Glottal Source Model Selection for Stationary Singing-Voice by Low-Band Envelope Matching

Fernando Villavicencio

Yamaha Corporation, Corporate Research & Development Center, 203 Matsunokijima, Iwata, Shizuoka, Japan

Abstract. In this paper a preliminary study on voice excitation modeling by single glottal shape parameter selection is presented. A strategy for direct model selection by matching derivative glottal source estimates with LF-based candidates driven by the Rd parameter is explored by means of two state-of-the-art similarity measures and a novel one considering spectral envelope information. An experimental study on synthetic singing-voice was carried out aiming to compare the performance of the different measures and to observe potential relations with respect to different voice characteristics (e.g. vocal effort, pitch range, amount of aperiodicities and aspiration noise). The results of this study allow us to claim competitive performance of the proposed strategy and suggest us preferable source modeling conditions for stationary singing-voice.

1 Introduction

The transformation of voice source characteristics represents a challenge of major interest in terms of expressive speech synthesis and voice quality control. A main task to achieve transformation is found in the modeling of the excitation (source) characteristics of the voice. However, a robust decomposition of the source and filter contributions represents a major challenge due to exisiting non-linear interactions limiting the robustness of an inverse filtering process.

Some works propose iterative and deterministic methods for voice decomposition such as [1] and [2] respectively. A recent strategy consists of approximating the glottal contribution by exhaustive search using the well-known LF model [3], [4]. Although the different techniques show promising results the performance is commonly sensitive to aspects of the voice that may significantly vary in continuous speech among individuals (e.g. first formant position, voice quality, voicing).

We aim to perform voice excitation modeling as an initial stage for future voice quality modification purposes on stationary singing-voice samples used for concatenative singing-voice synthesis. The controlled recording conditions (vocal effort, pitch, energy) of such signals allow us to delimit the analysis context of the main glottal source characteristics and to derive a simplified strategy to model them by selecting an approximative model.

Our study follows the works of [3] and [4] proposing derivative glottal signal modeling by selecting Liljencrants-Fant (LF) based models issued from a set of

T. Drugman and T. Dutoit (Eds.): NOLISP 2013, LNAI 7911, pp. 42–49, 2013.

glottal shape parameter (Rd) candidates. Furthermore, we propose a novel selection measure based on accurate spectral envelope information. This strategy, refered to as "normalized low-band envelope" (NLBE) is compared with the measures proposed in the referenced works based on phase and joint frequency-time information. An experimental study over a set of synthetic signals emulating the target singing samples was carried out seeking to observe the main relations between the signal's characteristics and the performance provided by the different selection measures.

This paper is structured as follows. In section 2 the proposed NLBE estimation is introduced. The synthetic data used for objective evaluation based on stationary singing-voice is described in section 3. In section 4 the results of the experimental study are reported. The paper ends at section 5 with conclusions and future work.

2 NLBE Glottal Source Model Selection

2.1 Rd Based Voice Quality Modeling

The Rd parameter allows us to quantify the characteristic trends of the LF model parameters (Ra, Rk, Rg) ranging from a tight, adducted vocal phonation ($Rd \approx 0.3$) to a very breathy abducted one ($Rd \geq 2.7$) [5]. Three main voice qualities are distinguished along this range: *pressed*, *modal* (or normal) and *breathy*. In [6] 0.8, 1.1 and 2.9 were found as approximative values of Rd for these voice qualities on baritono sung vowels. Similarly, our interest is focused on stationary singing preferably sung with modal voice. Accordingly, it can be expected that Rd estimates on the underlying glottal excitation are found close to the mentioned modal value keeping a slow and narrow variation over time. This principle was therefore considered in order to derive the glottal-model selection strategy described in the next section.

2.2 Normalized Low-Band Envelope Based Rd Estimation

One of the main features of the progress of the Rd parameter on the LF model is the variation of the spectral tilt of the resulting derivative glottal signal spectrum. Low Rd values produce flat-like spectra whereas higher ones show increasing slopes. Moreover, the low-frequency voiced band of voice source spectra is mainly explained by the glottal pulse contribution and studies have shown the importance of the difference between the two first harmonics ($H1 - H2$) as one of the main indicators of variations on its characteristics [7].

