TIME-FREQUENCY ANALYSIS OF THE GLOTTAL OPENING

Wolfgang Wokurek
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart, Azenbergstraße 12, 70174, Germany
wokurek@ims.uni-stuttgart.de

ABSTRACT

Simultaneous recordings of the laryngograph signal and speech recorded in an non-reverberating environment are investigated for acoustic evidence of the glottal opening within the microphone signal. It is demonstrated that the high resolution time-frequency analysis of the microphone signal by the smoothed pseudo Wigner distribution (SPWD) shows responses of the vocal tract to both, the glottal closure and the glottal opening. Thus, a convolution-based model for the relation between the laryngograph signal and the microphone signal is evaluated. It turns out, that the microphone signal may be viewed as filtered version of a power function of the laryngograph signal. Hence, such a nonlinear processed laryngograph signal may be an appropriate model for the acoustic excitation of the vocal tract.

1. INTRODUCTION

Acoustic excitation of voiced speech sounds is performed mainly by the closing of the vocal folds. However, the glottal opening contributes to the acoustic waveform of that sounds. A very successful model for the acoustics of speech production is the source filter model [6]. For voiced sounds its source signal is produced at the glottis, that is filtered by the resonances of the vocal tract. Unfortunately there is no direct access to the source signal (pressure or velocity) of the vocal tract.

At least three physical strategies (aerodynamical, optical, electrical) are available to acquire signals that are indirectly related to the glottal signal. The air stream through the nose and the mouth is recorded with the Rothenberg mask covering nose and mouth [10]. The air stream is an acoustic quantity but at this end of the vocal tract an inverse filtering technique is necessary to estimate the glottal signal. Optical methods are transillumination and high speed filming [4]. Both methods try to measure the open area of the glottis. Transillumination uses a fiberscope inserted through the nose to iluminate the gottis, and the light intensity passing through the glottis is measured by a photosensor externally attached to the larynx. High speed filming is performed through the mouth via a mirror located near the uvula or again through a fiberscope inserted through the nose. Finally, the electrical conductivity of the glottis is exploited by the laryngograph (electroglottography), which requires only a set of electrodes externally attached to the larynx [1], [2], [3], [5]. All these methods cause discomfort to the speaker. This study reports on experiments that exploit the microphone signal and the laryngograph signal — maybe the least discomfort ones.

Applications of good source signal estimates include: phonetic phonation style research, which might improve speech synthesis, automatic speech and speaker recognition, as well as speaker verification; diagnostics of speech pathology [7] e.g. dysphonia; and forensic applications.

The paper consists of two main parts. Section 3. focusses on the time frequency analysis of the glottal opening, and section 4. presents a nonlinear model for the estimation of the acoustic excitation from the laryngograph signal.

2. RECORDING

Vowels were spoken in a sound treated speaker cabin and registered simultaneously with a laryngograph and by a microphone. A reduction of the reverberations is important since delayed versions of the glottal closure impulse might be confused with the glottal opening impulse which is the main subject of this study. Analyses were performed on sampled versions of that signals (16 bit, 16 kHz).

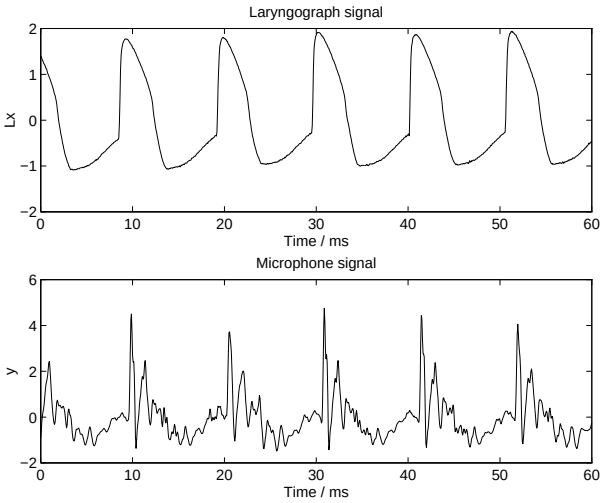

Figure 1. The laryngograph signal

The laryngograph uses a set of electrodes, contacted to the throat left and right to the larynx, and measures a signal of electrical conductivity that shows the vocal fold vibration. Figure 1 shows a sample waveform of the laryngograph signal. Glottal closing is indicated by a rapid increasing of the electrical conductivity. Usually, the opening phase of the glottal cycle is slower than the closing phase causing a slower conductivity decreasing. The laryngograph signal has a shape that is comparable to the volume velocity of the airstream through the glottis, which derivative is the pres-

sure signal. In conjunction, volume velocity and pressure form acoustic waves that excite the resonances of the vocal tract and the nasal cavity. Therefore the glottal opening is likely to produce an acoustic excitation of lower amplitude and bandwidth compared with the glottal closure.

