Deconvolution of the Glottal Pulse Using a Finite-Difference Solution of the Acoustical Klein-Gordon Equation

H. S. Kalsi
Dept. of Informatics
King's College London
E-mail: hardial.kalsi@kcl.ac.uk

E. R. Pike
Dept. of Physics
King's College London
E-mail: roy.pike@kcl.ac.uk

Z. Cvetković
Dept. of Informatics
King's College London
E-mail: zoran.cvetkovic@kcl.ac.uk

***Abstract*—Deconvolution of the glottal-pulse waveform from the speech signal remains an active field of research although dating back over half a century. In the main, existing approaches use classical inverse filtering frequency-domain methods to estimate both the vocal-tract and glottal-pulse waveforms. In this paper, we adopt a new approach which takes advantage of two relatively recent developments: firstly, the physical modeling of the speech process by means of the Klein-Gordon wave equation of relativistic quantum mechanics and, secondly, a finite-difference calculation of this equation to find the impulse response of the vocal tract. This approach allows accurate parameterisation of the impulse response which simplifies the blind deconvolution. Results show considerable improvement compared with existing algorithms when applied to synthetic speech where the ground truth is known.**

I. Introduction

Knowledge of the glottal pulse and the impulse response of the vocal tract is immeasurably useful in a number of speech-related problems, including synthesis [1], recognition [2] and laryngology [3]. A crucial requirement for authentic synthetic speech is knowledge of the glottal pulse [4], [5], and in turn the interaction between the glottal-pulse waveform and the vocal tract [6]–[8]. Research has shown that much of the naturalness of speech arises from the source signal and also how the tract influences it, and thus the naturalness of synthesised speech can be improved by incorporating this interaction. The vocal source gives attitude, emotion and stress to the speech signal, which can be described as naturalness [9], [10]. Consequently, using an adequate glottal pulse can lead to more natural sounding, human-like speech synthesis. In the case of laryngology, obtaining the waveform of a glottal pulse via deconvolution would be a non-invasive way of medical intervention and a rapid method for understanding how the vocal folds are operating. Dysphonia, the medical term for vocal disorders, which can signal life-threatening conditions, can be diagnosed by examining the vocal folds [11]. Unlike common invasive methods, deconvolution operations would simply require the patient to speak into a microphone, saving resources and eliminating the need for surgical procedures.

The speech production process, within a stationary part of a given vowel, historically has been described as a linear time-invariant (LTI) system due to the assumed linearity between its inputs (glottal pulse and impulse response) and output (speech signal), and the assumed time invariance of such inputs with respect to the output [8]. Consequently, human speech is the convolution of the two components: the source of excitation, namely the glottal-pulse train, and the impulse response of the vocal tract. If an excitation signal, *i.e.* glottal pulse train, *p(t)* is applied to an LTI system with impulse response *h(t)*, then the output signal *s(t)* is:

$$s(t) = (p * h)(t) = \int_{-\infty}^{+\infty} p(\tau)h(t-\tau)\,d\tau \qquad (1)$$

where $*$ denotes convolution.

Finding a glotal pulse train $p(t)$ from an observed speech signal $s(t)$ amounts to a blind deconvolution, which is an ill-posed problem. Consequently, reaching the desired solution requires supplementary constraints and control. Clearly, as much a priori information as possible must be known about the two functions which are the components of the convolution to obtain the desired, correct solution.

Constraints which are typically used stem from existing models for the glottal signal, including Klatt and Klatt [12], Rosenberg [13], Liljencrants-Fant (LF) [14] and CALM [15]. These models describe the speech source via different variables, however there are some consistencies which are seen through all of them, such as: the glottal flow is quasi-periodic, either positive or null and a continuous function of time. Such constraints are used in a number of existing methods to extract the glottal pulse including IAIF (Iterative Adaptive Inverse Filtering) [16], closed phase co-variance [17], more recent additions of Monte Carlo methods [18] and a number of others [19]–[24].

