SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

Transformation of voice quality in singing using glottal source features

João P. Cabral[1], *Alexsandro R. Meireles*[2]

[1]The ADAPT Research Centre, Trinity College Dublin, Ireland
[2]Federal University of Espírito Santo, Brazil

cabralj@scss.tcd.ie, meirelesalex@gmail.com

Abstract

Glottal activity information can be very important in several speech processing applications, such as in speech therapy, voice disorder diagnosis, voice transformation and text-to-speech synthesis. However, the use of algorithms for estimating glottal parameters from the speech signal is very limited in those applications because of problems with robustness and accuracy. In singing synthesis, the glottal source representation is also very important because it is closely related with the emotions and singing style. This paper proposes a robust method to estimate the voice quality parameters of the glottal source by using both the electroglottographic (EGG) signal and the acoustic recordings of singing voice for five vowels in three different voice qualities: modal, breathy and creaky. The analysis of the resulting measurements permitted to confirm that voice quality parameters of the glottal source are correlated with the type of voice. Moreover, another experiment was conducted to show that it is possible to transform the modal singing voice into breathy and creaky by using an analysis-synthesis method that incorporates a glottal source model.

Index Terms: singing, voice transformation, glottal source analysis, distorted voices, EGG signal

1. Introduction

The use of a distorted voice[1] (henceforth DV) in the production of speech sounds is widely used in speech communication. Some of the terms used for DVs are distorted tones, distortion, overdrive, drive, creak, fry, breathy voice, creaky voice, growl, grunt, scream, phaser, rattle, snarl, fry scream (cf. [1, 2, 3, 4, 5, 6, 7, 8, 9]). As can be seen, some of the terms for DVs are based on impressionistic views of the resulting sound and others are based on physiological descriptions of the vocal tract. Humans use these vocal effects since they are born and keep using them throughout their life for manifesting emotions such as anger, terror, fear, and sadness. Some common examples are the baby cry[2], the human screams[3], and the voice tremulations[4] in fear. Although the occurrence of DVs is pretty common in our daily life, much of the time we are not conscious of these uses. Ask a person, for example, if he/she ever produced a laryngeal DV setting such as in the song Enter Sandman by Metallica[5], she will probably say no and comment that it is typical of this metal genre. Nevertheless, we do use this DV setting when we are angry, for example.

Despite these considerations, most of the time we use modal voice with some instances of creaky voice in voice quality production. Yet, there are some professionals such as actors and singers that intentionally change their most common voice quality in order to try to evoke certain emotions on the listeners. Take for example the vocal performances of actors such as the American actor and comedian Mel Blanc[6], known as the man of 1,000 voices, and the Brazilian actor and comedian Chico Anysio[7] who created hundreds of characters with unique voices.

Although actors use a wide variety of distorted tones in their performances, unless they use a fixed DV or vocal setting (see [10]) for creating a character (refer to the voice quality productions of Mel Blanc and Chico Anysio), they mix these settings with his/her most common voice quality. Similarly, some singers use distorted tones in their songs to create different moods, but in certain musical genres such as death metal they use a DV throughout a song (see for example the use of the ventricular folds in songs by Sepultura[8], and Death[9]).

As we have pointed out, the use of DVs is very common and is used with certain specific intentions in theater and music, although its use is regularly associated with the voices in the rock and/or metal genre because of the greater variety of DVs found in these genres. Despite of that, it occurs in any musical style (country, soul, blues, jazz), even in classical music (see [11]).

As known by singers, much of the time the terminology to describe the DVs is based on impressionistic views that may vary a lot among vocal coaches. In singing literature we find many terms that sometimes have the same name and correspond to different sounds and also the opposite. That is why speech scientists need to describe the DVs according to their physiological and/or acoustical settings. As far as we know, the researcher who best describes the variety of DVs in singing is the Brazilian vocal coach Ariel Coelho. In his course [4][10], he proposes to describe all the DVs based on physiological settings.

2. Synthesis of singing

Although significant work can be found in the literature on singing synthesis, only recently attention has been given to the synthesis of DVs. Most of the time the authors deal with modal voices and the synthesis of certain aspects of the acoustics of singing such as vibrato, overshoot, voice fluctuations [12], and vibrato alone [13]. There are also certain authors that associate acoustics with EGG measurements ([14, 15, 16, 17, 18]).

