{H1H2} = \mathrm{E}\left[H1H2_{LF} - H1H2_{estim}\right] \quad (1)$$

$$E_{HRF} = \mathrm{E}\left[(HRF_{LF} - HRF_{estim})/HRF_{LF}\right] \quad (2)$$

$$E_{ST} = \mathrm{E}\left[ST_{LF} - ST_{estim}\right] \quad (3)$$

These measures are plotted for each method in the lower four rows of Fig. 2, as overall scores (left column) and in terms of their dependencies on R_d, f_0, and F_1 (moving to the right). KWRS tests on each measure (first column) showed a significant effect of the method factor, while Wilcoxon rank-sum tests assessed the difference between each pair of distributions relative to the GIF method, with all judged significantly different ($p < 10^{-16}$). Overall, GFM outperformed the other methods with lower *SD* and smaller perceptive parameter errors while IOP tended to outperform IAIF in all scores apart from E_{H1H2}. Note that while methods perform similarly for low-frequencies (*H1H2*), performance variation increases for mid- (*HRF*) and high-frequency (*ST*) regions. This confirms that RMS distance favours low frequency accuracy, and prompts a question as to which distance measure best matches the error we want to minimise.

KWRS tests applied for each measure and each method distribution for each dependency indicated a significant effect of each parameter for all measures and methods. In terms of sensitivity to R_d, IAIF tended to perform well for low values, eclipsed by IOP and then GFM as R_d, and hence spectral tilt, increased. This highlights the inability of IAIF to encompass varying spectral tilt, as only glottal flows with little spectral tilt were correctly estimated. Conversely, IOP-IAIF and GFM-IAIF better extracted high frequency information in the glottal waveform. However, they both appeared to attempt to extract spectral tilt even when there was none, explaining poorer performance for low R_d values. GFM was also more accurate in terms of parameter f_0, with IAIF and IOP being more sensitive to high and low values of f_0, respectively. Finally, F_1 interaction demonstrated a performance drop for low values, particularly for E_{H1H2}, suggesting that the glottal formant (tightly linked to this), was misdetected when F_1 was within the same order of magnitude as the glottal formant.

4.2. Evaluation on natural speech

The approach used in the literature for GIF evaluation on natural speech is to compute variations of glottal flow parameters

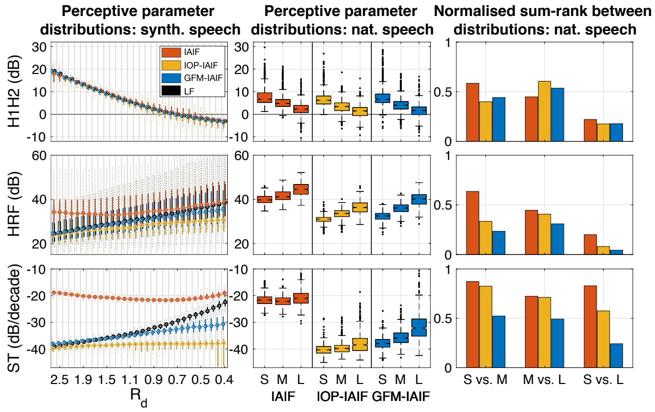

Fig. 3. Spectral parameter distributions for the three methods applied to synthetic speech (left) depending on R_d, and to natural speech (middle) depending on vocal effort (S: soft; M: medium; L: loud). Right: Piecewise comparison by normalised rank-sum distribution. Top to bottom: *H1H2*, *HRF*, and *ST*. Orange: IAIF; Yellow; IOP-IAIF; Blue: GFM-IAIF; Black (synth. speech only): LF ground truth.

and their consistency for changes in voice quality. We follow the approach of Drugman et al. [7], using recordings of 12 different vowels uttered at 3 vocal effort levels (soft, medium, loud), each repeated 26 times, by two German speakers in the de6 (male) and de7 (female) databases[1] [24] (i.e. total 1872 recordings, sampled at 16 kHz). Since ground truth is unavailable for natural speech, direct optimisation of the method parameters was not possible. However, the medians of optimum parameters for each method obtained on the synthetic speech evaluation were used: (For IAIF: $L_v = 14$; $L_g = 3$; $d = 0.98$; For IOP-IAIF: $L_v = 20$; $L_g = 3$; $d = 0.99$; For GFM-IAIF: $L_v = 16$; $L_g = 3$; $d = 0.99$). All methods used a 3$^{\text{rd}}$ order model in glottis fine estimation.

Fig. 3 displays the distribution of *H1H2*, *HRF*, and *ST* (in rows, from top to bottom) previously obtained for synthetic speech depending on R_d, and for natural speech (middle column) depending on vocal effort, for the three methods (Orange: IAIF; Yellow; IOP-IAIF; Blue: GFM-IAIF) and the LF ground truth (black) for synthetic speech. Assuming that the range R_d reflects the variation of vocal effort of natural speech, we observe that the evolution of perceptive parameters follows the same trend and range between synthetic and natural speech, across all methods. In particular, we observe the expected decrease in *H1H2* and increase in *HRF* and *ST* with increasing vocal effort [8, 18, 20]. Hence, this shows a consistency between synthetic and natural speech.

