Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework

Perfecto Herrera, Jordi Bonada
Audiovisual Institute, Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
{pherrera, jboni}@iua.upf.es http://www.iua.upf.es

Abstract

Periodic or quasi-periodic low-frequency components (i.e. vibrato and tremolo) are present in steady-state portions of sustained instrumental sounds. If we are interested both in studying its expressive meaning, or in building a hierarchical multi-level representation of sound in order to manipulate it and transform it with musical purposes those components should be isolated and separated from the amplitude and frequency envelopes. Within the SMS analysis framework it is now feasible to extract high level time-evolving attributes starting from basic analysis data. In the case of frequency envelopes we can apply STFTs to them, then check if there is a prominent peak in the vibrato/tremolo range and, if it is true, we can smooth it away in the frequency domain; finally, we can apply an IFFT to each frame in order to re-construct an envelope that has been cleaned of those quasi-periodic low-frequency components. Two important problems nevertheless have to be tackled, and ways of overcoming them will be discussed in this paper: first, the periodicity of vibrato and tremolo, that is quite exact only when the performers are professional musicians; second: the interactions between formants and fundamental frequency trajectories, that blur the real tremolo component and difficult its analysis.

1 Introduction

Long sustained notes become boring and uninteresting if their steady states have a strictly constant fundamental frequency. Because of that and other musical reasons to be found in music performance treatises, good performers invest a lot of time to developing proficiency in techniques for the continuous modulation of frequency and/or amplitude. This kind of modulations are respectively called *vibrato* and *tremolo* and its feasibility for every instrument depends on its sound generation mechanisms (for example, string instruments favor vibrato and otherwise reeds favor tremolo). The scientific study of vibrato can be traced backwards to the work by Seashore [1] who, notwithstanding his technological limitations, yielded a rough but valid characterization of that phenomenon. More recent studies ([2], [3], [4], [5]) have been developed with the help of modern analysis techniques and devices but we can conclude that, although we understand the basic facts about vibrato in different musical instruments, more research is needed on vibrato as a physical phenomenon (not to mention as a musical resource, indeed), specially on its temporal evolution and its way of change between consecutive notes. Anyway, if it is parametrically and/or procedurally possible to describe vibrato, it should be possible to manipulate it for musical, engineering, or acoustical purposes.

As Desain and Hoenig [6] noted, the shape of musically modulated signals is quite complex to be extracted without a solid model of analysis. One of those models could be the Spectral Modeling Synthesis (SMS) developed by Serra [7]. Recent software developments inside that framework ([8], [9], [10]) have made possible to segment a continuous signal such as a musical phrase or a long note into different *regions* that have different basic *features* or *parameters* both static and evolving along time (i.e. mean fundamental frequency, mean amplitude, amplitude tendency, noise profile, amplitude and fundamental frequency envelope, etc.); once those parameters have been extracted, it is possible to manipulate them separately in order to achieve delicate sound transformations during the re-synthesis stage. Consequently there are certain situations in which it could be useful to separate the contribution of modulation processes over a stable set of parameters in order to achieve a greater flexibility and better quality of synthesis and transformation.

The vibrato problem can be decomposed into three subproblems to be tackled: 1) Identification (or Detection and Parameterization); 2) Extraction; and 3) Re-synthesis. Considering that our target system is an off-line (non real-time) one in this paper we will focus on the first two points (see [11] for a synthesis oriented paper).

2 Frequency –domain strategy

From the frequency-domain point of view, vibrato detection in an off-line system assumes that a steady state has been correctly delimitated and parameterized in a previous stage of the analysis; that is to say that we have obtained a fundamental frequency track whose frequency is constrained in a range of less than a whole tone around an ideal "mean" (although the usual vibrato depth reported by different studies carried away with professional musicians is lesser than half a tone, it should be noted that not so well trained performers generate larger excursions from the nominal fundamental frequency). The fundamental frequency track obtained in SMS analysis is an envelope of data representing Hertzs along time, and has a number of values equal to the frame rate of analysis (typically we use 345 points per second so that each one of our *envelope frames* integrates information for such a temporal lapse); thus, that envelope will be the starting data for the process (for details see [9]).

