Inter- and Intra-speaker Variability of Glottal Flow Derivative using the LF Model

Markus R. Iseli and Abeer Alwan

Dept. of Electrical Engineering UCLA
Los Angeles, 90095

ABSTRACT

The vowels /a, i, u/ spoken by American English talkers with non-pathological voices are described by means of voice source model parameters using the Liljencrants-Fant (LF) model. The sampling frequency of the data is 8 kHz which matches approximately telephone bandwidth. After inverse filtering, trends of voice source characteristics depending on the LF parameters are analyzed and compared to literature and listening results.

Keywords: voice source, LF model, LF parameters.

1. INTRODUCTION

Non-pathological voice source characteristics have been studied by inverse filtering the speech waveform [11], analyzing the speech spectra [6], or by measuring the airflow at the mouth [10]. Knowing the voice source parameters can be beneficial for many speech processing applications, such as speaker identification [8], and speech synthesis. In [6], individual and gender variations in source parameters have been analyzed using measures from speech spectra and taking into account the influence of the vocal tract transfer function. In [2], the voice source measures were derived from the Liljencrants-Fant (LF) model parameters which describe the glottal flow derivative.

In this paper, a modified version of the algorithm in [8] was adopted. The algorithm involves inverse-filtering and course LF parameter estimation. Source - vocal tract interaction [4] is an important issue in inverse filtering techniques, and to get reliable data the inverse filtered data were checked manually before estimating the LF model parameters.

2. DATA DESCRIPTION

Acoustic waveforms of the vowels /a, i, u/ were extracted from /ba, bi, bu/ tokens of a CV database which was digitally recorded at 16kHz and contains recordings of two male (M1, M2) and two female (F1, F2) talkers of American English with non-pathological voices. For our purposes a downsampled version at 8kHz (to match 4kHz telephone bandwidth) was used and a 125 ms segment chosen in the middle part of the vowel, where the waveform was relatively steady, was analyzed. It was important that the influence of the preceding consonant on the vowel was as small as possible and that the tapering at the end of the utterance was not considered. Eight repetitions of each vowel were used for each talker which gives a total of one second of speech per vowel and talker and a total of four seconds of speech per vowel. In some of the following calculations it is assumed that every token of 125ms is wide-sense-stationary.

3. DATA PROCESSING

To obtain the LF model parameters for the voice source flow derivative, the recorded speech was inverse filtered. The vocal tract filter is modeled as an AR process (appropriate mainly for vowels) and the lip-radiation filter as an MA process. Lip-radiation acts like a differentiator on the flow velocity at the mouth thus transforming flow velocity to pressure. The AR and the MA filters can be interchanged without changing the speech production model and thus the source-flow derivative is represented by the flow source and the MA differentiator filter. To calculate the source-flow derivative, linear prediction analysis of the speech signal provides an estimate of the vocal tract transfer function which is used to inverse-filter the speech signal [8].

The next step is to match the inverse filtered signal with the LF model. Glottal flow and its derivative represented by an LF model with its 7 parameters[1] are shown in Figure 1 for one glottal cycle of duration T_0 where $F_0 = 1/T_0$ is the pitch period.

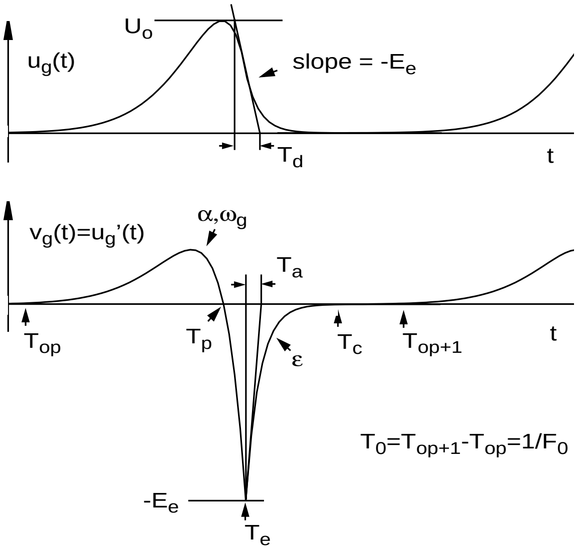

Figure 1: Glottal flow $u_g(t)$ and its derivative $v_g(t)$ with the LF model parameters.