A non-parametric Wilcoxon rank-sum test assessed the significance between different pairs of distributions (soft vs. medium; medium vs. loud; soft vs. loud) for natural speech. All pairs were assessed to be significantly different ($p < 10^{-4}$). The normalised rank-sum calculated from each pair is plotted in Fig. 3 (right column). Lower values denote more distinct distributions, and a greater likelihood that the parameter can discriminate between vocal effort types. We observe that all methods discriminated equally well the vocal efforts regarding *H1H2*. This is consistent with synthetic speech observations where the error in *H1H2* estimation was very small for all methods, thanks to the minimisation of the RMS distance. Regarding *HRF*, GFM-IAIF showed the best discriminative power, followed by IOP-IAIF, then IAIF. Similarly for synthetic speech, the range of *HRF* values is reduced for IOP-IAIF, and more compressed for IAIF, with high *HRF* values. Finally, GFM-IAIF *ST* distributions also show better spread with vocal effort. Again, both on synthetic and natural speech, IOP-IAIF tends to attribute too much spectral tilt for all vocal efforts, while IAIF tends to attribute little spectral tilt for all vocal efforts.

Overall, these results indicate that a stronger pre-emphasis has a role in the preservation of mid and high frequency features while a too weak pre-emphasis (i.e. 1st order in IAIF) does not attribute enough spectral tilt to the glottis. However a too strong pre-emphasis (i.e. the unconstrained order of IOP-IAIF) tends to attribute too much tilt, hence reducing performance.

5. CONCLUSION

This paper has presented and explored a new proposed method for glottal inverse filtering, GFM-IAIF, which ensures a third order filter in the pre-emphasis step, motivated by spectral glottis source models. Evaluation against standard IAIF and the recently-proposed IOP-IAIF on both synthetic and natural speech showed that while the low frequency region is equally well extracted by the three methods, the choice of a third order filter derived from GFM, led GFM-IAIF to provide the best estimation of both glottal formant and spectral tilt relative to voice quality variation. The performance gain is stronger at high frequencies, matching expectations from the literature [8, 9, 20].

One can note a discrepancy between these results and those from Mokhtari et al. [5]. However that paper estimated the glottal flow with GFM by deconvolving the speech signal with the gross estimated vocal tract instead of the fine estimated vocal tract; leading to a suboptimal version of the GFM method. Another difference is that their evaluation was made on a computational physics model. This contrasts to the evaluation in the current paper which uses methods well established in the literature, conducted on both synthetic and natural speech. GFM-IAIF has been shown to provide good discrimination with vocal quality, and it provides an intuitive way to describe voice quality as well as matching the input parameters of glottal-source synthesis models [11]. We aim to conduct further evaluation on those parameters in future.

[1] https://github.com/numediart/MBROLA-voices/

6. REFERENCES

[1] Gunnar Fant, *Acoustic Theory of Speech Production*, Mouton, 1970.

[2] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[3] Olivier Perrotin and Ian V. McLoughlin, "On the use of a spectral glottal model for the source-filter separation of speech," *arXiv:1712.08034*, December 2017.

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

[5] Parham Mokhtari, Brad Story, Paavo Alku, and Hiroshi Ando, "Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production," *Speech Comm.*, vol. 104, pp. 24–38, 2018.

[6] Parham Mokhtari and Hiroshi Ando, "Iterative optimal preemphasis for improved glottal-flow estimation by iterative adaptive inverse filtering," in *Proc. of Interspeech*, Stockholm, Sweden, August 21-24 2017, pp. 1044–1048.

[7] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.

[8] Donald G. Childers, "Vocal quality factors: Analysis, synthesis and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[9] Boris Doval, Christophe d'Alessandro, and Nathalie Henrich, "The spectrum of glottal flow models," *Acta Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[10] Olivier Perrotin and Christophe d'Alessandro, "Vocal effort modification in singing synthesis," in *Proc. of Interspeech*, San Francisco, CA, USA, September 8-12 2016, pp. 1235–1239.

[11] Lionel Feugère, Christophe d'Alessandro, Boris Doval, and Olivier Perrotin, "Cantor digitalis: Chironomic parametric synthesis of singing," *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.

[12] Christer Gobl and Ailbhe Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Comm.*, vol. 40, no. 1, pp. 189–212, 2003.

[13] Gilles Degottex, *Glottal source and vocal-tract separation: Estimation of glottal parameters, voice transformation and synthesis using a glottal model*, Ph.D. thesis, Univ. Pierre et Marie Curie (UPMC), Nov. 2010.

[14] Paavo 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.

[15] John Makhoul, "Linear prediction: A tutorial review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[16] Boris Doval, Christophe d'Alessandro, and Benoit Diard, "Spectral methods for voice source parameters estimation," in *Proceedings of Eurospeech*, Rhodes, Greece, September 22-25 1997, pp. 533–536.

[17] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "COVAREP – a collaborative voice analysis repository for speech technologies," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Florence, May 4-9 2014, pp. 960–964.

[18] Gunnar Fant, "The LF-model revisited. transformations and frequency domain analysis," Quarterly Progress and Status Report 2-3, Royal Institute of Technologies - Dept. for Speech, Music and Hearing, 1995.

[19] Dennis H. Klatt and Laura C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.

[20] Sirisha Duvvuru and Molly Erickson, "The effect of change in spectral slope and formant frequencies on the perception of loudness," *Journal of Voice*, vol. 27, no. 6, pp. 691–697, 2013.

[21] Gunnar Fant, Johan Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," Quarterly Progress and Status Report 4, Royal Institute of Technologies - Dept. for Speech, Music and Hearing, 1985.

[22] Bernard Gold and Lawrence Rabiner, "Analysis of digital and analog formant synthesizers," *IEEE Trans. Audio and Elect.*, vol. 16, no. 1, pp. 81–94, 1968.

[23] Manu Airaksinen, Tuomo Raitio, Brad Story, and Paavo Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE Trans. on Audio, Speech, and Signal Processing*, vol. 22, no. 3, pp. 596–607, March 2014.

[24] Marc Schroder and Martine Grice, "Expressing vocal effort in concatenative synthesis," in *Int. Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, August 3-9 2003, pp. 2589–2592.