# JOINT ESTIMATE OF SHAPE AND TIME-SYNCHRONIZATION OF A GLOTTAL SOURCE MODEL BY PHASE FLATNESS

*Gilles Degottex, Axel Roebel, Xavier Rodet*

IRCAM - CNRS-UMR9912-STMS, Analysis-Synthesis Team
1, place Igor-Stravinsky, 75004 Paris

**ABSTRACT**

A new method is proposed to jointly estimate the shape parameter of a glottal model and its time position in a voiced segment. We show that, the idea of phase flatness (or phase minimization) used in the most robust Glottal Closure Instant detection methods can be generalized to estimate the shape of the glottal model. In this paper we evaluate the proposed method using synthetic signals. The reliability related to fundamental frequency and noise is evaluated. The estimation of the glottal source is useful for voice analysis (ex. separation of glottal source and vocal-tract filter), voice transformation and synthesis.

***Index Terms—*** glottal source, joint estimate, phase flatness, phase minimization

## 1. INTRODUCTION

The source-filter model (eq. 1) is used in this paper to represent the voice production. This model is made of three main elements which are convolved in the time domain and therefore multiplied in the frequency domain. These elements are: the glottal source, the vocal-tract and the lips radiation. The glottal source is assumed to be produced by the periodic opening and closing of the glottis. This source has therefore a specific shape in time domain. Additionally, this shape has a time-position relatively to a time reference (ex. the center of an analysis window). Then, the Vocal-Tract Filter (VTF) transform this source. Finally, this transformed source is radiated outside of the mouth through the lips adding one more filtering effect. Analyzing a window of voiced signal, we estimate the shape parameter of a glottal model (a shape model of the glottal source) and we synchronize temporally this glottal model in the analysis window.

The time synchronization can be reduced to the detection of a maximum excitation instant. Physiologically, this instant corresponds more or less to the closure of the glottis. Therefore, it is called Glottal Closure Instant (GCI). Numerous GCI detection methods already exist [1, 2, 3]. The source model is often seen as a Dirac and thus, one of the best approaches is to flatten the phase of an LPC residual [1, 2] (minimize the absolute value of the phase as much as possible). As soon as the time-synchronization is assumed to be known, the shape parameter is then estimated [4, 5, 6]. Vincent *et. al* proposed to use the error of an ARX model to estimate the time-synchronization [3]. In his method, the glottal shape is also jointly approximated with a dictionary of shapes. We recently proposed a rough estimate of the shape which is time-independent [7]. We used this estimate to develop a fast and robust method detecting GCIs [8]. In this new paper, the proposed method is used as a refinement method. Estimates of the methods mentioned above can be used as starting values for the proposed method.

The glottal source is a causal/anti-causal mixed-phase signal [9]. Therefore, the Z-transform of the glottal source has roots (zeros or poles) outside of the unit circle. Since the production model is made of time convolutions, theses roots are kept in the final voiced signal. The impulse response of the VTF is assumed to be minimum-phase because it is a passive and lossy medium. Its roots are thus strictly inside the unit circle. Therefore, since the source is a mixed-phase signal and the VTF is a minimum-phase signal, the only property difference between the source signal and the VTF signal is in the phase. We propose to focus on this main difference of property: the phase of a voice model are optimized to fit the phase of the Fourier representation of an analysis window. The phase flatness criteria has already been proposed to detect GCIs [1, 2] (also known as phase minimization criteria). We propose to extend this idea to the estimate of a shape parameter of a glottal model.

Section 2 presents the spectral relations obtained from the source-filter model: VTF derivation, convolutive residual and phase flatness measure. The main sources of errors disturbing these computations are also discussed at the end of this section. Section 3 gives more technical details about the joint estimation method. Finally, in section 4, the reliability of this estimator is evaluated with synthetic signals. A comparison with Electro-Glotto-Graphic signals and an evaluation with real signals are planned for a future publication.

## 2. VOICE MODEL AND PHASE FLATNESS CRITERIA

First, this section presents the voice model and its elements. Then, we will see that the VTF can be retrieved from a glottal model and a function computing the minimum-phase spectrum of its argument. Our goal is to estimate the shape parameter $\theta$ and the time-synchronization $\phi$ of a glottal model with an observed voiced signal. Therefore, we express the convolutive residual as the ratio between the observed spectrum and the model spectrum. If the phase of this residual is flat (as small as possible), this residual is a Dirac in time domain, the model is equal to the observed spectrum and $(\theta, \phi)$ are optimal. A measure of phase flatness is therefore proposed to measure this optimality.

