Proceedings of the Institute of Acoustics

A STUDY OF GLOTTAL PULSES DERIVED BY INVERSE FILTERING

John N. Holmes

Speech Technology Consultant, 19 Maylands Drive, Uxbridge, UB8 1BH

ABSTRACT

A new interactive inverse filtering program has been developed, and used to derive glottal flow pulses for several male and female speakers. After the vocal folds first make contact in the closing phase, considerable additional air movement occurs as a result of the folds moving upwards while increasing their contact area over the next millisecond or so. These air flow shapes observed after the main glottal closure discontinuity can explained by surface movement of the vocal folds, but they are not generally well modelled by the exponentially decaying pattern of the LF model [1], attributed to "glottal leakage".

1. INTRODUCTION

The motivation for this study was to investigate the properties of voiced excitation that might be relevant for very-high-quality speech synthesis using a formant-based speech production model. Many workers have used formant synthesis with the simple four-parameter model of glottal flow (the LF model) developed by Fant, Liljencrants and Lin [1], but this paper will demonstrate that some aspects of that model are seriously inadequate for modelling the complexity of voiced excitation for some speakers.

The acoustic theory of speech production [2] shows that, subject to certain idealizing assumptions, the vocal tract between glottis and lips can be modelled by a filter whose transfer function has an infinite number of complex conjugate pole pairs and no zeros. The average frequency spacing of the complex conjugate poles is approximately 1 kHz for adult males. There are thus typically four pole pairs, corresponding to the lowest four formants of the speech signal, in the frequency range up to 4 kHz. The effect of the infinite number of poles above 4 kHz can be approximated by a fixed "higher pole correction" [3] or by a sampled-data filter whose frequency response inherently repeats itself indefinitely.

The conversion from volume flow at the lips to pressure waveform received by an external microphone can be represented by a simple differentiating action. Thus there is a simple overall model relating the microphone signal to glottal flow, which naturally suggests that one could reverse the action specified by the model to derive the glottal flow from the acoustic signal. All that appears to be required is an inverse filter to cancel the effect of the formants with suitably placed complex conjugate zeros, and an integrator to cancel the effect of the radiation at the lips. Some means for approximating the effect of higher poles must also be provided, either by some suitable band-limiting filter or by using a sampled-data inverse filter. Over the last 40 years several workers [4,5,6] have experimented with using an inverse filter in which the frequencies and bandwidths of the complex zeros are adjusted manually, while looking at the inverse filtered waveform, to minimize the ripple caused by the formants. However, there can be problems in the detail of the inverse filtering which are discussed in the next section. The results presented in [6], in which manual adjustment was compared with setting the inverse filter automatically using linear prediction analysis, show additional difficulties with automatic methods.

2. PROBLEMS WITH INTERACTIVE INVERSE FILTERING

2.1 Assumptions of the vocal tract model

Speech production theory assumes that the cross-dimensions of the vocal tract are so small that there is only plane-wave propagation along its length. Above 3 kHz the dimensions of the vocal tract can be large enough for significant departure from plane wave propagation to occur [7], introducing additional poles and zeros into the vocal tract transfer function. The possibility of non-plane-wave propagation at higher frequencies and the uncertainty about the effect of higher poles in the vocal tract response mean that the spectral magnitude of the inverse filtered pulse will be very unreliable above about 3 kHz, and may be significantly in error even at lower frequencies.

The model will also be seriously in error if any coupling between the oral and nasal cavities causes the vocal tract to have a side branch. To illustrate the difficulties that arise in nasalized vowels, Fig. 1a shows the effect of inverse filtering a non-nasalized vowel, whereas Fig. 1b shows the results for a phonemically equivalent nasalized vowel. It can be seen that there is an additional very low frequency ripple caused by the coupling to the nasal cavity, which the inverse filter has not been able to cancel. The presence of nasalization has also caused a substantial change in the bandwidth and frequency of the first formant.

