A STRATEGY FOR LF-BASED GLOTTAL-SOURCE & VOCAL-TRACT ESTIMATION ON STATIONARY MODAL SINGING

Fernando Villavicencio

Research & Development Division / Speech Technology Group
Yamaha Corporation
203 Matsunokijima, Hamamatsu, Shizuoka, Japan
fernando.villavicencio@music.yamaha.com

ABSTRACT

This paper presents a methodology for estimation and modeling of the glottal source and vocal-tract information. The strategy proposes a simplified framework based on the characteristics of stationary singing following a selection of glottal pulse model candidates driven by a single shape parameter. True-Envelope based models are applied, allowing efficient modeling of the observed filter information and accurate cancellation of the glottal source contribution in the spectrum. According to experimental studies on synthetic and real signals the methodology observes adequate approximation of the source and filter information, leading to natural resynthesis quality using synthetic glottal excitation. The proposed estimation framework represents a promising technique for voice transformation on stationary modal voice.

Index Terms— Speech analysis, speech synthesis, glottal source estimation, vocal-tract estimation

1. INTRODUCTION

Voice transformation represents a number of techniques allowing us to modify the perceived characteristics of the voice. A fundamental task is found in the manipulation of the excitation and filter characteristics according to the model of the speech production system. A robust decomposition of these elements represents a major challenge due to the limited information available to perform simultaneous estimation, and to potential non-linear interactions not considered in the inverse filtering process.

Some works propose iterative and deterministic methods for voice decomposition such as [1] and [2] respectively. Recent strategies ([3], [4], [5]) use the transformed Liljencrants-Fant (LF) glottal flow model [6] in the analysis framework. In particular, [4] and [5] propose an approximation of the glottal contribution by exhaustive search among LF-model candidates. Previously, LF modeling was considered in methods based on the estimation of a joint source-filter system, refered to as ARX-LF [7], [8].

We aim to manipulate the modal, stationary, monophonic singing voice samples used as corpora of the concatenative singing synthesizer VOCALOID [9]. These corpora are recorded following flat and stable characteristics (*e.g.* loudness, vocal effort, pitch), suggesting to focus the analysis of the excitation characteristics on a reduced acoustic context and to consider an approximation of both glottal and vocal-tract contributions at each speech epoch based on the information of the previous one. Using approximate information of the glottal source might not lead to perceived differences after a resynthesis process, as it can be extrapolated from works as [10]. We remark that continuous speech and expressive singing (including non-modal voice) do not observe, in general, the same acoustic characteristics and should be furthermore studied.

Our motivation is to derive a simplified source-filter estimation framework by reducing the glottal model search and optimization schemas of [4] and [5]. In addition, we consider True-Envelope (TE) based models seeking efficient modeling of the spectral information of both source and filter contributions. The glottal source estimation strategy was introduced in previous work [11]. This paper presents an extensive study including the estimation of the filter contribution and an evaluation on both synthetic and real data.

The paper is structured as follows. In section 2 the various techniques are described. Section 3 presents the proposed estimation strategy. The different matching functions for glottal model selection are described in section 4. The experiments on synthetic and real signals are presented in section 5. The paper ends at section 6 with conclusions.

2. TECHNIQUES

2.1. Glottal shape parameter (*Rd*) based source modeling

The glottal flow, which in a source-filter basis represents the main excitation contribution of voiced speech, is typically represented by its differentiated version, also called derivative glottal waveform. The LF model allows an approximation of this waveform in terms of four parameters (t_p, t_e, t_a, E_e) specifying its main time-domain characteristics. Furthermore, a set of R parameters R_a, R_g, R_k were derived based on observed correlations between t_p, t_e, and t_a. Finally, an analysis on the progression of the R parameters ranging over extreme phonations (*e.g.* from adducted to abducted voice) leads to a single-parameter Rd [6], denoting a progression of the main glottal-pulse shape properties, as shown in Fig.1.

The Rd parameter shows in $0.3 < R_d < 2.7$ its main range of variation. Three main voice qualities are typically distinguished along this range: *pressed*, *modal (or normal)* and *breathy*. In [12], 0.84, 1.19 and 2.9 respectively were found as average values for these voice qualities on baritone sung vowels. Accordingly, Rd estimates on modal stationary phonations might be expected around the corresponding value, while showing a smooth variation over time.

