ESTIMATION OF SOURCE PARAMETERS BY FREQUENCY ANALYSIS

*Luís C. Oliveira**

AT&T Bell Laboratories
Murray Hill NJ 07974 USA

ABSTRACT

The new generation of text-to-speech systems needs the ability to control the voice quality of the synthesized speech by varying the excitation source. This feature is fundamental to improve naturalness and to synthesize female or child voices. The variation of the voice quality is also important when trying to synthesize expression. The problem involves two aspects: the ability to control the source parameters of the speech synthesizer and the possibility of extracting these parameters from natural speech. This paper describes a source model based on the polynomial model for the glottal flow suggested by Rosenberg [9] that has an exact representation in the frequency domain, and an automated procedure to estimate its parameters from natural speech.

1. INTRODUCTION

Current and forseeable applications of speech synthesis require improved naturalness through changes in the voice quality of the synthesized speech and the ability of producing female and children voices. For this reason, text-to-speech systems are incorporating more realistic voice source models [7, 2]. One of the major reasons for this development was the inability of the previous models to synthesize a convincing imitation of a female voice.

Although the inclusion of more complex source models improves the naturalness, voice quality changes during an utterance and dynamic variations of the source parameters are required: breathiness tends to increase in unstressed syllables and in utterance-finals [7], the spectral tilt seems to be smaller for open vowels [6] and there are significant changes in the glottal pulse shape at onset and on termination of the voiced source [4]. However, the studies on the variations of the source parameters have been restricted to relatively small sets of speech material due to the difficulty of the analysis, requiring hand marking of the glottal events.

The first part of this work describes the modifications of the source model proposed by Rosenberg [9]. The modified model adds the control of the spectral tilt of the voice source and the level of aspiration noise. The second goal of this work was to devise a strategy for automatic estimation of the source parameters. A frequency domain method was selected, and the model parameters are estimated by a fitting it to the short-time Fourier transform of the natural inverse filtered signal.

*on leave from INESC/IST, Lisbon, Portugal

2. SOURCE MODELING

Several formulations have been suggested for the voice source model. Fujisaki and Ljungqvist[3] evaluated several models and concluded that the tested models reduced the LP prediction error by 3 to $4.2dB$, compared to a single pulse excitation. This suggests that a detailed modeling of the glottal is not very important. Since it is easier to find an exact frequency representation of a continuous signal in the time domain, we have decided to use the Rosenberg polynomial model.

2.1. Polynomial Model

We will assume the following general equation for the glottal flow for each fundamental period (T_0):

$$u_g(t) = \begin{cases} k_0 + k_1 t + k_2 t^2 + k_3 t^3 & \text{if } 0 \le t < T_{op} \\ 0 & \text{if } T_{op} \le t < T_0 \end{cases}$$

where T_{op} is the duration of the open glottis phase. The ratio between this value and T_0 is often referred as the *open quotient* ($OQ = T_{op}/T_0$).

The parameter k_0 is irrelevant since we will incorporate the radiation characteristic at lips (essentially a pole at the origin) on the glottal model and so we will be using the derivative of the glottal flow. The other parameters, k_1 to k_3, can be related by imposing conditions to the model.

The first condition imposes a zero DC value to the derivative of the glottal flow, $u'_g(t)$ – the signal that is going to be used as the excitation of the vocal tract filter. This also means that $u_g(0) = u_g(T_{op})$:

$$\int_0^{T_0} u'_g(t)dt = 0 \quad \Rightarrow \quad k_1 + k_2 T_{op} + k_3 T_{op}^2 = 0$$

To simplify the model we will assume a second condition that assumes a null derivative of the glottal flow at the origin:

$$u'_g(0) = 0 \quad \Rightarrow \quad k_1 = 0$$

These two conditions are equivalent to:

$$\begin{cases} k_1 = 0 \\ k_1 + k_2 T_{op} + k_3 T_{op}^2 = 0 \end{cases} \Leftrightarrow k_2 = -T_{op} k_3$$

and the equation for the glottal flow derivative becomes:

$$u'_g(t) = k_3 \left(-2T_{op} t + 3t^2\right) \quad 0 \le t < T_{op}$$

As a result of the second condition the maximum flow value occurs at a fixed position inside the open phase: $t_M = \frac{2}{3}T_{op}$. This is the major drawback of this simplification: it cannot model changes in the skewness of the glottal pulse.
The third condition imposes a high-frequency spectral envelope independent of the glottis open phase duration, which is equivalent to imposing a fixed discontinuity on the derivative of the glottal flow at $t = T_{op}$:

$$u'_g(T_{op}) = -G \quad \Rightarrow \quad k_3 = \frac{-G}{T_{op}^2}$$

The constant G will be adjusted to normalize the energy of the glottal flow waveform. By applying this last condition we get the Rosenberg model for the glottal flow:

$$u_g(t) = \begin{cases} \frac{G}{T_{op}^2}\left(T_{op}t^2 - t^3\right) & \text{if } 0 \le t < T_{op} \\ 0 & \text{if } T_{op} \le t < T_0 \end{cases}$$

