ONE-FORMANT VOCAL TRACT MODELING FOR GLOTTAL PULSE SHAPE ESTIMATION

*Yu-Ren Chien**

National Taiwan University
G.I. Communication Engineering
Taipei 106, Taiwan
yrchien@ntu.edu.tw

Axel Röbel

IRCAM - CNRS STMS
Sound Analysis-Synthesis Team
75004 Paris, France
axel.roebel@ircam.fr

ABSTRACT

This work considers the task of estimating the source and filter from human voice signals. Since the energy of voiced sound concentrates on discrete frequencies, a notable challenge with this task would be that higher pitches in the signal can make the harmonically related frequency response samples of the vocal tract filter an incomplete representation. In view of this, we propose to model the magnitude and phase response of the first formant as an alternative to the minimum phase property of the vocal tract filter. In particular, the magnitude response of the vocal tract filter sampled at the first three partials only, is sufficient for determining the phase response of the first formant. We verified our new method with glottal pulse shape parameter estimation experiments conducted on the CMU Arctic dataset, which showed that single-formant filter is an adequate alternative to minimum-phase filter in vocal tract modeling for glottal pulse shape estimation.

Index Terms— Vocal tract filter, Liljencrants-Fant model, formant, phase, glottal pulse shape

1. INTRODUCTION

Human voice exhibits a wide range of timbres. In the first place, the various vowels pronounced by an individual can be deemed as different timbres. Second, vocal timbres can also vary greatly among speakers or singers. Thirdly, a particular range of pitches may be sung by a male singer at some times as modal voice, and as falsetto voice [1] at other times.

A pitch-independent timbral representation of human voice is attractive for several reasons: To begin with, such a representation could be so closely related to vowels as to underlie a speech recognition or lyric alignment system. Secondly, it could also be applied to speaker or singer identification, where speakers or singers are identified according to specific personal voice qualities found in recordings. Third, falsetto detection is also a possible application of such a timbral representation. Finally, with an appropriate resynthesis procedure, the representation could support pitch transformation of human voice that preserves the vocal timbre.

This work aims to develop a system that estimates the source and filter from human voice signals, where the resulting source-filter model will serve as a pitch-independent timbral representation of the analyzed human voice. As shown in Figure 1, the input to the system is a human voice signal, and the desired output of the system includes the glottal pulse shape and the vocal tract filter.

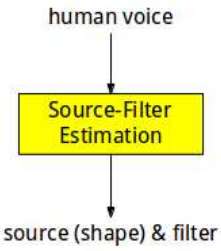

Fig. 1. The objective of this work.

The human voice signal is, by nature, an air pressure signal at the ears of the listener that depends on the particular airflow at the speaker's or singer's lips. It is common practice in acoustic phonetics [2] to model the relation between the airflow signal at the glottis and that at the lips by a linear system, which we call the *vocal tract filter*. In this model, the airflow signal at the lips is represented by the signal resulting from passing the glottal airflow signal through the vocal tract filter. Moreover, the vocal tract filter varies with vowel and with speaker/singer.

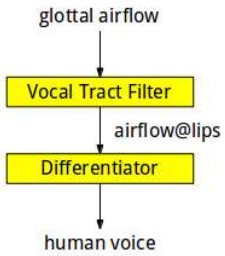

Fig. 2. Voice production model.

The dependency of the human voice signal on the airflow at the lips can be modeled by a differentiator [2], as depicted in Figure 2. Since linear systems commute, we have an alternative model, the source-filter model, for the production of human voice, as shown in Figure 3. In the source-filter model, the derivative of the glottal airflow signal, which we call the *glottal source signal*, goes through the vocal tract filter to give the human voice signal. The shape of glottal source signal is an essential determinant of voice quality; for example, the shape has been shown to differ significantly between

*This work was supported by the National Science Council Overseas Project for Postgraduate Research sponsored by the National Science Council of Taiwan under Grant: NSC102-2917-I-002-071. The authors are grateful for Stefan Huber's help with the experiments.

modal and falsetto registers [1].