Gentilucci and colleagues [19] are some of the authors that have synthesized a DV. They proposed a software tool to recre-

[1]The term Distorted Voice (DV) is used in this paper to refer to any variety of voice quality (see Laver, 1980, [10]) production, either laryngeal or supralaryngeal, that differs from modal phonation (eg. creaky voice, whispered voice, ventricular folds phonation).

[2]https://www.youtube.com/watch?v=6Lp-h1S3rMk

[3]https://www.youtube.com/watch?v=HRs9tI44Ejs

[4]https://www.youtube.com/watch?v=iQISI7DOVCY

[5]https://www.youtube.com/watch?v=CD-E-LDc384

[6]https://www.youtube.com/watch?v=ZeAM1vwEcFg

[7]https://www.youtube.com/watch?v=ZeAM1vwEcFg

[8]https://www.youtube.com/watch?v=ZeAM1vwEcFg

[9]https://www.youtube.com/watch?v=ZeAM1vwEcFg

[10]Unfortunately, his course is only available in Brazilian Portuguese.

10.21437/SMM.2019-7

ate or magnify in real-time a "distorted part" of the voice signal based on acoustic data, in order for a person to sing with less vocal effort. Also, Cosi and Tisato [20] have synthesized the overtone singing based on acoustic data.

As can be seen, the synthesis of DVs is in its beginning and there is much to do in the field. One important task is to develop a method for identifying the different DVs that occur in singing based on acoustic and physiological aspects. Then, it is also necessary to develop algorithms and software applications that can convert the modal voice into a DV. There is a potential market opportunity for this solution in the area of singing synthesis, since although there are software tools that convert the voice in singing environments (see VoiceSynth[11] and Virsyn[12]), as far as we know, there is no software that can change the modal voice into several types of DVs. Therefore, our initial work in this paper aims to contribute to advances in this promising field of research.

According with the speech production model, speech can be synthesized by passing the glottal source signal through the synthesis filter that represents the vocal tract. The radiation effect in the lips can be modelled by a simple differentiation operation. This linear source-filter model assumes that the source and vocal tract components are independent and it can be generalized to singing synthesis because the two processes are similar, e.g. [21]. The glottal source component is very important because it carries both linguistic (e.g. expressed in the perceived pitch) and non-linguistic information (e.g. perceived emotions, voice identity characteristics of the speaker, and voice qualities). In this work, the two components are estimated from singing recordings in order to analyze the voice quality correlates of the glottal parameters and to transform a modal voice into breathy and creaky voices by manipulating those parameters. We focus in these voices, because they are very common in speech and singing. Also, several studies have showed that important glottal parameters can be robustly estimated for these voice qualities and measured their acoustic correlates, so we can compare our findings with those those works.

3. The glottal source model

The Liljencrants-Fant (LF) model [22] is an acoustic model of the glottal source derivative, which is shown in Figure 1.

The LF-model is defined by six shape parameters: t_c, t_p, t_e, T_a, T_0, and E_e. The LF-model is often represented by a simplified version defined by five parameters, in which the instant of complete closure, t_c, is set to the period T_0. In this work, this simplified LF-model was used.

The LF-model can also be described by shape parameters related to voice quality properties [23]. The most important are the open quotient $OQ = (t_e + T_a)/T_0$, speed quotient $SQ = t_p/(t_e - t_p)$, and the return quotient $RQ = T_a/T_0$. OQ measures the relative duration of the glottal pulse, SQ is related to the symmetry of the glottal pulse and RQ is mainly correlated with the spectral tilt characteristic of the glottal signal. The typical range of the OQ is between 0.3 and 0.8, while SQ is between 1.5 and 4. The values of the RQ are much lower because the return phase is usually a small fraction of the pitch period. These parameters represent different quotients, which describe specific properties of the source signal. For example, the open quotient can also be defined by the reduced form $OQ_e = t_e/T_0$, in which the return phase is not included in the open phase.

[11] https://www.voicesynth.com/

[12] http://www.virsyn.de/de/Home/home.html

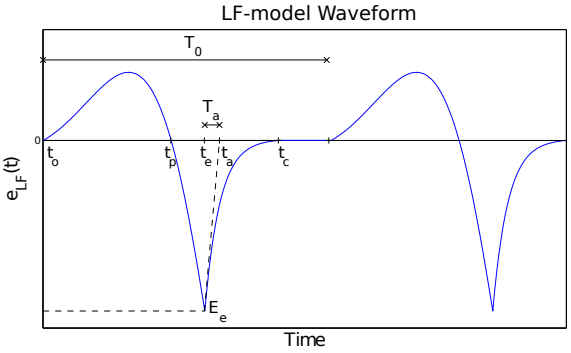

Figure 1: *Segment of the LF-model waveform and its shape parameters:* t_o, t_c, t_p, t_e, T_a, T_0, *and* E_e.

