ANALYSIS AND MODIFICATION OF EXCITATION SOURCE CHARACTERISTICS FOR SINGING VOICE SYNTHESIS

A. Roebel, S. Huber, X.Rodet

IRCAM-CNRS-UPMC STMS
1, pl. Igor-Stravinsky, 75004 Paris

G. Degottex

Computer Science Department, University of Crete,
71409 Heraklion, Greece

ABSTRACT

The present article investigates into the use of the LF glottal pulse model for singing synthesis and transformation. A recent estimator of the LF-glottal pulse shape parameter (rd) is used to analyze a small collection of professional singing examples and the results are discussed in the context of recent findings relating the rd shape parameter to other speech signal parameters (intensity and vibrato). We propose a rd shape parameter model for vibrato rendering and present an algorithm that allows modifying the glottal pulse shape parameter of a given speech signal and is used to enhance the vibrato generation in a speech to singing transformation system.

Index Terms— Glottal pulse parameter estimation, singing synthesis, speech transformation.

1. INTRODUCTION

The estimation of the glottal source parameters is an active research area [1, 2, 3, 4] that lays the ground for new synthesis and signal transformation techniques. The goal of the present study is to investigate whether recent glottal pulse parameter estimation and transformation techniques can be used to enhance the performance of algorithms for synthesis and transformation of singing voice. The long term objective of this research are algorithms that allow the modification of the characteristics of a singing signal or allow transforming speech into singing [5]. For this additional means for manipulation of the vocal tract transfer function (VTF) will have to be established, but these are not covered in the following.

An algorithm that allows the control of the source characteristics of a given speech or singing signal requires the following steps First, an estimate of the properties of the existing source characteristics has to be available. Second, the target parameter contour of the excitation source characteristics has to be created, and third the signal has to be transformed such that the desired characteristics of the excitation signal are respected.

Numerous algorithms have been proposed that allow estimating the glottal pulse component from a given speech signal [6, 7, 8, 1, 2, 3, 4]. If signal transformation is desired a parametric representation of the glottal waveform is advantageous because it allows controlling the properties of the excitation by means of parameters. Any algorithm that provides glottal pulse excitation parameters can in principle be used. In the present study we will use the algorithm presented in [4] for the analysis of the glottal pulse parameters.

Many studies have investigated into properties of the glottal source for the singing voice [9, 10, 11]. Still there does not exist a complete model that would allow to control the different speech parameters (f0, intensity, glottal pulse parameters, formant positions) in a coherent manner. This is partly due to the fact that the laryngeal mechanism that are used to produce the speech can not be deduced from the signal. Moreover, there are many individual strategies to control the voice production that are employed by different singers. We will propose a simple strategy that allows to create artificial rd contours for the intended application.

The transformation of excitation pulse characteristics is a the central reasons for the development of pulse parameter estimators. The use of a dedicated source model promises extended transformation controls as well as a higher degree of naturalness for transformations changing pitch [8, 12, 13, 14]. In the present study we will use the shape invariant phase vocoder [15] for signal modification.

The article is organized as follows. In section **2** we introduce the speech signal model and in section **3** we will discuss the glottal pulse parameter estimator to be used. In section **4** we discuss some recent findings related to pulse parameters for singing and relate these findings with results obtained with the rd estimator. In section **5** we introduce the signal transformation model and discuss different strategies for the transformation of the rd parameter. In section **6** we will present some of the results obtained with the proposed transformation algorithm in a speech to singing transformation system.

2. VOICE PRODUCTION MODEL

The voice production model for singing voice that will be used in the following is an extended source filter model given by

$$S(w) = (G_{rd}(w) + N(w))C(w)L(w)H(w, f0, D). \quad (1)$$

It consists of a representation of the Liljencrants-Fant (LF) glottal pulse model ($G_{rd}(w)$) that is parameterized by the shape parameter rd described in [16], the VTF denoted as $C(w)$ assumed to be minimum phase, an approximate representation of the lips radiation following $L(w) = jw$, the glottal noise with low pass characteristic $N(w) = \frac{F_{VU}^2}{(jw - F_{VU}(n))(jw + F_{VU}(n))}$ that is controlled by the time dependent voiced/unvoiced frequency boundary (VUF) represented by $F_{VU}(n)$ in rad. $H(w, f0, D)$ represents the harmonic structure parametrized with the fundamental frequency $f0$ and the delay between pulse sequence and frame center in terms of the phase delay D of the fundamental.