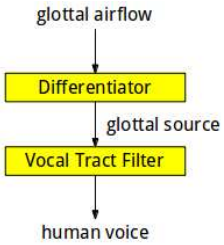

Fig. 3. Source-filter model.

The task of source-filter analysis of human voice signals, as detailed above, is challenging in that only extremely limited information is available from the observed signals. For any given fundamental frequency of the voice, only the harmonic frequency components in the voice signal provide clues for the analysis, while all other frequency components are invariably zero in the absence of noise. As a consequence, it is crucial to take advantage of *a priori* information, typically in the form of a source model or a filter model that closely follows the reality in human voice production with a small number of parameters, in the analysis. As will be discussed in Section 2, all existing approaches to source-filter estimation fail to adopt accurate models for technical reasons.

As will be described in Section 3, we estimate the glottal pulse shape by sampling the pulse shape space and finding among the resulting pulse shapes the one that, along with the observed human voice signal, represents a frequency response with the lowest deviation from our vocal tract filter model [3, 4]. This representation of frequency response is only populated at partial frequencies of the human voice, and is particularly sparse when the pitch is high. To measure the deviation from model, one would need a filter model with a small number of parameters, so that part of the tested frequency response can be used to determine the filter parameters, and the rest of it can be checked against the filter. To this end, we propose to model the vocal tract over the first formant frequency range only, with a three-parameter formant filter. With this filter model fitted to the tested magnitude response at the first three partials, we measure the deviation by the phase errors in the tested phase response. Experimental results will be presented in Section 4, providing comparison with a recent approach that models the vocal tract with a minimum-phase filter in particular.

2. RELATED WORK

The spectral envelope of audio signal has been used widely as a pitch-independent timbral representation of the human voice. Examples of this type of representation include the mel-frequency cepstral coefficients (MFCC) [5] and linear predictive coding (LPC) [6]. It can be regarded as a simplified source-filter model where the source has a fixed, flat spectral envelope, and the overall spectral envelope represents the frequency response of the filter. This simplification can be inappropriate for some applications in that a glottal pulse shape is represented in the frequency domain by some particular amplitude ratios among the partials, and its spectral envelope should actually stretch along the frequency axis as the pitch increases. In contrast, we represent the glottal pulse shape by a parametric model from acoustic phonetics.

The estimation of source and filter from singing voice signals has been investigated in [7], where the glottal pulse shape is essentially represented by the KLGLOTT88 model, and the vocal tract is modeled by linear predictive coding (LPC). Note that linear predictive coding implements a discrete-time all-pole filter with poles not necessarily in complex-conjugate pairs, while the single-tube resonator is a continuous-time filter characterized by complex-conjugate pairs of poles [8]. The LPC vocal tract model also underlies a recent approach proposed in [9]. To ensure the best possible accuracy in modeling, we adopt the continuous-time formant filter and the transformed Liljencrants-Fant model in this work.

Another type of vocal tract model that has been investigated in the literature is the minimum-phase filter [10, 11, 3, 4]. In [3, 4], the glottal pulse shape is estimated by sampling the pulse shape space and finding among the resulting pulse shapes the one that, along with the observed human voice signal, represents a discrete-time frequency response with the lowest deviation from the minimum phase. To measure this deviation, it is assumed that the impulse response of the discrete-time vocal tract filter deviates significantly from zero only within the first $2N_p$ samples, i.e., it has an effective duration no greater than $2N_p$, where N_p denotes the number of partials below the Nyquist frequency. The tested frequency response (which is sampled at partial frequencies) then approximates the discrete Fourier transform (DFT) of this "finite impulse response," and determines (with its magnitude) a minimum-phase filter, against which the tested phase response can be checked. Even so, the assumption may fail when the impulse response of the vocal tract filter has a large effective duration, or when the fundamental frequency is high. Moreover, even if the partials below the Nyquist frequency are sufficient for determining the minimum-phase filter, the higher portion of these partials could have been corrupted by noise or nonstationarity in the signal. In stark contrast, the single-formant filter adopted here does not involve the validity of the tested frequency response as a DFT spectrum, nor does it depend on the higher partials in the signal; in this respect, the proposed filter model is believed to solve the problems with the minimum-phase filter model.

