Modelling Aspiration Noise During Phonation Using the LF Voice Source Model

Christer Gobl

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland

cegobl@tcd.ie

Abstract

This paper presents a technique for modelling the aspiration noise produced during phonation. The method employs the widely used LF voice source model, which is here extended to include a turbulence noise source for the generation of aspiration. Drawing on speech production theory and on empirical data, the overall amplitude level as well as the within-pulse modulation of the noise are determined by the specific shape of the LF model waveform. The main advantage of this approach is that it simplifies the control of the glottal source, as the temporal dynamics of the variation in noise level need not be explicitly set, but is generated automatically. Informal listening tests also suggest that the proposed modelling technique may contribute to the naturalness of synthesised speech, particularly with regard to breathy voicing.

Index Terms: aspiration noise, glottal, voice source, LF model

1. Introduction

Many of the parameters used in the modelling of speech acoustics are interrelated and do not vary independently. The main objective of the present research is to explore the relationship between glottal pulse characteristics and the turbulence noise component during phonation, and to formulate a model by which the noise component can be predicted from voice source parameters.

To analyse the glottal source, source-filter decomposition based on inverse filtering is often used, e.g. [1]. Such analysis mainly provides data on the quasi-periodic voice source, but little or no information about the aspiration noise. Being able to model the noise component on the basis of voice source data would enhance the analysis as well as simplifying the resynthesis of the parameterised speech signal. It may also contribute to the naturalness of synthetic speech, particularly for voice qualities involving significant audible aspiration noise, such as breathy voice and whispery voice.

One of the most comprehensive and widely used formant synthesisers is the Klatt synthesiser, KLSYN88 [2]. It has a large number of control parameters facilitating very detailed modelling of the speech output. When the parameters are appropriately set, very high naturalness can be achieved. However, the large number of parameters can make it difficult to control. As there are few constraints amongst the parameters, great care is required to avoid the risk of an unnatural output.

The HLsyn synthesiser [3] provides a smaller number of ‘high-level’ quasi-articulatory parameters, and from these, the settings are automatically generated for the ‘low-level’ acoustic parameters of KLSYN88. Thus, the HLsyn parameters simplify the control, as they impose natural constraints on the KLSYN88 parameters. On the other hand, some of the HLsyn parameters are not easily obtainable from the acoustic speech signal, e.g., the size of the anterior and posterior glottal opening, subglottal pressure, etc. Without this information, it can be challenging to use HLsyn as a tool for resynthesis from acoustic analysis.

The current research builds on work presented in [4]. The aim is to explore how acoustic data, together with knowledge on speech acoustics, aerodynamics and physiology, can be used to calculate otherwise relatively inaccessible information, in this case aspiration noise. Although not being the focus here, underlying physiological parameters can thereby also be estimated.

2. Theoretical framework

The main theoretical framework for this modelling is Fant’s source-filter theory of speech production [5], which considers the speech signal as the response of the vocal tract filter to one or more sound sources. During phonation, the source is regarded as a volume velocity source, and we shall here define the voice source as the airflow through the glottis, $U_g(t)$. This signal excites the natural modes of the vocal tract, producing oral airflow at the lips. The radiated sound pressure is approximately proportional to a differentiation of the oral airflow signal. Therefore, the differentiated glottal flow (glottal flow derivative) relates more closely to the radiated sound pressure, and is often used when modelling the voice source.

For the noise modelling we rely on Stevens’ detailed theoretical account of turbulence noise generation in speech, as outlined in [6], and on experimental data from [7] and [8].

Considering the glottal resistance, but ignoring the viscous losses, we can derive the following approximation for how the glottal airflow U_g depends on the glottal area A_g and the transglottal pressure drop P_{tg}

$$U_g(t) = \sqrt{\frac{2\,P_{tg}}{\rho}} \cdot A_g(t) \qquad (1)$$

Assuming a constant transglottal pressure drop, it suggests that glottal flow and glottal area are proportional. Obviously, Eq. 1 does not consider the effects of the supraglottal and subglottal load or the glottal inductance, but for the modelling developed here we shall make use of this approximation.

Turbulence noise produced in the glottis or in the vicinity of the glottis is known as aspiration noise. Most of the noise is produced above the glottis at the ventricular folds and along the surface of the epiglottis, and this noise can be treated as a sound pressure source. Noise produced within the glottis is considerably weaker and can be modelled as a volume velocity source [6].