### 2.1. Voice production model

Within a given window, the voiced acoustic waveform is assumed to be a stationary periodic signal of fundamental frequency $f_0$. Therefore, it can be fully represented by a discrete spectrum $S_k$ where the $k^{th}$-bin represents the $k^{th}$-harmonic partial of the Fourier Transform of the observed signal. From this Fourier representation, we can thus express the voice production model as follow:

$$S_k = e^{jk\phi} \cdot G_k \cdot C_{k-} \cdot L_k \tag{1}$$

$e^{jk\phi}$ define the time position $\phi$ of the glottal shape. $G_k$ is a mixed-phase spectrum defining the shape of the glottal source. $C_{k-}$ is the VTF, a minimum-phase filter sampled by $f_0$ (the minimum-phase property is denoted by the negative sign). Finally, $L_k$ is the filter corresponding to the lips radiation. This filter is usually modeled by a time derivative [10] and therefore $L_k = jk$.

### 2.2. Vocal-Tract Filter derivation

First, we define $\mathcal{E}_-(.)$ as a function computing the minimum-phase spectrum of its argument through the power cepstrum [11]. From equation (1), by division in frequency domain (deconvolution in time), one can express the Vocal-Tract Filter estimate $C^{\theta}_{k-}$ with a given glottal model $G^{\theta}_k$ parametrized by $\theta$ and the lips radiation model $L_k = jk$:

$$C^{\theta}_{k-} = \mathcal{E}_-\big(\frac{S_k}{G^{\theta}_k \cdot jk}\big) \tag{2}$$

A more general case using a minimum-phase envelope estimate is discussed in [7]. Because $\mathcal{E}_-(.)$ is computed from the amplitudes only, this function has the property of distributivity on multiplication and division. Additionally, the lips radiation model can be either removed from the numerator or multiplied by the denominator:

$$C^{\theta}_{k-} = \frac{\mathcal{E}_-(S_k/jk)}{\mathcal{E}_-(G^{\theta}_k)} = \frac{\mathcal{E}_-(S_k)}{\mathcal{E}_-(G^{\theta}_k \cdot jk)} \tag{3}$$

Practically, in both cases, the value at frequency zero has to be extrapolated to compute the minimum-phase spectrums with $\mathcal{E}_-(.)$ because this value is set to zero by the lips radiation in $S_k$. From our experiments, the second solution is more stable and is thus used in the following presentation.

### 2.3. Convolutive residual and phase flatness

Since $C^{\theta}_{k-}$ is expressed as a function of $\theta$, we can now derive the computation of the convolutive residual, the ratio of the observed spectrum by the model spectrum:

$$R^{(\theta,\phi)}_k = \frac{S_k}{e^{jk\phi} \cdot G^{\theta}_k \cdot C^{\theta}_{k-} \cdot jk} = \frac{S_k \cdot \mathcal{E}_-(G^{\theta}_k \cdot jk)}{e^{jk\phi} \cdot G^{\theta}_k \cdot \mathcal{E}_-(S_k) \cdot jk} \tag{4}$$

which can be written as:

$$R^{(\theta,\phi)}_k = e^{-jk\phi} \cdot \frac{S_k}{\mathcal{E}_-(S_k)} \cdot \frac{\mathcal{E}_-(G^{\theta}_k)}{G^{\theta}_k} \cdot \frac{\mathcal{E}_-(jk)}{jk} \tag{5}$$

The amplitudes of $S_k$, $G^{\theta}_k$ and $jk$ are flatten by their respective minimum-phase spectrums. Therefore, $R^{(\theta,\phi)}_k$ is an all-pass filter whatever the parameters are: $|R^{(\theta,\phi)}_k| = 1 \ \forall k \ \forall\theta \ \forall\phi$. As desired, the problem is focused on the phase difference between the observed signal and its model.