Since we are going to use the model on a digital computer the conditions are better applied to the discrete-time equation. Following the same procedure we get the discrete-time equation for the glottal flow derivative:

$$u'_g(n) = \begin{cases} \frac{G\left[(2N_{op}-1)n - 3n^2\right]}{N_{op}^2 - 3N_{op} + 2} & \text{if } 0 \le n < N_{op} \\ 0 & \text{if } N_{op} \le n < N_0 \end{cases} \qquad (1)$$

2.2. Flutter and Diplophonia

The voice source generator also includes also the possibility of modeling some irregularities in the periodic component, namely the jitter in the pitch period and the diplophonic double pulsing. We use the same approach described in [7], where the amplitude of the fluctuations in the fundamental frequency are controlled by a flutter coefficient, and the diplophonia coefficient controls the delay and attenuation of the second pulse in every period pair.

2.3. Modeling the Glottis Closure: Spectral Tilt

In the evaluation of glottal models formulations, Fujisaki & Ljungqvist [3] concluded that models that provided detailed modeling of the glottal closure performed the best. Since the Rosenberg polynomial model did not incorporate these features we added a decaying exponential during the closed glottis phase. This was accomplished by filtering the polynomial model by a first-order low-pass filter, as suggested in [7].

2.4. Aspiration Noise

The turbulence generated at the glottis is an important characteristic of the breathy and whispery voice qualities. The correct modeling of this phenomena is a current topic of research and usually requires a more detailed modeling of the vocal apparatus: glottal opening area, impedance of the vocal tract at the glottis, etc. Since this information is not available in our system the effect was incorporated in a minimalistic form by adding flow modulated flat-spectrum noise to the periodic component $u'_v(n)$. The shape of the modulation function does not seem to be very important in the naturalness of the resulting speech and we are using the flow wave for adding noise during the open phase of the glottis. The results of the perceptual test described in [5] showed that the aspiration noise in breathy vowels should consist of noise bursts synchronous with the glottal cycle.

	Natural		TTS	
	Male	Female	Male	Female
New preferred	75.6%	76.6%	68.4%	73.5%
Average rating new	2.9	2.7	3.0	2.6
Average rating old	2.9	2.2	2.3	2.1

Table 1. Results of the evaluation of the waveform synthesizer

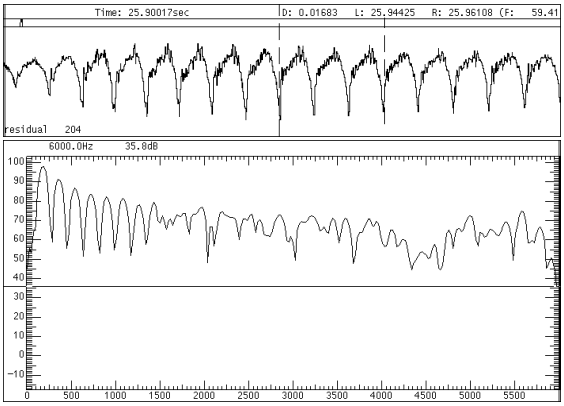

Figure 1. Inverse filtered signal and the magnitude of the Fourier transform (Hanning window)

2.5. Evaluation With Fixed Parameters

A preference test was performed between the new model and the one used previously used in the AT&T Bell Labs TTS system [10]. Four situations were studied using parameters generated by the TTS system with male and female voices, and parameters extracted from utterances spoken by the male and female informants whose voice was used for building the TTS system tables. The variable parameters were: the LP reflection coefficients, the energy, the voicing decision and the fundamental frequency value in the voiced regions. The amplitude controls of the source generators varied according to the voicing decision, and the open quotient, OQ, varied proportionally to the fundamental period. The average open quotient and the remaining source parameters were manually adjusted and kept fixed in each of the four situations.
For each situation, two versions of 200 sentences were synthesized using the new and the old synthesizers. A panel of 8 listeners was asked to select a preferred version for each sentence and to rate the strength of preference on a 1 to 6 scale. The results are presented in table 1, showing an overall preference of 73.5% for the new version.