The amplitude spectrum of a turbulence noise source is thought to be essentially flat, with levels falling at low and high frequencies [5, 6, 9]. The spectrum may also depend on the glottal area, and how this can be modelled is discussed in [9].

The noise intensity can be assumed to be proportional to the cross-sectional area of the constriction, i.e. A_g in the case of aspiration noise. This is based on the fact that, all else being equal, the intensity of the source is proportional to the area over which turbulence is produced, and this area is in turn directly determined by the size of the jet, which is given by the constriction area [6]. As pressure is proportional to the square root of intensity, the radiated sound pressure of the aspiration noise source is assumed to be approximately proportional to the square root of the glottal area. Empirical data and modelling experiments [10, 11] have supported this relationship, although Rothenberg's data [12] point to a more complex relationship.

If we assume proportionality between glottal flow and glottal area as in Eq. 1, it follows that the strength of the noise source is also approximately proportional to the square root of the glottal airflow.

However, according to Eq. 1, the flow also depends on the transglottal pressure. The intensity of a noise source produced in speech has been reported to vary in proportion to the sixth power of the flow, e.g., [6, 9]. Thus, the radiated sound pressure would be proportional to the third power of the flow, and given Eq. 1, it follows that the noise source amplitude is also proportional to $P_{tg}{}^{1.5}$.

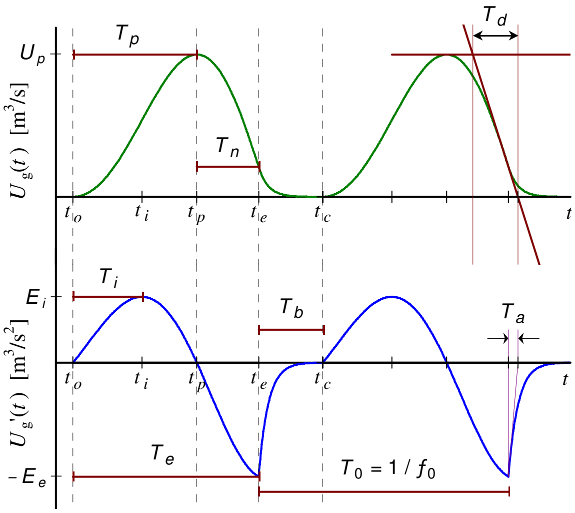

Figure 1. *LF model pulses (bottom) and corresponding glottal flow pulses (top).*

3. Voice source modelling

Acoustic modelling of the human voice source often involves a parametric model with the capability of capturing salient characteristics of the glottal flow pulse. One such model is the well-established LF model [13], which is typically defined in terms of differentiated glottal flow, but can equally well be used to model the true glottal flow. Fig. 1 shows two LF model pulses, both for flow (top) and flow derivative (bottom).

Except for the voice fundamental frequency f_0, there are no 'standard' parameters for describing the voice source. However, in addition to f_0, two to three parameters are normally used to describe the pulse shape, and one amplitude parameter is used to define the strength of the glottal excitation.

The typical set of parameters used with the LF model for describing the source would be f_0, R_a, R_g, R_k, and E_e: $R_a = T_a/T_0$, $R_g = T_0/(2T_p)$ and $R_k = T_n/T_p$, describe the pulse shape, and E_e is the excitation strength, defined by the maximum discontinuity in the flow waveform (see Fig. 1). For the generation of the LF model waveform (see further Section 3.1 below), the parameters are converted into the necessary parameters of the model, as described in detail in [14].

Data on how these voice source parameters vary in running speech show a high level of interconnection: thus, systematic covariation of values for different parameters is often observed, e.g., [8, 15, 16].

To capitalise on this natural correlation between the typical LF parameters, Fant [16] has proposed a global pulse-shape parameter R_d. This parameter captures some of the main characteristics of the pulse by taking some of the covariation into account. A similar parameter has been proposed by Alku et al. [17], the 'normalized amplitude quotient' NAQ, but in the present modelling we shall be using the R_d parameter.

Note that the R_d parameter is based on the *declination time* of the glottal pulse [18], $T_d = U_p/E_e$ where U_p is the peak glottal flow (see Fig. 1 and Eq. 2 – note that T_d is here expressed in seconds), normalised to the fundamental period.

$$R_d \;=\; 1000\,\frac{U_p}{E_e}\,\frac{f_0}{110} \;=\; \frac{1}{0.11}\,\frac{T_d}{T_0} \qquad (2)$$

R_d and T_d have been shown to not only capture some of the natural covariation between source parameters, but also to be efficient in disambiguating voice qualities on a tense to lax continuum. On the basis of the R_d value, a system for predicting other source parameters, such as R_a and R_k, has also been elaborated [8, 16].