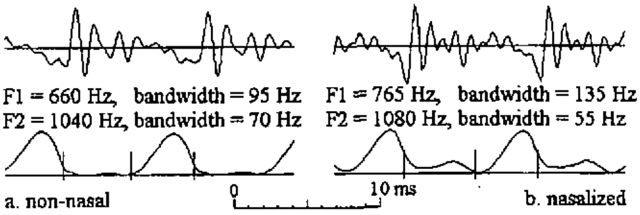

Fig. 1. Inverse filtering comparison of non-nasal and nasalized vowels.

2.2 Low-frequency phase distortion

If inverse filtering is to be used to derive the glottal flow, it is important that there should be no significant distortion in the recording of the speech signal. Distortion in the recording process itself can be avoided by feeding the signal directly into a computer using an analogue-to-digital converter. Modern condenser microphones have an extremely flat frequency response, but because the microphone and its amplifier cannot respond down to zero frequency some low-frequency phase distortion is unavoidable.

It was advocated in [8] that low-frequency phase distortion could be quantified by supplying a calibration signal with any speech recordings, but this suggestion has not been generally adopted. In consequence, when using available speech databases, the amount of low-frequency phase distortion is not known precisely. However, it is fairly easy to provide a variable amount of phase-distortion correction to recorded signals, by using a simple all-pass filter with one real pole and one real zero in its transfer function. Its phase characteristic needs to be reversed by reversing the signal in time before passing it through the filter, which is trivially easy for signals stored in a computer. The only problem is to decide on the pole and zero co-ordinates to produce the right amount of phase correction. It is, however, quite easy to get an approximate setting for the phase correction merely by adjusting it to get the most plausible glottal pulse shape, after the formant ripple has been cancelled as well as is possible.

Rothenberg [9] has avoided the low-frequency phase distortion problem and the need for an integrator by covering the mouth with a special mask that has instrumentation for measuring the volume velocity directly, providing a response down to zero frequency. Unfortunately the measurement system has a bandwidth of less than 2 kHz, and the inconvenience of requiring the speaker to wear special equipment means that it is difficult to collect large amounts of data. However, an important advantage of Rothenberg's method is that it can provide calibration for the magnitude of typical glottal flows, and can give a guide of the sort of pulse shapes to be aimed for when adjusting a phase correction filter. Thus Rothenberg's findings can be used to validate the results of inverse filtering of the pressure waveform.

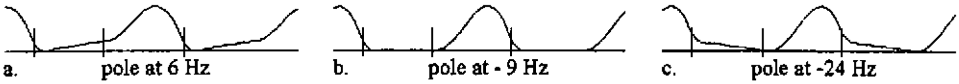
a. pole at 6 Hz b. pole at - 9 Hz c. pole at -24 Hz

Fig. 2. Effects of different amounts of low-frequency phase correction.

Fig. 2 illustrates a typical inverse-filtered glottal pulse with three different amounts of low-frequency phase correction. Fig. 2b has the pole of the correcting filter at -9 Hz, which was judged to be the most plausible setting for this signal, but it is not necessarily correct. Varying this correction by 15 Hz in each direction (Figs. 2a and 2c) makes a very noticeable difference to the shape of the derived pulse, but fortunately the perceptual effect of such small phase errors would be undetectable in speech synthesis.

2.3 Low-frequency noise

Random variations of air pressure at sub-sonic frequencies will normally be present in any room used for speech recordings. Although these variations are very small compared with the signal level in the speech frequency range, the use of an integrator to derive the glottal flow will greatly increase their relative amplitude. Over the duration that needs to be displayed to show a small length of glottal waveform, the main effect of this low-frequency noise will be to add an approximately linear function to the displayed waveform. The effect can be cancelled, therefore, simply by subtracting a linear function whose parameters are chosen to make the derived glottal waveform sit in contact with a horizontal baseline.

