Analysis of the Voice Source in Different Phonation Types: Simultaneous High-Speed Imaging of the Vocal Fold Vibration and Glottal Inverse Filtering

Hannu Pulakka1, Paavo Alku1, Svante Granqvist2,3, Stellan Hertegård5, Hans Larsson2,3, Anne-Maria Laukkanen4, Per-Åke Lindestad3, Erkki Vilkman5

1: Helsinki University of Technology, Espoo, Finland
2: Royal Institute of Technology, Stockholm, Sweden
3: Karolinska Institute, Huddinge University Hospital, Stockholm, Sweden
4: University of Tampere, Finland
5: University of Oulu, Finland

Hannu.Pulakka@hut.fi

Abstract

Glottal flow waveforms estimated by inverse filtering acoustic speech pressure signals were compared to glottal area functions obtained by digital high-speed imaging of the vocal fold vibration. Speech data consisted of breathy, normal and pressed phonations produced by two male and one female subjects. The results yield both qualitative and quantitative information about the relationship between the glottal flow and the corresponding area function. It was shown, for example, that a distinct knee in the glottal flow waveform in the opening phase corresponds to the abrupt opening of the vocal folds in normal and pressed phonation. In addition, the obtained quantitative data corroborates known theoretical considerations according to which the shape of the glottal flow is more asymmetric than the corresponding area function.

1. Introduction

Two information signals are of vital importance in understanding human voice production: the glottal volume velocity waveform, which serves as the acoustic excitation of voice, and the glottal area function, which describes how the area between the vibrating vocal folds behaves as a function of time. The glottal volume velocity waveform can be estimated using a non-invasive technique called inverse filtering [1], [2]. The idea behind inverse filtering is to first form a model for the vocal tract transfer function. The effect of the vocal tract is then canceled from the speech waveform by filtering this through the inverse of the model. Estimation of the glottal area function, on the other hand, can be done using photoglottography (PGG), which involves illuminating the glottis from above or below and monitoring the light intensity on the other side to measure the glottal area [3], [4]. Currently, PGG is widely replaced with digital high-speed imaging in the estimation of the glottal area function [5]-[8]. In this method, a sequence of digital images of the illuminated vocal folds is recorded via an endoscope at a high time resolution. Frame rates of several thousand frames per second can be achieved [6]. Thus, many digital images can be obtained of the vocal fold vibration during each glottal cycle.

In this study, glottal flow waveforms obtained by inverse filtering acoustical speech pressure waveforms are compared to glottal area functions derived from simultaneous high-speed digital imaging. The goal of the study is to assess, both quantitatively and qualitatively, the relationship between the glottal flow and the synchronized glottal area function when speech is produced in different phonation types.

2. Materials and methods

Two males and one female, all of healthy voices, served as subjects. They produced sustained /ae/ vowels in three modes of phonation: breathy, normal, and pressed. The duration of each phonation was approximately one second. The following signals were recorded simultaneously: (1) high-speed video image sequence of the vocal fold vibration, (2) acoustic speech pressure waveform, and (3) electromyography (EGG).

The high-speed image sequences of the vocal folds were obtained via a rigid endoscope and recorded by a Weinberger high-speed camera at the frame rate of about 1900 frames per second. The resolution of each frame was 256 by 64 pixels. The acoustic speech pressure signal was captured by a condenser microphone at the distance of 10 cm from the mouth of the subject. The acoustic signal, the EGG, and an additional synchronization signal were sampled simultaneously with a computer sound card at the sampling rate of 16 kHz and stored on a hard disk.

The synchronization of the image sequence with the acoustic signal and the EGG was arranged by recording the state of a two-position foot pedal on the synchronization channel. The pedal was pressed at the end of each recording, which also stopped the high-speed imaging. Thus, the instant of the last image could be determined from the instant of the pedal pulse in the synchronization signal. Additionally, the high-speed video signal was summed to the synchronization channel to enable accurate measurement of the video frame rate afterwards. However, there is some uncertainty about the exact timing of the last image frame relative to the pedal pulse. The standard deviation of the time shift between the image sequence and the other signals was measured to be 0.39 frames in a similar recording setting [7].