2.2. True-Envelope estimation for efficient spectral modeling

A fundamental aspect of our strategy relies on a precise spectral features extraction. This is achieved using accurate spectral envelope information. TE estimation provides efficient fitting of the spectral envelope based on an iterative cepstral smoothing of the amplitude

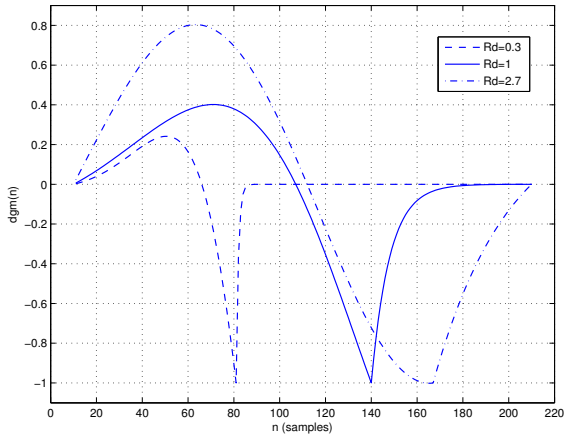

Fig. 1. LF-based derivative glottal pulse for different Rd values.

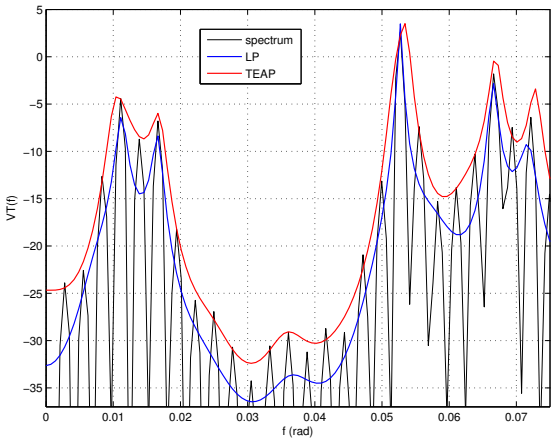

Fig. 2. Example of VTF fitting using Linear Prediction (LP) and TEAP modeling (same order used in both systems).

spectrum [13]. Thus, similarly to [4], we use True-Envelope (TE) based models for both features-modeling and inverse filtering purposes.

TE is used at the glottal source spectra cancellation step in a similar way as in [4], as explained in the next section. However, we keep a physical motivation by using all-pole modeling for the VTF fitting. Note that although Linear-Prediction (LP) is the common solution for this task it shows poor matching of the spectral peaks due to the biasing effect caused by harmonics [14]. This may distort the observed vocal-tract information and the glottal excitation after inverse filtering. Issues related to envelope fitting were already addressed in previous work [8]. We therefore use the True-Envelope based all-pole modeling presented in [15], that we refer here as the TEAP model. This technique uses the envelope estimations obtained from TE as a target spectrum for the autocorrelation matching criteria of an autoregressive filter. Basically, it follows the strategy introduced in [16] using interpolated spectrum information for all-pole modeling.

The cepstral order of the True-Envelope, 0_{TE}, can be set according to the fundamental frequency such as $0_{TE} = F_S/(2f_0)$ for optimal fitting [17] (F_S denotes the samplerate). This value, when applied as the order of the all-pole system provides generally maximal precision. A comparison between LP and TEAP fitting of a spectrum featuring the observed VTF information is shown in Fig.2.

2.3. Vocal tract filter derivation and inverse filtering

In our processing framework the signal is windowed pitch-synchronous in a narrow-band basis (4 speech epochs) centered at the Glottal Closure Instants (GCI). In detail, s_k will denote the $k-th$ frame from signal $s(n)$ centered at gci_k ($s_k = s(n)\ for\ n = [gci_{(k-2)}, gci_{(k+2)}]$). Both derivative glottal flow and VTF information are extracted from each s_k, as described in this section.