2.4 Ripple components in the effective glottal flow

When the glottis is closed, the source waveform into the vocal tract is often assumed to be precisely zero. Within the limitations of the vocal tract model, it would then be possible to cancel any formant ripple exactly, to produce zero response during the closed-glottis period. When the glottis is open the effect of its acoustic impedance on the complete vocal system will be to modify both the formant frequencies and the bandwidths. The settings of the inverse filter needed to cancel formant ripple when the glottis is open will not then be correct for when it is closed. However, the closed-glottis settings will still be correct for cancelling the transfer function from glottis to lips, and the ripple components that will be apparent when the glottis is open will be genuine components of the glottal flow. These ripple components are thus not an indication of any deficiency of the inverse filtering.

However, if the purpose of inverse filtering research is to study properties of voiced excitation signals, it is not sensible to regard the formant ripple components of glottal flow as part of the excitation, because they depend on the vocal tract configurations for particular vowels. It would obviously be more useful to model the glottal-area-dependent formant modifications within the vocal tract model, and to use the underlying glottal flow shape without the ripple components.

2.5 Vocal fold surface movements when the glottis is closed

When the glottis is completely closed there is obviously no air flow between the vocal folds. However, it is clear from high-speed motion pictures of the folds during phonation [10] that the vocal fold surface tissue shows considerable movement during the closed phase, so will cause some air movement at the bottom of the pharynx. This type of movement has been confirmed by Baer [11] in a study of phonation on excised dog larynxes. Baer's experiments show that the vocal folds make their initial contact at a low position within the glottis. The folds subsequently come into contact higher up and they simultaneously move upwards over the next millisecond or so. There is thus an additional effective flow component immediately after first contact of the folds, caused partly by the overall upward movement and partly by the ejection of any small wedge of air that might lie between the folds at the moment of contact. Baer estimates that the effective air displacement rate caused by these surface movements just after glottal closure is around 20 – 30 ml / s, compared with perhaps 500 ml / s maximum flow through the glottis. Although these observations refer to the general pattern of his results on a large number of separate larynxes, he also reported considerable differences of detail from one larynx to another. There is every reason to believe that human larynxes, which are similar in general form and dimensions to the dog larynxes used by Baer, will show a similar range of variation.

It can be seen in high-speed pictures [10] that there are also significant ripples flowing across the surface of the vocal folds immediately after closure. In terms of volume displacement they are very small compared with the peak flow in a typical glottal pulse, but almost all of the power in the actual flow through the glottis is at the fundamental frequency. At typical F1 and F2 frequencies the power is so much less that these ripple components can cause significant modification to the overall spectrum. In consequence, when adjusting the inverse filter to minimize closed-phase ripple it is often not possible to get very good cancellation throughout the closed phase and it is frequently necessary to choose the least-disturbed part of the waveform for making the adjustments.

3. INVERSE FILTERING PROGRAM

A new inverse filtering program has been written to run on a PC with interactive control from the keyboard and from the mouse. It uses a sampled-data inverse filter with a sampling rate of 8 kHz. The positions of the inverse filter zeros are shown as equivalent s-plane co-ordinates, and the inverse filtered waveform for a selected 20 ms region of the input signal is also displayed. Up to five complex zeros can be provided in the inverse filter, but any of the available zeros can be optionally excluded. All of the zeros can be moved in frequency and bandwidth by dragging with the mouse. As any zero is being moved the inverse filtered waveform is continuously updated to take into account the change of inverse formant characteristics. It is thus extremely easy to get almost optimum cancellation of formant ripple within a few seconds by manipulating the zero co-ordinates in turn. For very fine adjustment of the zeros it is better to use the four arrow keys, which can increment or decrement the frequency or bandwidth in 5 Hz steps. This size of step will usually produce a just-noticeable difference to the inverse filtered waveform for signals for which good formant cancellation is possible.

The inverse filtered waveform can be shown in three different modes. The first mode only applies the zeros to the signal, so the waveform represents the differentiated glottal flow. Pressing the "I" key will integrate this waveform to show the volume velocity pulse. The "D" key will differentiate the waveform, to show the twice-differentiated flow. This condition makes it easier to adjust the parameters of the higher formants, by preventing the waveform from being dominated by the low-frequency components.

An adjustable low-frequency phase corrector of the type described in Section 2.2 is included to compensate for any observed phase distortion. In the integrated mode the low-frequency noise problem is successfully dealt with by subtracting a linear function with the appropriate slope.

