Vibrato: detection, estimation, extraction, modification

S. Rossignol[1,2], P. Depalle[3], J. Soumagne[2], X. Rodet[1] and J.-L. Collette[2]

rossigno@ircam.fr,depalle@music.mcgill.ca,soumagne@ese-metz.fr,rod@ircam.fr

[1]IRCAM,1 place Igor-Stravinsky, 75004 PARIS, FRANCE

[2]Supélec 2, rue Édouard Belin, 57078 METZ, FRANCE

[3]Music Faculty, McGill University, MONTRÉAL, QUÉBEC, CANADA

Abstract

This paper deals with vibrato detection, vibrato extraction from f_0 trajectory, and vibrato parameter estimation and modification. Vibrato detection and extraction are aimed at being a first step for note segmentation of singing voice signals. The aim is also to characterize sounds with the descriptor: "presence of vibrato" or "absence of vibrato". Changing vibrato parameters, that is to say its magnitude and its frequency, is also one of the possible musical applications. It requires firstly to detect the presence of vibrato. In order to do that, several approaches are possible: we can analyse directly the sound signal or its f_0 trajectory. For each approach, several techniques exist: some of them are described in the paper: the "spectrum modelling" method, the "spectral envelopes distortion" method, the "AR prediction" method, the "analytic signal" method and the "minima – maxima detection" method. Their performance are compared. Secondly, the parameterization is completed: if there is vibrato, the parameters of the vibrato, that is to say its frequency and its magnitude, are given. Thirdly, the vibrato is extracted from f_0 trajectory to obtain a no-vibrato melodic evolution. This "flat" fundamental frequency is useful for segmentation of musical excerpts into notes, but can also be used for sound modification or processing.

1 Introduction

This work on vibrato detection and extraction is a part of a segmentation and indexation tool described in [RRS+99]. This tool is a first step of a more ambitious project whose goal is to segment and labelize any sound (monophonic or polyphonic, harmonic or inharmonic, etc.) as automatically as possible.

Several methods to detect vibrato have been tested. The two first methods presented are based directly on the sound signal. The first one, contrary to the second one, provides all the parameters of the vibrato. These two methods work with 20 up to 30 ms long windows. The three other methods we present are based on the f_0 trajectory and require 300 ms long windows. They give the vibrato parameters. Finally, the vibrato can be removed on the f_0 trajectory using data obtained with the last method.

Each vibrato detection method provides its own answer to the question: is there any vibrato? Data fusion methods are used for combining several opinions on a given question. The aim is to obtain a more precise and/or certain decision than any local decision. Each opinion is provided by a sensor. Then, in our case, each vibrato detection method can be considered as a sensor. The same issue is possible to answer the question: What are the vibrato frequency, magnitude and phase?

2 Vibrato detection

2.1 Method based on "spectrum modelling"

Let us consider a periodic and centered sound signal composed of the first H harmonics of its fundamental frequency f_0. This signal is modulated by a vibrato which frequency, magnitude and phase are respectively f_v, A_v and ϕ_v. The signal model $M(t)$ is:

$$M(t) = \sum_{h=1}^{H} B_h \cos\left[2\pi h f_0 t + \frac{hA_v}{f_v}\sin(2\pi f_v t + \varphi_v) + \varphi_h\right]$$

where B_h is the magnitude of the partial h and φ_h its phase. $M(t)$ can be written as:

$$M(t) = \sum_{h=1}^{H}\sum_{n=-\infty}^{+\infty} J_n\left(\frac{hA_v}{f_v}\right) B_h \cos[(2\pi f_v t + \varphi_v)n + 2\pi h f_0 t + \varphi_h]$$

where $J_n\left(\frac{hA_v}{f_v}\right)$ is the Bessel function of the first kind and the nth order. The Fourier spectrum, computed on a windowed signal frame, is:

$$\hat{M}(f) = \sum_{h=1}^{H}\sum_{n=-\infty}^{+\infty} J_n(\frac{hA_v}{f_v})\times \{\hat{W}(f - hf_0 - nf_v)(a_h + ib_h)\exp(in\varphi_v) + \hat{W}(f + hf_0 + nf_v)(a_h - ib_h)\exp(-in\varphi_v)\}$$

where $a_h = \frac{B_h}{2}\cos(\varphi_h)$ is the real part of the complex magnitude of the partial h, $b_h = \frac{B_h}{2}\sin(\varphi_h)$ the imaginary part of the complex magnitude of the partial h, and $\hat{W}$ the Fourier Transform of the window.