Because the model represents the derivative of the glottal flow, a constant flow at the glottis, as occurring when a talker has a glottal leakage, cannot be detected easily. The 3 phases of the glottal cycle are the closed phase, the open phase and the return phase. The closed phase is described as the phase with no flow changes and begins with time T_c. When the glottis starts to open at T_{op}, the interaction between the glottis and vocal tract load is high until a more constant flow has been achieved. At time T_p the glottal flow reaches its maximum and at time T_e, the negative amplitude E_e of the flow derivative shows the maximum rate of flow decrease. Although the flow is reduced after time T_p, the return phase defined in the literature does not begin until after time T_e. The open and return phases are modeled by the equations

[1] The LF model is inherently a five-parameter model as for synthesis parameters T_a and T_c are not needed.

$$v_g(t)=\begin{cases}\frac{-E_e}{\sin(\omega_g(T_e-T_{op}))}e^{\alpha(t-T_e)}\sin(\omega_g(t-T_{op})) & \mathrm{T}_{op}\le t\le T_e\\ \frac{E_e}{\varepsilon T_a}[e^{\varepsilon(T_e-t)}-e^{\varepsilon(T_e-T_c)}] & T_e<t<T_c\\ 0 & \text{else}\end{cases} \qquad (1)$$

The growth factor α determines the ratio of E_e to the peak height of the positive portion of the glottal flow derivative and parameter ω_g can be calculated by the formula

$$\omega_g = \pi / (T_p - T_{op}). \qquad (2)$$

Parameter T_a is the effective duration of the return phase, whereas ε is an exponential time constant and is not considered as a separate parameter because for small enough T_a, εT_a is close to 1 [3].

Though a closed form frequency response representation of these equations does not exist it can be shown that a change of parameters used in the open phase equation does affect more the lower frequencies whereas the higher frequencies are more affected by the return phase which can be described as a first order low-pass filter with cut-off frequency at $F_a=1/(2\pi T_a)$. The following sections describe the process of inverse filtering and LF parameter matching for our example.

3.1 Inverse Filtering

The process of inverse filtering is mainly following the algorithm used in [8], and in [11]. In [11] a flow chart can be found. The data was high-pass filtered using an order 30 FIR filter with cut-off frequency at 60 Hz. To obtain a good estimate of the vocal tract configuration the speech data were analyzed during glottal closure where there is no interaction between the voice source and vocal tract. We experienced problems when automatically determining the glottal closure by F_1 modulation, especially for the vowels /i/ and /u/. This may have been due to either an incorrect guess of F_1 and/or F_1 being close to F_0. In our processing we used the information of the instant of maximal glottal excitation T_e after which the return phase begins. The instants of maximal glottal excitation were determined by picking the negative peaks of the whitened acoustic waveform. For the whitening a linear prediction (LP) analysis of order 12 was made. Since the glottal closure will occur shortly after T_e, the speech signal was inverse filtered with the covariance method of linear prediction of order p=9 with the first analysis window beginning one sample after T_e. The analysis window was chosen to have minimal length of p+3 and was shifted sample by sample. LP analysis was preceded by an adaptive pre-emphasis of the signal [9], where the pre-emphasis coefficient was determined by a first-order autocorrelation LP. The resulting inverse filtered signal was then visually inspected and for each token the best window location was used. Maximal absolute ratio of negative maxima to positive maxima, minimal mean of the signal, and, of course, visual similarity to the expected LF model shape were the selection criteria. As Fant [2] mentioned; "Conventional LF parameter extraction from inverse filtering is far from an exact procedure and is subject to variations in individual strategies and, thus, runs the risk of being more an art than a science." Figure 2 shows inverse filtered signals of interest for the three vowels. The effect of a second glottal pulse in male voices [6] could be observed for M2 speaking the vowel /a/.

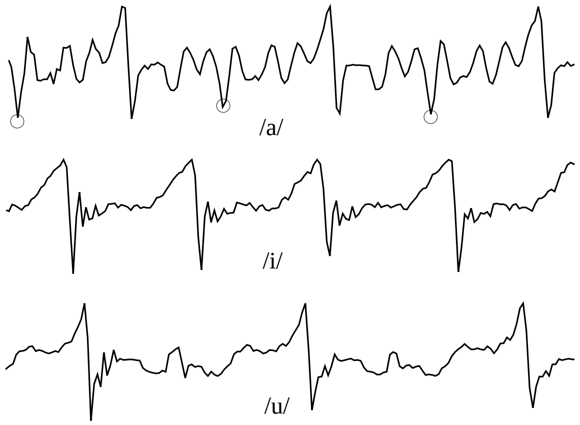

Figure 2: Inverse filtered signals for vowels /a/, /i/, and /u/ for talker M2. Second pulses for /a/ are marked with a circle.