The ability to remove any one of the zeros from the inverse filter has an important use for illustrating how the voicing is exciting any particular formant. Whenever there is significant excitation of a formant caused by movement of the surface of the vocal folds during glottal closure, this will be in evidence in the form of changes in intensity or phase of the formant ripple displayed on the screen.

When any stage of the inverse filter is excluded, the mouse and arrow keys are used to control the frequency and bandwidth of a synthetic formant. The synthetic formant waveform can be moved about on the screen, and if required it can be superimposed on the natural formant waveform, but in a different colour. For getting a close waveform match to the natural formant the amplitude and phase of the synthetic waveform can also be controlled. It has been found that it is easy to adjust the parameters of this synthetic formant to get the best match possible to any chosen section of the natural waveform. This technique makes small errors in the formant parameters more obvious, so enabling parameter values to be obtained more precisely than is possible by moving the inverse filter zero directly. While making these adjustments, the best approximation to the true formant parameters can be obtained by choosing the part of the natural formant waveform that matches best to an exponential decay.

4. EXPERIMENTS AND DISCUSSION

4.1 Speech material

The aim of this study was to compare glottal pulse shapes for many speakers under similar conditions. The Scribe database of British English [12] includes a large number of male and female speakers, all speaking the same narrative passage. It was recorded in an anechoic chamber under carefully controlled conditions. For this paper the data on the Scribe 0 CD-ROM (covering speakers from the south-east region of England) was chosen, and the first six speakers of each sex were used. For convenience the speakers will be referred to by the names of the sub-directories where the signals are stored on the CD-ROM. The names for the male and female speakers begin with M and F respectively. For speaker MAC four vowels were chosen for study, with different formant frequencies. For all other speakers only one vowel was used, from the word "craft" in the first sentence of the passage about sailing.

4.2 Results

Fig. 3 shows a wide band spectrogram of 0.1 s of speech signal taken from the second half of the word "sailor" in the first sentence of the passage spoken by MAC. A portion of duration 20 ms has been chosen for detailed analysis, for which the results are presented in Fig. 4. Fig. 4a shows the chosen 20 ms of speech waveform, which was taken from the neutral vowel in the second syllable of the word. Figs. 4b, 4c and 4d show the inverse filtered glottal flow and its first two time-derivatives. The estimated points of glottal opening and glottal closure are marked. Although the volume velocity pulse appears to be very smooth, it is obvious from looking at the derivative waveforms that there is a lot of detailed structure in the high frequency components. The peak in the second derivative waveform which occurs at glottal closure is about 0.55 ms before the flow waveform reaches the baseline. There is a smaller second peak in Fig.4d, which occurs as the baseline is reached. This waveform shows components at frequencies near to the second formant during the closed-glottis region. In Fig. 4e the inverse-filter zero cancelling the second formant has been removed. The resultant waveform shows that the amplitude of F2 noticeably increases two cycles after glottal closure and suddenly reduces again another two cycles later. There are then three cycles during which F2 appears to be decaying smoothly, followed by another increase of amplitude at the point of glottal opening. The synthetic formant waveform shown in the top part of Fig. 4e has been the adjusted to match the smoothly decaying part of the natural F2.