R_d and T_d seem to capture aspects of the quasi-periodic voice source, but can they also be used to make predictions about the aspiration noise? The strength of the aspiration noise depends on aerodynamic factors and on the glottal configuration, factors that also shape the quasi-periodic part of the glottal source. It therefore seems reasonable to expect features of the aspiration noise to be reflected in the glottal pulse characteristics.

3.1. The LF model

The LF model [13] is a parametric voice source model, which is often referred to as a four-parameter model. Apart from f_0, four parameters define the pulse: three parameters for the pulse shape and one for the amplitude of the pulse.

To generate the LF waveform, we also need to set one further 'parameter'. The model requires the size of the positive area of the differentiated flow pulse be the same as the size of the negative area. In other words, the area over the whole duration of an LF pulse should be zero. This ensures that the maximally closed phase of the glottal flow pulses is not offset with respect to the zero-flow line. This 'area balance' is implicitly set on the basis of the values of the other parameters.

As indicated by Eqs. 3 and 4 and Fig. 1, the model has two parts: the *open phase*, $T_e = t_e - t_o$, and the *return phase*, $T_b = t_c - t_e$. The open phase is defined by a sinusoidal function that increases exponentially in amplitude from glottal opening to the main excitation, where the amplitude reaches the value $-E_e$ (Eq. 3). The frequency of the sine function is $F_g = \omega_g/(2\pi)$ and the rate of the amplitude increase is determined by α. E_0 is a scale factor that is used to achieve area-balance.

$$E(t) = U'_g(t) = E_0\, \mathrm{e}^{\alpha t} \sin \omega_g t \quad \text{for } t_o \le t \le t_e \tag{3}$$

The return phase is defined by an exponential function, as shown in Eq. 4. Note that the time point t_c coincides with the time point of the glottal opening of the following pulse: thus, the exponential function returns asymptotically not to zero, but to a specific positive value, so that zero is reached at time point t_c. The main parameter of the return phase is T_a which measures the 'effective' duration of the return phase. As can be seen in Fig. 1, T_a is the duration from t_e to the point where a tangent, fitted at the starting point of the return phase, crosses the zero line.

$$E(t) = \frac{-E_e}{\varepsilon T_a}\left(\mathrm{e}^{-\varepsilon(t-t_e)} - \mathrm{e}^{-\varepsilon T_b}\right) \quad \text{for } t_e < t < t_c \tag{4}$$

The time-constant of the exponential function of the return phase is ε^{-1}, which can be determined iteratively from T_a and T_b, using the following equation

$$\varepsilon = \frac{1}{T_a} \cdot \left(1 - \mathrm{e}^{-\varepsilon T_b}\right) \tag{5}$$

The LF model can also be expressed in terms of glottal airflow (upper part of Fig. 1). The glottal flow for the two parts respectively is given by Eqs. 6 and 7.

$$U_g(t) = \frac{E_0\, \mathrm{e}^{\alpha t} \sin\left(\omega_g t - \arctan \dfrac{\omega_g}{\alpha}\right)}{\sqrt{\alpha^2 + \omega_g^{\,2}}} + \frac{E_0\, \omega_g}{\alpha^2 + \omega_g^{\,2}} \tag{6}$$

$$U_g(t) = \frac{E_e}{\varepsilon^2 T_a}\left[\mathrm{e}^{-\varepsilon(t-t_e)} + \varepsilon\, \mathrm{e}^{-\varepsilon T_b}\left(t - \left(t_c + \frac{1}{\varepsilon}\right)\right)\right] \tag{7}$$

As indicated earlier, the LF model pulse can be uniquely determined by six parameters, e.g., α, ω_g, E_e, T_e, T_a and T_b (where $T_e = t_e - t_o$, see Fig. 1). Apart from these six parameters, E_0 and ε also need to be computed in order to generate the LF pulse: ε is determined by T_a and T_b, and E_0 is calculated iteratively together with α to achieve area-balance.

4. Extending the LF model to incorporate aspiration noise

White Gaussian noise was used as the basic aspiration noise source. The amplitude spectrum is essentially flat, and no spectral shaping was included in the current simulations, for reasons of simplicity.