3. METHOD

Given a human voice signal represented by the sequence of short-time spectra $\{\mathbf{s}^{(i)}\}_{i=1}^{L}$, we perform the estimation of glottal pulse shape at the corresponding (uniformly spaced) sequence of time positions in the signal, $\{t_i\}_{i=1}^{L}$, with a frame rate equal to the minimum fundamental frequency in the signal. To limit our exploration in the pulse shape space to a finite number of pulse shapes, we sample the space in advance, giving the set of pulse shape parameter values $\Gamma = \{\gamma_m\}_{m=1}^{N_s}$. For each analysis time position t_i, we find the pulse shape $\hat{R}_d^{(i)}$ that, along with the observed human voice spectrum $\mathbf{s}^{(i)}$, represents a frequency response with the lowest deviation from our vocal tract filter model:

$$\hat{R}_d^{(i)} = \arg \min_{R_d \in \Gamma} D\left(R_d, \mathbf{s}^{(i)}\right), \tag{1}$$

where $D(\cdot)$ measures the deviation.

3.1. Sampling the Space of Glottal Pulse Shapes

The glottal source signal can be approximated by the transformed Liljencrants-Fant model, which is a three-parameter signal model [12, 7]. The three parameters are the fundamental period T_0, the closure excitation magnitude E_e, and the pulse shape parameter R_d.

The model has the following form:

$$g(t;T_0,E_e,R_d) = \begin{cases} E_0 e^{\alpha t}\sin(\omega_g t), & 0 \le t \le T_e; \\ -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_0-T_e)}], & T_e \le t \le T_0, \end{cases} \quad (2)$$

where t denotes the time in seconds, and the dependent variables E_0, α, ω_g, T_e, ϵ, and T_a can be derived as functions of (T_0, E_e, R_d). Since T_0 and E_e have no effect on the pulse shape, we let the space of glottal pulse shapes be represented by $\{R_d : 0.3 \le R_d \le 5.5\}$ and assume $T_0 = 1$ and $E_e = 1$ throughout the rest of this paper. The upper limit of an R_d value is 2.7 in standard transformed LF model; here, the raised value of the limit depends on the extension formulas given in [4]. Moreover, when the values of T_0 and E_e are fixed, R_d is mapped to T_e by an invertible function.

As for the sampling, we let $\gamma_1 = 0.3$ and $\gamma_{N_s} = 5.5$. The intermediate values $\gamma_2, \cdots, \gamma_{N_s-1}$ are such that the open quotient values (T_e/T_0) corresponding to $R_d = \gamma_1, \cdots, \gamma_{N_s}$ form a uniform sampling in the open quotient space.

3.2. Measuring the Deviation From Vocal Tract Model

As formulated in (1), we estimate the pulse shape by optimizing an objective $D(\cdot)$ over the set of pulse shapes Γ. Each shape hypothesis R_d implies a frequency response of the vocal tract filter that remains to be validated. We test the validity of the frequency response by measuring its deviation from our vocal tract filter model.

3.2.1. Tested Frequency Response

As the first step in the procedure of deviation measurement, a harmonic representation of the human voice signal is extracted from the observed spectrum $\mathbf{s}^{(i)}$. To this end, we estimate the fundamental frequency contour with a monophonic version of the pitch estimator in [13], which gives a pitch estimate for each unvoiced or silent time position as well as each voiced time position. Let the complex spectrum values interpolated at the first 3 partials be denoted by Y_1, Y_2, and Y_3. With the observed signal regarded as the output of a filter and the glottal source signal (as specified by the hypothesis R_d) regarded as its input, the frequency response of the filter can be calculated at the 3 partials:

$$H_p = \frac{Y_p}{X_p},\ p = 1, 2, 3, \quad (3)$$

where X_1, X_2, and X_3 denote the corresponding Fourier coefficients of the glottal source signal specified by R_d.