To derive the VTF information, in contrast with [3], the glottal source contribution is not cancelled by pre-emphasis filtering. Looking for higher precision we proceed in a similar way as in [4], given a LF model $\check{g}$ of the derivative glottal waveform for s_k we compute its spectral representation in the form:

$$E\check{g} = TE(20\log_{10}|\check{G}|), \tag{1}$$

where $\check{G}$ denotes the DFT of $\check{g}$ and $TE(\cdot)$ the operator representing the True-Envelope estimator. Next, the glottal contribution is cancelled on S_k (DFT of s_k) using the linear representation as follows:

$$S_v = \frac{S_k}{10^{(E\check{g}/20)}}, \tag{2}$$

with S_v denoting the DFT of the vocal-tract related spectrum. Finally, the VTF is computed in terms of the TEAP estimator:

$$V = TE_{ap}(20\log_{10}|S_v|). \tag{3}$$

Conversely, given a system V, the derivative glottal waveform g can be extracted from s_k by inverse filtering:

$$g = V^{-1} * s_k. \tag{4}$$

3. ITERATIVE SOURCE-FILTER EXTRACTION STRATEGY

3.1. Conditions for analysis: stationary modal voice

The motivation of the proposed glottal and vocal tract estimation strategy relies on the assumption of three fundamental conditions with regards to the modal singing signals of interest:

- *Modal vocal effort:* The main glottal shape characteristics can be sufficiently approximated by LF-modeling near reported *modal* Rd parameter values.
- *Stationarity:* the source and filter characteristics vary smoothly. An evolution of the glottal shape between epochs does not represent a difference larger than an assumed ΔRd.
- *Voicing:* the level of turbulence or aspiration noise is low enough to neglect a masking of significant VTF features after cancellation of the glottal contribution on the spectrum.

Although these three conditions may not be fulfilled following the particular characteristics of an individual voice they are commnly observed among modal singing, in particular, in the corpora of interest described in the introductory section.

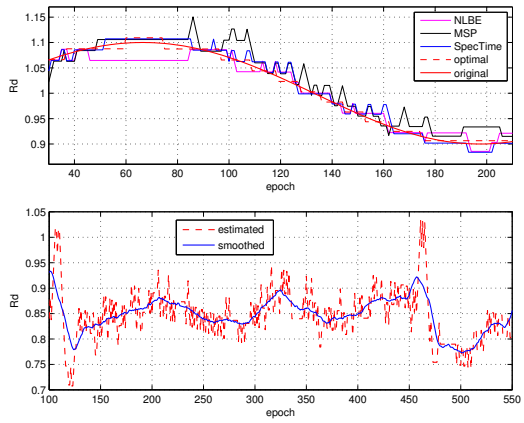

Fig. 3. Example of Rd parameter estimation on a synthetic signal using all matching measures (top). Example of estimation on a real signal before and after smoothing (bottom).

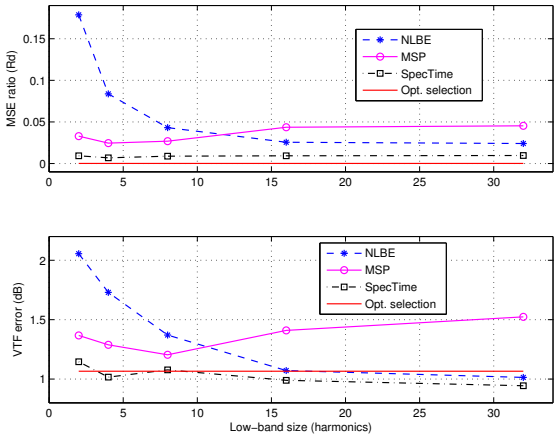

Fig. 4. Performance of Rd and VTF estimation on synthetic data as a function of the matching band (number of harmonics I).

3.2. Estimation framework

The analysis framework is based on the assumption that the vocal-tract configuration does not change between sucessive epochs. Accordingly, at each s_k, the derivative glottal extraction g is obtained by inverse filtering using $V_{(k-1)}$ (the estimation of V at frame $k-1$) according to Eq.4. Then, a representative LF model is selected from the $\breve{g}_c$ candidates derived from the set of Rd values:

$$Rd_C = [Rd_{k-1} - \Delta Rd, \quad Rd_{k-1}, \quad Rd_{k-1} + \Delta Rd]. \qquad (5)$$

The selection is done after matching the candidates with g in terms of any of the measures described in the following sections. Note that Rd_{k-1} corresponds to the value selected for $s_{(k-1)}$ and that ΔRd is set heuristically according to the expected maximal deviation of the glottal shape between epochs. Values in the range of $\Delta Rd = [2.5\%, 10\%]$ of Rd_{k-1} observed adequate results on stationary singing after performing resynthesis using the estimated Rd values to generate the synthetic glottal flow.