However, no existing set of constraints used in glottal pulse deconvolution is sufficient to disambiguate the problem completely, in a manner which is verifiably accurate, and thus lead a to a unique solution. We propose a method for constraining the problem by using physical modeling of the vocal tract and excitation source, via the acoustical Klein-Gordon equation, as the foundation of our work. This gives an impulse response which is paramaterized by some unknown, purely physical

parameters which when found, for a given vowel signal, accurately estimates the vowel's impulse response and in turn it's glottal waveform.

II. Vowel Production Process

The vocal tract curvatures which give rise to different vowels can be modeled as a collection of potential barriers and wells, such as those seen in the relativistic form of the Schrödinger equation known as the Klein-Gordon equation [25]. Consequently the speech wave is actually defined by the varying curvatures (second spatial-derivatives) of the area along the vocal tract. The result is a form of the Klein-Gordon equation named the acoustical Klein-Gordon (AKG) equation, given by (2):

$$\left\{\frac{1}{c^2}\partial_{tt} - \partial_{xx} + U(x)\right\} u(x,t) = 0 \tag{2}$$

where c is the speed of sound, u is the pressure as a "wave function" and $U(x)$ is the potential function representing the curvature of the tract. This second-order differential equation models the propagation of pressure in space and time in the vocal tract, and thus the solution of this propagation at the mouth gives the speech signal.

There are only 6 positions from the glottis to the mouth where a curvature of the vocal tract must be present in order to produce any of the 27 IPA (International Phonetic Alphabet) spoken vowel sounds of all the world's languages [26]. For example, the Schwa is the simplest of the 27 spoken vowel sounds and it is produced by having an open mouth with no curvature of the vocal tract, much like sound propagation through an open-ended pipe.

The impulse response of the vocal tract is thus completely specified by only three parameters: the length of the vocal tract, the amount of opening which is present at the mouth (hereon called the mouth barrier - commonly known as *lip radiation*) and the curvature of the vocal tract (the set of potential barriers and wells, $U(x)$, in the AKG equation). An example of $U(x)$ is shown in Figure 1. The vocal-tract curvature which results from the collection of potentials is shown in red.

In the AKG equation, the lip radiation is governed by the height of the potential barrier at the mouth. This regulates the scattering of the acoustic potential function as it meets the vocal tract-mouth barrier and acts as a differentiator [27]. In the AKG equation, the mouth acts as an exact differentiator (where the output signal at the mouth is the precise differential of the signal prior to it leaving the vocal-tract) only for a precise barrier height. The potential setup of each vowel is known via the AKG equation [26], however the two unknowns for a recorded speech signal are the vocal-tract length and the value for the mouth barrier. Once these are found, the impulse response of the vocal tract for a recorded speech signal is completely specified. This paper focuses on the Schwa vowel due to the simplicity of the corresponding curvature function. Generalisations to other vowels are straightforward.

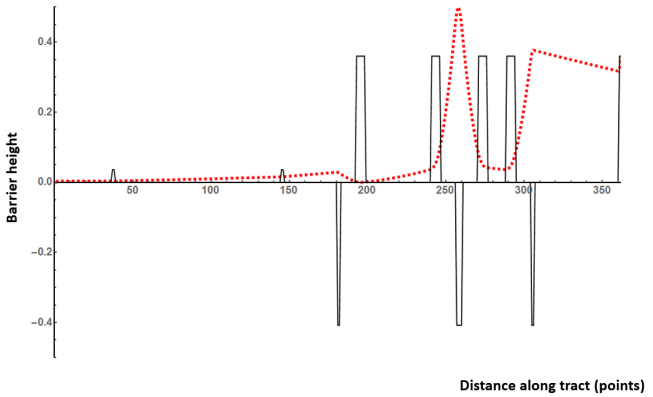

Fig. 1. An example of the potential function U(x) (black) and its corresponding tract curvature (red, dashed)