The glottal area waveforms were detected semi-automatically from the high-speed image sequences using a custom-made computer program “High Speed Tool Box” [8]. No attempts were made to calibrate the area measurements to yield absolute glottal areas. Area functions were computed for each of the nine phonations (3 subjects, 3 phonation types) by
positioning the analysis window approximately in the middle of each utterance. Five glottal cycles located in the analysis window (see Figures 1 and 2) were selected for closer examination. The semi-automatically detected area functions were checked manually by comparing the area values with the image sequences. Manual corrections were made when necessary.

The acoustic signal was shifted relative to the EGG and high-speed image sequence to compensate for the sound propagation delay from the vocal folds to the microphone. The computation of the delay was based on the mouth-to-microphone distance (10 cm) added to the length of the vocal tract (17 cm for the male subjects and 15 cm for the female subject). Comparison of the inverse filtered flow with the glottal area function indicates that this compensation may not provide as accurate synchronization as it was supposed to do. The positioning of the rigid endoscope into the subject’s mouth might have caused some inconsistency in the mouth-to-microphone distances.

The acoustic pressure signal was inverse-filtered using the Iterative Adaptive Inverse Filtering (IAIF) method [9], [10]. The order of the all-pole filter to model the vocal tract and the pole position of the filter that models the lip radiation effect were adjusted manually for each utterance to obtain appropriate inverse filtered flow waveform with least ripple during the closed phase. It is worth noticing that inverse filtering was applied to the microphone pressure signal and, therefore, the DC flow could not be obtained.

Both the inverse filtered flow waveforms and the area waveforms were parameterized to obtain quantitative data for comparison. The following time-based quotients [11] were calculated: Firstly, the relative closing phase was quantified with the Closing Quotient (ClQ), which is defined as the length of the glottal closing phase divided by the fundamental period. Secondly, the symmetry of the open phase was quantified with the Speed Quotient (SQ), which is defined as the ratio between the lengths of the glottal opening and closing phases. Thirdly, the relative open phase was quantified with the Open Quotient (OQ) defined as the length of the open phase divided by the length of the fundamental period. The instants of opening and closing as well as the instants of the maximal signal value within the glottal cycle were easy to identify both from the flow and the area waveforms and they were located manually. Many of the estimated glottal flows showed two kinds of glottal openings (see Figure 1): Primary opening (denoted by square) when the flow starts to increase from its minimum value, and secondary opening (denoted by diamond) when the flow shows a distinct knee indicating the beginning of a more rapid increase. For these cases, two versions of the Speed Quotient and Open Quotient were calculated for the flow waveform by using either the primary opening (for SQ1 and OQ1) or the secondary opening (for SQ2 and OQ2) as the instant of glottal opening. All the quotients were first calculated for each glottal cycle and then averaged over the analysis window.

The EGG signal is available but the current study focuses only on the comparison of the high-speed video images and the glottal flows estimated by inverse filtering.

3. Results

The obtained results are discussed in the following separately for each phonation type. The corresponding quotients are shown in Table 1.

3.1. Breathy phonation

All the estimated glottal flows in breathy phonation were smooth in shape and they showed no indication of a clear closed phase (i.e., values of OQ for breathy phonation in Table 1 are close to unity for flows). The closed phase was absent also in the glottal area functions, for which the value of OQ was 1.00 in all the three samples analyzed. During the open phase of the glottal cycle, the flow waveforms were more asymmetric (i.e., skewed to the right) than the area functions as indicated by the parameters values: SQs of flow was larger than SQs of area, whereas ClQ of flow was smaller than ClQ of area. See Figure 2 for an example of breathy phonation.

3.2. Normal phonation

In normal phonation, there was a clear closed phase in all the glottal flows estimated by inverse filtering. The relative open phase of the flow, especially when measured by OQ1, was larger than that of the area as shown in Table 1. Similarly to breathy phonation, the flow waveform was more skewed to the right than the area function. In particular, the asymmetry of the glottal flow open phase, when measured from the instant of primary glottal opening with SQ1, was much larger in the flow than in the area. The difference in the asymmetry of the flow and area was also more pronounced in normal phonation than in breathy phonation. See Figure 1 for an example of normal phonation.

3.3. Pressed phonation

The relative closing phase of the flow reached its shortest value in pressed phonation as shown by the values of CIQ for pressed phonation given in Table 1. Moreover, the relative open phase of the glottal flow (as measured by OQ2) and also the relative open phase of the area function were shortest in this phonation type. When the opening of the flow was considered to occur at the instant when the waveform started to rise from its minimum value, the pressed samples showed the largest differences in the OQ values between the flow and the area.