We propose to measure the similiarity between Rd-modeled derivative glottal candidates and extracted ones by comparing their spectral envelope within a low-frequency band after normalization of the average energy. The spectral envelope is estimated pitch synchronous in a narrow-band basis (4 pulses) centered at the glottal closure instant. The envelope model correspond to the one described in [8] seeking to use accurate envelope information. Note that by following this strategy

we aim to approximate the main glottal characteristics within a small *Rd* range rather than estimate accurate *Rd* values. Moreover, assuming a smooth variation of the vocal phonation a simple candidates selection is proposed by exclusively considering a small deviation of *Rd* between succesive epochs.

The method is described as follows. Let $S(f)$ be the narrow band spectrum of the speech frame $s(t)$ (4 periods) and $A_{vt}(f)$ the system representing its corresponding vocal tract transfer function. As usual, the derivative glottal signal $dg_e(t)$ is extracted by analysis filtering according to

$$DG_e(f) = S(f)/A_vt(f) \tag{1}$$

Following, a *Rd* candidate is used to generate an excitation sequence $dg_{rd}(t)$ of same length (*Rd* fixed, gain $Ee = 1$). The spectral envelopes $Edg_e(f)$ and $Edg_{rd}(f)$ are estimated from $dg_e(t)$ and $dg_{rd}(t)$ respectively using *optimal* True-Envelope estimation [8] in order to observe accurate $H1 - H2$ information.

The matching is limited to the low-band defined within the range $[f0, Mf0]$, where M represents a number of harmonics fully considered as voiced. The normalization gain G_{dB} is computed as the difference between the average energy of $Eg_e(f)$ and $Eg_{rd}(f)$ within the mentioned low-frequency band

$$G_{dB} = \frac{1}{K}\sum_{f=f0}^{Mf0} Edg_e(f) - \frac{1}{K}\sum_{f=f0}^{Mf0} Edg_{rd}(f) \tag{2}$$

note that G_{dB} represents an estimation for $dg_{rd}(t)$ of the actual gain Ee . The matching error is defined as the mean square error between the envelope of the extracted excitation and the one of the normalized *Rd* model, according to

$$Error_{nlbe} = \frac{1}{K}\sum_{f=f0}^{Mf0} (Eg_e(f) - [Eg_{rd}(f) + G_{dB}])^2 \tag{3}$$

where K represents the number of frequency bins within $[f0, Mf0]$. The corresponding Rd_{nlbe} value for $s(t)$ is selected following the candidate observing the smallest error.

For comparison, the *Mean Squared Phase* (MSP) measure described in [3] and the joint spectral-time cost function proposed by [4] (labeled as *SpecTime*) were also used as selection cost measures. Note that the harmonic phase information for MSP compuation was obtained from the closest DFT bin to the harmonic frequencies and that the DFT size N was set equal to the frame length. We note that a potential lack of precision of the harmonic information given the DFT size may limit the performance of the MSP and *SpecTime* measures.

3 Synthetic Data

3.1 "Emulating" Stationary Singing-Voice Samples

The synthetic data consist of short units (1 sec length) aiming to emulate stationary singing samples of individual vowels. To generate the LF-based pulses

sequence a small sinusoidal modulation (5% of maximal deviation) over time was applied around the central values of $f0$ and Rd selected for test seeking to reproduce a smooth variation of the glottal excitation. The modulation of Ee was derived from that of Rd (double negative variation). These criteria follow the basic correlations between these features mentioned in [5]. An example of the resulting parameters evolution used for synthesis is shown in Figure 1.