For the noise modulation, the theoretical framework elaborated in Section 2 is used: as the airflow changes during the course of the glottal cycle, the noise amplitude should vary in proportion to the square root of the glottal flow.

In terms of the LF model, the flow is given by Eqs. 6 and 7. However, we also need to take the dc flow U_{dc} into account, as there is typically some amount of glottal leakage during the so-called closed phase [7]. Generally, we do not have access to the level of dc flow. Neither do we know the precise relationship between the ac and dc components of the flow. As the turbulence noise level and modulation are determined by the overall rate of glottal flow, we clearly need to obtain this information. We shall here attempt to estimate U_{dc} as well as U_{ac} from the voice source parameters.

We also need to obtain the pulse amplitude of the LF pulse: by evaluating Eq. 5 for $t = T_p$, we can derive Eq. 8 for the peak flow value U_p of the LF model pulse.

$$U_p = \frac{E_0\, \omega_g}{\alpha^2 + \omega_g^{\,2}}\left(\mathrm{e}^{\alpha\pi/\omega_g} + 1\right) \tag{8}$$

As the U_p value is scaled differently from U_{dc}, it is necessary to scale the waveform of the LF model pulse by a factor of U_{ac}/U_p, so that flow values match.

To scale the noise with regard to the transglottal pressure drop P_{tr} is particularly important at voice termination. If the transglottal pressure is assumed to be constant, this tends to result in an artificially high level of noise prior to voice termination, followed by an unnaturally abrupt cessation of the noise source.

To derive a value proportional to P_{tr} on the basis of the LF parameters, we use the empirical relationships between subglottal pressure P_s, E_e and f_0 found for speech data having f_0 values in the low to mid frequency range [8]. Given a constant f_0, E_e was found to be proportional to $P_s^{1.1}$, which also means that P_s is proportional to $E_e^{0.9}$. Given a constant E_e, P_s was found to be proportional to $f_0^{0.7}$. Thus, we shall assume P_s to be proportional to $E_e^{0.9} f_0^{0.7}$.

As discussed in Section 2, we need to scale the noise source by a factor proportional to $P_{tg}^{1.5}$. If we replace P_s with P_{tg}, we get the following relationship: $P_{tg}^{1.5} \sim E_e^{1.35} f_0^{1.05}$.

Given U_{ac}, U_p, U_{dc}, E_e and f_0, the modulated aspiration noise source $S_{AH}(n)$ is calculated as follows:

$$S_{AH}(n) = E_e^{1.35} f_0^{1.05} \sqrt{\frac{U_{ac}}{U_p} U_g(n) + U_{dc}} \times AH \cdot rand(n) \tag{9}$$

where $rand(n)$ is the random number generator producing white Gaussian noise, AH is a parameter controlling the overall strength of the white noise, and $U_g(n)$ is the sampled version of the LF flow waveform (Eqs. 6 and 7).

As noted earlier, R_d and T_d are voice source parameters that can capture many of the essential characteristics of the glottal pulse, and they are used here to estimate U_{dc} and U_{ac}. To predict U_{dc}, we use the absolute value of the declination time T_d, which is likely to be more suitable than the f_0-normalised R_d value, as we would expect U_{dc} to decrease as a function of increasing f_0.

Based on data from [7] for 25 male and 20 female speakers producing vowels with 'normal', 'loud' and 'soft' voice, correlation analysis was carried out for the parameters shown in Table 1. The results suggest a high correlation for U_{dc} and T_d ($r = 0.92$), and linear regression analysis provides the following relationship for predicting U_{dc}.

$$U_{dc} = 83\,T_d + 34 = 83\frac{U_p}{E_e} + 34 \qquad (10)$$

For the estimation of U_{ac}, we use the f_0-normalised R_d value, as it shows a high correlation with U_{ac}. The inverse of R_d is used for the regression analysis: U_p has been shown to vary with the square root of E_e [16], and one might therefore expect a linear relationship between U_{ac} and $1/R_d$. The relationship between U_{dc} and the inverse of R_d ($r = 0.91$) is as follows:

$$U_{ac} = \frac{379}{R_d} - 91 = 42\frac{E_e}{U_p}\cdot T_0 - 91 \qquad (11)$$

Note that U_{dc} is updated for every pulse, but U_{ac}/U_p should be calculated only once, as U_{ac}/U_p is simply used to rescale the LF pulses. Note also that the correlation analysis relates to vowel steady-states: the square root relationship between U_p and E_e is often not valid for voice onsets and offsets. Therefore, U_{ac}/U_p is here obtained from R_d values from the first steady state vowel, by calculating the mean value for four consecutive pulses.