The VTF information of s_k, noted V_k, is updated using the selected $\breve{g}_c$ according to Eq.2 and Eq.3. The procedure is repeated for the succesive epochs. A slight modification is considered for initialization: firstly, the number of candidates in Rd_C is increased to explore a larger range within an assumed *modal* interval (*e.g.* $Rd_C = [0.6, 1.3]$). Following, g and V are extracted for each g_c applying Eq.2, Eq.3, and Eq.4 straightforwardly. The initial conditions $Rd_{k=1}$ and $V_{k=1}$ are then chosen according to the closest glottal waveform match. Although this initialization criterion lacks an optimization step for V it was shown to converge near the actual source and filter conditions in experiments with synthetic signals.

4. GLOTTAL WAVEFORM MATCHING

The modeling of the derivative glottal source is performed by selecting a LF model $\breve{g}_c$ of the set described by Eq.5. The selection follows the minimum error between the candidates and g, the excitation extracted from s_k after inverse filtering using $V_{(k-1)}$. To complement the study presented in [11] we evaluate the same error measures, described in the following subsections.

4.1. Mean Square Phase matching (MSP)

This refers to a residual-phase flatness measure inspired by [4] and considered as indicative of the matching between original and synthetic waveforms. A synthetic version of s_k denoted by s'_c is obtained for each candidate $\breve{g}_c$ given V. The MSP error computation is described as follows:

$$s'_c = V * \breve{g}_c, \qquad (6)$$

$$R_c = S_k / S'_c, \qquad (7)$$

$$e_{msp,c} = \frac{1}{I} \sum_{i=1}^{I} (\angle R_{c,p(i)})^2, \qquad (8)$$

where S'_c denotes the DFT of s'_c and $p(i)$ is the bin index in R_c closest to the $i-th$ harmonic. Note that I is the total number of harmonics considered for matching, suggested in [4] to be found in the range $[2, 4]$.

4.2. Joint harmonic and time-domain matching (SpecTime)

A similarity measure between glottal waveforms based on spectral and time-domain information is proposed in [5]. The spectral part corresponds to:

$$e_s = \{0.5 - |cor(\log|G_{p(i)}|, \log|\breve{G}_{c,p(i)}|)|\} \cdot w_s, \qquad (9)$$

where G and $\breve{G}_c$ represent the DFT versions of g and $\breve{g}_c$ respectively, and $p(i)$ the harmonic bins as previously described. The operator $cor(\cdot)$ represents the Pearson correlation between the harmonic amplitudes. The time-domain part is derived similarly:

$$e_t = \{0.5 - |cor(g, \breve{g}_c)|\} \cdot w_t. \qquad (10)$$

Following [5] the weights w_s and w_t are set to 0.6 and 1 respectively. Finally, the total matching error $e_{st,c}$ corresponds to

$$e_{st,c} = e_s + e_t. \qquad (11)$$

4.3. Normalized low-band envelope matching (NLBE)

A novel measure based on the differences between the spectral envelopes Eg and $E\check{g}_c$ (obtained from g and $\check{g}_c$ respectively) was introduced in [11]. The MSE is computed after normalization of the average energy as follows:

$$e_{nlbe,c} = \frac{1}{L}\sum_{f=f_0}^{I\cdot f_0}(Eg_f - [E\check{g}_{c,f} + Gg])^2, \quad (12)$$

where f_0 denotes the fundamental frequency and $I \cdot f_0$ the matching cut-off frequency, limited by I as in the previous measures. L denotes the number of bins covering the matching band. The term Gg denotes the energy bias between the envelopes computed as:

$$Gg = \frac{1}{L}\sum_{f=f_0}^{I\cdot f_0}[Eg_f - E\check{g}_{c,f}]. \quad (13)$$

Note that Gg corresponds to an estimation (in dB) of the LF gain parameter Ee for $\check{g}_c$, computed as

$$\check{E}e_c = 10^{(Gg/20)}. \quad (14)$$

This is an alternative to an approximation based on the minima of g, as applied in [4], [5], and [3].

For comparison purposes, the DFT size was fixed to the length of s_k for all analysis. Note that the va-lues observed at the $p(i)$ positions may not accurately represent the actual harmonic-peak amplitudes, limiting, eventually, the precision of MSP and SpecTime measures.