3. ESTIMATION OF THE SOURCE PARAMETERS

After the selection of the source model, the next goal was to devise a strategy to compute the model parameters from natural speech. Fig. 1 shows the advantage of using the frequency domain: the spectral representation of the inverse filtered signal is characterized by a periodic component in the low frequencies, with the main lobes of the window spectrum located at harmonics of the fundamental frequency, and an almost random component in the higher frequencies region.

3.1. Inverse Filtering

The correct determination of the source parameters requires a good estimate of the glottal waveform. One approximation of this signal can be computed using an estimate of the vocal tract transfer function for inverse filtering the natural speech, provided

that it was recorded without phase distortion. In this work we have adopted a vocal tract model that uses the ressonances estimated by pitch-synchronous linear prediction analysis. This method, described in [10], uses an accurate epoch finder and locates the linear prediction analysis window synchronous to the glottal activity. The results should be interpreted keeping in mind the limitations of the adopted vocal tract model.

3.2. Frequency Representation of the Source Model

The four parameters of the selected source model have different effects in the frequency domain: the *open quotient* (OQ) changes the relative amplitudes of the first and second harmonics; the *spectral tilt* (a_{st}) controls the spectral slope of the periodic component; the *voice amplitude* (A_v) controls the amplitude of the periodic component; and the *aspiration amplitude* (A_h) changes the amplitude of the random component, when the spectrum has a mixed behavior: periodic in the low frequencies and random in the higher frequencies;
The voiced source model can be expressed by the equation:

$$u'_{vh}(n) = A_v u'_v(n) + A_h u'_h(n)$$

and the magnitude of its Fourier transform,

$$\left|U'_{vh}\left(e^{j\omega}\right)\right| = \underbrace{A_v \left|\tilde{U}'_g\left(e^{j\omega}\right)\right| \left|\frac{1-a_{st}}{1-a_{st}e^{j\omega}}\right|}_{periodic} + \underbrace{A_h \left|U'_h\left(e^{j\omega}\right)\right|}_{random}$$

The derivative of the glottal flow is the periodic repetition of:

$$u'_g(n) = \frac{G}{N_{op}^2 - 3N_{op} + 2}\left[\left(2N_{op}-1\right)n - 3n^2\right] w_r(n)$$

where $w_r(n)$ is the rectangular window. The Z-transform of $u'_g(n)$ can be expressed as:

$$U'_g(z) = \frac{G}{N_{op}^2 - 3N_{op} + 2}\left[-\left(2N_{op}+2\right) z\frac{d}{dz}W_r(z) - 3z^2\frac{d^2}{dz^2}W_r(z)\right] \quad (2)$$

where $W_r(z) = (1 - z^{-N_{op}})/(1 - z^{-1})$ is the Z-transform of the rectangular window. By making $z = e^{j\omega}$ and computing the magnitude of eq. 2 we get the magnitude of the Fourier transform:

$$\begin{aligned}\left|U'_g\left(e^{j\omega}\right)\right| = \frac{G}{\left|N_{op}^2 - 3N_{op} + 2\right|}\ \Big[\ \Big[\ (N_{op}^2 - 3N_{op} + 2) + \\ (-2N_{op}^2 + 2N_{op} + 4)\cos(\omega) + (N_{op}^2 + N_{op})\cos(2\omega) - \\ (2N_{op}+2)\cos(N_{op}\omega) + (2N_{op}-4)\cos\left[(N_{op}+1)\omega\right]\Big]^2 + \\ \Big[\ (-2N_{op}^2 + 2N_{op} + 4)\sin(\omega) + (N_{op}^2 + N_{op})\sin(2\omega) - \\ (2N_{op}+2)\sin(N_{op}\omega) + \\ (2N_{op}-4)\sin\left[(N_{op}+1)\omega\right]\ \Big]^2\Big]^{\frac{1}{2}} \frac{1}{\left|2\sin\left(\frac{\omega}{2}\right)\right|^3}\end{aligned}$$

If the glottal flow derivative is considered a periodic function, $\tilde{u}'_g(n)$, it can be expanded in a discrete-time Fourier series with the Fourier coefficients, a_k related to the Fourier transform of $u'_g(n)$ by:

$$a_k = \frac{1}{2\pi} A_v U'_g\left(e^{j\omega_k}\right) \frac{1-a_{st}}{1-a_{st}e^{j\omega_k}} \quad (3)$$

3.3. Analysis Procedure

Having an equation for the frequency representation of the source model being used, the problem now is to find a strategy for fitting the equation to the spectrum of the inverse filtered signal to simultaneously estimate the four source parameters: open quotient, spectral tilt, voicing amplitude and aspiration amplitude.