The shape parameters of the LF-model have been widely used to study the voice quality of speech signals [23, 24, 25, 26, 27]. The relations between the LF-parameters and two major types of voice, breathy and creaky, are summarized below:

- *Breathy*: Highly abducted phonation, which results in a high OQ. Typically, there is a slow glottal closure (high RQ). There is an high symmetry of the glottal pulse that corresponds to a small SQ.
- *Creaky*: Very adducted phonation (short glottal open interval) [25], with a small OQ and low RQ (short return phase). The asymmetry of the glottal pulse is also large (as well as the SQ) when compared with the modal voice (neutral voice quality).

4. Analysis-synthesis method

4.1. Glottal source analysis

Glottal source analysis is a complex and difficult problem because the glottal source signal can not be directly measured. A common approach in speech processing applications is to estimate the source signal from recorded speech and then extract the glottal parameters from the resulting signal. This can be done by using inverse filtering techniques such as pre-emphasis inverse filtering [28], iterative adaptive inverse filtering (IAIF) [29], and closed-phase inverse filtering [30]. The problem is that existing techniques can not separate well the two components, because they rely on assumptions such as the linearity of the source-filter model and other approximations which limit the accuracy and robustness of the source separation. It is also possible to estimate source parameters directly from the speech signal. The fundamental frequency ($F_0 = 1/T_0$) is a popular acoustic feature that can be robustly estimated from speech, but other glottal parameters that are correlated with voice quality can also be measured from speech, e.g., the R_d shape parameter of the LF-model [31]. However, such parameters only represent particular characteristics of the glottal source and, in general, do not permit sufficient control over glottal parameters for transformation of different voice qualities.

Electroglottography is a popular non-invasive measurement of vocal fold motion that is used to obtain robust estimation of two glottal parameters, the glottal opening (GOI) and closing (GCI) instants. The GCI corresponds to the time instant t_e in the LF-model, while GOI is the t_o. The glottal parameter es-

timation approach in this work is based on a two-way channel analysis. First, the glottal opening and closing instants are estimated using the EGG signal. Then, the other parameters of the LF-model are estimated from the speech signal by using inverse filtering and automatic parameterization of the resulting glottal source signal. The idea is to use the accurate parameter estimates from EGG to constrain and consequently improve the speech analysis method.

The EGG analysis method used in this work is similar to that described in [32]. The GCI and GOI are estimated by detecting the strongest negative peak in the derivative of the EGG (DEGG) and the highest positive peaks on the DEGG, respectively. This is done with the peak detection algorithm called *peakdet* [33], which is available in the *covarep* toolbox [34].

The other LF-model parameters (t_p, T_a, and E_e) are estimated from the speech/singing signal as described in [35]. First, an estimate of the glottal source derivative is computed using the IAIF technique [29]. The GCIs are then estimated from the speech signal by using the RAPT algorithm [36, 37] and they are aligned with the epochs estimated from EGG using an algorithm developed in previous work [38]. After this process, the epochs that are not aligned are removed as well as those that do not satisfy constraints on the pitch period (within the range of 50 Hz to 500 Hz). The GCIs permit to perform the analysis pitch-synchronously on each segment delimited by contiguous epochs. For each pitch cycle, the LF-model waveform is fitted to the glottal source derivative using a non-linear optimization algorithm. This LF-parameter estimation method is described in more detail in [35].

Another important component of the glottal source signal is the aspiration noise. The main characteristic of the aspiration noise is the time-modulation effect that shapes its energy envelope. It is important to model this noise component well to produce high-quality voice in synthesis or voice transformation applications. For example, in the breathy voice, typically, the vocal folds do not close completely producing aspiration noise. In this work, this noise component is estimated and parameterized as in [39]. Basically, it uses an Harmonic/Stochastic Model (HSM) to extract the stochastic component from the speech signal and then obtain the aspiration noise by inverse filtering to remove the vocal tract component from the noise signal. This method also parameterizes the aspiration noise using a triangular function to represent its amplitude envelope and the Harmonics-to-Noise Ratio (HNR).