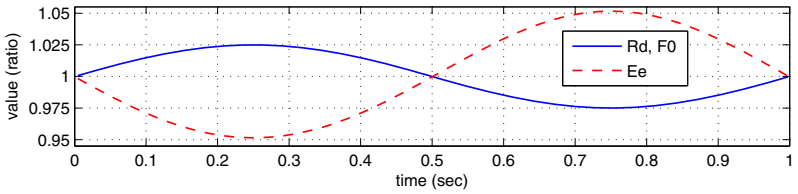

Fig. 1. Evolution of the sythesis LF parameters normalized by their average value

The vocal tract function (VTF) correspond to a *True-envelope* all-pole system [9] estimated after manual LF modeling on central segments of 5 stationary sung vowels of 6 singers (3 males, 3 females), resulting in 30 different VTFs of varying order ($[83 - 170]$). The original VTF information was kept unchanged for both synthesis and extraction purposes in order to exclusively compare the selection performance of the different measures. The *aspiration* (excitation) noise corresponds to the two-components modulated white-noise model proposed in [6]. The synthetic signals were generated by convolution of the filter and source parts after the sumation of the LF and noise contributions in an overlapp-add basis. Note that given the large filter orders it was applied a zero-padding to the source frames of the same frame length in order to ensure a reasonable underdamping on the synthesized waveforms. The samplerate was fixed to $44.1KHz$.

3.2 Aperiodicities Synthesis

Beyond the degree of aspiration noise other common excitation phenomena are $T0$ *aperiodicities* in the form of pitch and energy frame-to-frame variations (known commonly as *jitter* and *shimmer* respectively). The characteristic of these variations is random with reported maximal values of $\approx 2\%$ in pathological voices (e.g. *harsh, hoarse*) [10]. Although these phenomena mainly concerns non-modal voice it may be found in intended "modal" phonations of individuals with voices observing some natural degree of *roughness*. Following, shimmer and jitter were also considered in the synthesis framework and applied jointly as random frame-by-frame variations of Ee and $f0$.

3.3 Experiments

We aimed to evaluate the proposed modeling strategy using the different selection measures on a data set including varied filter and excitation characteristics.

Accordingly, 3 different pitch ranges corresponding to the musical notes $A2$, $G3$ and $F4$ $(110, 196, 350Hz)$ were considered to build the synthetic data seeking to explore a reasonable singing range. Moreover, several Rd ranges and amounts of aspiration noise and aperiodicities arbitrarly selected were also considered, resulting in about 750 different test signals.

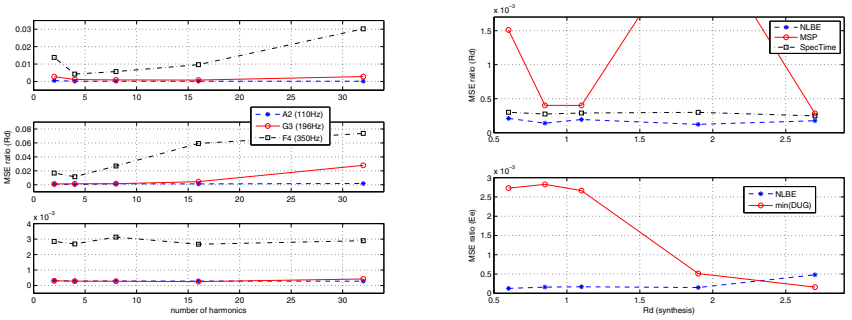

Fig. 2. Rd selection performance as a function of the low-band length (for matching) and the pitch range on a modal region ($Rd = 1.1$) for NLBE (left, top), MSP (left, middle) and *SpecTime* (left, bottom) selection cost measures. Evaluation over a set of Rd values (top, right) and estimation of the LF gain parameter Ee (bottom, right).

4 Results

The set of candidates Rd_c tested for Rd selection at each voice epoch consisted of the previous selected value and 2 neighbouring ones limited to a potential deviation Rd_{step} (arbitrarily set to 2.5%). We used this criterion instead of a fixed Rd step due to the non-linear evolution of the spectral envelope gaps observed along the Rd scale. The selection performance was quantified by means of the MSE ratio (normalized error) between the actual and selected Rd values according to the NLBE, MSP and *SpecTime* cost functions. Two Ee estimation strategies were also compared and evaluated similarly: a proposed one using the gain parameter G_{dB} of NLBE and the standard strategy consisting on a direct computation from the negative peak of the derivative glottal signal, labeled as min(DUG).