3.2 Matching the LF Parameters

Each inverse filtered pitch period was matched with the LF model and the seven parameters $\lambda = \{T_{op}, T_e, T_c, T_a, \alpha, \omega_g, E_e\}$ had to be estimated minimizing the squared error between the inverse filtered signal $v_{if}(t)$ and $v_g(t, \lambda)$ from (1). To reduce the complexity of the system we approximated the total least squared error with an error E1 and E2, which represent the open phase error and the return phase error, respectively. Equations (3) and (4) are written in time discrete notation: T_x becomes N_x.

$$E_1(n)=\frac{1}{2}\sum_{n=N_o}^{N_e}\left(v_{if}-\frac{-E_e}{\sin(\Omega_g(N_e-N_{op}))}e^{\alpha(n-N_e)}\sin(\Omega_g(n-N_{op}))\right)^2 \qquad (3)$$

$$E_2(n)=\frac{1}{2}\sum_{n=N_e+1}^{N_c-1}\left(v_{if}-\frac{E_e}{\varepsilon N_a}[e^{\varepsilon(N_e-n)}-e^{\varepsilon(N_e-N_c)}]\right)^2 \qquad (4)$$

With the Newton-Raphson algorithm for non-linear equations ([12]) α and Ω_g were obtained from (3), E_e and ε from (4). The initial value for Ω_g was calculated using (2). Instead of N_a, ε was estimated and N_a calculated with

$$N_a=(1-e^{\varepsilon(N_e-N_c)})/\varepsilon\,. \qquad (5)$$

N_e was chosen as the negative peak instant. Because all times are discrete they were found iteratively (N_{op}, N_c). Range checking prevented the parameters from diverging.

4. RESULTS

The correctness of the results was checked by comparing them to published mean values and to spectral characteristics of the speech spectra. [6] describes glottal characteristics using spectral measures and [1] finds relations between these measures and LF model parameters. The time related parameters F_0 to F_3, T_{op}, T_p, T_e, T_a, and T_c are analyzed. As no direct correspondence to perceptual measures was found for parameters ω_g and α they are omitted in this section.

4.1 F_0 and Formant Frequencies

Tables 1-3 summarize the F_0, F_1, F_2, and F_3 values for the four talkers. Note, that also F_0 is context dependent, which is not shown here.

	M1	M2	F1	F2
	123-235	110-205	140-200	190-242

Table 1: Mean pitch frequencies F_0 of the talkers (in Hertz).

	M1	M2	F1	F2
/a/	703-765	671-796	781-953	796-984
/i/	234-359	218-359	312-359	390-421
/u/	250-296	328-359	312-390	406-437

Table 2: Range of F_1 frequency (in Hertz).

	M1	M2	F1	F2
/a/	906-1078	1125-1234	1171-1312	1000-1187
/i/	1890-2140	2046-2218	2687-2890	2578-2718
/u/	1046-1171	1187-1312	1796-2000	1109-1531

Table 3: Range of F_2 frequency (in Hertz).

	M1	M2	F1	F2
/a/	2390-2562	2265-2437	2765-3078	2593-2921
/i/	2968-3156	2640-2890	3375-3609	3250-3546
/u/	1921-2031	2125-2265	2750-2937	2421-2656

Table 4: Range of F_3 frequency (in Hertz).

4.2 Open Quotient

The open quotient (OQ) is defined as the ratio of the open phase to the total length of the glottal cycle OQ = $(T_e\text{-}T_{op}) / T_0$. The estimates of the parameters T_e together with the pitch period T_0 are reliable for most of the periods. The beginning of the open phase T_{op} is more difficult to estimate. Table 5 shows mean and standard deviation for OQ values.

	M1	M2	F1	F2	Avg.
/a/	45(17)	37(15)	57(11)	55(11)	**51(15)**
/i/	41(17)	52(13)	55(13)	58(8)	**53(14)**
/u/	30(10)	62(5)	49(14)	52(10)	**48(14)**
Avg.	**39(16)**	**50(16)**	**54(13)**	**56(10)**	

Table 5: Open quotient for LF model (in percent); the value in parenthesis is the standard deviation.