4.2. Vocal tract spectrum estimation

The vocal tract filter is estimated using the Glottal Spectral Separation (GSS) method [40]. It consists of removing the spectral effects of the source model from the speech/singing signal $s(t)$, i.e. $H(w) = S(w)/|E_{LF}(w)|$, where $|E_{LF}(w)|$ is the amplitude spectrum of one period of the LF-model signal and $S(w)$ is the FFT spectrum of $s(t)$. Finally, the amplitude spectrum of the vocal tract filter is calculated by computing the spectral envelope of $|H(w)|$. The spectral envelope is computed using the analysis method of the STRAIGHT vocoder [41]. For unvoiced speech/singing, the spectrum is represented by the spectral envelope of STRAIGHT, without performing the first LF-model separation step. In this work, all the recorded singing samples are voiced (correspond to vowels).

4.3. Synthesis

Voiced sounds are synthesized from the GSS parameters based on the method described in [40], which consists of multiplying the amplitude spectrum of a mixed excitation signal by the vocal tract spectrum. Then, the singing waveform is obtained by computing the inverse FFT of the spectrum of the synthesized signal. The synthesis method used in this work is slightly different because it does not perform the synthesis in the frequency domain. Instead, it converts the FFT coefficients to coefficients of an FIR filter using the autocorrelation function. Then, singing is synthesized by passing the mixed excitation model through this linear-phase vocal tract filter. The reason for using the filtering operation is that it produces smooth variations between frames. The excitation of voiced speech is produced by adding two pitch cycles of the LF-model signal with the noise signal that is scaled in energy using the HNR parameter. Finally, the synthetic short-time signals are overlapped-and-added using asymmetric Hanning windows that add to one, in order to obtain smooth transitions between speech frames.

Table 1: *Mean values of glottal source parameters and HNR, measured for the three voice qualities.*

	F0 (Hz)	**OQ**	**SQ**	**RQ**	**HNR** (dB)
Modal	86.5	0.44	2.32	0.040	23.23
Breathy	152.9	0.58	2.28	0.088	11.25
Creaky	71.1	0.34	2.62	0.017	26.5

5. Voice transformation experiment

5.1. Singing database

The second author (male) recorded five sung vowels in three different voice qualities: Modal, breathy and creaky. In addition, the EGG signals were recorded simultaneously with the singing signal. This data collection was made with an EG2-PCX[13] which allows the synchronized acquisition of EGG and voice signals (both sampled at 44.1 kHz).

5.2. Glottal parameter measurements

Both the voice and EGG signals were downsampled to 16 kHz. The analysis of glottal source parameters was performed as described in Section 4.1. First, the GCIs were estimated from the EGG signal and then manually verified. The manual GCI correction generally consisted of removing the GCIs that were incorrectly detected in silence regions. Then, the GOIs were estimated using the peakdet algorithm, by constraining the detection of the GOI for each segment delimited by two consecutive GCIs. From the EGG analysis, $F_0^i = 1/T_0^i$ was obtained by calculating the duration of the frame i delimited by two consecutive epochs (this duration corresponds to T_0^i). The OQ parameter was also calculated for each pitch cycle as $OQ_e^i = t_e^i/T_0^i$, where t_e^i is the duration between GOI and GCI for the frame i. The other voice quality parameters of the LF-model, $SQ^i = t_p^i/(t_e^i - t_p^i)$ and $RQ^i = T_a^i/T_0^i$, were calculated by using t_e^i from EGG analysis and the additional intervals t_p^i and T_a^i estimated in the speech analysis stage, respectively.

The average values of the LF-parameters measured for the different voice qualities are given in Table 1. Meanwhile, Table 2 shows the deltas variations of the mean values of the glottal parameters which were calculated from Table 1 to transform a modal voice into breathy and creaky respectively. The scale factors of the HNR parameter were also calculated for transfor-

[13] http://www.glottal.com/Electroglottographs.html

Table 2: *Delta variation factors of the mean values of the glottal parameters to transform modal into breathy and creaky voices.*

	F0(Hz)	**OQ**	**SQ**	**RQ**
Breathy	66.46	0.14	-0.04	0.05
Creaky	-15.35	-0.10	0.30	-0.02