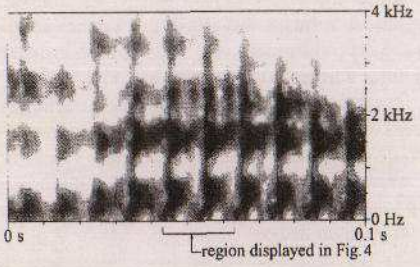

Fig. 3. Wideband spectrogram of speech for analysis

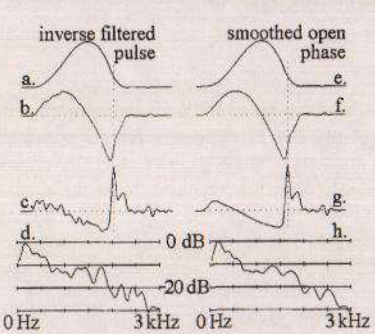

Fig. 5. Glottal pulse spectra

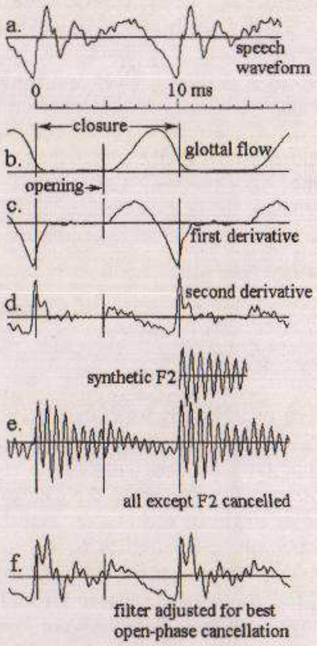

Fig. 4. Inverse filtering of region marked in Fig. 3.

It seems most likely that for this talker the peak in the second derivative waveform is a consequence of some discontinuity of movement of the vocal folds approximately 0.55 ms after they first make contact. This type of discontinuity is consistent with the motion observed by Baer and described in Section 2.5.

Fig. 4d shows quite a lot of ripple in the second derivative of the glottal flow during the open-glottis phase. The zeros for F1 and F2 were adjusted to try to minimize this ripple and the results are shown in Fig. 4f. Although the open-phase ripple has thereby been reduced, this readjustment has completely disturbed the cancellation of the formants in the closed-glottis region.

One cycle of the glottal pulse shown in Figs. 4b, 4c and 4d has been excised, and is plotted in Figs. 5a, 5b and 5c. The twice-differentiated pulse of Fig. 5c was padded with zeros to four times its length and Fourier analysed. The resultant spectrum up to 3 kHz is shown in Fig. 5d. The spectrum shape up to about 600 Hz is largely attributable to the general shape of the glottal pulse, and would be fairly similar for most male speakers. However, there is a region of lower spectral intensity between 750 and 1150 Hz which requires explanation. The waveform of Fig. 5c shows noticeable ripple during the open glottis interval (caused mainly by shift of the F2 frequency) and also activity in the closed period, as discussed above. The open-phase ripple was smoothed by hand to produce the waveforms shown in Figs. 5e, 5f and 5g. This modification has only smoothed out some of the fine spectral detail (Fig. 5h), thus leading to the conclusion that this spectral dip is caused by the closed-phase activity. The fact that the waveform of Fig. 5c shows two peaks approximately 0.55 ms apart would, of course, be expected to produce a spectral dip around 900 Hz and a peak at about 1800 Hz.

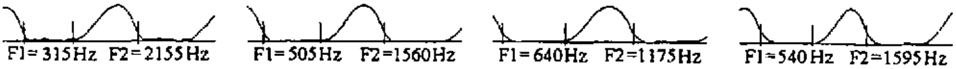

Fig. 6. Glottal pulses derived from four different vowels for speaker MAC.

Fig.6 shows the glottal flow for four vowels from speaker MAC, which have very different formant frequencies. All four flow pulses show similar features, although there are some differences in duration, associated with different values of fundamental frequency. Examination of the twice-differentiated glottal pulses and the spectra showed very similar features in all cases, with some form of double peak near glottal closure and a spectral dip somewhere in the range between 700 and 1200 Hz.

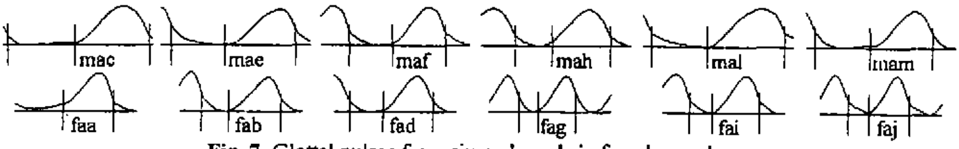

Fig. 7. Glottal pulses from six male and six female speakers.