3. GLOTTAL PULSE PARAMETER ANALYSIS

In the following we shortly describe the pulse parameter estimator algorithm that has been described fully in [4]. Due to space constraints we discuss on the following only some details that differ from the original paper, notably the algorithm toi be used for the estimation of the sinusoidal parameters.

The algorithm constructs a sinusoidal model of a frame of the speech signal under investigation. This sinusoidal model is transformed into a harmonic model describing a single pitch period using a sample-rate that falls onto the harmonic grid at position $2K + 2$. The model is assumed to be noise free up to harmonic K. According to eq. (1) the voice production model can then be simplified into

$$S(k) = G_{rd}(k)C(k)L(k)e^{jkD} \quad k \in [0, 1, \ldots, K]. \quad (2)$$

We construct $S(k)$ from a signal frame by means of finding the parameter set having minimum error using the procedure described in [17]. The detection of voiced and unvoiced frames follows the description in [15], but the method proposed in [17] could have been used as well. For the estimation of the fundamental frequency we follow [4]. The algorithm proceeds by means of testing the minimum phase property of the VTF spectrum that is obtained for a sufficiently dense grid of rd values by means of

$$\hat{C}_{rd}(k) = \frac{S_k}{G_{rd}(k) \cdot jk}. \quad (3)$$

For the correct rd parameter the residual will represent the minimum phase transfer function VTF and the algorithm selects the rd by means of minimizing the deviation of the residual from a minimum phase transfer function taking into account an arbitrary delay D. The simplification of $L(k)$ into jk introduces as error an additional constant factor that does not affect the results. Note, that the evaluation of the minimum phase property of the VTF becomes problematic if the duration of the impulse response of the VTF is close to or above the period. In these cases ambiguous solutions arise that we avoid by means of manually constraining the RD solution space. Other sources of ambiguous solutions are the fact that the LF model does not fit the real excitation source or the non stationarity of the VTF. Note, that the non-stationarity of the pitch that is encountered in real world speech or singing signals has a comparatively low impact on the rd estimation. We are currently investigating into different means that enhance the robustness of the analysis and will publish our findings elsewhere.

4. GLOTTAL PULSE PARAMETER MODEL

The glottal pulse parameters of the singing voice have been studied by many authors. Recent studies that are particularly relevant for the present discussion are [9, 8, 10, 11]. The first two studies have investigated into the open quotient (Oq) of the source excitation signal, using the derivative of electroglottographic (EGG) signals. The parameter Oq describes the part of the period during which the glottal folds are open. The study concluded that the Oq depends mainly on the laryngeal mechanism that is used to produce the sound, the pitch, and the vocal intensity. It has been reported that for the laryngeal mechanism 1 (normal voice) Oq was found to be in the range 0.3 to 0.8, while for mechanism 2 (falsetto or head register) the range was slightly larger within 0.5 to 0.95. With respect to the vocal intensity the study reports that a negative correlation between intensity and Oq is common for mechanism 1, but rare for mechanism 2. On the contrary the dependency on the pitch seems limited to the laryngeal mechanism 2.

The authors of [11] investigate into the relation of the glottal pulse parameters and vibrato. The result of this study supports the idea that the singing signal spectrum can be modeled precisely without taking into account variations of the Oq during vibrato cycles. However the detailed study in [9] reports that Oq was modulated for vibrato singing for most of the singers that were investigated. The Oq and F0 contours are reported to be in opposite phase. In [8] the rd parameter is used to control voice quality of the singing voice. Moreover a nonlinear function is established that links the glottal flow derivative at the glottal closure instance E_c to the rd parameter. In [16] the rd parameter is related to the vocal effort and an example is given where the rd parameter scales with the maximum of the glottal flow U_0 that is itself related to the amplitude of the fundamental.

To summarize the findings we conclude that the rd parameter changes significantly between singers. We cannot expect to construct a model that covers the behavior of the rd parameter in all situations. Nevertheless, given the relation between vocal effort and rd [16] it seems plausible that the rd parameter is related to the gesture that configures the vocal tract and the glottis. We base our rd transformation model on the hypothesis that the rd parameter will change with the signal intensity as well as with the vibrato extent because both are related to the vocal effort. The reduction of the rd parameter should be accompanied by an increase in intensity and therefore a means to modify the intensity in a coherent manner with the rd parameter contour is required.