4.1 Effect of the Low-Band Length and the Rd Range

We were firstly interested to observe the performance on signals corresponding to a modal range ($Rd = 1.1$) in terms of the low-band length (number of harmonics) considered on the cost functions and the effect of the pitch range. The results are shown in Fig. 2 (left), note that for clarity a different axis scaling was applied on the plots. As expected, it can be seen the negative effect of increasing pitch on the Rd identification performance. A smaller fundamental period may represent

a larger overlapping between pulses, and therefore, a larger mixing of the spectral information. In general, it was found that by using 4 harmonics it was already possible to achieve the lower error regions accross the different measures. NLBE provided the lowest average error on low-pitched data although all methods showed comparable performance and stability ($NLBE = 2.0e-4$, $MSP = 4.6e-4$, $SpecTime = 2.9e-4$). Accordingly, aiming to focus on preferable modeling conditions only the low-pitch ($A2$) data set and the 4 harmonics limit as low-band criterion were kept for the following experiments.

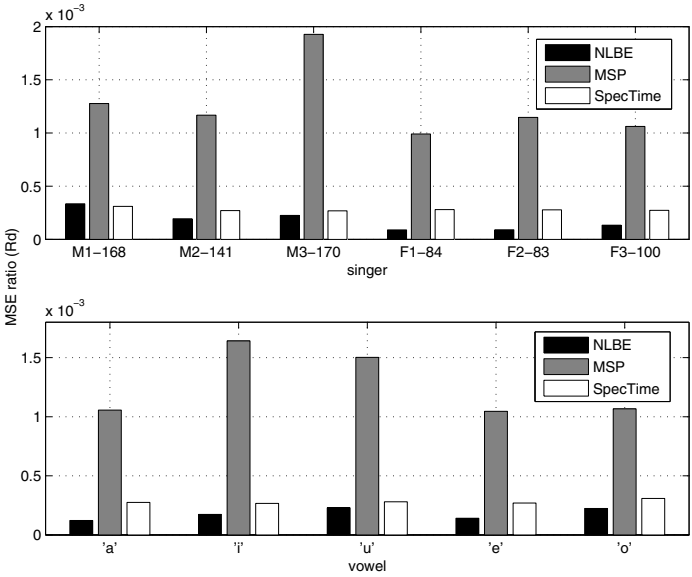

Fig. 3. *Rd* selection performance per singer (top) and vowel (bottom) case for synthetic data covering different *Rd* regions and single pitch range ($A2$)

In Fig. 2 (right) are also shown the results when using several *Rd* ranges on the synthetic signals. There was not a significant effect of the glottal shape (*Rd* range) on the selection performance besides an irregular evolution on the MSP selection (some values on the plot are out of range). NLBE and *SpecTime* showed higher and more stable performance. Concerning *Ee* estimation (bottom), there was some dependency of the direct computation with respect to *Rd*, showing maximal errors of about $\approx 5\%$ of the parameter value on low *Rd* signals.

4.2 Effect of the VTF Charateristics

Figure 3 shows the results of the previous experiment per singer (top) and vowel (bottom) case. The scores suggest some dependency of the performance accross

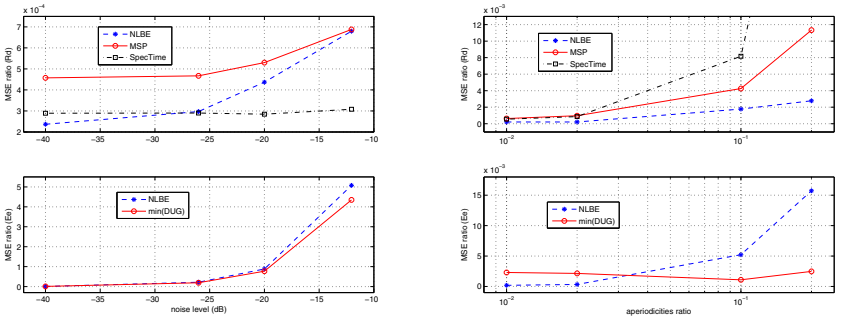

Fig. 4. *Rd* selection and *Ee* estimation performance as a function of the amount of aspiration noise (left) and $T0$ aperiodicities (right) on a modal region ($Rd = 1.1$)