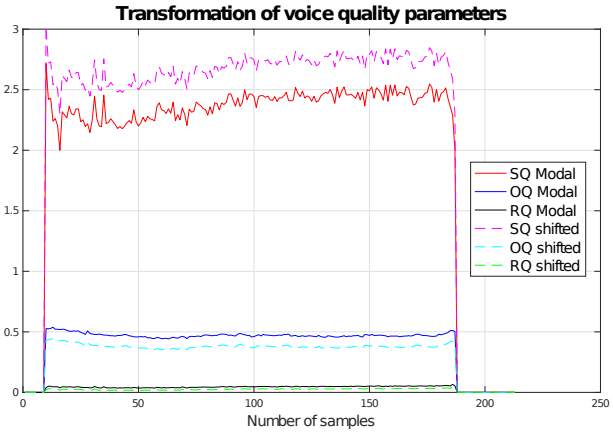

Figure 2: *Trajectories of glottal parameters estimated for the vowel /a/ sung in modal voice and the transformed trajectories (by shifting) for synthesis of the same vowel with creaky voice.*

mation of modal into breathy and creaky voices. They are 0.48 and 1.14 respectively.

5.3. Transformation of voice quality

The vectors of the glottal parameters values for each vowel sung with a modal voice were transformed by adding the constant delta values of Table 2 to the respective vectors. Figure 2 shows an example of the voice quality parameters estimated for the vowel /a/ of modal voice and the trajectories obtained after transformation to synthesize the vowel with creaky voice. Next, the new values of the time-domain parameters of the LF-model (t_e, t_p, and T_a) were calculated using the transformed voice quality quality parameters. The LF-model waveform was generated from the glottal parameters (T_0, t_e, t_p, T_a and the unchanged E_e) for each frame and added to the noise component. The energy of the noise component is changed using the scaling factor of the HNR parameter calculated to transform modal voice into the given target voice quality. Finally, the synthetic singing waveform is generated by passing the excitation signal through the vocal tract filter. The recorded and synthesized samples have been made available at `https://www.scss.tcd.ie/~cabralj/samples-voice-transformation.html`.

6. Discussion

The variations of the glottal parameters between the modal voice and the two other voice qualities are in concordance with previously reported results obtained for speech [23, 24, 25, 26, 27]. This result is expected because there is a close relationship in the production of speech and singing sounds.

The authors of this paper performed an informal perceptual evaluation of the synthetic speech and found that the perceptual quality of the vowels synthesized with the LF-model is very natural. By adding the noise component of the excitation to the periodic LF-model signal, the synthetic voice sounds even closer to the recorded voice. However, the noise also introduces some perceptual artifacts which are stronger in the synthesized breathy voice. A better noise model is needed to reduce this distortion.

The transformation of modal into creaky sounds surprisingly very good, given that only the LF-model parameters are transformed without taking into account the variation of other possible properties of the signal correlated with this voice quality. For example, it is expected that changes in the vocal tract and other source characteristics, such as F0 perturbations and amount of aperiodicity, also reflect the variations in the voice quality. The synthesized breathy voice is not so close to the recorded breathy voice as we expected. We believe that it is necessary to transform additional parameters which are important for this type of voice, especially parameters related with the aspiration noise model. Nevertheless, the perceived voice quality of the transformed signal is clearly different from the original modal voice and there is no clear perceptual distortion without using the noise component of the excitation.

Although this preliminary study is limited to a few samples of isolated vowels and only one speaker, it is supportive of our hypothesis that glottal parameters are important for synthesis of different types of singing voice. As future work, we plan to extend the parameter transformations to take into account other relevant features such as the vocal tract parameters, parameters representing dynamic changes of the F0 contour (e.g. jitter and F0 range), and additional noise component parameters (e.g. parameters of the noise energy envelope). Also, we are conducting ongoing experiments to evaluate the quality of the synthetic voice transformations and recording more data for further analysis and evaluation.

7. Conclusion

This paper estimated the voice quality parameters of the glottal source by using both the EGG signal and the acoustic recordings of singing voice for five vowels with three different voice qualities: modal, breathy and creaky. From our own perceptual evaluation of samples produced with the analysis-synthesis method that incorporates a glottal source model, we verified that this can be used as a very powerful tool to work with the DVs used in speech/singing. As we have discussed here, the breathy voice needs further improvements with additional parameters so as to sound closer to the quality of the recorded voice. After the adjustments for transforming the modal voice into breathy and creaky voices, we intend in the future to explore other challenging voice transformations that include voice qualities such as vocal fry, ventricular folds phonation, vocal folds combined with ventricular folds phonation, and noisy phonation.

