Singing Voice Vibrato

Peter Sciri, Alois Sontacchi
Inst. of Electronic Music and Acoustics, Univ. of Music and Performing Arts, Graz, Austria,
Email: {sciri,sontacchi}@iem.at

Introduction

The human singing voice has many aspects unique to it such as its wide tonal range, the rich repertoire of timbral colorizations or the possibilities regarding dynamics and phrasing while singing. The entity of all those attributes yields an instrument with very special properties. As one of those properties, vibrato can be singled out being one of the most artistic and virtuoso features of the singing voice.

Aim of this work is to gain further insight on the mechanisms involved in vibrato production. A signal processing framework allows for analysing and modifying natural voice recordings and obviates the need for additional contemporary EGG or laryngographic recordings.

F0 Tracking

As proposed and discussed by numerous authors, the fundamental perceptual component of vibrato is the variation of pitch. Additionally, variations of the amplitude and the vocal tract filter contribute to vibrato [1]. In [2] though, Arroabarren pointed out that the change of the spectral envelope of the glottal source (GS) and the vocal tract transfer function (VTR) do not change significantly during vibrato. Hence, a representation of the frequency and amplitude modulation of the acoustic signal is sufficient.

The *YIN* F0 estimator [3] uses a nomalized difference function that is applied to the autocorrelation. A 'best local estimate' criterion and parabolic interpolation lead to highly accurate results at subsample precision.

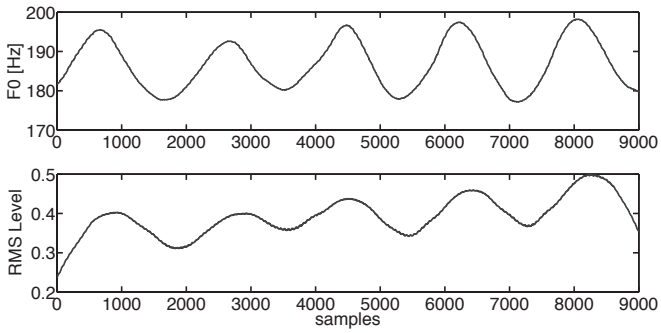

Figure 1: Temporal evolution of F0 and RMS level of a 9000 sample (@11.025kHz) baritone recording

Amplitude Tracking

Pitch modulation is accompanied by a variation of intensity - in terms of musicology one would speak of *tremolo*.

To visualize the temporal evolution of the amplitude an RMS Level detector of window length τ as denoted in eq. 1 is applied.

$$a_{RMS}[n] = \sqrt{\frac{1}{\tau}\sum_{t=-\frac{\tau}{2}}^{\frac{\tau}{2}} x[n+\tau]^2} \qquad (1)$$

As can be seen in figure 1, the pitch modulation occurs in phase with the amplitude modulation.

Inverse Filtering

Central consideration of this work is the separability of glottal source and vocal track response as it has been dealt with in literature for years (further reading e.g. in [4]). The problem of blindly deconvolving the speech signal, hence dividing it into an excitation signal and a system response is commonly achieved by some kind of autoregressive modelling. One problem that arises when fitting the model by simply shifting a window with a fixed hop size is that the assumption of a white excitation signal has to hold. In practical cases, i.e. for vocal signals, this is not the case.

To unveil the VTR that is not corrupted by the spectral information of the GS, the analysis frame has to be restrained to a time interval in which only one (preferably inpulsive/white) excitation occurs at the beginning and no further excitation takes place for the rest of the frame. In other words, the frame would contain only the impulse response of the system. As the major excitation of a glottal cycle occurs at the *instant of glottal closure* (GCI), this assumption is fulfilled in the closed phase (CP) of the glottal cycle. Though the vocal folds do not necessarily close entirely - meaning a constant air flow may occur also during the CP - this DC bias is not of further interest for the analysis.

Detection of the Instant of Glottal Closure

Crucial aspect of this model is the correct determination of a single closing instant. First step to reveal a series of GCIs is to compute an initial candidate set by applying low order LPC to a voice speech segment and investigating the error signal. It is going to expose peaks that correspond to the maximum excitations, which occur when the vocal folds close. To determine these peaks within an analysis frame a method proposed in [5] exploits the properties of the group delay function of a minimum phase system. This creates a set of initial candidates that are refined in two postprocessing steps:

1. *Elimination of False Positives*: a (short-time) dynamic programming algorithm calculates costs according to the pitch deviation (as priorly estimated by the F0 detection) and waveform similarity (normalized crosscorrelation) for consecutive subsets of four GCI candidates

2. *Glottis Cycle Prototyping*: align cycles GCI synchronously as shown in figures 2(a)–(b) and calculate a cycle prototype. Crosscorrelation with the prototype reveals the temporal displacement (which has shown to be zero-mean) of every single instant.

Once a reasonable set of GCIs is found, high-order *constrained closed phase covariance linear prediction* [6] is used to perform source-filter decomposition which grants minimal influence of the GS spectral envelope and subglottal coupling. The calculated coefficients can by used to perform either PSOLA [4] using the GCIs as pitch marks or - physically more reasonable - using an LTV Lattice structure and interpolating the PARCOR coefficients [4] in between the GCIs.