3.3.1. Harmonic Peak Picking

The analysis procedure starts by locating all the local maxima of the inversed filtered signal spectrum. The second step is the selection of the harmonic peaks: the largest peaks in the vicinity of multiples of the fundamental frequency. The procedure stops when the largest peak is too far from the harmonic frequency and the frequency of the last harmonic peak plus half the fundamental frequency, is defined as the harmonic spectrum cut-off frequency, F_{hc}.

3.3.2. Removal of the Window Spectrum: SLS Analysis

The numeric computation of the spectrum of a periodic waveform requires the usage of some window to truncate the signal. The resulting representation, the short-time spectrum, is the convolution of the periodic signal spectrum with the spectrum of the window. In figure 1 the periodic spectrum is not composed of pulses with the amplitude of the Fourier coefficients a_k, but of lobes resulting from the convolution. The Hanning window minimizes the interference between adjacent harmonics, due to its low leakage (the spectral envelope decreases with $1/\omega^3$).
The problem of finding the real amplitude of the harmonic pulses from the short-time spectrum was addressed in the context of the harmonic modeling of voiced speech in [1]. The method was later extended to the unvoiced regions and named as the Stationary Least Squares (SLS) Analysis [8].
The analysis assumes a sinusoidal representation for the signal to be estimated, $\hat{s}(t) = \sum_{k=-L}^{L} a_k e^{j\omega_k t}$, with $\omega_{-k} = -\omega_k$ and $a_{-k} = a_k^*$. In the harmonic region of the spectrum the frequencies of the exponentials are located at multiples of the fundamental frequency, ω_0: $\omega_k = k\omega_0$.
The complex amplitudes, a_k, are estimated by minimizing the weighted least squares criterion:

$$\int_{-\infty}^{+\infty} w^2(t)\left|s(t) - \hat{s}(t)\right|^2 dt$$

3.3.3. Non-Linear Fitting of the Spectral Envelope

Having the Fourier coefficients, a_k (eq. 3), determined by the SLS analysis, we can now estimate the parameters of the periodic model: A_v, N_{op} and a_{st}.
By using the N_{hp} harmonic frequencies, ω_k, and amplitudes, a_k, a non-linear fitting can be performed by using the Levenberg-Marquardt method to minimize the equation:

$$\chi^2(A_v, N_{op}, a_{st}) = \sum_{k=1}^{N_{hp}} \left(|a_k| - A_v \frac{|1-a_{st}|}{2\pi} \frac{\left|U'_g\left(e^{j\omega_k}\right)\right|}{\left|1-a_{st}e^{j\omega_k}\right|} \right)^2$$

The method is first applied with the first few harmonics (usually 3) to find the open quotient. Next, using all the harmonic peaks found, the spectral tilt and the voiced amplitude are estimated together.
As discussed in section 3.1., the inverse filtered signal is only an approximation of the glottal waveform, due to the simplified

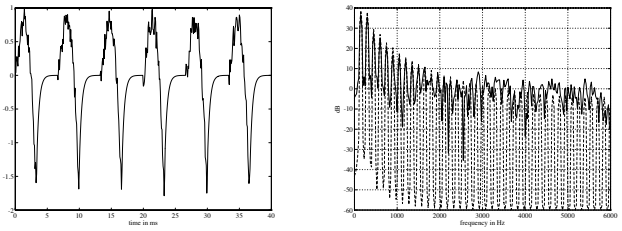

Figure 2. The source model for voiced speech

model. In regions where this model is not valid the fitting method can produce invalid solutions for the parameters. When this happens the solution is discarded.

3.3.4. Aspiration Noise Amplitude

At this point of the analysis procedure, all the parameters of the periodic component have been determined. It is now necessary to know the amplitude of the the random component, A_h. Fig. 2 shows the time and frequency representation of the source model for voiced speech: the dashed line is the periodic component alone and the continuous line is the complete signal. The low-pass characteristic of the glottal waveform makes the random component predominant in the higher frequencies. This suggests the determination of aspiration amplitude by the average difference between the short-time spectrum of the inverse filtered signal and the model for the periodic component, in the random region of the spectrum ($F > F_{hc}$).