3.2.2. Vocal Tract Filter Model for Low Frequencies

For the true vocal tract filter, we assume that its frequency response is primarily shaped by the first formant at the first 3 partials, with all other formants or anti-formants located far above the 3 partials. Consider the frequency response of the continuous-time formant filter [2]:

$$H(\omega; c, \sigma_0, \omega_0) = \frac{c}{[j\omega - (\sigma_0 + j\omega_0)][j\omega - (\sigma_0 - j\omega_0)]}, \quad (4)$$

where ω is the angular frequency in radians per second, $c > 0$ is a gain factor introduced to compensate for the fixed value of E_e in the source model, $\sigma_0 < 0$ controls the bandwidth, and ω_0 is the formant frequency. Its magnitude and phase are given by

$$|H(\omega; c, \sigma_0, \omega_0)| = \frac{c}{\sqrt{(\omega^2 - \omega_0^2 - \sigma_0^2)^2 + 4\sigma_0^2\omega^2}}, \quad (5)$$

and

$$\angle H(\omega; c, \sigma_0, \omega_0) = \begin{cases} -\arctan\frac{-2\sigma_0\omega}{\sigma_0^2+\omega_0^2-\omega^2}, & \text{if } \omega^2 < \sigma_0^2 + \omega_0^2; \\ -\frac{\pi}{2}, & \text{if } \omega^2 = \sigma_0^2 + \omega_0^2; \\ -\pi + \arctan\frac{-2\sigma_0\omega}{\omega^2-\sigma_0^2-\omega_0^2}, & \text{if } \omega^2 > \sigma_0^2 + \omega_0^2. \end{cases} \quad (6)$$

When $\omega^2 \ll \sigma_0^2 + \omega_0^2$, we have

$$|H(\omega; c, \sigma_0, \omega_0)| \approx \frac{c}{\omega_0^2 + \sigma_0^2}, \quad (7)$$

and

$$\angle H(\omega; c, \sigma_0, \omega_0) \approx \frac{2\sigma_0\omega}{\sigma_0^2 + \omega_0^2}, \quad (8)$$

which means that the effect of higher formants or anti-formants on the overall frequency response at the first 3 partials can be approximated by a linear-phase all-pass filter. Therefore, we can reasonably model the vocal tract with the single-formant filter for the first 3 partial frequencies.

3.2.3. The Deviation

To measure the deviation of the tested frequency response (3) from the single-formant filter (4), we fit the filter to the tested magnitude response, and evaluate the phase errors in the tested phase response with respect to the fitted filter.

To fit the single-formant filter to the tested magnitude response, we solve the following system of equations:

$$|H(\omega_p; c, \sigma_0, \omega_0)| = |H_p|,\ p = 1, 2, 3, \quad (9)$$

where ω_p denotes the angular frequency of the pth partial. A solution exists if and only if c_s, A_s, and B are all positive and $B < \sqrt{A_s}$, where

$$c_s = \frac{p_{1,2} - p_{1,3}}{q_{1,3} - q_{1,2}}, \quad (10)$$

$$p_{1,2} = \omega_1^2\omega_2^2, \quad (11)$$

$$p_{1,3} = \omega_1^2\omega_3^2, \quad (12)$$

$$q_{1,2} = \frac{\frac{\omega_1^2}{|H_2|^2} - \frac{\omega_2^2}{|H_1|^2}}{d_{1,2}}, \quad (13)$$

$$q_{1,3} = \frac{\frac{\omega_1^2}{|H_3|^2} - \frac{\omega_3^2}{|H_1|^2}}{d_{1,3}}, \quad (14)$$

$$d_{1,2} = \omega_1^2 - \omega_2^2, \quad (15)$$

$$d_{1,3} = \omega_1^2 - \omega_3^2, \quad (16)$$

$$A_s = p_{1,2} + q_{1,2}c_s, \quad (17)$$

$$B = \frac{\frac{c_s}{|H_1|^2} - (\omega_1^2 - \sqrt{A_s})^2}{4\omega_1^2}. \quad (18)$$

The solution is

$$c = \sqrt{c_s}, \quad (19)$$

$$\sigma_0 = -\sqrt{B}, \quad (20)$$

$$\omega_0 = \sqrt{\sqrt{A_s} - B}. \quad (21)$$

Note that the 3 positivity conditions ensure that the fitted filter has real coefficients, while the other inequality condition ensures that the poles are complex conjugates of each other.

