AN ALGORITHM FOR GLOTTAL VOLUME VELOCITY ESTIMATION

Ashraf Alkhairy

MIT, Room 36-547
50 Vasser Street
Cambridge, MA 02139

ABSTRACT

We present a new method for the estimation of the glottal volume velocity from voiced segments of the radiated acoustic speech pressure waveform. Our algorithm is based on spectral factorization of the signal and is a general purpose procedure. It does not suffer from residual effects or assume constraining models for the vocal tract and the glottal source, as is commonly the case with existing methods. The resulting estimate of the glottal volume velocity is accurate and can be used for modeling and synthesis purposes.

1. INTRODUCTION

Determination of the volume velocity at the glottis for voiced regions of speech is desirable for several reasons. In formant synthesis, high quality speech may only be produced with accurate parameterization of the glottal waveform [1]. In the area of modeling, descriptors of the glottal volume velocity are needed for the development of accurate aerodynamic and acoustic models based on articulatory and physiologic information. Characteristics of the excitation at the glottis may also be used in the determination and verification of the speaker. Also, departures of the glottal airflow attributes from normal may be helpful in the determination of laryngeal pathologies.

The glottal waveform may not be measured directly, and can only be estimated based on observed signals. Due to the importance of the subject matter, a large number of methods have been proposed during the past thirty years for the estimation of the glottal waveform.

In one of the approaches, the glottal waveform is estimated by speaking into a reflectionless uniform tube, while holding the vocal tract in the neutral position [2]. While this method is simple, it is not general-purpose. The large variations in the glottal waveform shape among speakers, phonemes, contexts, phonation and speaking styles make it necessary that the volume velocity at the glottis be estimated for a variety of situations [3].

Another approach that utilizes special-purpose equipment is that of Rothenberg [4]. In this method, the volume velocity at the lips is measured using a mask, and the glottal excitation is estimated by removing the vocal tract resonances using inverse filtering. The procedure allows for computation of the glottal volume velocity that is amplitude calibrated and has a correct dc offset. However, it is only useful for measurements of airflow upto 1 kHz [5], and its success is highly dependent on the accuracy and applicability of the all-pole vocal tract model, with a fixed number of formants.

Most other methods, including the one proposed in this paper, are based solely on the measurement of the radiated pressure waveform using a microphone at some distance from the lips. In these methods, the dc offset cannot be calculated accurately due to the radiation characteristic which has a zero in the neighbourhood of the origin. Also, amplitude caliberation is difficult in such approaches.

Existing pressure waveform based methods share a common approach, and are described as inverse filtering methods [6, 7, 8, 9, 10, 11, 12, 13, 14]. The glottal waveform is obtained by cancelling the effects of a predetermined number of formants. Variations of the basic approach occur with regard to the structure of the glottal waveform imposed. Some methods model the glottal waveform as the output of an all zero system, while others impose parameterized time domain models. A few of the methods allow vocal tract models with anti-formants. A large number of approaches assume that the volume velocity at the glottis has a well defined closed phase, which does not occur in many cases. These methods need to estimate the region of the closed phase which cannot be done precisely in an automated environment. The existing methods require human intervention and are very sensitive to the restrictive models imposed [15].

2. PROBLEM FORMULATION

The physical quantities of interest during the production of speech are: the volume velocity at the glottis, $u_G(t)$, the volume velocity at the lips, $u_L(t)$, and the pressure waveform measured by the microphone, $p(t)$.

For a voiced signal, the glottal volume velocity may be written as $u_G(t) = \sum_i \delta(t - t_i) * g(t)$, where $g(t)$ denotes the glottal waveform if there were only a single pitch period. We have assumed here that the glottal waveform does not change over the period under consideration.

In many applications we can assume that the speech system is represented by a linear time invariant system over a two pitch period, and that $g(t)$ has an extent less than that. In such cases, $P(s) = D(s)G(s)H(s)$, where $D(s)$ is the Laplace Transform of the impulse train, $G(s)$ is the Laplace Transform of $g(t)$ and $H(s)$ is the transfer function between $g(t)$ and the pressure signal. This transfer function may be written as $H(s) = V(s)R(s)$, where $V(s)$ is the transfer function from the glottal volume velocity to the volume velocity at the lips, and $R(s)$ is the radiation characteristic, represented by the transfer function from the lip volume velocity to the pressure at the microphone.

The problem of glottal volume velocity estimation may be formulated as follows: Given $p(t)$, determine $g(t)$. Clearly, this is not possible without additional constraints on $G(s)$ or $H(s)$. These constraints should reflect reality, while being generic enough to be widely applicable. In the next section we discuss these constraints.

3. SOURCE-FILTER CHARACTERIZATION

Most sounds in English and other languages have articulatory configurations for which $V(s)$ can be well modelled by an all-pole system. In addition, since the vocal tract is a physical system, $V(s)$ must be stable with poles in the left half plane.