3. TIME FREQUENCY ANALYSIS

Both the laryngograph and the microphone signal are analyzed with the same time-frequency analysis method, the smoothed pseudo Wigner distribution (SPWD) of the analytical signal [8], [9]. To achieve a better representation of the high frequency components of both signals, a standard preemphasis of the high frequency ($H(z) = 1 - \alpha z^{-1}$, $\alpha = 0.97$) is applied. The positively valued regions of the SPWD are displayed in logarithmic scale, showing a dynamic range of 35dB (of the preemphasized signal). Time and frequency smoothings are adjusted to produce a time resolution of 1.5ms and a frequency resolution of 150Hz.

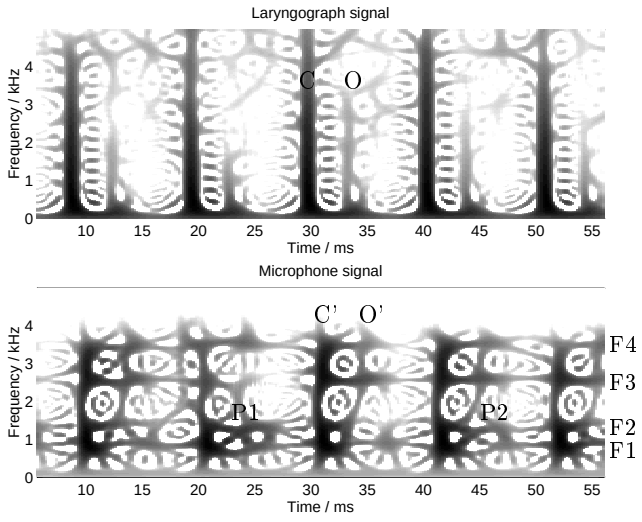

Figure 2. SPWD analysis

Figure 2 shows SPWD analyses of five pitch periods of the steady state segment of the vowel [ɑː] uttered by a male speaker. The SPWD of the laryngograph signal (upper diagram) clearly shows five stronger broadband pulses (represented as vertical bars) followed by weaker pulses. The stronger pulses are interpreted as the main acoustic excitation by the glottal closing (label C). The weaker pulses follow after approximately 3ms and are caused by the glottal opening (label O). The same structure is found in the SPWD of the microphone signal (lower diagram). The propagation of the acoustic wave from the glottis to the microphone (0.35m at a speed of $343\,\mathrm{m\,s^{-1}}$) causes a delay of about 1ms between the speech and the laryngograph signal. Both the closure (label C') and the opening (label O') are present in all five pitch periods of the acoustic signal. Additionally the first four formants (labels F1 – F4) are shown as damped oscillations following each excitation either from the glottal closure or from the glottal opening.

The question arises under what conditions and how the remains of the glottal opening is visible in the SPWD of the microphone signal. Firstly the glottal opening should be fast enough to produce a sufficiently strong acoustic excitation. Secondly the analysis method should be capable of representing it. Although the SPWD analysis method is limited through the presence of interference terms in multicomponent signals, it is used here because of its capability of simultaneous high time and frequency resolutions.

Within one pitch period, time smoothing reduces the interference between different formants. With a time resolution of 1.5ms these interference is suppressed for formants that are separated by at least $\frac{1}{1.5\mathrm{ms}} = 670\mathrm{Hz}$. Only vowels with front tongue position e.g. [iː, ɪ, eː, ɛ] have such a good separation — and therefore lack of interference — between the first two formants. Unfortunately, vowels with high or back tongue position e.g. [ɑː, a, oː, uː] have less distance between their first and the second formant. Hence, the glottal opening most likely is buried by these first two formants and by their interference.

The interference betweend adjacent pitch periods is governed by the frequency resolution of 150Hz. The interference phenomena are limited tho those signal components that are grasped with the same analysis window of about $\frac{1}{150\mathrm{Hz}} = 6.7\mathrm{ms}$ duration. Therfore, interference between glottal closures is expected for fundamental frequencies above 150Hz.