Method	BDL-A	BDL-B	JMK-A	JMK-B	SLT-A	SLT-B	Mean
F1	0.63513	0.61094	0.06314	0.09787	0.21776	0.24906	0.31232
Min-φ	0.50852	0.46195	0.20102	0.19639	0.10467	0.21843	0.28183

Table 1. Experimental results in Pearson's r coefficient.

Once the formant filter is fitted to the tested magnitude response, we can check the tested phase response against the fitted filter. Since the glottal closure instant is unknown, there is an unknown linear-phase bias in the Fourier coefficients $\{X_p\}_{p=1}^{3}$ with respect to the true glottal source signal. In addition, there is an unknown linear-phase bias in our vocal tract model because higher formants or anti-formants are not included in the model. As a result, a tested phase response that deviates from the phase response of the fitted filter by an amount proportional to frequency would indicate that the tested frequency response is perfectly valid for a vocal tract filter. This justifies the measurement of deviation by the nonlinearity in, instead of the magnitude of, the phase errors found in the tested phase response with respect to the fitted filter [4]:

$$D\left(R_d, \mathbf{s}^{(i)}\right) = \begin{cases} \sum_{l=0}^{2} \frac{1}{3} \sum_{p=1}^{3} \left(\frac{\Delta^{-l}\Delta^2 \angle R_p^\theta}{\pi} \right)^2, & \text{fitting successful;} \\ 3, & \text{otherwise.} \end{cases} \tag{22}$$

Here Δ^{-l} denotes the lth-order antidifference operator, Δ^2 denotes the second-order forward difference operator, $R_0^\theta = 1$, and

$$R_p^\theta = \frac{H_p}{H(\omega_p; c, \sigma_0, \omega_0)}, \; p = 1, 2, 3. \tag{23}$$

4. RESULTS AND DISCUSSIONS

4.1. Test Data

To collect human voice data for testing a source-filter estimator, it would be desirable, if not impossible, to record the glottal airflow in addition to the sound pressure radiated from the lips, so that the ground-truth pulse shape could be obtained directly from the airflow. In practice, it is virtually impossible to measure the airflow through one's vocal folds, and the feasible measurement is the electroglottograph (EGG) [14]. We test our glottal pulse shape estimator on the CMU Arctic dataset [15], which contains both EGG measurements and acoustic recordings of the speech of two male subjects (BDL and JMK) and one female subject (SLT). The speech of each subject is composed of two parts (part A and part B), each part made up of about 500 phrases. The open quotient of glottal pulse is estimated from the EGG measurements with the DECOM method [16] as a function of time, which serves as the ground truth of glottal pulse shape.

4.2. Performance Measure

In our experiments, some of the time positions $\{t_i\}_{i=1}^{L}$ are voiced, while the others are unvoiced or silent. We perform voicing detection by detecting glottal closure instants (GCIs) with the SIGMA algorithm [17]. After linear interpolation on a 250-hertz time grid, the pulse shape estimates within each voiced segment are converted to open quotient values and compared with the EGG-derived open quotient values within the voiced segment with Pearson's r coefficient. An average r coefficient, weighted by segment length, is calculated over all the voiced segments in a phrase.