In order to find an impulse response, an analytical form of the Schwa vowel can be obtained by using the frequency Green's function for the AKG equation. This Green's function for the pressure wave at the exterior boundary of the barrier at the mouth, $G(k)$, can be calculated [26] and has the form:

$$\frac{4e^{i\Delta(k-\kappa)}k\kappa}{(-e^{-2i\Delta\kappa}(k-\kappa)^2 + (k+\kappa)^2)(1 - \frac{e^{-2iak}(1-e^{-2i\Delta\kappa})(k^2-\kappa^2)}{e^{-2iak}((k-\kappa)^2+(k+\kappa)^2)})}$$

where k is the wave number in free space, κ is the wave number in the piecewise-constant mouth potential barrier height, Δ is the barrier width and a is the length of the vocal-tract. This function calculates the response to a single-frequency excitation of volume velocity at the glottis, $v(t) = e^{2\pi i\nu t}$, where ν is frequency in Hz, thus $k = \frac{2\pi\nu}{c}$. Using this substitution for k, the function $\mathcal{G}(\nu)$ is found where $\mathcal{G}(\nu) = G(\frac{2\pi\nu}{c})$. Using the symmetries of the Green's function, the impulse response of the vocal tract is then the real part of the inverse Fourier transform of $\mathcal{G}(\nu)$:

$$h(t) = \int_{-\infty}^{+\infty} (\cos(2\pi\nu t)\mathcal{G}_{re}(\nu) - \sin(2\pi\nu t)\mathcal{G}_{im}(\nu))\, d\nu \ , \tag{3}$$

where $\mathcal{G}_{re}(\nu)$ and $\mathcal{G}_{im}(\nu)$ are the real and imaginary part of $\mathcal{G}(\nu)$, respectively. An example of such an impulse response for a Schwa vowel is shown in Figure 2. A Schwa sound is a result of the convolution between such an impulse response and a pulse train $p(t)$. The next section describes a method for recovering the glottal pulse from a recorded speech signal $s(t)$, by performing deconvolution using impulse responses $h(t)$ that correspond to the given vowel, and selecting the positive solution $p(t)$ which has the minimal ℓ^1 norm.

A. Deconvolution Methodology

Real speech was recorded using a condenser microphone at a sample rate of 44.1 kHz. An example of a recording, when cropped to a length of three glottal-cycles, is shown in Figure 3.

The proposed deconvolution algorithm consisted of the following successive steps:

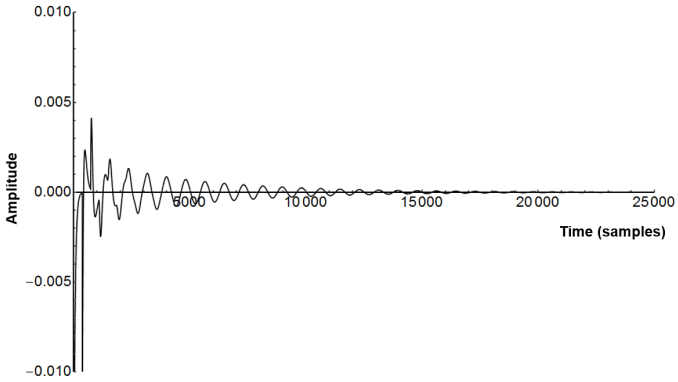

Fig. 2. An example of an impulse response for the Schwa vocal-tract generated using the Green's function solution of the AKG equation

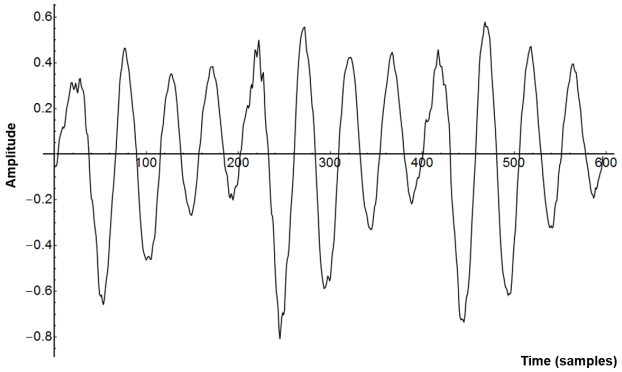

Fig. 3. Example of recorded speech signal for a Schwa vowel.

1. The vocal tract length was estimated, from the recorded signal. For a Shwa, this was done by calculating the average time between zero crossings in the signal and their following peaks – in this time the pulse travels the length of the vocal tract. In order to remove spurious local peaks caused by noise a Gaussian filter was applied as a preprocessing step. An example of the filtered signal is shown in Figure 4.