8. Acknowledgments

The first author is supported by the Science Foundation Ireland (Grant 13/RC/2106) as part of ADAPT (www.adaptcentre.ie), at Trinity College Dublin, and by a New Horizons grant from the Irish Research Council entitled “The COG-SIS Project: Cognitive effects of Speech Interface Synthesis” (Grant R17339). The second author is supported by Espírito Santo Research Foundation (FAPES, grant 221/2017).

9. References

[1] Cross, M., "The Zen of Screaming," Director: Denise Korycki. 1 DVD and 1 CD (145 min). NTSC, Color. New York: Loudmouth Inc, 2007.

[2] Sadolin, C., "Complete Vocal Technique." Denmark: CVI Publications, 2012.

[3] Vendera, J., Raise your voice: the advanced manual. Ohio, USA: Vendera Publishing, 2013.

[4] Coelho, A. "Drives Vocais - Memórias de Acesso". Accessed in June 29, 2019. Available at `https://pages.hotmart.com/v5564914p/drives-vocais-memorias-de-acesso/`, 2019.

[5] Borch, D. Z., Sundberg, J., Lindestad, and P-Å, Thalén, M. "Vocal fold vibration and voice source aperiodicity in 'dist' tones: a study of a timbral ornament in rock singing," Logop. Phoniatr. Vocology, vol. 29, no. 4, 147–153, 2004.

[6] Guzman, M., Barros, M., Espinoza, F., Herrera, A., Parra, D., and Muñoz, D. "Laryngoscopic, acoustic, perceptual, and functional assessment of voice in rock singers," Folia Phoniatr. Logop., vol. 65, no. 5, 248–56, 2014.

[7] Caffier, P. P., Ibrahim, N. A., Ropero R. M. del M., Wienhausen, S., Forbes, E., and Seidner, W. "Common vocal effects and partial glottal vibration in professional nonclassical singers," Journal of Voice, vol. 32, no. 3, 340–346, 2018.

[8] Izdebski, K., Blanco, M., Di Lorenzo, E., and Yan, Y. "High speed digital phonoscopy of selected extreme vocalization (Conference Presentation)," In: SPIE BIOS, Optical Imaging, Therapeutics, and Advanced Technology in Head and Neck Surgery and Otolaryngology, San Francisco, USA, vol. 10039, no. 9, 2017.

[9] Sundberg, J., The science of the singing voice. Dekalb, IL: Northern Illinois University Press, 1987.

[10] Laver, J., The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press, 1980.

[11] Edgerton, M., The 21st-Century Voice: Contemporary and Traditional Extra-Normal Voice. Lanham: Scarecrow Press, 2005.

[12] Saitou, T., Masashi, U., and Masato, A., "Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis". Speech Communication, vol. 46, no. 3, pp. 405-417, 2005.

[13] Meron, Y. and Hirose, K. "Synthesis of vibrato singing," ICASSP '00., 2000.

[14] Deshpande, P. S. and Manikandan, M., S., "Effective Glottal Instant Detection and Electroglottographic Parameter Extraction for Automated Voice Pathology Assessment," IEEE Journal of biomedical and health informatics, 22(2), 2018.

[15] Bouzid, A. and Ellouze, N., "Voice source parameter measurement based on multi-scale analysis of electroglottographic signal," Speech Communication, vol. 51, pp. 782–792, 2008.

[16] Bernardoni, N. H., DAlessandro, C., Doval, B., and Castellengo, M., "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," Journal of the Acoustical Society of America, vol. 117, no. 3, pp. 1417–1430, 2005.

[17] Sapienza, C. M., Stathopoulos, E. T., and Dromey, C., "Approximations of Open Quotient and Speed Quotient from Glottal Airflow and EGG Waveforms: Effects of Measurement Criteria and Sound Pressure Level," Journal of Voice, vol. 12, no. 1, pp. 31–43, 1998.

[18] Chen, Y., Robb, M., and Gilbert, H. R., "Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation," Journal of Speech, Language, and Hearing Research, vol. 45, pp. 821-829, October, 2002.