The radiation transfer function, $R(s)$, is also representative of a physical system, and thus its poles must lie in the left half plane. The zero, resulting from differentiation, is theoretically supposed to be at the origin. For practical purposes however, it is expected to lie in the left half plane. Consequently, $H(s)$ can be modelled by a minimum phase system.

The glottal transfer function, $G(s)$, may be characterized as follows. A first approximation to $g(t)$, is that of the impulse response of a two-pole acausal system. It should be noted that $G(s)$, as a system, does not exist in reality, but is simply a mathematical model. Thus constraints of causality are not applicable here.

From the above discussions, we may conclude that $H(s)$ is a minimum phase system and $G(s)$ is a maximum phase system, at least as a first approximation. This observation forms the basis for the development of our method to estimate $g(t)$.

As can be seen, our characterization does not make a number of over-simplifications made by existing approaches. These include the following: the glottal waveform includes a closed phase, the representations of $H(s)$ or $G(s)$ are lumped systems; $H(s)$ or $G(s)$ are all-pole/all-zero systems; the number of poles and zeros are fixed a priori. The last assumption is particularly disadvantageous to existing methods because only the number of formant frequencies can be fixed a priori. Others related to the glottal pulse, additional resonances of the vocal tract, and behavior at extreme frequencies after sampling are usually unknown.

4. METHOD

Two issues need to be addressed in the estimation of $g(t)$ based on the characterization of the source and filter discussed in the previous section. First, $p(t)$, which is quasi-periodic, is measured rather than $p_u(t) = h(t) * g(t)$. Thus a procedure has to be developed to remove the effects of periodicity, thereby estimating $p_u(t)$.

Second, a robust algorithm is needed to estimate $g(t)$ from $p_u(t)$ with proper time and frequency resolution. Both parts of our method are discussed in the next section. All the processing is conducted pitch-synchronously, allowing us to estimate, possibly differing, glottal pulses for each pitch period.

5. ALGORITHM

To process the speech signal digitally, we need to sample the recorded pressure waveform, $p(t)$. The sampling rate should be large enough to include all frequencies of interest in analysis, allow the computation of pitch pulse locations with high resolution, and perform phase unwrapping accurately. The latter two requirements are essential for a sucessful implementation of our algorithm. We have observed that a sampling frequency of 10 kHz is sufficient for our requirements and is also high enough for analysis.

The pulse locations are computed using a peak detection algorithm. Zero crossings in the region of highest change from negative to positive values represent the pitch markers. The algorithm is applied to the sampled pressure data from each pitch period independently, where the length of a pitch period is generally less than 128 samples.

The periodicity in the pressure waveform is removed using the following steps:

1. Compute $P[k] = \text{FFT}(p[n])$, of length N, where N denotes the length of the pitch period for the segment under consideration.

2. Resample $P[k]$ to length 256. This results in the removal of the periodicity, while at the same time providing enough time and frequency resolution of the resulting signal.

3. Compute $p[n]$, the real part of the inverse FFT of the resampled $P[k]$, to determine the aperiodic representation of the pressure waveform segment and remove imaginary parts resulting from computational inaccuracies.

The second stage in our algorithm is the computation of the maximum phase glottal waveform, $g[n]$, from $p[n]$. This is accomplished using the following methodology:

1. Compute $\hat{p}[n]$, the complex cepstrum of $p[n]$. The approach we use is that of assigning the log of the magnitude spectrum to the real part of the cepstrum FFT, and the unwrapped phase to the imaginary part. Prior to the computation, $p[n]$ is circularly shifted to have no phase discontinuity at π.

2. Select the acausal part of $\hat{p}[n]$. Any real number may be assigned to the value at $n = 0$, because it only effects the scaling of the recovered glottal waveform. For convenience, we choose a value of zero.

3. Compute the inverse cepstrum of the acausal part. This represents the glottal waveform. No adjustments for the circular shifts are required.

As can be inferred from the above description, the algorithm utilizes stable procedures, thereby making it robust. Furthermore, it is very easy to program, and does not require human interference.

6. EXAMPLES

The method presented in this paper was tested on three vowels and four consonants spoken by three speakers. In all cases, the estimated glottal waveform conformed to the general properties of the volume velocity at the glottis, and the radiation zero was estimated as a part of the minimum phase signal. The results support the characterization on which the method is based, and suggests that the algorithms performs well even in cases

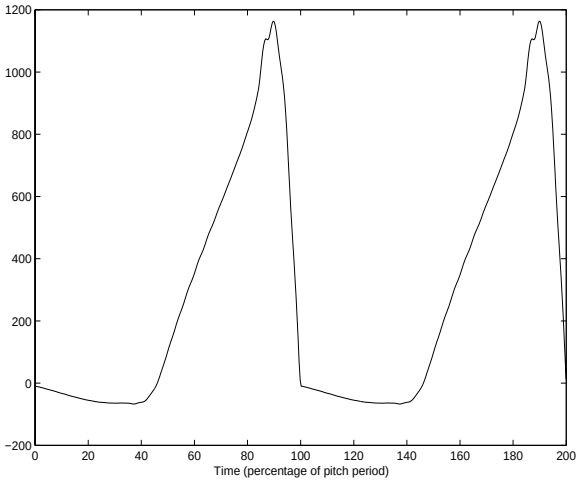

Figure 1: Glottal Waveform for /a/

where the assumptions concerning first approximations are not strictly applicable.