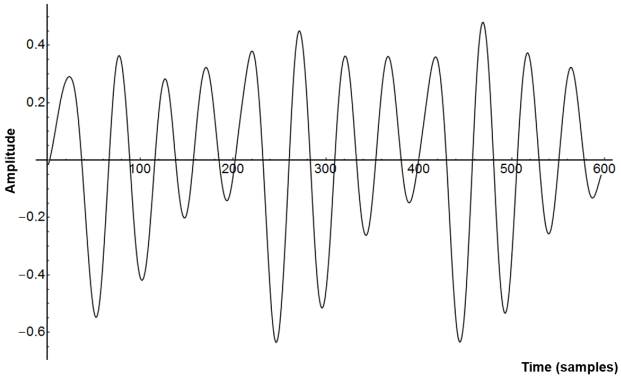

Fig. 4. Example of recorded speech signal for a Schwa vowel after spurious local peaks were removed.

2. Impulse responses $h(t)$ were found for varying values of the mouth-barrier reflection coefficient. For any given value of the coefficient and the estimated vocal tract length, the corresponding impulse response can be created by implementing formula (3) directly, however in this work finite-difference time-domain (FDTD) operations were used instead, which decreased the computational cost substantially. As the AKG equation is a hyperbolic partial differential equation, FDTD calculations can be simplified by transforming it into the following characteristic coordinates: $y = \frac{t-x}{2}$ and $s = \frac{t+x}{2}$. The major advantage of using the FDTD method in these characteristic coordinates is that it provides a rapid calculation time. The hyperbolic form of the acoustical Klein-Gordon equation is given by (4):

$$\partial_y \partial_s u(y,s) = -\frac{1}{16} u(y,s) U(y,s) \tag{4}$$

The discretized form of this is:

$$u[1+p,1+q] = u[p,q] + (u[p,1+q] + u[1+p,q])(1 - \frac{1}{32} U[p,q]) \tag{5}$$

The impulse response of the vocal tract is then obtained as the solution of this equation at the location of the lips when the excitation is the Dirac function. Figure 5 illustrates a simplified version of how the discretized AKG equation is solved. The first column is the input derivative of the glottal input. The second column is the glottal input itself (1.00 at the first point imitates an impulse input). The column labeled Mouth is where the mouth barrier is present and Mic is where a microphone would be placed. A potential of 0.3 is present halfway along the tract and a potential of 0.5 at the mouth. The input at the glottis propagates towards the mouth and where there is a potential barrier, some pressure reflects back towards the glottis and some continues to propagate towards the mouth.

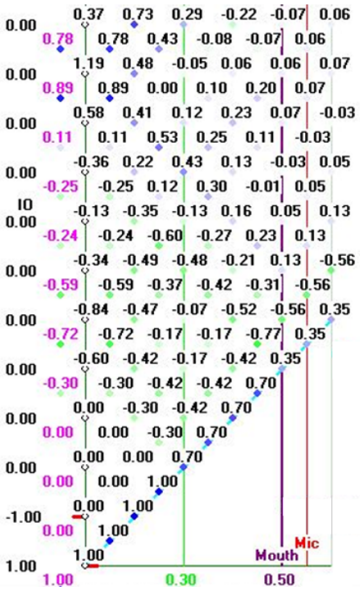

Fig. 5. Example of FDTD calculations for a vocal tract with curvature present in the vocal tract.

The mouth-barrier reflection coefficient was varied from 0 to 1 with an increment of order 10^{-4}. To account for any error in the calculated tract length (despite it working to a high accuracy on synthetic signals), this process was repeated for tract lengths of +/- 10% of the estimation.

3. Impulse responses $h(t)$ were deconvolved from the speech signal in order to produce candidate glottal pulses $p(t)$. The deconvolution was performed by dividing the discrete Fourier transforms of $s(t)$ and $h(t)$ and finding the inverse discrete Fourier transform of the result. In this manner one candidate glottal pulse $p(t)$ was generated for each $h(t)$ obtained in the previous step.