Fixing the negative excitation amplitude of the glottal pulse derivative at glottal closure instant (E_c) does not provide a convincing energy modification because the energy of the fundamental can decrease strongly with decreasing rd [16, Fig. 10]. Experimental investigation suggested that the strong decrease of the amplitude of the fundamental is not observed in real sounds. The nonlinear function linking E_c with rd in [8] requires knowledge about the absolut energy level of the speech signal and is therefore difficult to use for signal transformation. We tried to normalize the amplitude of the fundamental of the excitation pulse such that changing the rd parameter keeps the amplitude of the fundamental constant ($U_o == const$ in [16, Fig. 11]). In that case the signal energy did grow too rapidly with decreasing rd. By means of normalizing the amplitude of the fundamental to Oq^2 the increase in energy was sufficiently compensated to create a perceptually stable result and the energy modification that resulted from glottal pulse modification appeared quite natural in all cases we tried and it was therefore used for the following tests. In the context of adding vibrato to stationary singing signals we propose to reduce the rd parameter during the vibrato section to introduce a notion of vocal effort. The implementation that provided rather convincing results derives the rd contour from the contour describing the vibrato extent such that rd and vibrato gesture share a common movement. Note, that we do not take into account the modulation of the rd with the vibrato, because the variation seems to be too small to be perceptually important [18]. Further investigation is required whether adding this rd modulation is perceptually important.

To validate the results obtained with the proposed rd estimation algorithm a small collection of singing excerpts has been analyzed and the results have been compared to the findings mentioned above. The analysis window size covering a signal frame is adapted to the fundamental frequency such that the window always contains 4 periods. The rd estimates have been obtained by means of evaluating the minimum phase property of $\hat{C}_{rd}(k)$ using 7 to 22 partials. The difference related to the number of partials used for evaluation was found to be negligible. For the calculation of the minimum phase spectra an harmonic model with $K = 32$ partials is established. Note, that by means of using the fact that the minimum phase signal is constructed by a windowing operation in the cepstral domain, we can determine the error we have to accept due to the fact that the harmonic spectrum is not covering the complete band. The DFT transform of the window that is applied in the cepstral domain is relatively narrow, and therefore, the error we make in using only 32 partials seems acceptable.

We demonstrate an example of the rd estimation results using a tenor signal with vibrato. The signal waveform, its fundamental frequency and the estimated rd parameter are shown in fig. 1. The rd strongly decreases from its start value to the range 0.6 to 0.8 during the first few 100ms of each note onset. This behavior is consistent with the expected increase in vocal effort. The rd changes in phase opposition with the F0 during the vibrato cycles. This observation is consistent with the findings described in [9]. The rd parameter has a glitch around 4.2s. These kind of glitches are related to the ambiguities of the estimator mentioned in section **3**.

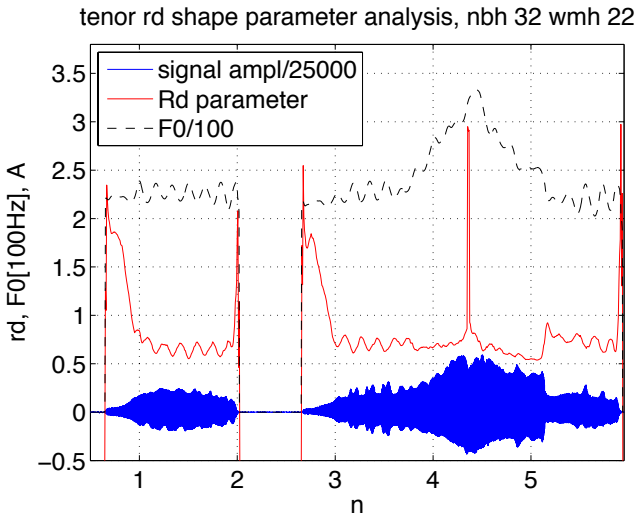

Fig. 1. Rd shape parameter analysis for a tenor signal with vibrato.

5. GLOTTAL PULSE TRANSFORMATION

The transformation of the glottal pulse parameters has been integrated into the shape invariant phase vocoder presented in [15]. 3 parameters are needed to segregate the deterministic and stochastic excitation components from the speech signal: The rd shape parameter, the VUF and the F0. We use the approach described in [14] and will briefly summarize the main ideas. For more details we refer to the original paper. The rd and $F0$ parameters determine the spectral shape of the deterministic part of the excitation signal. From the VUF we derive the noise level $(N(w)L(w))$ including the lips radiation from the derivative of the glottal pulse spectrum. Due to the low pass characteristics of the noise level including the lips radiation is assumed to be constant and equal to $(L(F_{VU})G_{rd}(F_{VU}))$ above the VUF. Following our signal model eq. (1), the division of the signal spectrum by the excitation spectrum leaves us with a spectrum that contains a white excitation with the VTF. The VTF will be estimated from this spectrum using the true envelope estimator [19, 14]. In contrast to [14] the model described so far will not be used to re-synthesize a synthetic signal but to transform the input signal. The experimental evaluation is performed using a system for speech to singing transformation. The following extensions of the standard approach using transposition with spectral envelope preservation have been investigated.