Finally, if we assume the real shape of the glottal source can be correctly represented by our chosen glottal model and there is only an error of parametrization $(\Delta\theta, \Delta\phi) = (\theta^*, \phi^*) - (\theta, \phi)$, the observed spectrum can be replaced by the voice production model:

$$R^{(\theta,\phi)}_k = e^{-jk\phi} \cdot \frac{e^{jk\phi^*} G^{\theta^*}_k C_{k-} L_k}{\mathcal{E}_-(e^{jk\phi^*} G^{\theta^*}_k C_{k-} L_k)} \cdot \frac{\mathcal{E}_-(G^{\theta}_k)}{G^{\theta}_k} \cdot \frac{\mathcal{E}_-(jk)}{jk} \tag{6}$$

The lips radiation $L_k$ is equal to $jk$ and can thus be eliminated. Moreover, if the cepstral coefficients of the VTF above the quefrency $q_0$ $(= T_0)$ are negligible, $\mathcal{E}_-(C_{k-}) \approx C_{k-}$ and the VTF is also eliminated from the equation. Additionally, defining the spectrum error $X^{\Delta\theta}_k = X^{\theta^*}_k / X^{\theta}_k$:

$$R^{(\theta,\phi)}_k \approx e^{jk\Delta\phi} \cdot G^{\Delta\theta}_k / \mathcal{E}_-(G^{\Delta\theta}_k) \tag{7}$$

$G^{\theta}_k$ is not linear-phase. Consequently, the position error $e^{jk\Delta\phi}$ can not compensate the shape error because $G^{\Delta\theta}_k$ and its minimum-phase spectrum are not linear-phase. Finally, if the phase of $R^{(\theta,\phi)}_k$ tends to zero ($R^{(\theta,\phi)}_k$ tends to a Dirac), $\Delta\phi$ tends to zero and the phase of $G^{\Delta\theta}_k$ tends to zero. Consequently, it means that $(\theta, \phi)$ tends to $(\theta^*, \phi^*)$.

To measure the phase flatness of the convolutive residual, we propose to use the Mean Squared Phase (MSP):

$$MSP(\theta, \phi, N) = \frac{1}{N} \sum_{k=1}^{N} \big(\angle R^{(\theta,\phi)}_k\big)^2 \tag{8}$$

where $N$ is the maximum number of harmonics taken into account in the measure and $\angle(.)$ is the function computing the angle of the given complex number. $MSP(\theta, \phi, N)$ can thus be minimized to optimize the position and the shape parameter of the glottal model.

### 2.4. Main sources of error

*Condition on the VTF*: Due to the periodicity of the glottal source, the VTF is sampled by the harmonics. Therefore, the cepstral coefficients of the VTF above $q_0$ $(= T_0)$ have to be negligible. If it is not the case, the minimum phase computed by $\mathcal{E}_-(S_k)$ in eq. (3) does not correspond to the real ones. Lower $f_0$, better the VTF representation (see sec. 4 for the consequences on the method).

*Noise level*: In high frequencies, the noise level exceeds the harmonic level. Therefore, the sampling rate has to be small enough to avoid this noise. The problem has to be constrained to the smallest sufficient frequency band.

*Minimum-phase reconstruction*: To obtain minimum-phase estimates as close as possible to the reality, the sampling rate has to be as high as possible. Indeed, theoretically, all frequencies up to infinity are needed to reconstruct the perfect minimum phase of a spectrum. This condition is opposed to the previous one. Therefore, a balance should be struck between these two sources of bias. Since the consequences of this problem is not yet evaluated precisely, we kept the sampling rate at $32kHz$.

## 3. METHOD

The parameters are estimated in an optimization context: for each parameters value $(\theta, \phi)$, the convolutive residual is computed with equation (4). Then, the corresponding error is computed with the phase flatness measure given by equation (8).

The Discrete Fourier Transform (DFT) is used to compute the spectrum of a voiced segment with a *hanning* window function. A window with a length of only one period is not suitable since the effect of the windowing function has to be negligible compared to the features we want to extract from the DFT. Therefore, in this context of voice analysis, 3 periods are used. From this spectrum, the harmonics $S_k$ have to be estimated: the amplitude and phase of the $k^{th}$-harmonic partial are obtained by estimating the amplitude and phase of the nearest peak of $k \cdot f_0$ in the DFT of the voiced segment.