4. The final glottal pulse $p(t)$ was selected. The primary constraint used in this method was positivity. The glottal-pulse waveform is believed to be completely positive as there cannot be a negative flow leaving the glottis. Positive $p(t)$ solutions were then assessed using ℓ^1-norm minimisation. The minimum ℓ^1-norm positive solution was selected as the optimal glottal-pulse candidate. This was chosen as minimisation of the ℓ^1-norm is a widely accepted method of providing spare results in signal analysis [28] and more specifically, Glottal Inverse Filtering [29]. Sparsity in the glottal signal arises from the periodic closed-phase where no pressure leaves the glottis due to it being closed.

III. Results

Before applying the proposed method to real speech, it was applied to a synthetic Schwa vowel. In the case of synthetic vowels, we know both the impulse response and glottal pulse which were used to create it. Figure 6 shows the known glottal input, the extracted pulse via the IAIF method and the extracted pulse via the proposed method. The proposed method produced results with a RMSE of 5.83×10^{-3}. Conversely, the IAIF algorithm pulse had a RMSE of 4.40×10^{-2} – approximately an order of magnitude larger. Similar comparisons were observed in other tests performed with synthetic speech.

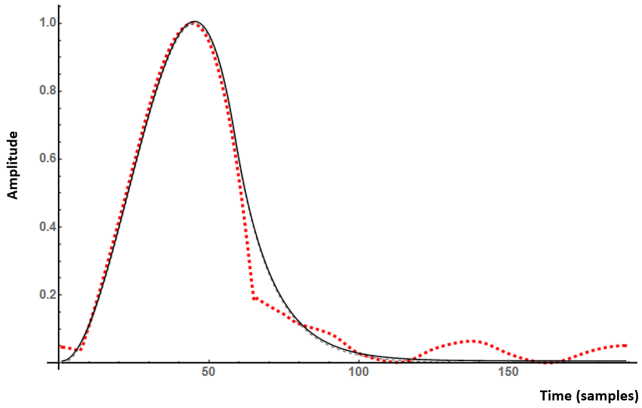

Fig. 6. Testing method using a synthetic Schwa vowel. Known glottal input (grey, dashed), extracted pulse via IAIF (red, dashed) and extracted pulse via proposed method (black).

The proposed method was then applied to numerous real, spoken Schwa vowels. An example of a glottal pulse extraction

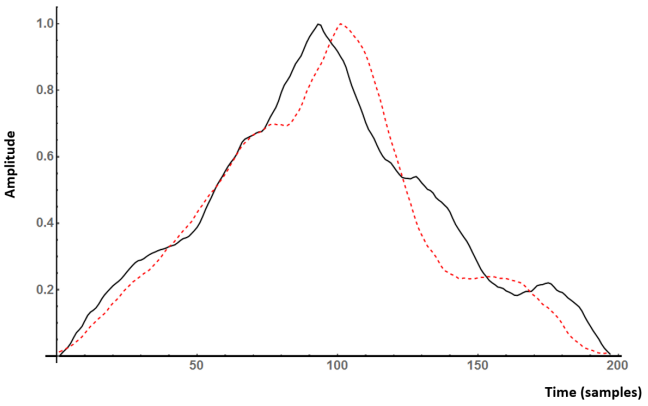

Fig. 7. Recovered glottal pulse for spoken Schwa vowel with recovered pulse via IAIF method (dotted)

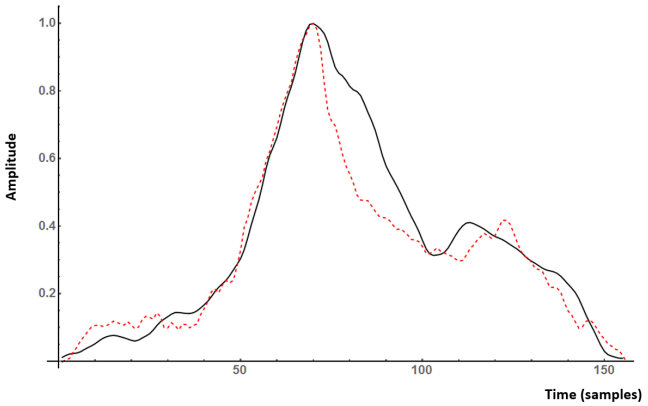

Fig. 8. Recovered glottal pulse for spoken 'ah' vowel with recovered pulse via IAIF method (dotted).