Modifying Vibrato

Inverse filtering yields the glottal source derivative as it is caused by the lip radiation effect. To perform synthesis or cancellation of vibrato on a physically and musically meaningful basis, the glottal volume velocity waveform is computed by compensating this effect. This is commonly achieved by (leaky) integration.

As an example for the capabilities of our framework we want to present vibrato cancellation. Figure 2(a) shows pitch-aligned cycles of the volume velocity waveform taken from a baritone recording with an average pitch of 186Hz and a frequency modulation of ±8Hz (note the variation of pitch period visible from the grey area at the top which is unallocated matrix space). It allows for an interesting observation: the closed phase (up to approx. 24 on the y-axis) and the first half of the open phase (approx. at sample 43) remain completely unchanged during vibrato. Only the following *release phase* that ends with the subsequent glottal excitation exposes time variation and hence causes the frequency modulation. Restraining modifications to this very specific time interval as a consequence yields an appropriate simulation of the physical process of vibrato production.

Recomposing a voice signal involves application of the lip radiation effect (e.g. by taking the derivative of the modified volume velocity waveform) and using one of the above mentioned time-variant filter techniques to reapply the vocal tract transfer function.

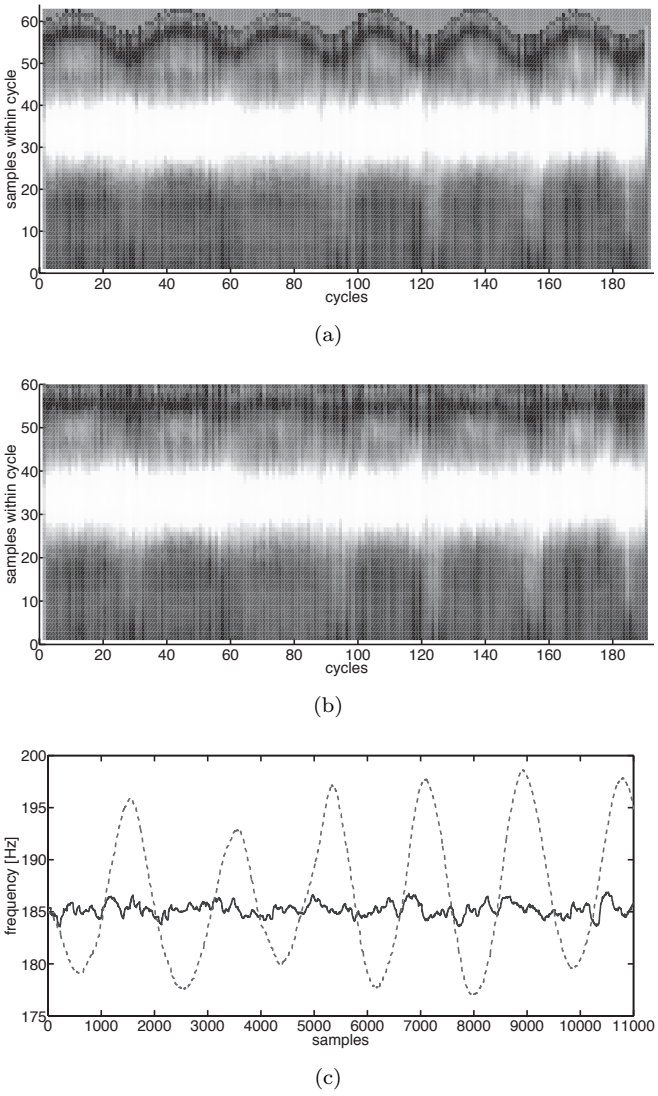

Figure 2: Vibrato cancellation: (a) shows pitch-aligned glottal volume velocity wavefrom cycles of a baritone vibrato (±8Hz at an average pitch of 186Hz) whereas in (b) the vibrato has been remove with the described method (note that the area from 0-43 samples within a cycle has not been exposed to modification); (c) displays the F0 measurement before (dashed) and after vibrato cancellation (solid).

References

[1] Sundberg, J.: Acoustic and psychoacoustic aspects of vocal vibrato, Speech Transmission Lab. Quart. Progress and Status Rep., Vol. 35, Nr. 2-3, 1994

[2] Arroabarren, I.: On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14, July 2006

[3] de Cheveigne, A.: YIN, a fundamental frequency estimator for speech and music, J.A.S.A, Vol. 111, Nr. 4, 2002

[4] Rabiner, L and Schafer, W: Theory and Application of Digitial Speech Processing, Pearson Education, Inc., 2011, ISBN-13: 978-0-13-705085-7

[5] Brookes, M.: A quantitative assessment of group delay methods for identifying glottal closures in voiced speech, IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14, March 2006

[6] Alku, P.: Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering, J.A.S.A, Vol. 125, Nr. 5, 2009