In this paper, the Liljencrants-Fant (LF) glottal model [12] is used. This model define the time-derivative of the glottal shape $G^{\theta}_k \cdot jk$. This model is controlled by 3 shape parameters $(O_q, \alpha_m, t_a)$,

the fundamental frequency $f_0$ and the excitation amplitude $E_e$. We assume $f_0$ to be known *a priori*. Numerous methods can be used to compute $f_0$ from the voiced signal (ex. *YIN* [13], *Swipep* [14] or by harmonic matching [15]). Since the proposed method works with the phase only, it is not necessary to estimate $E_e$. Finally, instead of using $\theta = (O_q, \alpha_m, t_a)$ we use the relaxing parameter $\theta = Rd$ which is a value on a meaningful curve in the shape parameter space [16, 5].

Theoretically, since only two variables $(Rd, \phi)$ are estimated, only the first two harmonics are necessary to find a solution ($N = 2$ in eq. 8). However, the glottal model definitely does not corresponds perfectly to the real glottal source. Therefore, a mean solution with all available harmonics is preferable. $N$ is thus set to $\lfloor VUF/f_0 \rfloor$, where $VUF$ is a Voiced/Unvoiced Frequency [17].

### 3.1. Shape estimate with an *a priori* time-synchronization

In this section, we will show that the estimation results are not satisfying if the time-synchronization $\phi$ is not jointly optimized with the shape parameter $Rd$. Indeed, the time-synchronization can be assumed to be known thanks to numerous methods [1, 2, 8]. Using one of these estimates, the convolutive residual is computed with $R_k^{(Rd,0)}$ and the problem is reduced to a one-dimensional constrained optimization. We used a Brent's method to find the global minimum.

The $Rd$ estimation error related to a time-synchronization error $\Delta\phi$ has been computed with synthetic signals of fundamental frequency $128Hz$ (see section 4 for more information about the synthesis). Considering a time-synchronization error $|\Delta\phi| < 2.5\%$ of the period, the standard deviation of the $Rd$ error is $\approx 0.75$. With $|\Delta\phi| < 10\%$ of the period, the standard deviation of the $Rd$ error is $\approx 1.25$. Therefore, the error of the $Rd$ estimate increases substantially when $|\Delta\phi|$ increases. Such a sensibility to a time-synchronization error is therefore unsatisfying and $Rd$ and $\phi$ have to be be jointly estimated.

### 3.2. Joint estimate of shape and time-synchronization

Using equation (8), the error function corresponding to a linear-phase deviation is a deep and narrow valley embedded in a noisy neighborhood (fig. 1 bottom right). In such a context, the search of the global minimum is unrealistic. However, the high frequency behavior of the error function comes from the high frequencies of the convolutive residual. Therefore, to smooth the error function, $R_k^{(Rd,\phi)}$ is first limited to the $2^{nd}$ harmonic ($N = 2$). Then, a Sequential Quadratic Programming algorithm (SQP) is used to find the minimum of the error function from a starting point [7, 8] (fig. 1 top left). Then, $N$ is increased one harmonic by one harmonic up to $\lfloor VUF/f_0 \rfloor$ while using the SQP algorithm at each incrementation to refine $(Rd, \phi)$ obtained at the preceding step (fig. 1).

## 4. EVALUATION WITH SYNTHETIC SIGNALS

The results of such an estimator are difficult to validate. A measurement of the glottal flow (usually associated to the acoustic source of the vocal-tract) could be compared to the glottal model estimates, but the measurement of such a flow is not yet possible *in vivo*. Therefore, we mainly evaluate the proposed estimator with synthetic signals.

The synthetic signal (eq. 9) is controlled by the LF shape parameter $Rd^*$, the position of the GCI $\phi^*$, the fundamental frequency $f_0$ and two zero-mean Gaussian noises $N^{\sigma_g}$ and $N^{\sigma_a}$ of standard-deviation $\sigma$ corresponding to glottal noise and additive noise respectively. The periodic behavior of the glottal source is produced by