Interference will happen between formant oscillations and the 'new' excitation due to both glottal closure and glottal opening. In particular the latter may supply a hint to a weak glottal opening that is not displayed explicitly as impulse like vertical bar. Such a glottal opening may be the reason of simultaneous 'disturbance' of all formant oscillations (i.e. deviations from the horizontal line) as shown in figure 2 labels P1, P2. However, the display of the glottal opening is blurred, if it occurs close to a glottal closure or if it is too weak.

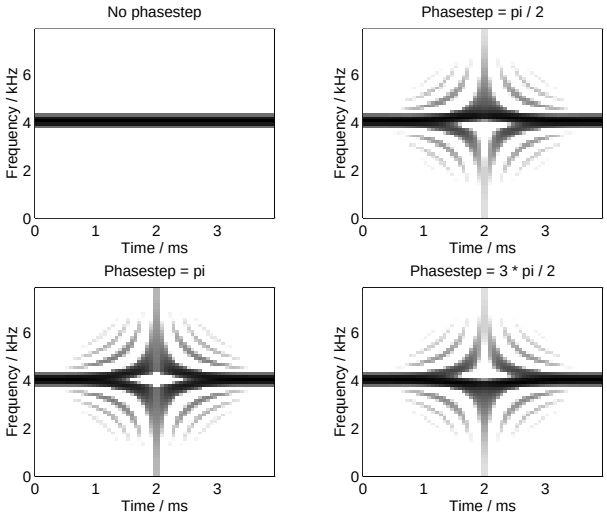

Figure 3. The phase step effect

To show this effect more clearly, figure 3 shows SPWD's of an oscillation that is subject to a sudden phase step at the center of the displayed time segment. Different amounts of the phase step cause different positions of a 'hole' disturbing the horizontal line of the oscillation. Here, the sudden phase step is used as a model for the 'out of phase' excitation of a formant oscillation due to excitation by glottal closure and glottal opening.

A speculation might be that the glottal opening will be found in the SPWD of stressed syllables but not in that of unstressed syllables. However, testing German vowels in both positions suggested independence between word

stress and the visibility of the glottal opening in the SPWD. Clearly, as a feature of the vocal fold vibration it is more directly related to phonation and only indirectly to the phonological or the syntactic structure of speech.

4. MODEL

So far, the observations mention that there might be a tight relation between the laryngograph signal $L_x(n)$ and the excitation signal $x(n)$ of a source filter model for the microphone signal $y(n)$. For simplicity, the nonlinear function

$$f_\beta(x) = \text{sign}(x)\ |x|^\beta \tag{1}$$

of the laryngograph signal is assumed to be the excitation signal

$$x(n) = f_\beta(L_x(n)) \quad . \tag{2}$$

Since the stationary segments of vowels are considered, a causal linear time invariant filter with impulse response $h(n)$ is used as a model for the vocal tract, the radiation (and the recording equipment). Hence, the microphone signal is

$$y(n) = \sum_{k=D}^{D+N-1} h(k)\ x(n-k) \quad , \tag{3}$$

where the delay due to sound propagation is explicitly modelled by D. N determines the duration of the impulse response.

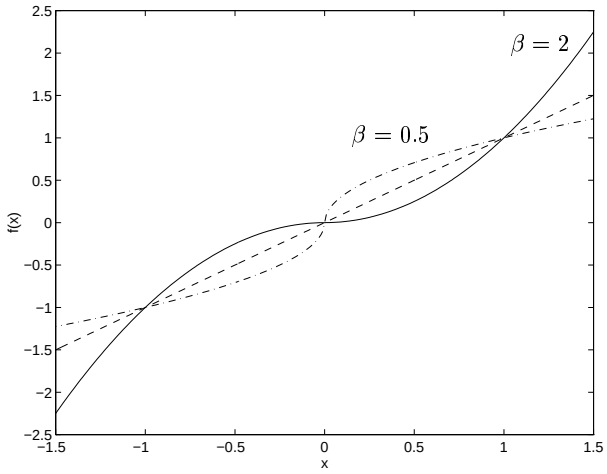

Figure 4. The nonlinear function

Figure 4 shows the nonlinear function Eq.(1) for three values of the nonlinearity parameter $\beta = 0.5, 1, 2$. The function may express the one to one connection with $\beta = 1$, reduction of slopes and peaks with $\beta < 1$ as well as increasing of slopes and peaks with $\beta > 1$. Obviously for every positive parameter $\beta > 0$ the inverse function of $f_\beta(.)$ is

$$(f_\beta)^{-1}(.) = f_{(\beta^{-1})}(.) \tag{4}$$

a nonlinearity of the same type with the reciprocal value of the nonlinearity parameter β^{-1}.