the different VTFs. We claim this might be explained not only by differences on the low-frequency features but also by the filter order differences (specified at each singer label) that may affect the waveform underdamping length and thus, the amount of overlapping between waveforms. Note the lower performance of MSP among all filter cases. It was already mentioned that our short DFT size criterion may limit the precision of the phase information required by MSP.

4.3 Effect of Aspiration Noise and Aperiodicities

An increasing level of noise on the excitation reduces the maximal voiced frequency affecting, eventually, the glottal information. Figure 4 (left) shows the results for different amounts of *aspiration* noise added to the LF component before the synthesis convolution. As expected, there was a significant drop in the performance at important noise levels in most of the results excepting a surprising stability showed by the *Rd* selection from *SpecTime*. The results confirm the difficulties of modeling *aspirated* and *breathy* voices. Note however that reasonable scores could be keep until moderate amounts of noise ($\leq -25dB$).

Conversely, *SpecTime* was the most sensitive measure with respect to $T0$ aperiodicities, as shown in Figure 4 (right). The aperiodicities scale denotes the maximal deviation percentage related to the mean values of Ee and $f0$ applied frame-by-frame. In general, the drop in the performance might be explained by the degradation of the harmonic structure at the low-band due to the random variations of energy y frequency applied to the fundamental component. NLBE shows the best performane, however, all results, including Ee estimation seem to be robust enough to cover aperiodicities amounts reaching the mentioned levels of pathological voices ($\leq 2\%$). The results above this value might be mainly relevant to study some extreme vocal phonation cases.

5 Conclusions and Future Work

This paper presented an experimental comparison of methods for glottal model selection on a large synthetic set of stationary singing signals. The results showed evidence that a proposed selection strategy based on low-frequency spectral envelope matching provides comparable estimation performance to recent techniques based on phase, amplitude and time-domain information.

The experiments showed relations between different voice characteristics and the glottal selection performance, suggesting preferable source modeling conditions. Furthermore, studies should be done to extend the study to real singing-voice. The author is currently studying the perfomance of the overall direct glottal modeling strategy in a joint source-filter estimation framework.

References

1. Alku, P.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication 11, 109–118 (1992)
2. Drugman, T., Bozkurt, B., Dutoit, T.: Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. Speech Communication 53, 855–866 (2011)
3. Degottex, G., Röbel, A., Rodet, X.: Joint estimate of shape and time-synchronization of a glottal source model by phase flatness. In: Proc. of ICASSP, Dallas, USA, pp. 5058–5061 (2010)
4. Kane, J., Yanushevskaya, I., Chasaide, A.N., Gobl, C.: Exploiting time and frequency domain measures for precise voice source parameterisation. In: Proc. of Speech Prosody, Shanghai, China, pp. 143–146 (May 2012)
5. Fant, G.: The lf-model revisited. transformations and frequency domain analysis. STL-QPSR Journal 36(2-3), 119–156 (1995)
6. Lu, H.-L.: Toward a High-Quality Singing-Voice Synthesizer with Vocal Texture Control, Ph.D. thesis, Stanford University (2002)
7. Henrich, N.: Etude de la source glottique en voix parlée et chantée, Ph.d. thesis, Université Paris 6, France (2001)
8. Röbel, A., Rodet, X.: Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation. In: Proc. of DAFx, Spain (2005)
9. Villavicencio, F., Röbel, A., Rodet, X.: Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation. In: Proc. of ICASSP (2006)
10. Kreiman, J., Gerratt, B.R.: Perception of aperiodicity in pathological voice. Journal of the Acoustical Society of America 117, 2201–2211 (2005)