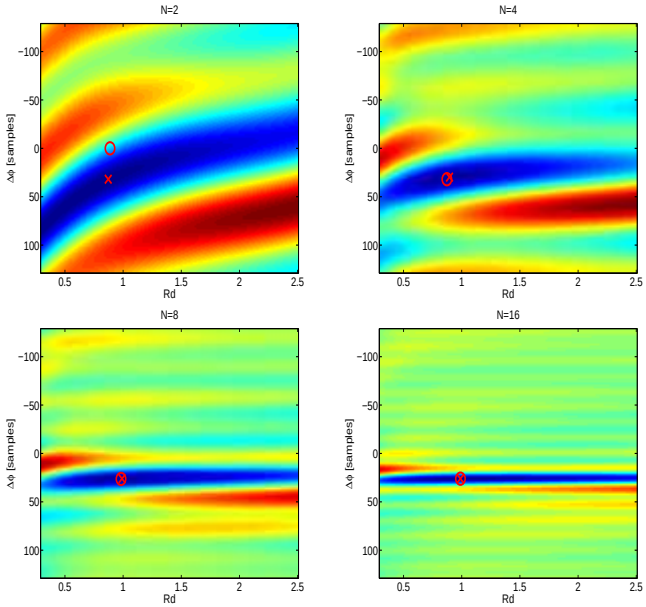


**Fig. 1**. Error surface evolution while increasing the number of harmonics $N$ (darker the color, smaller the error). Starting values are indicated with a circle and SQP final step with a cross. Optimal values are $\Delta\phi = 25$ samples and $Rd = 1$

repeating $G^{Rd^*}(\omega)$ every $1/f_0$. 11 different VTFs $C_-^p(\omega)$ are used to model vowels $p$ covering the vocalic triangle. Theses VTFs are obtained from an articulatory model proposed by Maeda [18]. The vocal-tract area function is first computed from articulatory parameters. Then, the transfer function $C_-^p(\omega)$ is computed from the reflexion coefficients corresponding to the area function. The length of the acoustic tube is fixed to $17cm$ and the opening of the glottis is set to $2.5cm^2$ to simulate a loss in this simulated acoustic tube.

$$\begin{aligned} E(\omega) &= e^{j\omega\phi^*} \cdot G^{Rd^*}(\omega) \cdot \Big[\sum_{k\in\mathbb{N}} e^{j\omega k/f_0}\Big] + N^{\sigma_g}(\omega) \\ S(\omega) &= E(\omega) \cdot C_-^p(\omega) \cdot j\omega + N^{\sigma_a}(\omega) \end{aligned} \quad (9)$$

### 4.1. Error related to the fundamental frequency $f_0$

In a first test, the error of the estimation related to $f_0$ is evaluated. For each $f_0$ value, the estimation error is computed for the 11 different VTFs $C_-^p(\omega)$ and a random delay $\phi^*$ in $[-0.1 \cdot T_0; 0.1 \cdot T_0]$. The mean and standard-deviation of theses errors are then computed and shown in figure 2. To focus on the influence of $f_0$, glottal and additive noises are set to zero in equation (9).

For both $Rd$ and $\phi$, the variance of the error increases with $f_0$ since the sampling of the filter response $C_-(\omega)$ by $f_0$ does not provide enough information to reconstruct the minimum phase of $S(\omega)$ perfectly (see sec. 2.4). Both plots show that the method is very satisfying for a range of $f_0$ used in adult voice ($\approx 100 - 250Hz$). Compared to the shape estimate with an *a priori* $\phi$ (sec. 3.1), a joint method is therefore necessary.

### 4.2. Error related to the noise levels $\sigma_g$, $\sigma_a$

A second test evaluates the error of the estimation related to glottal and additive noises. $\sigma_g$ then $\sigma_a$ varies between $-50dB$ and $10dB$. $\sigma_g$ is relative to the standard-deviation of the source signal ($E(\omega)$

without noise) while $\sigma_a$ is relative to the standard-deviation of the speech signal. When one noise is tested, the other one is set to zero. Figure 2 (right plots) shows the mean and standard-deviation of the $Rd$ and $\phi$ errors. For each $\sigma$ value, the error is computed 4 times with the 11 different VTFs $C_-^p(\omega)$ and a random delay $\phi^*$ in $[-0.1 \cdot T_0; 0.1 \cdot T_0]$. To focus on noises influence, $f_0$ is fixed to $128Hz$.

For noise level under $-12dB$: the $Rd$ error is nearly unbiased, its standard-deviation is smaller than 0.5; the time-synchronization is slightly biased and its standard-deviation is smaller than $5\%$ of the period. Consequently, the reliability of the method is satisfying for many applications. Moreover, in this experiment the $VUF$ was fixed to $4kHz$ to focus on noise influence. The error should be smaller when using a $VUF$ estimate [17].