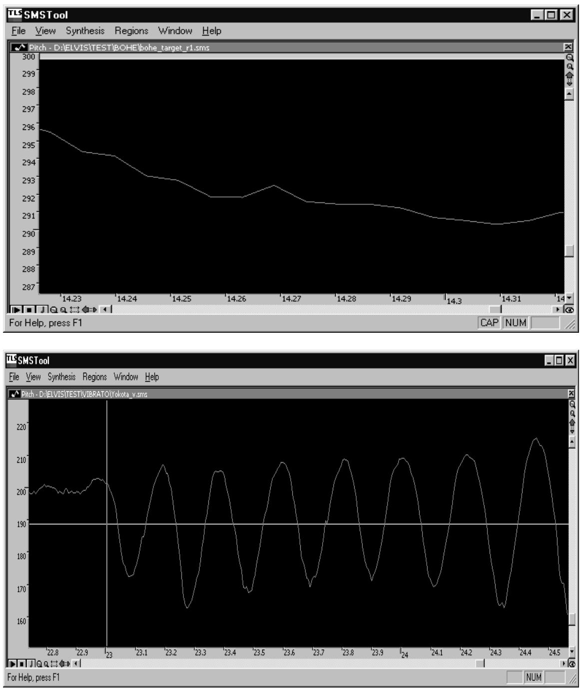

Figure 1. Two fundamental frequency tracks: a) from a steady state portion of sound without vibrato; b) from a steady state portion of a sound with vibrato.

The vibrato detection proceeds as follows: the discrete fundamental frequency track is first transformed into a 0-centered track by computing the global mean and subtracting it from every fundamental frequency value in the original track. This smoothed and 0-centered track is then windowed. A window size of 128 points or 0.37 seconds (more than 2 times the lowest period that is expected to be found for a vibrato) with a 50% of window overlap has been proved suitable for our purposes. Different kind of windows (Blackman-Harris, Kaiser, Hamming) have been tested yielding no significative differences. For each window its FFT is then computed and spectral peaks are calculated with parabolic interpolation. In case the analyzed region has vibrato we get a prominent peak around 5/6 Hz (in fact it is the most prominent peak detected). As expected, such a clear and stable peak is not present when the region has no vibrato. The vibrato detection process concludes with the extraction and storage of the rate and the depth of vibrato as high-level parameters of the analyzed frame (in fact, they will be later pooled with the values for all the *envelope frames* and global mean values will be extracted for a whole region).

At this point, the vibrato extraction proceeds. Different algorithms could be implemented, as for example a similar one to the SMS low-level analysis (i.e. by additive synthesis and subtraction of the harmonic part), but it is more economic and easy to "crop" the prominent peak (and sometimes the second one) of every envelope. Then the IFFT of the altered spectrum is computed so that we get a signal without the modulation components, that is, more stationary than the original one.

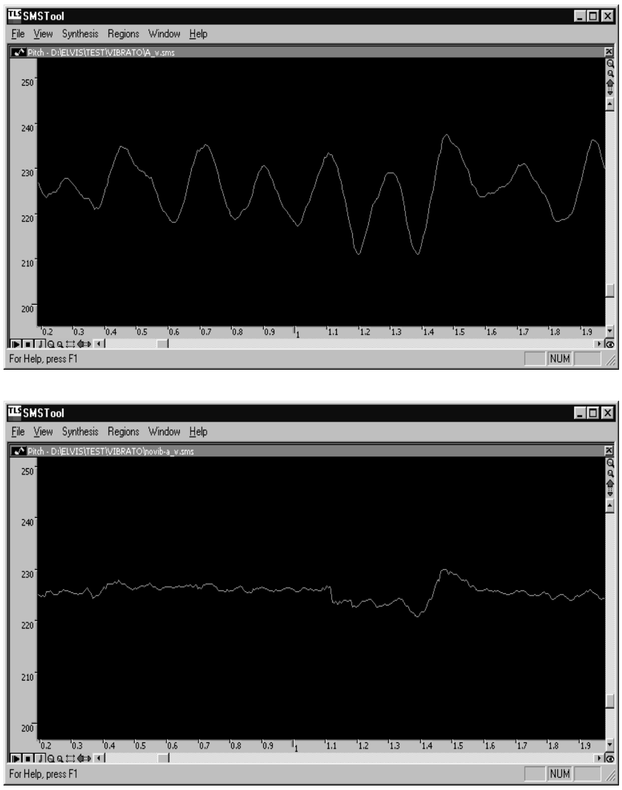

Figure 2. Comparison of a fundamental frequency track of a steady state portion of sound: a) with its original vibrato; b) the same fundamental track after vibrato extraction in the frequency-domain.