For M1 and M2 for the vowels /u/ and /a/, respectively seem low. There is an interpretation for the low value in M2's case. M2 has two glottal pulses per period and therefore has a reduced opening time per pulse. Another explanation could be, that the second pulse induces a higher variance in parameter change within only a few pitch periods. In fact, the amplitudes of the first and second harmonics (H1, H2), for talker M2's /a/ changed the most when shifting the analysis window by only one frame. The OQ for the female talkers is on average bigger than that for male which agrees with [6]. F2 has a bigger OQ than F1, the same holds for M2 and M1. Three talkers have lowest OQ values for the vowel /u/.

In an attempt to relate the OQ from spectral domain measures we used the equation

$$H1^* - H2^* = -6 + 0.27\exp(5.5OQ_i) \tag{6}$$

proposed in [1] and reformulated it to

$$OQ = \log((H1^* - H2^* + 6) / 0.27) / 5.5 \tag{7}$$

where H1*, H2* are the amplitudes of the first and second harmonics of the spectrum, corrected by the amount by which the vocal tract transfer function, especially the first formant, amplifies them [5]. The equation was only used when the difference between F_0 and F_1 [13], and F_1 and F_2 was higher than 200Hz and when the term (H1*-H2*+6) was above zero. Comparing Table 5 with Table 6 shows the similarity of the results.

	M1	M2	F1	F2	Avg.
/a/	40-48	53-64	42-64	65-66	**53**
/i/	N.A.	50	N.A.	51-65	**57**
/u/	N.A.	47-58	N.A.	50	**52**

Table 6: Range of OQ (in percent) calculated from spectral harmonics.

4.3 Spectral Tilt

The Spectral Tilt (ST) is closely related to the LF parameters of the return phase ST ~ T_a = 1 / F_a: the shorter the time T_a, the higher the low-pass filter cut-off frequency Fa, hence, the lower the ST. Female voices usually have a larger T_a than men, especially if the voice is breathy. Mean estimates (+/- 0.5ms) of the parameter T_a are shown in Table 7, the values are generally low. The reason for this is the sensitivity to small signal perturbations when fitting the return phase in the time domain. As [1] mentioned, frequency domain fitting for quantifying F_a can give more accurate estimates. From Table 7 it can be seen that for female speakers the ST for /i/ is highest and for /a/ lowest, which correlates with the results in [7]. Interestingly, for male speakers T_a behaves almost contrarily: /i/ has a very low T_a. For females a higher ST compared to males can be observed which confirms the theory that females have usually breathier phonation [6]. Again, talker M2 has a very low spectral tilt for /a/, which can be due to the second impulse observed in his inverse filtered waveform (Figure 2).

	M1	M2	F1	F2	Avg.
/a/	0.24	0.10	0.20	0.32	**0.23**
/i/	0.15	0.14	0.26	0.48	**0.31**
/u/	0.37	0.20	0.13	0.47	**0.30**
Avg.	**0.25**	**0.15**	**0.2**	**0.42**	

Table 7: Mean values of LF parameter T_a in milliseconds.

The frequency domain measure H1*-A3* characterizes ST. It is the amplitude difference between the corrected amplitudes of the first harmonic H1 and the third formant A3. The correction for A3 is different from the one for H1 and H2. It can be found in [5]. This measure also yields a higher ST for females than for males, a higher ST for F2 compared to F1, and a higher average ST for the female talker for /u/ compared to /a/. Combining the information that F2 has a higher ST and a bigger OQ than F1 leads to the assumption that F2 should have a breathier voice and indeed, this could be confirmed by listening to the token. Talker M2 has the lowest ST of all talkers.

	M1	M2	F1	F2	Avg.
/a/	6.7	0.1	4.1	8.6	**4.1**
/i/	N.A.	0.8	N.A.	19.3	**16.6**
/u/	N.A.	N.A.	7.2	12.6	**9.9**
Avg.	**2.8**		**10.9**		

Table 8: Mean values for the frequency domain measure H1*-A3* (in dB), as an indicator for ST.