4.3. Comparison With Minimum-Phase Filter

Tests of the proposed estimator gave an average correlation coefficient, unweighted among phrases, for each part in the testset, as listed in the row "F1" in Table 1. Here, the performance varies widely among different speakers, with the average r coefficient exceeding 0.6 for speaker BDL, and below 0.3 for speakers JMK and SLT. To investigate the effect of vocal tract model on the performance, we repeated the tests for a variant of the estimator that resulted from substituting the minimum-phase filter [4] for the single-formant filter, which gave the results in the row "Min-ϕ" in Table 1. For both methods, the correlation coefficients are much lower for speakers JMK and SLT than for speaker BDL. The advantage of the single-formant filter over the minimum-phase filter could be observed from the results for speaker BDL, where the minimum-phase filter gave an r coefficient close to 0.5 on average. Note that the minimum-phase filter outperforms the single-formant filter for speaker JMK. Inspection of the spectrogram of a phrase from the speaker showed that the third partial of the voice is often so weak that no reliable sinusoidal parameters can be extracted from it. Since only 3 partials are used in fitting the single formant filter, any corrupted partial in the three can have a dramatic impact on the estimation.

5. CONCLUSION

A low-frequency-band vocal tract filter model for glottal pulse shape estimation has been presented. The model has only 3 parameters and can thus be uniquely determined by filter gain values at 3 different frequencies. Since only a small number of spectral observations are available for glottal estimation from a high-pitched voice signal, this model is believed to be particularly suitable for analyzing female or singing voice signals. Experiments on speech data show that the proposed model yields pulse shape estimates that are as accurate as those obtained with the minimum-phase filter model. In the future, it would be interesting to evaluate the performance of the estimator on singing voice data.

6. REFERENCES

[1] Gláucia Laís Salomão and Johan Sundberg, "What do male singers mean by modal and falsetto register? An investigation of the glottal voice source," *Logopedics Phoniatrics Vocology*, 2009.

[2] Gunnar Fant, *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*, Mouton, The Hague, 1970.

[3] Gilles Degottex, Axel Roebel, and Xavier Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[4] Stefan Huber, Axel Roebel, and Gilles Degottex, "Glottal source shape parameter estimation using phase minimization variants," in *Interspeech*, Portland, United States, September 2012.

[5] Paul Mermelstein, "Distance measures for speech recognition—psychological and instrumental," in *Pattern Recognition and Artificial Intelligence: Proceedings of the Joint Workshop on Pattern Recognition and Artificial Intelligence, Held at Hyannis, Massachusetts, June 1-3, 1976*, C.H. Chen, Ed. 1976, Academic Press Rapid Manuscript Reproduction, Acad. Press.

[6] John D. Markel and Augustine H. Gray, *Linear prediction of speech*, Springer-Verlag, New York, 1976.

[7] Hui-Ling Lu, *Toward a High-Quality Singing Synthesizer with Vocal Texture Control*, Ph.D. thesis, Stanford University, 2002.

[8] Gunnar Fant, *Speech Acoustics and Phonetics: selected writings*, Springer, Dordrecht, 2005.

[9] Alan O. Cinneide, *Phase-distortion-robust voice-source analysis*, Ph.D. thesis, Dublin Institute of Technology, 2012.

[10] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Zeros of Z-transform (ZZT) decomposition of speech for source-tract separation," in *Proc. ICSLP, International Conference on Spoken Language Processing, Jeju Island (Korea)*, 2004.

[11] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.

[12] Gunnar Fant, "The LF-model revisited. Transformations and frequency domain analysis," *STL-QPSR*, 1995.

[13] Chunghsin Yeh, Axel Roebel, and Xavier Rodet, "Multiple fundamental frequency estimation and polyphony inference of polyphonic music signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18-6, pp. 1116–1126, 2010.

[14] F.L.E. Lecluse, M.P. Brocaar, and J. Verschuure, "The electroglottography and its relation to glottal activity," *Folia Phoniatrica et Logopaedica*, vol. 27, pp. 215–224, 1975.

[15] John Kominek and Alan W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004.

[16] Nathalie Henrich, Christophe d'Alessandro, Boris Doval, and Michèle Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–1332, March 2004.

[17] M. R. P. Thomas and P.A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 8, pp. 1557–1566, Nov 2009.