Pitch modification with VTF preservation: Transposition with spectral envelope (formant) preservation is expected to be improved when the excitation source is treated separately from the VTF. To achieve this the VTF is determined as described above, and instead of preserving the spectral envelope, now the VTF is preserved.

Modification of pulse parameters: Recorded speech signals often have a relatively high rd value (in our example $rd \in [2-4]$). Moreover, the central segments of the syllables have to be time stretched significantly to create the target notes. After time stretching these segments the resulting notes are somewhat static, they don't contain any expressive modulation. A first solution for this problem was to add vibrato to these segments. The typical vibrato has a maximum extent of 50cents and a vibrato rate of 5Hz. The vibrato extent follows a sinusoidal contour covering exactly one half of a period of the sinusoid. To improve over this existing approach a means to add an evolution of the rd parameter was intended.

To achieve the transformation of the glottal pulse parameters a transformed excitation spectrum is derived from the desired rd parameter and the estimated noise level N. The combination of the pulse and noise spectrum predicts a modification of the VUF, but in the present implementation this change is not yet taken into account. The signal spectrum is then modified by means of inverse filtering the original excitation spectrum and filtering with the new one. The time dependent filtering is performed in spectral domain.

6. EXPERIMENTAL RESULTS

We experimentally evaluated the 2 methods based on the glottal pulse parameters in the context of a system for speech to singing transformation. The system uses recorded speech signals that are annotated semiautomatically describing syllables, phonemes and pauses. All vowels are segmented further into three sub segments containing a start phase, the stable center, and the end phase. For real speech there is not always a stable center, but the annotation tries to find the segment that is most appropriate. A target melody is created and syllables are assigned to notes. For each syllable there is one center segment selected that is used to create the final note. This segment is stretched to ensure proper note length and receives vibrato and optionally rd modification. All other segments of the syllable will be transposed but not time stretched, to preserve as much as possible the realism of onsets and plosives. Note changes in sequences of connected vowels are handled using a strategy similar to [5]. Transposition required to achieve the notes sequences were within the range -500 to 1200 cents and time stretching of the stable centers into notes requires factors in the range from 4 to 20. For comparison a baseline system is used that consists of the shape invariant phase vocoder with spectral envelope preservation using the true envelope estimator but without the excitation model. The results have been evaluated subjectively by 4 expert listeners with professional background in music. We summarize their comments here below. Sound examples are available here [20].

Pitch modification with VTF preservation: The subjective evaluation confirms that for the strong transposition that is required within a speech to singing system the use of a separate source model increased the naturalness of the transformed note. The improvement was qualified as noticeable but not dramatic.

Modification of pulse parameters: We replaced the rd contour during the vibrato segments by means of a sinusoidal contour reproducing the same half period of a sinusoid that was used as well for the vibrato extent. The peak reduction of the rd parameter was either 0.6 or 1. The listeners were given three results for comparison. The results obtained using the *pitch modification with VTF preservation* as well as the two results using modified rd contours. The listeners were asked to produce two different rankings of the results using two different criteria. A first ranking was considering signal quality related to artefacts, and a second ranking considering expressivity. The rankings provided by the 4 listeners are listed in table (1)

Ranking of expressivity				Ranking of artefacts			
Pers Pers	orig rd	rd -0.6	rd -1	Pers Pers	orig rd	rd -0.6	rd -1
A	3	2	1	A	1	2	3
B	3	2	1	B	1	2	2
C	3	1	2	C	2	1	3
D	3	2	1	D	1	2	3

Table 1. Results of comparing the three algorithms by 4 different expert listeners (A-D). The tables contain the ranking established using the two criteria using rank 1 for the best algorithm. rd contour transformation is consistently ranked better in expressivity but at the same time introduce new artefacts. See details in the text.