Table 1. *Mean values taken from [7]. Note that female data for 'soft voice' were not available.*

	Loud (male)	Normal (male)	Soft (male)	Loud (female)	Normal (female)
U_{ac} (cm^3/s)	380	260	193	180	140
U_{dc} (cm^3/s)	110	120	152	90	90
E_e (cm^3/s/ms)	481.1	279.6	134.15	248.9	164
f_0 (Hz)	125	111.1	111.1	200	200
T_0 (ms)	0.008	0.009	0.009	0.005	0.005
T_d (ms)	0.79	0.93	1.44	0.72	0.85
R_d	0.90	0.94	1.45	1.31	1.55

5. Summary and conclusions

A modelling technique is formulated for the generation of aspiration noise during phonation: the noise level is automatically controlled and is modulated as a function of the LF model glottal pulse. The principles for the modelling draw on acoustic theory as well as on experimental data, and should help to simplify the analysis and synthesis of the glottal source in running speech. Informal listening tests further suggest that it may contribute to a more natural sounding synthesis of breathy voicing. Currently, more formal perception tests are being elaborated to establish its perceptual effects.

Note that the empirical expressions for estimating U_{dc} and U_{ac} are only based on the mean values from five categories of data. Further data analysis should help to refine the predictions, by including perhaps also information from other voice source parameters such as R_a, R_g, etc. This is likely to be crucial for better estimates during voice onset and offset.

An envisaged future direction will be to explore, and to contrast with the current modelling, how aspiration noise would be modulated to be consistent with the findings presented by Rothenberg [12]. The effects on the noise due to variation in the spectral shape of the noise, according to suggestions in [9], are also to be considered.

6. References

[1] Gobl, C. and Ní Chasaide, A. "Techniques for analysing the voice source," in W. J. Hardcastle and N. Hewlett (Eds.) *Coarticulation: Theory, Data and Techniques*, Cambridge University Press, Cambridge, 300-320, 1999.

[2] Klatt, D. H. and Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, 87, 820-857, 1990.

[3] Hanson H. M. and Stevens K. N., "A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn, *J. Acoust. Soc. Am.*, 112, 1158-1182, 2002.

[4] Gobl, C., "Aspiration noise generation based on glottal pulse characteristics", *Proceedings of the 9th Western Pacific Acoustics Conference*, Seoul, Korea, 2006.

[5] Fant, G., *The Acoustic Theory of Speech Production* (Mouton, Hague, 2nd edition, 1970).

[6] Stevens K. N., *Acoustic Phonetics*, The MIT Press, Cambridge, Massachusetts, 1998.

[7] Holmberg, E. B., Hillman, R. E., and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *J. Acoust. Soc. Am.*, 84, 511-529, 1988.

[8] Fant, G., "The voice source in connected speech," *Speech Communication*, 22, 125-139, 1997.

[9] Ananthapadmanabha T. V. and Prasad M. G., "A note on jet noise component in phonation", unpublished research report.

[10] Stevens K. N., "Airflow and turbulence noise for fricative and stop consonants," *J. Acoust. Soc. Am.*, 50, 1180-1192, 1971.

[11] Shadle C., "The Acoustics of Fricative Consonants," Technical Report 506, Research Laboratory of Electronics, MIT, Cambridge, Massachusetts, 1985.

[12] Rothenberg M., "Glottal noise during speech," *STL-QPSR,* Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 2-3, 1-10, 1974.

[13] Fant, G., Liljencrants, J., and Lin Q., "A four-parameter model of glottal flow," *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 4, 1-13, 1985.

[14] Gobl, C., "The Voice Source in Speech Communication: Production and Perception Experiments Involving Inverse Filtering and Synthesis," Doctoral thesis, KTH, Stockholm, Sweden, 2003.

[15] Gobl, C., "Voice source dynamics in connected speech," *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 1, 123-159, 1988.

[16] Fant, G., "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 2-3, 119-156, 1995.

[17] Alku, P., Bäckström, T., and Vilkman, E. "Normalized amplitude quotient for parameterization of the glottal flow," *J. Acoust. Soc. Am.*, 112, 701-710, 2002.

[18] Fant, G., "Glottal source and excitation analysis", *STL-QPSR*, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 1, 85-107, 1979.