Figure. 3 (top) shows an example of the results for the estimation of Rd on a synthetic signal by the different matching functions, an optimal selection according to ΔRd and the actual value. All measures lead closely to the real values. The noisy nature of the estimations may result in perceived degradations after resynthesis. This is alleviated by applying simple mean filter smoothing. An example of this process is shown in the same figure (bottom) with the result of the estimation on a real sustained sung vowel. These examples correspond to the corpora used in our objetive and subjective evaluation, described in the following section.

5. EXPERIMENTS

We firstly carried out an objective evaluation on synthetic data due to the impossibility of accessing the actual source and filter information in real signals if only the acoustic signal is available.

5.1. Synthetic data

To build a synthetic corpus, representative VTFs were extracted after manual setting of the LF model parameters to cancel the source contribution in the spectrum. A VTF was computed over a selected segment of sustained sung vowels recorded individually in studio (samplerate: $F_s = 44100Hz$). The samples correspond to the five vowels of Japanese sung by 10 singers (four males, six females), resulting in fifty different VTFs.

These VTFs were used to synthesize short samples (1 second length) keeping the VTF unchanged in the synthesis filter. To generate the excitation sequence, a sinusoidal modulation (one cycle) was applied to the LF parameters (Rd, Ee, and the fundamental period t_0) seeking to reproduce a smooth variation of the glottal characteristics on the excitation. The average f_0 was set according to the original sample used for VTF extraction with a modulation of 2.5% over time. The mean Rd value was fixed to 1 with a modulation (opposite to f_0) of 30% to cover a resonable interval near the assumed modal range. Finally, the gain value Ee was set arbitrarily to 1 and modified according to Rd following the correlations reported in [6].

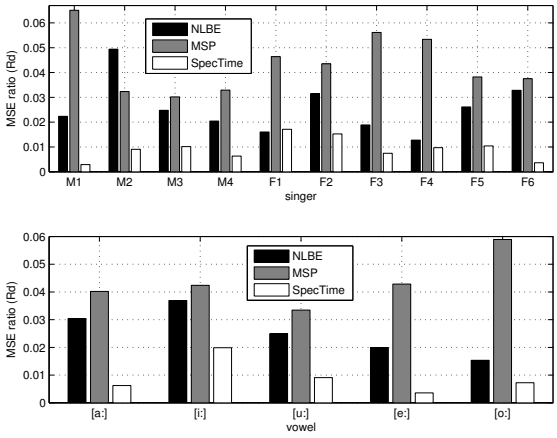

Fig. 5. Performance of the glottal source estimations with all matching measures accross the different singers and vowels.

The synthesis framework is based on the PSOLA technique [18] with a slight modification: zero padding of size $2t_0$ is applied to the derivative glottal signal to perform source-filter synthesis. The purpose of this is to include, to some extent, the damping characteristics of the synthesized voice epochs. The synthesis linear filter is set according to the autoregressive systems defined by the corresponding V_k. The synthetized waveforms are allocated in order to properly match the synthesis GCIs without applying any additional windowing in the overlap-add process. This strategy was found to provide natural resynthesis quality.

5.2. Objective evaluation on synthetic data

We evaluate both Rd and VTF estimation performance over the synthetic set. An evaluation is done, firstly, in terms of the matching cut-off frequency (number of harmonics I). Then, the number of harmonics was fixed and we looked into the different VTF cases (singers and vowels). For analysis, ΔRd is fixed to $\pm 5\%$ of the previous selected value and the actual GCI positions were kept. The Rd estimation performance was quantified by the normalized MSE between the actual and selected Rd values for all measures. The spectral distortion error between the original and the estimated VTFs was used as performance measure of the VTF estimation.

The results are shown in Fig.4 including, for comparsion, the case of optimal Rd selection (the closest to the actual value given ΔRd). SpecTime shows the best scores for Rd estimation and no dependency on the matching band. NLBE improves with increasing number of harmonics, showing slightly lower performance than SpecTime. MSP shows bigger overall errors, increasing with the size of the matching band. Similar trends are observed regarding the VTF estimation (bottom). Given the small overall error values ($\sim \pm 15\%$ for Rd, $\sim 1dB$ for VTF), the performance can be considered as adequate for source and filter approximation purposes.

Fig.5 shows the results per singer (top) and vowel (bottom) for

Rd estimation. The singers are ordered for increasing f_0 and labeled with M (male) or F (female). The error has no significant dependency on f_0. However, the performance is relatively degraded for the vowel [i:]. This is commonly attributed to the proximity of the first formant to the fundamental component (f_0).