[19] Gentilucci, M., Ardaillon, L., and Liuni, M., "Composing Vocal Distortion: A Tool for Real-Time Generation of Roughness," Computer Music Journal, vol. 42, Issue 4, Winter, pp. 26–40, 2018.

[20] Cosi, P. and Tisato, G., "On the magic of overtone singing," Voce, Parlato. Studi in onore di Franco Ferrero, pp. 83–100, 2003.

[21] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing – Acoustical analysis and physiological interpretations," the Fourth F.A.S.E. Symposium on Acoustics and Speech, vol. 2, pp. 50–70, 1981.

[22] Fant, G., Liljencrants, J., and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, vol. 26, no. 4, pp. 1–13, 1985.

[23] Childers, D. G., "Glottal Source Modelling for Voice Conversion," Speech Communication, vol. 7, no. 6, pp. 697–708, 1995.

[24] Fant, G., "The LF-model revisited. Transformations and frequency domain analysis," STL-QPSR Technical Report, pp. 119–156, Royal Institute of Technology, 1995.

[25] Gobl, C., "A preliminary study of acoustic voice quality correlates," STL-QPSR, KTH, Sweden, 1989.

[26] Keller, E., "The analysis of voice quality in speech processing," Lecture notes in computer science, vol. 3445, pp. 54–73, 2005.

[27] Alku, P., Strik, H., and Vilkman, E, "Parabolic spectral parameter: a new method for quantification of the glottal flow," Speech Communication, vol. 22, no. 1, pp. 67–79, 1997.

[28] Cabral, J. P. and Oliveira, L. C., "Pitch-synchronous time-scaling for prosodic and voice quality transformations", Proc. of INTERSPEECH, pp. 1137–1140, 2005.

[29] Alku, P. and Vilkman, E., "Estimation of the glottal pulseform based on discrete all-pole modelling", Proc. of ICSLP, Japan, 1994.

[30] Wong, D. Y., Markel, J. and Gray, Jr. A. H., "Least squares glottal inverse filtering from acoustic speech waveform," IEEE Trans. Acoustics, Speech and Signal Processing, vol. 27, no. 4, pp. 350–355, 1979.

[31] Degottex, G., Roebel, A. and Rodet, X., "Function of phase-distortion for glottal model estimation," Proc. ICASSP, pp. 4608–4611, 2011.

[32] Cabral, J. P., "Estimation of the asymmetry parameter of the glottal flow waveform using the Electroglottographic signal," INTERSPEECH, 2018.

[33] Mazaudon M. and Michaud A., "Tonal Contrasts and Initial Consonants: A Case Study of Tamang, a Missing Link in Tonogenesis," Phonetica, vol. 65, pp. 231–256, 2008.

[34] Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S., "COVAREP – A collaborative voice analysis repository for speech technologies," ICASSP, pp. 960–964, 2014.

[35] Cabral, J. P., Renals, S., Richmond, K., and Yamagishi, J., "Glottal Spectral Separation for Parametric Speech Synthesis," INTERSPEECH, pp. 1829–1832, 2008.

[36] Talkin, D., "A robust algorithm for pitch tracking (RAPT)," In chapter 14 of Speech Coding and Synthesis, Elsevier Science, W. B. Kleijn and K. K. Paliwal, pp. 495–518, 1995.

[37] Talkin, D. and Rowley, J., "Pitch-synchronous analysis and synthesis for TTS systems," Proc. of ESCA Workshop on Speech Synthesis, pp. 55–58, 1990.

[38] Cabral, J. P., Kane, J., Gobl, C., and Carson-Berndsen, J., "Evaluation of glottal epoch detection algorithms on different voice types," INTERSPEECH, pp. 1989–1992, 2013.

[39] Cabral, J.P. and Carson-Berndsen, J., "Towards a Better Representation of Glottal Pulse Shape Characteristics in Modelling the Envelope Modulation of Aspiration Noise," Advances in Nonlinear Speech Processing (NOLISP 2013), Lecture Notes in Computer Science, vol. 7911, 2013.

[40] Cabral, J. P., Renals, S., Richmond, K., and Yamagishi, J., "Glottal Spectral Separation for Speech Synthesis," IEEE Journal of Selected Topics in Signal Processing: Special Issue on Statistical Parametric Speech Synthesis, vol. 8, no. 2, pp. 195–208, 2014.

[41] Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, vol. 27, pp. 187–207, 1999.