The determination of the $P = 2H + 4$ unknown parameters $\underline{x} = [a_h,\ b_h,\ f_0,\ f_v,\ A_v,\ \phi_v]$ of an observed signal s is performed by an iterative minimization of the squared error ϵ between the observed spectrum $\hat{S}$ and its model $\hat{M}$.

Henceforth, the value of f_0 is supposed to be known, as the method is very sensitive to errors on this parameter. MATLAB simulations give results shown on figures 1 and 2. The values of the actual parameters are: $H = 3$, $f_0 = 400\ Hz$, $A_v = 20$, $f_v = 5$, $\varphi_v = 1$, $B_1 = 1$, $B_2 = 0.5$, $B_3 = 1/3$, $\varphi_1 = 1.0$, $\varphi_2 = 0.7$, $\varphi_3 = 1.6$. And the initial conditions for the iterative algorithm are: $A_{v(0)} = 19.0$, $f_{v(0)} = 5.25$, $\varphi_{v(0)} = 0.95$, $B_{1(0)} = 1.05$, $B_{2(0)} = 0.475$, $B_{3(0)} = 0.2375$, $\varphi_{1(0)} = 1.05$, $\varphi_{2(0)} = 0.665$, $\varphi_{3(0)} = 1.68$. The error on the parameter $param$ (which actual value is $param$) at the kth iteration is: $e(k) = 10\log_{10}\{|(param(k) - param)/param|\}$.

As $\epsilon_{(k)} = (\underline{M}_{(k)}^H - \underline{S}^H)(\underline{M}_{(k)} - \underline{S})$ and $\underline{x}_{(k-1)}$ is supposed to be known at step k, the analytical expression of the derivatives $\frac{\partial \underline{M}_{(k-1)}}{\partial x_i}$[1] are computed. Then: $\underline{x}_k = \underline{x}_{k-1} + \underline{\alpha}_{(k-1)}$, with $\underline{\alpha}_{(k-1)} = (\underline{H}_{(k-1)}^H \underline{H}_{(k-1)})^{-1} \underline{H}_{(k-1)}^H (\underline{S} - \underline{M}_{(k-1)})$ and $\underline{H}_{(k-1)} = \underbrace{\left[\frac{\partial \underline{M}_{(k-1)}}{\partial x_1} \cdots \frac{\partial \underline{M}_{(k-1)}}{\partial x_P} \right]}_{P} \Big\} N$. To obtain figures 1 and 2, we use $f_e = 44100\ Hz$, a window w of size $T = 0.04\ s$, and a 4096 bins FFT. It can be seen that the method converges toward the actual values in 10 iterations. This method has not been tested on real sound signals. In practice it is also required to consider the influence of the tremolo and of the harmonics of the vibrato.

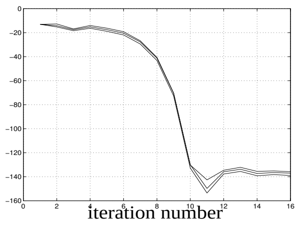

Figure 1: Error $e(k)$ on $A_{v(k)}$, $f_{v(k)}$ and $\varphi_{v(k)}$

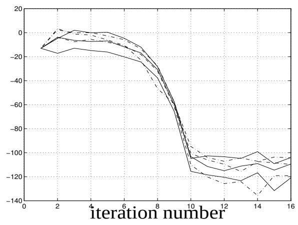

Figure 2: Error $e(k)$ on $B_{1(k)}$, $B_{2(k)}$, $B_{3(k)}$, $\varphi_{1(k)}$, $\varphi_{2(k)}$ and $\varphi_{3(k)}$

2.2 Method based on "spectral envelopes distortion"

Distortion of the spectral envelope of a sound due to vibrato is very typical. This distortion can easily be shown using two temporal windows centered at the same temporal position. The length of the first one is chosen sufficiently small (20 ms) in order to neglect vibrato influence while the length of the second one is large enough (60 ms) to exhibit the distortion due to vibrato.