4.1. Inverse Filter

The assumption of an autoregressive model for the microphone signal $y(n)$ allows the application of LPC techniques and inverse filtering to calculate the excitation signal $x(n)$. Using different estimators with various model orders produced estimated excitation signals that mostly had a shape so different to $L_x(n)$, that it seemed unreasonable searching for a β to fit the nonlinear function Eq.(4).

A closer look revealed that mostly one or more formants or residuals of them were reminiscent in the inverse filtered signal $x(n)$. In the case of a weak first formant often the fundamental frequency was canceled and the first formant dominated the residual signal. And in that rare cases where the formants were sufficiently compensated, the shape was different — maybe due to group delay differences.

4.2. Deconvolution

The following FFT based deconvolution procedure led to reasonable results with many of the tested recordings. The aim of the algorithm is to estimate the N point impulse response with D points delay $h(n)$ in Eq.(3), given the nonlinearity parameter β, the nonlinear processed laryngograph signal $x(n)$ (Eq.(2)) and the microphone signal $y(n)$. Both given signals are transformed to its discrete time Fourier transform (DTFT)

$$X(k) = \underset{n\to k}{\text{FFT}_L}(x(n)), \quad k, n = 0, \cdots, L-1, \quad L = 2^l \tag{5}$$

$$Y(k) = \underset{n\to k}{\text{FFT}_L}(y(n)) \tag{6}$$

using the L-point radix-2 fast Fourier transform. A transfer function is estimated by

$$\tilde{H}(k) = \begin{cases} \frac{Y(k)}{X(k)} & : \text{ if } |X(k)| > \epsilon \\ 0 & : \text{ else} \end{cases} \quad , \quad \epsilon > 0 \tag{7}$$

and transformed back to time domain by inverse fast Fourier transform

$$\tilde{h}(n) = \underset{k\to n}{\text{FFT}_L^{-1}}(\tilde{H}(k)) \quad . \tag{8}$$

To reduce the aliasing that is almost certainly introduced by Eq.(7) and the cyclic convolution properties of the DTFT, the block length L is selected according to

$$\frac{L}{2} = 2^{l-1} > \max(\text{length}(x), \text{length}(y)) \quad , \tag{9}$$

which would be sufficient for an finite impulse response of $\frac{L}{2}$ points. Note that the sampled signals are filled up with zeros to the transform length L.

The N point impulse response with D points delay $h(n)$ is extracted by

$$h(n) = \begin{cases} \tilde{h}(n) & : \text{ if } D \le n < D+N-1 \\ 0 & : \text{ else} \end{cases} \tag{10}$$

which may be interpreted as orthogonal projection onto the space of all D points delay impulse responses of length N.

The inverse filtering is performed using the DTFT of the impulse response Eq.(10)

$$H(k) = \underset{n\to k}{\text{FFT}_L}(h(n)) \tag{11}$$

and the allready calculated DTFT spectrum Eq.(6) by

$$\hat{X}(k) = \begin{cases} \frac{Y(k)}{H(k)} & : \text{ if } |H(k)| > \epsilon\,|Y(k)| \\ 0 & : \text{ else} \end{cases} \tag{12}$$

yielding an estimate of the acoustic excitation signal

$$\hat{x}(n) = \underset{k \to n}{\mathrm{FFT}_L^{-1}}(\hat{X}(k)) \quad . \tag{13}$$

Using the inversion of the nonlinear transform Eq.(2), the acoustic excitation $\hat{x}(n)$ is transformed to the hypothetic laryngograph signal

$$\hat{L_x}(n) = f_{\frac{1}{\beta}}(\hat{x}(n)) \tag{14}$$

that should coincide with the laryngograph signal $L_x(n)$, if the assumed nonlinearity Eq.(2) were correct and the impulse response $h(n)$ were the impulse response of the vocal tract. For the purpose of optimizing the nonlinearity paramenter β, the mean squared error between the laryngograph signal $L_x(n)$ and the hypothetic laryngograph signal $\hat{L_x}(n)$

$$e = \frac{1}{\mathrm{length}(L_x)} \sum_{n=0}^{\mathrm{length}(L_x)-1} \left(L_x(n) - \hat{L_x}(n)\right)^2 \tag{15}$$

will be minimized.

5. EVALUATION

The stationary vowel segments of [iː, eː, ɑː, oː, uː] uttered by two male speakers is the speech material for the evaluation of the proposed model and of the deconvolution procedure. Both sampled signals, the laryngograph and the microphone signal are normalized by its root-mean-square values.