3 Time-domain strategy

In the time-domain there are several robust techniques for fundamental frequency estimation [12] that could be suitable for vibrato extraction. Besides that, time-domain strategies offer important advantages such as the option of using shorter windows. In such a scenario, we could find practical situations that only demand to get rid of vibrato, but not necessarily to characterize it in full detail. Given that constraint, a filtering strategy seems quite suitable to be approached (on the other hand, see [13] for a time-domain -although not real-time- complete solution without using filters).

In order to design an appropriate filtering algorithm for this task we have to take into account the fact that the value to be given to the filter at every point of time should be centered around a conventional 0 (in this case the mean fundamental frequency). For an off-line system such a central value could be the mean frequency of the steady state, but in a real-time system that center must be approximated from the past behaviour of the fundamental frequency track. If much previous information is used in computing such an approximation then we will lose the temporal trends for the pitch, but if only very recent values are taken into account then we will lose the low frequency modulations that we are addressing to (in fact this is something like a paradox, because losing low frequency modulations is what we are trying to achieve!). An acceptable solution should be a number of points spanning more than a common vibrato cycle, and at least one of the shortest vibrato cycle we could find. After some trial and error we settled into a filter buffer that takes into account the preceding 80 envelope points (about one vibrato cycle of 4.5 Hz) and does not blur the mid-term variations of the fundamental. If the system does not yet have 80 data points it uses the mean of the available points. We feel, nevertheless, that this mechanism should be exhaustively refined in order to obtain better results as we can see from *Figure 3*.

After we get the discussed mean value, we can apply a filter to the incoming data. Because both the vibrato rate and depth will be constrained, we have implemented a 6-order Butterworth high-pass filter that effectively eliminates frequencies lower than 10 Hz from the fundamental frequency track. The selection of the filter was done with the help of MATLAB, and finally we opted for a filter defined along the following parameters: (passband=.25 radians (approx. 21Hz.), stopband=.11 radians (approx. 9 Hz), passband ripple = 3 dB, stopband attenuation = 40 dB).

Although this strategy does not allow to characterize vibrato at the filtering stage, the blackboard-like model implemented in the SMS analysis framework facilitates that vibrato parameters can be extracted later on by picking the relevant information from other concurrent analysis modules (of course there is an arguable time-resolution payoff).

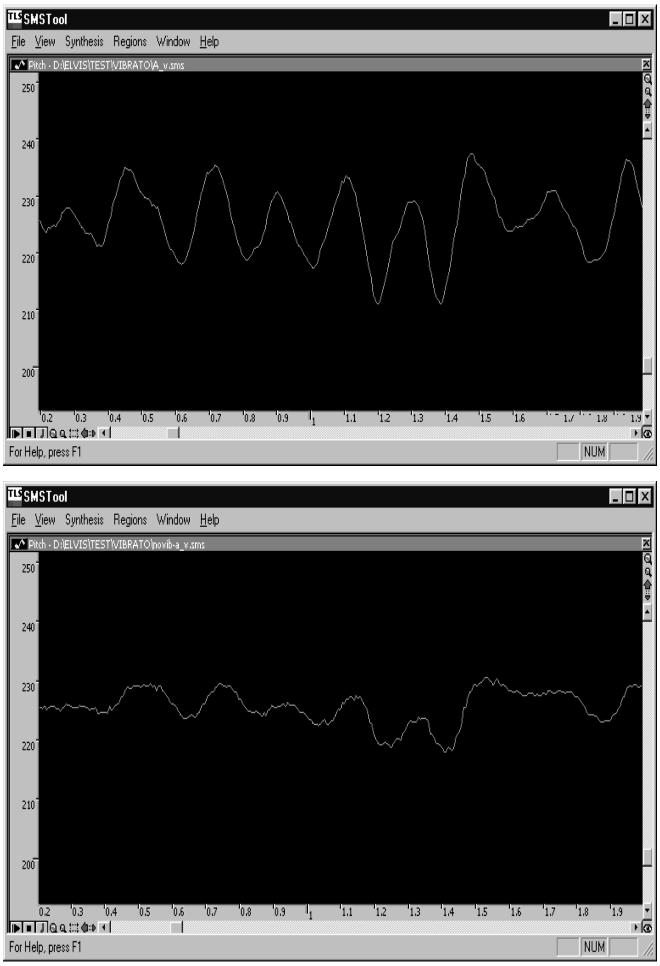

Figure 3. Comparison of an original fundamental track of a steady state portion of sound: a) with original vibrato; b) after time-domain vibrato extraction.