for a Schwa vowel is shown in Figure 7. For comparison, the result obtained with the IAIF method is overlaid. The results for real, spoken Schwa vowels show a great deal of similarity however some differences can be seen. Something which was seen in all extracted pulses using the proposed method were 'bumps', or distinct changes in the slope, seen periodically (approximately 24 samples in Figure 7). In fact, the period of these bumps was found to be the time taken to traverse two lengths of the vocal tract at the speed of sound. Consequently the hypothesis is that these 'bumps' are the periodic interactions of the glottis with the pressure in the vocal tract as it oscillates up and down the vocal tract. Experimentation with the AKG equation shows us that when there is a curvature close to the glottal-end of the vocal tract, the glottis experiences returning pressures from the tract at a higher rate. These high frequency oscillations which are present near to the glottis are potentially causing the glottis to shut faster than if there was no curvature in the tract such as in the Schwa vowel. Some preliminary work has shown a correlation between distances of curvature in the tract from the glottis and the period of the closed-phase. We conjecture that the bumps seen in the results reflect some form of 'source-tract

interaction'.

In all Schwa results, both the IAIF method and the proposed method had a distinct lack of closed phase. This is somewhat surprising as most current glottal-pulse models, e.g. [12]–[15], have a variable, yet distinct closed-phase. Some recent studies have started to bring up this issue with the assumed closed-phase [30].

Figure 8 shows an example of an 'ah' vowel. For this vowel, the curvature is different and thus the potential setup is somewhat different. Consequently, in the proposed method the potential setup was changed, and the same methodology used in the case of the Schwa was repeated.

IV. Conclusion

Using a physical model of the vocal tract and glottis, namely the acoustical Klein-Gordon equation, a parametrized form of the impulse response of the vocal tract was created. Glottal pulses were deconvolved from synthetic signals which showed a considerable improvement on the IAIF method. Pulses extracted from real speech signals showed considerable similarity to those obtained via IAIF however had some clear differences. Unlike most existing glottal models, results showed no distinct closed-phase for the Schwa vowel.

The method proposed can be extended to any of the other 27 IPA vowels by replacing the potential setup described in this paper for the Schwa vowel with that of the vowel of interest. An example of this has been shown in Figure 8.

Some interesting questions have been raised about the closed-phase in existing glottal models. Future work will increase the database of spoken vowels and repeat testing. With a larger number of examples, the hypothesis of the periodic 'bumps' can be explored further.

References

[1] Raitio, Tuomo, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku. "HMM-based speech synthesis utilizing glottal inverse filtering." IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 153-165, 2011.

[2] Fujisaki, Hiroya, and Mats Ljungqvist. "Proposal and evaluation of models for the glottal source waveform." ICASSP'86, Vol. 11, pp. 1605-1608, 1986.

[3] Li, Zhe, Hani Bakhshaee, Leah Helou, Luc Mongeau, Karen Kost, Clark Rosen, and Katherine Verdolini. "Evaluation of contact pressure in human vocal folds during phonation using high-speed videoendoscopy, electroglottography, and magnetic resonance imaging." Meetings on Acoustics, vol.11:060306, 2013.

[4] Iliev, Alexander I., Michael S. Scordilis, Joo P. Papa, and Alexandre X. Falco. "Spoken emotion recognition through optimum-path forest classification using glottal features." Computer Speech and Language, vol. 24, no. 3, pp. 445-460, 2010.

[5] Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion recognition from speech: a review." International journal of speech technology, vol. 15, no. 2, pp. 99-117, 2012.

[6] Childers, Donald G., and C. K. Lee. "Vocal quality factors: Analysis, synthesis, and perception." the Journal of the Acoustical Society of America, vol. 90, no. 5, pp. 2394-2410, 1991.

[7] Rothenberg, Martin. "Acoustic interaction between the glottal source and the vocal tract." Vocal fold physiology, pp. 305-323, 1981.

[8] Fant, Gunnar, and Qiguang Lin. "Glottal sourcevocal tract acoustic interaction." The Journal of the Acoustical Society of America, vol. 81, no. S1, pp. S68-S68, 1987.