4. Conclusions

The study provides useful information that helps in interpreting the characteristics of the glottal flow estimated by inverse filtering. In particular, the flow waveforms and corresponding area functions for normal and pressed phonations of the two male subjects show an interesting phenomenon during the opening phase: The flow starts to rise from the minimum level at the instant of primary opening but the high-speed images show no evidence of glottal opening at that moment. However, the secondary opening in the flow waveform occurs very close to the instant of glottal opening indicated by the area function (see Figure 1). Therefore, these comparisons of the flow and the area function seem to suggest that the existence of a distinct knee in the flow pulse during the opening phase (indicated by diamonds in Figure 1) is an
acoustical consequence of the abrupt opening of the glottis. However, the flow between the primary and secondary openings, which is by no means negligible as shown in the middle panel in Figure 1, is probably due to the vertical piston movement of the vocal folds. Both male subjects show this behavior in normal and pressed phonation. The flow pulse of the female subject does not have any clear secondary opening moment but the flow increases smoothly from the end of the closed phase until the maximum flow is reached. However, the OQ values of the female subject for normal phonation suggest that the flow begins to rise before the actual glottal opening occurs in the high-speed image sequence.

The quotients computed from the material indicate that the flow pulses are in general more skewed to the right than the corresponding area functions. This result is in line with known theoretical considerations of the relationship between the flow and the area function according to which a more asymmetric shape of the glottal flow is due to the interaction of the vocal tract and the glottal source and due to the acoustic mass of the supra-glottal airways [12, 13].

In interpreting the present data it is, however, worth pointing out that the temporal resolution of the glottal area function is relatively low. Especially in the case of female speech and in the case of male utterance of pressed phonation, there are only a few image frames available to describe the area function during the open phase of the glottal cycle. Thus, the reliability of the results deteriorates in these cases. Furthermore, the glottal area detection of pressed vowels is more difficult and more unreliable than in normal phonation due to the narrowing of the lower part of the vocal tract, which sometimes partly hides the vocal folds from the camera. Finally, the synchronization of the high-speed video images with the other signals as well as the compensation of the sound propagation delay from the vocal folds to the microphone should be more accurate to be able to make more detailed conclusions of the temporal differences between the inverse filtered flow waveform and the corresponding area function.

5. Acknowledgements

This study was supported by the Academy of Finland (project no. 205962).

6. References


<table>
<thead>
<tr>
<th>Phonation type</th>
<th>Subject</th>
<th>F0 (Hz)</th>
<th>Flow</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathy</td>
<td>Male 1</td>
<td>110</td>
<td>0.50</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Male 2</td>
<td>100</td>
<td>0.54</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>Female 1</td>
<td>160</td>
<td>0.41</td>
<td>1.33</td>
</tr>
<tr>
<td>Normal</td>
<td>Male 1</td>
<td>110</td>
<td>0.25</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td>Male 2</td>
<td>97</td>
<td>0.23</td>
<td>2.53</td>
</tr>
<tr>
<td></td>
<td>Female 1</td>
<td>160</td>
<td>0.28</td>
<td>2.22</td>
</tr>
<tr>
<td>Pressed</td>
<td>Male 1</td>
<td>109</td>
<td>0.17</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>Male 2</td>
<td>101</td>
<td>0.20</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td>Female 1</td>
<td>173</td>
<td>0.13</td>
<td>2.28</td>
</tr>
</tbody>
</table>

Table 1: Estimated parameters. See Section 2 for a description of the parameters.

Figure 1: Sound pressure, glottal flow and glottal area of normal phonation by the subject Male 1. The instants determined for the parametrization of the pulses are indicated in the figure: primary opening (square), secondary opening (diamond), maximum flow or area (circle), and closure (triangle). The instants of high-speed imaging frames are shown as thin vertical lines. Vertical axes are arbitrary.

Figure 2: Sound pressure, glottal flow and glottal area of breathy phonation by the subject Male 2. The instants determined for the parametrization of the pulses are indicated in the figure: opening (square), maximum flow or area (circle), and closure (triangle). The instants of high-speed imaging frames are shown as thin vertical lines. Vertical axes are arbitrary.