Citation for published version (APA):

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher’s definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher’s website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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:

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

where $\ast$ 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], Liljencranz-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):

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

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 of all the world’s languages [26]. For instance, 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 spoken vowel sounds of all the world’s languages.

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 shown in Figure 1. The vocal-tract curvature which results from the collection of potentials is shown in red.

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:

\[
4e^{i\Delta(k-\kappa)}k\kappa \left( -e^{-2i\kappa} (k-\kappa)^2 + (k+\kappa)^2 \right) \left( 1 - \frac{e^{-2i\kappa}(1-e^{-2i\kappa})(k^2-\kappa^2)}{e^{-2i\kappa}(k-\kappa)^2+(k+\kappa)^2} \right)
\]

where \( k \) is the wave number in free space, \( \kappa \) is the wave number in the piecewise-constant mouth potential barrier height, \( \Delta \) 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 \( \nu \) is frequency in Hz, thus \( k = \frac{2\pi \nu}{c} \). Using this substitution for \( k \), the function \( G(\nu) \) is found where \( G(\nu) = G(\frac{2\pi \nu}{c}) \). 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 \( G(\nu) \):

\[
h(t) = \int_{-\infty}^{+\infty} (\cos(2\pi\nu t)G_{re}(\nu) - \sin(2\pi\nu t)G_{im}(\nu)) d\nu,
\tag{3}
\]

where \( G_{re}(\nu) \) and \( G_{im}(\nu) \) are the real and imaginary part of \( G(\nu) \), respectively. An example of such an impulse response for a Schwa vowel is shown in Figure 2.

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:
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.

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):

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

The discretized form of this is:

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

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.
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 $\ell^1$-norm minimisation. The minimum $\ell^1$-norm positive solution was selected as the optimal glottal-pulse candidate. This was chosen as minimisation of the $\ell^1$-norm is a widely accepted method of providing sparse 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 $5.83 \times 10^{-3}$. Conversely, the IAIF algorithm pulse had a RMSE of $4.40 \times 10^{-2}$ – approximately an order of magnitude larger. Similar comparisons were observed in other tests performed with synthetic speech.

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