Transformation of a synthetic rd contour is systematically evaluated to increase expressivity. The stronger rd reduction is always qualified to increase expressivity, however, one of the graders considered the increased expressivity as too much and therefore he preferred the rd reduction with smaller peak value. The improved expressivity comes however at the price of increased artefacts. The change in the rd parameter increases the energy in the frequency band that is right above the VUF. In this segments sinusoids should be created (corresponding to an increase in the VUF) but the transformation of the noise is not yet implemented. Especially for frames with low VUF the transformation of the noise into sinusoids reflecting the coherent increase of the of the VUF seems to be required.

7. CONCLUSIONS

The article presents work on the use of recent glottal pulse estimators and extended source filter models for transformation of speech into singing. A recent algorithm for estimation of the parameters of the LF glottal pulse model has been applied to singing signals providing results that seem consistent with findings that have been obtained with EGG measurements. More experiments are certainly needed to validate the algorithm. Notably the robustness to model errors and noise requires more investigation. A signal transformation algorithm has been described that is based on a recent shape invariant phase vocoder and that has been extended to incorporate the possibility to preserve and control the glottal pulse parameters. Finally, a method has been proposed that allows to coherently change the signal energy with the rd parameter changes. rd parameter contours have been derived from the vibrato extent contours. The rd transformation algorithm has been evaluated in the context of a problem requiring the transformation of speech into singing. The results are promising and show that the modification of the rd parameter contours results in an increased expressivity of the synthetic singing. Further studies are required to reduce artefacts related normally to signals segments with a low VUF boundary, and to improve the realism of the rd contours especially when no vibrato is present.

8. ACKNOWLEDGEMENTS

The authors would like to thank Prof. L. Fagnan for providing singing excerpts and Prof C. d'Allessandro for helpful comments.

9. REFERENCES

[1] P. Alku, M. Airas, and B. Story, "Evaluation of an inverse filtering technique using physical modeling of voice production," in *Proc 8th Intern. Conf. Spoken Language Proc.*, 2004.

[2] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *Jour. of the Acoustic Society of America*, vol. 125, no. 5, pp. 3289–3305, 2009.

[3] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2011.

[4] G. Degottex, A Röbel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[5] T. Saito, M. Unoki, and M. Akagi, "Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, pp. 405–417, 2005.

[6] D.G. Childers and C.K. Lee, "Vocal quality factors: analysis, synthesis, and perception," *Journal Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[7] G. Fant, "Some problems in voice source analysis," *Speech Communication*, vol. 13, no. 1-2, pp. 7–22, 1993.

[8] H.-L. Lu, *Toward a high-quality singing synthesier with vocal texture control*, Ph.D. thesis, Stanford University, 2002.

[9] N. Henrich, *Etude de la source glottique en voix parlée et chantée.*, Ph.D. thesis, Université Paris 6, 2001.

[10] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *Journal Acoust. Soc. Am.*, vol. 117, no. 3, pp. 1417–1430, 2005.

[11] I. Arroabarren and A. Carlsena, "Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices," *Jour. of the Acoustic Society of America*, vol. 119, no. 4, pp. 2483–2497, 2006.

[12] K.I. Nordstrom and P.F. Driessen, "Variable pre-emphasis lpc for modeling vocal effort in the singing voice," in *Proc. 9th Int. Conf. on Digital Audio Effects (DAFx)*, 2006, pp. 157–160.

[13] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an ARX-LV source-filter decomposition and hnm modeling," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007, vol. 4, pp. 525–528.

[14] G. Degottex, A. Röbel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5128–5131.

[15] A. Röbel, "Shape-invariant speech transformation with the phase vocoder," in *Proc. International Conf. on Spoken Language Processing (InterSpeech)*, 2010, pp. 2146–2149.

[16] G. Fant, "The lf-model revisited. transformations and frequency domain analysis.," *Quarterly Progress and Status Report, Dept of speech, music and hearing, KTH*, vol. 36, no. 2-3, pp. 119–156, 1995.

[17] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[18] N. Henrich, G. Sundin, D. Ambroise, C. d'Alessandro, M. Castellengo, and B. Doval, "Just noticeable differences of open quotient and asymmetry coefficient in singing voice," *Journal of Voice*, vol. 17, no. 4, pp. 481–494, 2003.

[19] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modeling with unknown model order," *Pattern Recognition Letters, Special issue on Advances in Pattern Recognition for Speech and Audio Processing*, pp. 1343–1350, 2007.

[20] A. Röbel, "Expressive vibrato synthesis demo," `http://anasynth.ircam.fr/home/english/media/expressive-vibrato-synthesis-singing`.