4 Interaction between vibrato and spectral shape

If we examine the amplitude track of a region with vibrato it seems that there also are cyclic modulations around an ideal "mean value". Although if could be tempting to consider them as examples of a concomitant "tremolo" and therefore to proceed with that track as we did with the frequency, we should be warned that superficially similar expressive resources as vibrato and tremolo could have different musical meanings and uses, and do not need to be associated. It should also be noted that (at least in human singing) amplitude variations follow a pattern not as regular as frequency does. In fact the main factor for the observed variations in amplitude are, other than a tremolo process, the interaction between the vibrato process and the resonances of the vocal tract [14], [15]. Therefore, our strategy for eliminating those amplitude modulations goes as follows: in the frequency-domain case, an spectral envelope for the steady region is computed; then, we proceed by re-calculating the "right" amplitude value for every track (or partial) frame by frame. By "right amplitude" we mean the amplitude that the track should have, considering the trajectory correction induced by the vibrato-suppression procedure (for example, let's suppose that the original fundamental frequency track entered a resonance region; after vibrato suppression its amplitude would be still reflecting its presence in such a resonance region, but in fact the track is not there anymore, so we will interpolate –from the spectral envelope- the amplitude corresponding to the current spectral location for that track).

On the other hand, in the time-domain case we are just starting to implement an "incremental-resolution spectral envelope extracting algorithm" much in the vein of the spectral tracings used by [11], and similar to the one apparently used by humans [16]. Such an algorithm, whereby the correction of the amplitudes is done frame by frame (as explained before), computes a spectral envelope that gets increasing resolution as we get more frames from the basic analysis.

5 Conclusions

In this paper we have presented a two-fold approach for managing vibrato inside an specific analysis/synthesis framework like SMS. Although the higher level attributes extracted in the analysis process allow both the satisfactory characterization of vibrato and also its removal from a steady state portion of a sound in the frequency domain, there will be practical situations in which only removal will be mandatory and then we can apply a simpler time-domain strategy. Nonetheless more research is needed, and it shall be pursued for us, in order to refine the current algorithms, and, afterwards, achieve a flexible and acceptable synthesis of vibrato notes.

Sound examples related to this paper can be found at: http://www.iua.upf.es/~perfe/papers/dafx98poster-soundexamples.html.

References

[1] C. E. Seashore. *Psychology of Music*. New York: McGraw-Hill, 1938. (Reprint: Dover, New York, 1967).

[2] J. Sundberg. "Vibrato and vowel identification". *Arch. Acoust*. 2, 257-266, 1977.

[3] E. Prame. "Measurements of the vibrato rate of ten singers". *J. Acoust. Soc. Am.* 96 (4), pp. 1979-1984, 1994.

[4] H. Honing. "The vibrato problem, comparing two solutions". *Computer Music Journal*, 19 (3), 1995.

[5] P. Desain and H. Honing. "Towards algorithmic descriptions of continuos modulations of musical parameters". *Proceedings of the ICMC*, 1995.

[6] P. Desain and H. Honing. "Modeling continuos modulations of music performance". *Proceedings of the ICMC*, 1996.

[7] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. Ph.D. Dissertation, Stanford University, 1989.

[8] X. Serra and others. "Integrating Complementary Spectral Models in the Design of a Musical Synthesizer". *Proceedings of the ICMC*, 1997.

[9] X. Serra and J. Bonada. "Sound Transformations Based on the SMS High Level Attributes". *Proceedings of the Digital Audio Effects Workshop (DAFX98), 1998.*

[10] P. Cano. "Fundamental Frequency Estimation in the SMS Analysis". *Proceedings of the Digital Audio Effects Workshop (DAFX98), 1998*

[11] R. Maher and J. Beauchamp. "An investigation of vocal vibrato for synthesis". *Applied Acoustics*, 30, pp. 219-245, 1990.

[12] W. Hess. *Pitch determination of speech signals*. Berlin: Springer-Verlag, 1983.

[13] S. Rossignol and others. "Feature Extraction and Temporal Segmentation of Acoustic Signals". *Proceedings of the ICMC*, 1998.

[14] Y. Horii. "Acoustic analysis of vocal vibrato: a theoretical interpretation of data". *J. of Voice*, 3 (1). 36-43. 1989.

[15] M. Mellody and G. H. Wakefield. "Modal distribution analysis of vibrato in musical signals". *Proceedings of SPIE Conf.*, 1998.

[16] J. H. Ryalls and P. Lieberman. "Fundamental frequency and vowel perception". *J. Acoust. Soc. Am.* 72 (5). 1631-1634, 1982.