5.3. Subjective evaluation on real data

Finally, experiments were conducted on real signals in order to study the perceived quality after source-filter resynthesis. The decomposition was applied to the five vowels of one of the singers of the corpora previously described. For simplicity, the estimates given by SpecTime were considered for evaluation according to the results of the objective evaluation. Three resynthesis cases: a) PSOLA (no source-filter resynthesis nor time-scale modification), b) source-filter resynthesis with estimated features and c) same as b) with smoothed parameters, were compared with the original recorded samples. A group of 20 professionals in audio were asked to evaluate the perceived quality in terms of the MOS scale (1= very degradated, 2=degradated, 3=degradations present, 4=slightly degradated, 5=clean). The purpose of including *transparent* PSOLA resynthesis is to discriminate degradations mainly due to distorted GCI estimates, computed with a strategy based on [19].

The results are shown in the table below (the standard deviation is included). Surprinsingly, the original excerpts were not always considered as fully natural/clean. A reason of this may be found in the difficulty to perceptually evaluate the naturalness of sustained voice in a short duration context.

Type	Original	PSOLA	Estimated	Smoothed
MOS	**4.3**$\pm$0.8	**3.6**$\pm$0.7	**2.7**$\pm$0.7	**3.6**$\pm$0.8

As expected, resynthesis with non-smoothed features shows the lowest scores. This is mainly due to degradations coming from jumps in the glottal model parameters between epochs. The scores obtained from resynthesized signals using smoothed parameters are similar to those of PSOLA synthesis, showing the convience of the simple smoothing strategy to avoid perceived degradations. This allow us to claim comparable resynthesis naturalness after following the proposed glottal excitation and vocal-tract filter estimation methodology.

6. CONCLUSIONS

We presented in this work a simplified strategy for source-filter estimation based on glottal-shape parameter modeling. The results of experimental studies on synthetic and real signals show adequate performance of the proposed methodology, showing natural resynthesis quality after simple optimization of the estimated parameters. Three different measures of the derivative glottal waveform similarity were compared, showing best results from the time and harmonic information based method (SpecTime).

Further investigation into latest improvements of the MSP measure and efficient subjective evaluation of sustained singing voice should be conducted. Informal experimentation showed promising results for Voice Transformation purposes. The definition of a whole transformation framework is currently under study by the author.

REFERENCES

[1] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

[2] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, 2011.

[3] J. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 195–208, 2014.

[4] G. Degottex, A. Röbel, and X. Rodet, "Joint estimate of shape and time-synchronization of a glottal source model by phase flatness," in *proc. of ICASSP*, Dallas, USA, 2010, pp. 5058–5061.

[5] J. Kane, I. Yanushevskaya, A. N. Chasaide, and C. Gobl, "Exploiting time and frequency domain measures for precise voice source parameterisation," in *proc. of Speech Prosody*, Shanghai, China, May 2012, pp. 143–146.

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

[7] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an arx model," in *Proc. of IEICE'95*, 1995.

[8] D. Vincent, O. Rosec, and Chon, "A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling," in *Proc of ICASSP'07*, 2007.

[9] H. Kenmochi and H. Oshita, "Vocaloid commercial singing synthesizer based on sample concatenation," in *Proc. of INTERSPEECH'07*, Antwerp, Belgium, 2007.

[10] N. Henrich, G. Sundin, D. Ambroise, M. d'Alessandro, C. Castellengo, and B. Doval, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *Journal of Voice*, vol. 17, 2003.

[11] F. Villavicencio, "Glottal source model selection for stationary singing-voice by low-band envelope matching," in *Advances in Nonlinear Speech Processing*. 2013, vol. 7911, Elsevier.

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

[13] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication (in Japanese)*, vol. 62, no. 4, pp. 10–17, 1979.

[14] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[15] F. Villavicencio, A. Röbel, and X. Rodet, "Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation," in *proc. of ICASSP*, 2006.

[16] H. Hermansky, H. Fujisaki, and Y. Sato, "Spectral envelope sampling and interpolation in linear predictive analysis of speech," in *Proc. of ICASSP '84*, 1984.

[17] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *proc. of DAFx*, Spain, 2005.

[18] H. Valbret, E. Moulines, and Tubach J.P., "Voice transformation using psola technique," in *Proc. of ICASSP '92*, 1992, vol. 1, pp. 145–146.

[19] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: a quantitative review," *IEEE TASLP*, 2012.