3.3.5. Stationary Blocks of Fundamental Periods

In the source analysis procedure that has been described, the stationarity of the inverse filtered signal was assumed. In general this assumption is false, but in short segments the signal can have a quasi-stationary behaviour.

Since the method requires at least 2.5 glottal cycles to be able to locate the harmonic peaks on the short-time spectrum, its is necessary to avoid abrupt changes of the signal inside the analysis window. To prevent this, the inverse filtered signal is scanned to group clusters of glottal cycles with slow varying differences in durations. Each cluster is then divided in overlapping analysis blocks containing from 3 to 5 cycles.

3.4. Example

Fig. 3 shows the result of the analysis of the sentence "Only lawyers love millionaires" spoken by a female speaker. The noticeable peak in the aspiration amplitude accounts for the turbulence in the voiced fricative "v" of the word "love" that was marked as voice by the pitch analyzer. Another interesting region is the glottal effects in "lawyers" producing a deep decrease of the open quotient and spectral tilt.

4. CONCLUSIONS

The source model of the current Bell Labs TTS system was improved to incorporate a better modeling of the glottal closure, the modeling of the turbulence produced at the glottis and the possibility of having multiple types of excitation of the vocal tract filter. The synthesized speech produced by the new version was preferred with both male and female voices and using natural and TTS generated parameters.

A frequency representation of the adopted source model was presented and used as the basis for a new automated method for the estimation of the source parameters. The frequency-

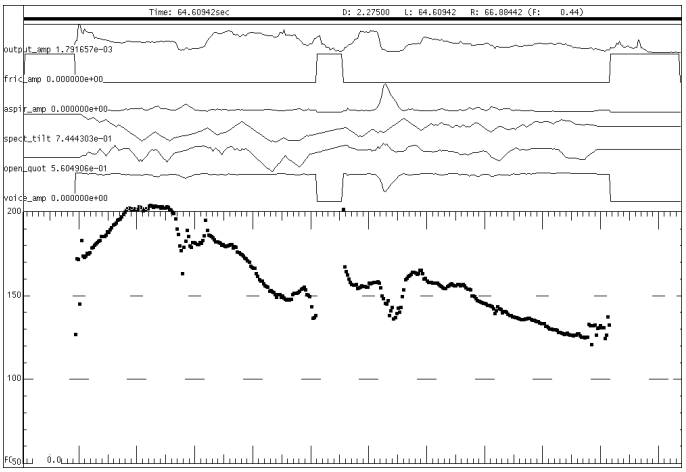

Figure 3. Output of the analysis procedure (top-down: RMS, A_f, A_h, a_{st}, OQ, A_v and F_0)

domain approach allows a better decomposition of the periodic and random components of the inverse filtered signal. Although a more systematic evaluation of the method is required, the time-varying source parameters estimated from natural speech generated high-quality re-synthesized speech on the new synthesizer.

The automated analysis method can be used on large amounts of speech material to find relations between the source parameters dynamics and the phonetic and linguistic context for developing rules for text-to-speech systems.

REFERENCES

[1] L. B. Almeida and J. M. Tribolet. Nonstationary spectral modeling of voiced speech. *Trans. ASSP*, ASSP-31(3):664–678, Jun. 1983.

[2] R. Carlson, B. Granström, and I. Karlsson. Experiments with voice modelling in speech synthesis. *Speech Communication*, 10(5-6):481–489, Dec. 1991.

[3] H. Fujisaki and M. Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *ICASSP*, pages 1605–1608, Tokyo, 1986.

[4] C. Gobl. Voice source dynamics in connect speech. *STL – QPSR*, 1:123–159, 1988.

[5] D. J. Hermes. Synthesis of breathy vowels: Some research methods. *Speech Communication*, 10(5-6):497–502, Dec. 1991.

[6] I. Karlsson. Female voices in speech synthesis. *J. of Phon.*, 19:111–120, 1991.

[7] D. H. Klatt and L. C. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *JASA*, 87(2):820–857, 1990.

[8] J. Marques and L. Almeida. Sinusoidal modeling of voiced and unvoiced speech. In *Eurospeech*, Sep. 1989.

[9] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *JASA*, 49(2 (Part 2)):583–590, 1971.

[10] D. Talkin and J. Rowley. Pitch-synchronous analysis and synthesis for tts systems. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 55–58, Autrans, France, Sep. 1990.