[9] Gobl, Christer, and Ailbhe N. "The role of voice quality in communicating emotion, mood and attitude." Speech communication, vol. 40, no. 1, pp. 189-212, 2003.

[10] Holmes, J. "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer." IEEE transactions on Audio and Electroacoustics, vol. 21, no. 3 pp. 298-305, 1973.

[11] Dejonckere, P. H., C. Obbens, G. M. De Moor, and G. H. Wieneke. "Perceptual evaluation of dysphonia: reliability and relevance." Folia Phoniatrica et Logopaedica,vol. 45, no. 2, pp 76-83, 1993.

[12] Klatt, Dennis H., and Laura C. Klatt. "Analysis, synthesis, and perception of voice quality variations among female and male talkers." The Journal of the Acoustical Society of America, vol. 87, no. 2, pp. 820-857, 1990.

[13] Rosenberg, Aaron E. "Effect of glottal pulse shape on the quality of natural vowels." The Journal of the Acoustical Society of America, vol. 49, no. 2B, pp 583-590, 1971.

[14] Fant, Gunnar, Johan Liljencrants, and Qi-guang Lin. "A four-parameter model of glottal flow." STL-QPSR, vol. 26, no. 4, pp 1-13, 1985.

[15] Doval, Boris, Christophe d'Alessandro, and Nathalie Henrich. "The voice source as a causal/anticausal linear filter." In ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis. 2003.

[16] Alku, Paavo. "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering." Speech communication, vol. 11, no. 2-3, pp. 109-118, 1992.

[17] Wong, D., J. Markel, and A. Gray. "Least squares glottal inverse filtering from the acoustic speech waveform." IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 4, pp. 350-355, 1979.

[18] Auvinen, Harri, Tuomo Raitio, Manu Airaksinen, Samuli Siltanen, Brad H. Story, and Paavo Alku. "Automatic glottal inverse filtering with the Markov chain Monte Carlo method." Computer Speech and Language, vol. 28, no. 5, pp. 1139-1155, 2014.

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

[20] Tze Wei Chu, Derek, Kaiwen Li, Julien Epps, John Smith, and Joe Wolfe. "Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics." The Journal of the Acoustical Society of America, vol. 133, no. 5, pp. 358-362, 2013.

[21] Airaksinen, Manu, Tuomo Raitio, Brad Story, and Paavo Alku. "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction." IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 22, no. 3, pp. 596-607, 2015.

[22] Cabral, Joao P., Korin Richmond, Junichi Yamagishi, and Steve Renals. "Glottal spectral separation for speech synthesis." IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 2, pp. 195-208, 2014.

[23] Sahoo, Subhasmita, and Aurobinda Routray. "A novel method of glottal inverse filtering." IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1230-1241, 2016.

[24] Bozkurt, Baris, Boris Doval, Christophe d'Alessandro, and Thierry Dutoit. "Zeros of z-transform representation with application to source-filter separation in speech." IEEE signal processing letters, vol. 12, no. 4, pp. 344-347, 2005.

[25] Forbes, Barbara J., E. Roy Pike, and David B. Sharp. "The acoustical KleinGordon equation: The wave-mechanical step and barrier potential functions." The Journal of the Acoustical Society of America, vol. 114, no. 3, pp. 1291-1302, 2003.

[26] Forbes, Barbara J. "A potential-function analysis of speech acoustics." PhD thesis, King's College London, 2000.

[27] Schafer, Ronald W., and Lawrence R. Rabiner. "Digital representations of speech signals." Proceedings of the IEEE, vol. 63, no. 4, pp. 662-677, 1975.

[28] Hurley, Niall, and Scott Rickard. "Comparing measures of sparsity." IEEE Transactions on Information Theory, vol. 55, no. 10, pp. 4723-4741, 2009.

[29] Airaksinen, Manu, Tom Bckstrm, and Paavo Alku. "Automatic estimation of the lip radiation effect in glottal inverse filtering." Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[30] Dias, Sandra de Oliveira. "Estimation of the glottal pulse from speech or singing voice." PhD thesis, University of Porto, 2014.