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, referred 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].

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 $R_{p_a}$, $R_{p_b}$, $R_k$ 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 $0.3 < Rd < 2.7$ 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

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, referred 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].

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 $R_{p_a}$, $R_{p_b}$, $R_k$ 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 $0.3 < Rd < 2.7$ 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
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 criterion 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, $\Omega_{TE}$, can be set according to the fundamental frequency such as $\Omega_{TE} = F_S/2f_0$ 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 $\hat{g}$ of the derivative glottal waveform for $s_k$ we compute its spectral representation in the form:

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

where $\hat{G}$ denotes the DFT of $\hat{g}$ 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:

$$S_v = \frac{S_k}{10(\log(30))},$$  

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

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

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

$$g = V^{-1} \ast s_k.$$  

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 present a difference larger than an assumed $\Delta Rd$.

- **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 commonly observed among modal singing, in particular, in the corpora of interest described in the introductory section.
3.2. Estimation framework

The analysis framework is based on the assumption that the vocal-tract configuration does not change between successive 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 \( \tilde{g}_c \) candidates derived from the set of \( Rd \) values:

\[
Rd_C = [Rd_{k−1} − \Delta Rd, Rd_{k−1}, Rd_{k−1} + \Delta Rd], \tag{5}
\]

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 \( \Delta Rd \) 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 \( \tilde{g}_c \) according to Eq.2 and Eq.3. The procedure is repeated for the successive 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 \( \tilde{g}_c \) 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 as \( s'_c \) is obtained for each candidate \( \tilde{g}_c \) given \( V \). The MSP error computation is described as follows:

\[
s'_c = V * \tilde{g}_c, \tag{6}
\]

\[
R_c = s_k / S'_c, \tag{7}
\]

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

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:

\[
e_s = \{0.5 − |cor(\log |G_{p(i)}|, \log |\tilde{G}_{p(i)}|)| \} \cdot w_s, \tag{9}
\]

where \( G \) and \( \tilde{G} \) represent the DFT versions of \( g \) and \( \tilde{g} \), 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:

\[
e_t = \{0.5 − |cor(g, \tilde{g}_c)| \} \cdot w_t. \tag{10}
\]

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

\[
e_{st,c} = e_s + e_t. \tag{11}
\]
4.3. Normalized low-band envelope matching (NLBE)

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

$$e_{nlbe,c} = \frac{1}{L} \sum_{f=f_0}^{f \cdot f_0} (E_{g,f} - [E_{\hat{g},f} + G_g])^2,$$

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 $G_g$ denotes the energy bias between the envelopes computed as:

$$G_g = \frac{1}{L} \sum_{f=f_0}^{f = f_0} |E_{g,f} - E_{\hat{g},f}|.$$

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

$$\hat{E}_{e,c} = 10^{(G_g/20)}.$$

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 values 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 $\Delta Rd$ 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 objective 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 (sample rate: $F_s = 44100Hz$). 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].

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 synthesized 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, $\Delta Rd$ 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 comparison, the case of optimal $Rd$ selection (the closest to the actual value given $\Delta Rd$). 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
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 degraded, 2=degraded, 3=degradations present, 4=slightly degraded, 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). Surprisingly, 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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Original</th>
<th>PSOLA</th>
<th>Estimated</th>
<th>Smoothed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>4.3 ± 0.8</td>
<td>3.6 ± 0.7</td>
<td>2.7 ± 0.7</td>
<td>3.6 ± 0.8</td>
</tr>
</tbody>
</table>

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