4.4 Parameter R_d

The parameter R_d is one of the most effective parameters to quantify the characteristics of the voice source waveform with a single numerical value ([1], [2]). Low R_d indicates extreme tight, adducted phonation, with low OQ and high ST whereas high R_d stands for very breathy, abducted phonation with high OQ and low ST. Its basic formula is

$$R_d = (U_o/E_e)(F_0/110), \qquad (8)$$

the parameters of which can be seen in Figure 1. R_d can also be calculated using the approximation

$$R_d \approx (1/0.11)(0.5+1.2R_k)(R_k/(4R_g)+R_a) \qquad (9)$$

With $R_a=T_a/T_0$, $R_g=T_0/(2T_p)$, and $R_k=(T_e-T_p)/T_p$, equation 9 combines 4 important LF parameters: T_0, T_p T_e and T_a. With these parameters from the model matching, R_d was calculated. The results are shown in Table 9. Note that the maximal time-resolution using 8kHz sampling frequency is 125µs.

	M1	M2	F1	F2	Avg.
/a/	249(128)	138(29)	259(64)	463(141)	**313(164)**
/i/	196(52)	224(104)	348(253)	697(331)	**435(330)**
/u/	358(145)	220(112)	205(105)	710(436)	**395(330)**
Avg.	**269(134)**	**191(96)**	**264(158)**	**608(330)**	

Table 9: R_d mean and standard deviation (in microseconds).

As in [2], R_d shows higher values for females than for males and similar to the OQ measurement, the values for /i/ are the highest for female talkers. Talker M2 again has a very low value. For the spectral domain the following formula holds.

$$R_d \quad = (H1^* - H2^* + 7.6) / 11.1. \qquad (10)$$

	M1	M2	F1	F2	Avg.
/a/	431(40)	448(737)	545(195)	1017(37)	**540(443)**
/i/	-	336(0)	-	902(170)	**821(264)**
/u/	-	347(101)	-	173(165)	**338(264)**
Avg.	**431(40)**	**418(493)**	**461(313)**	**661(341)**	

Table 10: R_d mean and std from the spectral domain (in µs).

Again, females have higher R_d, F2 has the highest R_d and M2 the lowest of all four talkers.

5. CONCLUSIONS AND DISCUSSION

A semi-automatic procedure for extracting voice source LF parameters was presented and the reliability of the results was estimated. The closed phase of the glottis and the instant of maximal glottal excitation were determined by hand. After inverse filtering the LF-parameters were estimated. The measures OQ, ST and the perceptually important parameter R_d were calculated. The results showed reliable trends in all three measures.

REFERENCES

1. G. Fant, “The voice source in connected speech,“ *Speech Commun.*, pp. 125-139, 1997.
2. G. Fant, “The LF-model revisited. Transformation and frequency domain analysis,“ *Speech Trans. Lab. Q. Rep.*, Royal Inst. of Techn. Stockholm, vol. 2-3, pp. 121-156, 1995.
3. G. Fant and Q. Lin, “Frequency domain interpretation and derivation of glottal flow parameters,“ *Speech Trans. Lab. Q. Rep.*, Royal Inst. of Techn. Stockholm, vol. 2-3, pp. 1-21, 1988.
4. G. Fant, “Glottal flow models and interaction,“ *J. Phonet.*, vol. 14, pp. 393-399, 1986.
5. H. M. Hanson, “Glottal characteristics of female speakers: Acoustic correlates,“ *J. Acoust. Soc. Amer.*, vol. 101, pp. 466-481, Jan. 1997.
6. H. M. Hanson and E. F.Chuang, “Glottal characteristics of male speakers: Acoustic correlates and comparison with female data,“ *J. Acoust. Soc. Amer.*, vol. 106, pp. 1064-1077, Aug. 1999.
7. I. Karlsson, “Voice source dynamics for female speakers,“ *Proc. ICSLP*, pp. 69-72, 1990.
8. M. D. Plumpe and T. F. Quatieri, “Modeling of the glottal flow derivative with application to speaker identification,“ *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 569-585, Sept. 1999.
9. M. D. Plumpe, “Modeling of the glottal flow derivative waveform with application to speaker identification,“ S.M. thesis, Mass. Inst. Techn., Cambridge, MA, Feb. 1997.
10. M. Rothenberg, “A new inverse-filtering technique for deriving the glottal air flow waveform during voicing,“ *J. Acoust. Soc. Amer.*, vol. 53, pp.1632-1645, 1973.
11. D. Y. Wong, J. D. Markel, and A. H. Gray, “Least squares glottal inverse filtering from the acoustic speech waveform,“ *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 350-355, Aug. 1979.
12. W. H. Press et.al., *Numerical recipes in C: The art of scientific computing*. New York: Cambridge Univ. Press, Jan. 1993.
13. Private communication with H. M. Hanson.