Fig. 7 shows glottal pulses for six male and six female speakers, all from the same phoneme. All of the new speakers show much more departure from the baseline during the closed-glottis period than occurs for speaker MAC. This effect would depend on the phase correction, but the unchanged recording environment should require the same settings. Speakers MAE and MAM showed some signs of a double peak in the twice-differentiated waveform, and there were dips in the pulse spectrum around 1500 Hz and 1000 Hz respectively. For speakers MAF, MAH and MAL there was no sign of a double peak, and no noticeable dip in the pulse spectrum. None of the female speakers showed any sign of the double peak at closure, and in general their closures were less sharp and the baseline showed greater curvature than the males, implying a larger relative up and down movement of the vocal folds. All speakers show a clear indication that the vocal folds make contact before the effective glottal flow has finished.

The above types of behaviour have been observed by many other workers, and in particular by Fant, Liljencrants and Lin [1], who developed a four-parameter model of glottal flow (the LF model) to describe their observed shapes. However, they modelled the region immediately after the first contact of the vocal folds as an exponential decay, and attributed the effective flow in this region to "dynamic leakage", assuming that the vocal folds do not meet completely along their entire length. Although it is true that speakers may not always achieve complete closure of the glottis, in view of the discussion in Section 2.5 it seems far more likely that this trailing shape after glottal closure is more often caused by the vocal fold surface movements reported by Baer and seen on high-speed films. For those speakers who show the double peak in the twice-differentiated flow, the exponential decay model is not satisfactory because it is unable to represent the resultant spectral dip and its effect on excitation of the formants. The parameters of the LF model can be used to make a very close fit to the spectrum of a natural glottal pulse up to around 500 Hz, and by adjusting the time scale of the exponential decay it is also possible to achieve the right balance between high and low frequency components. However, the LF model provides no possibility of representing the complexity of formant excitation in the F2 and high F1 ranges that is commonly observed in human speech. It seems likely that better modelling of these detailed aspects of glottal flow will be necessary for very-high-quality formant synthesis, and the results in this paper suggest that the requirements will differ for different speakers. Further investigation of these aspects is planned.

6. REFERENCES

[1] G. FANT, J. LILJENCRANTS and Q. LIN, 'A four-parameter model of glottal flow' *Speech Transmission Laboratory QPSR 4*, pp.1-13, Royal Institute of Technology, Stockholm, (1985)

[2] G. FANT, 'The acoustic theory of speech production' Mouton & Co., The Hague (1960)

[3] G. FANT, 'Acoustic analysis and synthesis of speech with applications to Swedish', *Ericsson Technics*, **1**, pp.3-108 (1959)

[4] R. L. MILLER, 'Nature of the vocal cord wave' *J. Acoust. Soc. Am.*, **31**, pp.667-677 (1959)

[5] J. N. HOLMES, 'An investigation of the volume velocity at the larynx during speech by means of an inverse filter' *Proc. IV Int. Congr. Acoust.*, Copenhagen (1962)

[6] M. J. HUNT, J. S. BRIDLE and J. N. HOLMES, 'Interactive digital inverse filtering and its relation to linear prediction methods' *Proc. IEEE ICASSP*, Tulsa, pp.15-18 (1978)

[7] J. N. HOLMES, 'Formant synthesizers, cascade or parallel?' *Speech Communication*, **2**, pp.251-273 (1983)

[8] J. N. HOLMES, 'Low frequency phase distortion of speech recordings' *J. Acoust. Soc. Am.*, **58**, pp.747-749 (1975)

[9] M. ROTHENBERG, 'A new inverse-filtering technique for deriving the glottal air-flow waveform during voicing' *J. Acoust. Soc. Am.*, **53**, pp.1632-1645 (1973)

[10] D. W. FARNSWORTH, 'High-speed motion pictures of the human vocal cords' *Bell Lab. Record*, **18**, pp.203-208 (1940)

[11] T. BAER, 'Investigation of phonation using excised larynxes' Ph.D. Thesis, Cambridge, Mass. (1975)

[12] 'The SCRIBE database' Prepared by the Speech Research Unit, DERA, Malvern, UK (April 1992)