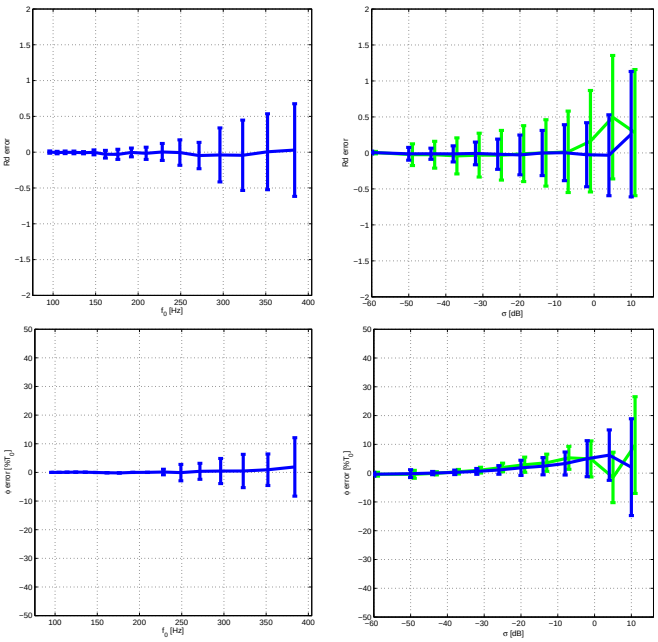

**Fig. 2**. Left plots: $Rd$ and $\phi$ error related to $f_0$. Right plots: Errors related to $N^{\sigma_g}$ (in green (gray in B&W)) and $N^{\sigma_a}$ (in blue (black in B&W)): The disturbing parameter on the horizontal axis, mean and standard-deviation of the estimation error on the vertical axis.

## 5. CONCLUSION

We have argued that the main difference between the glottal source and the Vocal-Tract Filter is their mixed-phase and minimum-phase property. Accordingly, a glottal model estimation method has been proposed. We have shown that the idea of phase flatness used in popular GCI detection methods can be generalized to estimate the shape of a glottal model. We have seen that the $Rd$ estimate is very sensitive to the time-synchronization. Therefore, shape and time-synchronization have to be jointly estimated. A quantitative evaluation of the method with synthetic signals shows that the method is reliable enough for fundamental frequencies corresponding to adult voices and robust for glottal and additive noises.

## 6. ACKNOWLEDGMENTS

This research is partly supported by the "Affective Avatar" ANR project and by a grant of Centre National de la Recherche Scientifique (CNRS).

## 7. REFERENCES

[1] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 325–333, 1995.

[2] A. Kounoudes, P. A. Naylor, and M. Brookes, "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," *ICASSP*, 2002.

[3] D. Vincent, O. Rosec, and T. Chonavel, "Estimation of lf glottal source parameters based on an arx model," *Interspeech*, 2005.

[4] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of z-transform (zzt) decomposition of speech for source-tract separation," *ICSLP*, 2004.

[5] H.-L. Lu, *Toward a High-quality Singing Synthesizer with Vocal Texture Control*, Ph.D. thesis, Stanford, 2002.

[6] R. Fernandez, *A Computational Model for the Automatic Recognition of Affect in Speech*, Ph.D. thesis, Massachusetts Institute of Technology, 2004.

[7] G. Degottex, A. Roebel, and X. Rodet, "Shape parameter estimate for a glottal model without time position," in *SPECOM*, 2009, pp. 345–349.

[8] G. Degottex, A. Roebel, and X. Rodet, "Glottal closure instant detection from a glottal shape estimate," in *SPECOM*, 2009, pp. 226–231.

[9] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," *VOQUAL*, 2003.

[10] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer Verlag, Berlin, 1976.

[11] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice–Hall, 1978.

[12] G. Fant, J. Liljencrants, and Q.-G. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[13] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, April 2002.

[14] A. Camacho, *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, Ph.D. thesis, University of Florida, USA, December 2007.

[15] C. Yeh and A. Roebel, "A new score function for joint evaluation of multiple f0 hypothesis," in *DAFx*, Naples, Italy, Octobre 2004, pp. 234–239.

[16] G. Fant, "The LF-model revisited. transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[17] S.-J. Kim and M. Hahn, "Two-band excitation for hmm-based speech synthesis," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 378–381, 2007.

[18] S. Maeda, "An articulatory model of the tongue based on a statistical analysis," in *Meeting of the Acoustical Society of America*, 1979.