We illustrate the glottal waveforms over two pitch periods for the vowels /a/, /u/ and /i/ in Figures 1, 2 and 3 respectively. In addition to expected time-domain properties, the spectral tilt for the pulses are 48 dB per decade. This is in agreement with the value of 40 dB per decade expected for a glottal waveform based on theoretical models.

7. CONCLUSION

This paper presents a simple, general-purpose method for accurate computation of the glottal volume velocity. The structure of the method allows it to be extended to nasal or nasalized speech segments. We are in the process of exploring modification of the method for such sounds.

8. REFERENCES

[1] D. Klatt and L. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," The Journal of the Acoustical Society of America, vol. 87, pp. 820–857, 1990.

[2] M. Sondhi, "Measurement of the glottal waveform," The Journal of the Acoustical Society of America, vol. 57, pp. 228–232, 1975.

[3] H. Hanson, "Glottal characteristics of femal speakers: Acoustic correlates," The Journal of the Acoustical Society of America, vol. 101, pp. 466–481, 1997.

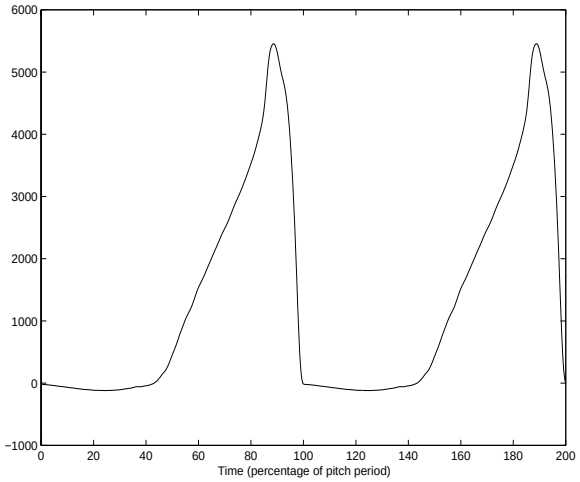

Figure 2: Glottal Waveform for /u/

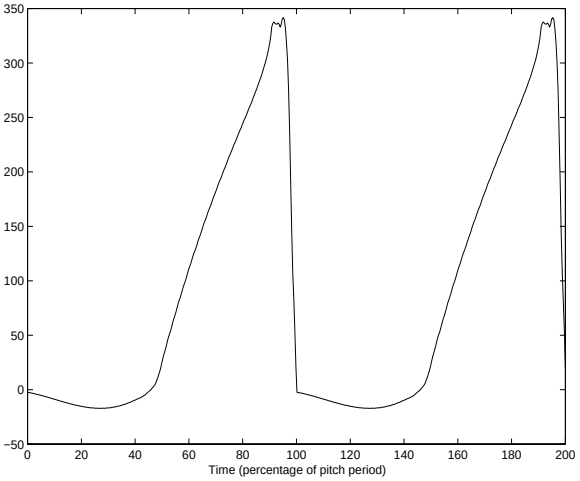

Figure 3: Glottal Waveform for /i/

[4] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform," The Journal of the Acoustical Society of America, vol. 1, pp. 1632–1645, 1976.

[5] P. Badin, S. Hertegard, and I. Karlsson, "Notes on the rothenberg mask," STL-QPSR, vol. 1, pp. 1–7, 1990.

[6] J. Miller, "Nature of the vocal cord wave," The Journal of the Acoustical Society of America, vol. 31, p. 667, 1959.

[7] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, pp. 350–355, 1979.

[8] V. B. M. Matausek, "A new approach to the determination of the glottal waveform," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 6, pp. 616–622, 1980.

[9] S. B. D. Veeneman, "Automatic glottal inverse filtering from speech and electroglottographic signals," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-33, no. 2, pp. 369–376, 1985.

[10] J. Picone, D. Prezas, and W. Hartwell, "Joint estimation of the lpc parameters and the multipulse excitation," Speech Communication, vol. 5, pp. 253–268, 1986.

[11] P. Milenkovic, "Glottal inverse filtering by joint estimation of an ar system with a linear input model," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-34, no. 1, pp. 28–41, 1986.

[12] B. Lim and C. Un, "Dual rls lattice joint process estimation algorithm for a time-varying arma speech model," Speech Communication, vol. 10, pp. 303–306, 1991.

[13] A. Krishnamurthy, "Glottal source estimation using a sum-of-exponentials model," IEEE Transactions on Signal Processing, vol. 40, pp. 682–686, 1992.

[14] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication, vol. 11, pp. 109–118, 1992.

[15] G. Fant, "Some problems in voice source analysis," Speech Communication, vol. 13, pp. 7–22, 1993.