Sound	β_{opt}	e_{min}
iː	1.12	0.22
eː	1.43	0.69
ɑː	1.67	0.11
oː	1.06	0.31
uː	1.38	0.36
iː	1.74	0.44
eː	1.64	0.27
ɑː	1.69	0.15
oː	1.66	0.32
uː	1.63	0.29

Table 1. Optimized $\beta \in [0.5, 2]$

An impulse response of 0.3ms delay ($D = 5$) and 6ms duration ($N = 95$) is selected to give a good subjective overall performance of the deconvolution algorithm after experimenting with different delays and durations. For each individual vowel the nonlinearity parameter β_{opt} is sought in the range $[0.5, 2]$ that minimizes the error Eq.(15). Table 1 shows these optimal parameters and the minimal error for each sound. Figure 5 shows both time signals, the laryngograph signal and its approximation by the nonlinear model for the [ɑː] sound of the second speaker. The first five rows correspond to the first speaker. Note, that his optimal parameters scatter from 1 to 1.4, whereas they are much more concentrated around 1.7 for the second speaker.

6. CONCLUSION

A first result of this work is that in some cases the glottal opening may be observed by high resolution time-frequency analysis of the acoustic speech signal. This observation was made, comparing SPWD plots of speech and of the laryngograph signal. Secondly, this survey results in a simple model

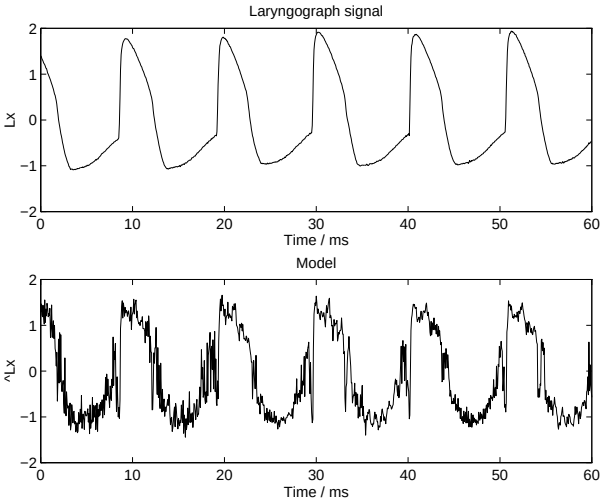

Figure 5. The laryngograph signal and its model

relating a nonlinear function of the laryngograph signal to the acoustic speech signal by a filter. This may be exploited to improve inverse filtering techniques to get indirect access to the acoustic excitation signal of the vocal tract.

Extensions of this brief study should include more speakers, female voice and different phonation types to allow a better understanding of the instant and the strength of the glottal opening and on the conditions that allow its observation in the microphone signal.

REFERENCES

[1] R.J. Baken, "Electroglottography", *Journal of Voice*, Vol. 6, No. 2, pp. 98-110, 1992.

[2] D.G. Childers, A.K.Krishnamurthy, "A critical review of electro-glottography", *Crit. Rev. Biomed. Engin.*, vol 12, pp. 131-161, 1985.

[3] R.H Colton, E.G. Conture, "Problems and Pitfalls of Electroglottography", *Journal of Voice*, Vol 4, No 1, pp. 10-24, 1990.

[4] T. Baer, A. Löfqvist, N.S. McCarr, "Laryngeal vibrations: a comparison between high-speed filming and glottographic techniques", *Journal of the Acoustical Society of America*, vol. 73, pp. 1304-1308, 1983.

[5] J.H. Essling, "Laryngographic study of phonation type and laryngeal configuration", *J. of the Int. Phonetic Association*, vol. 14, No. 2, pp. 56-73, 1984.

[6] G. Fant, "Acoustic Theory of Speech Production", Mouton, The Hague, 1960.

[7] A.J. Fourcin, "Normal and pathological speech: phonetic, acoustic and laryngographic aspects", Shingh, W. et al. (eds): Functinal surgery of the larynx and pharynx, pp. 31-51, 1993.

[8] F. Hlawatsch, and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations", *IEEE Signal Processing Magazine*, vol. 9, pp. 21-67, 1992.

[9] G. Kubin, F. Hlawatsch, and W. Wokurek, "High resolution time-frequency analysis of speech signals", Int. Conf. on Phonetic Sciences, 1987, Tallin, Estonia.

[10] M. Rothenberg, "Some relations between glottal air flow and vocal fold contact area", *ASHA Reports*, vol. 11, pp. 88-96, 1981.