In practice, spectral envelopes are estimated using the method described by Serra in [Ser89]. First, the local maxima of the magnitude spectrum are detected. Some rules are used to eliminate the peaks which are too small compared to their neighbours. Then, the spectral envelope is drawn by joining the detected peaks. The spectral envelope obtained with the first window is called a^I, and the second one is called a^{II}.

Now, we observe that the relative difference $d(f) = \frac{a^I(f) - a^{II}(f)}{a^I(f)}$ between the two envelopes grows quite linearly with frequency f (see figure 3) when there is some vibrato effect, otherwise a noisy curve is obtained. This near linear behavior can be explained using frequency modulation standard literature (see [Ven90]). The Carson's band is defined as the band which comprises 98 % of the energy of a modulated sinusoid. For sinusoidal modulation, this band is equal to $B = 2(f_{mod} + \Delta f)$, where f_{mod} is the modulation frequency and Δf the magnitude of the modulation. In our case, the Carson's band is equal to $B_i = 2(f_v + iA_v)$ for the partial i. Then $B(i)$ grows linearly with i.

The vibrato detection can be achieved by the estimation of $d(f)$ as a first order polynom: $\alpha f + \beta$. If there is vibrato, α is great, otherwise it is very small. The f_0 trajectory for a flute (without vibrato: it can be noticed that there is a very low vibrato on the longest note) and for a singing voice (with vibrato) excerpts are shown on figures 4 and 5. Results on the flute and on the singing voice excerpts are shown in figure 6.

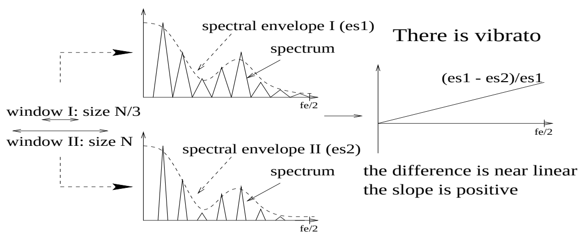

Figure 3: Linear growth of $d(f)$

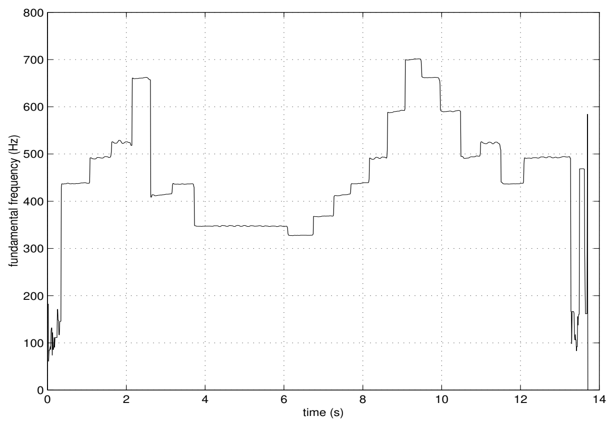

Figure 4: f_0 for the flute excerpt

2.3 Method based on "AR prediction"

In this method, the vibrato is extracted directly from the f_0 trajectory. Notes are usually shorter than one second. Standard values of vibrato periods $[0.1\ s\ \ldots\ 0.3\ s]$ limit the number of vibrato periods to a maximum of two or three periods per note, that is to say during a stationary

[1] The derivatives $\frac{\partial \underline{M}_{(k-1)}}{\partial x_i}$ are not given here, but one can easily found them.

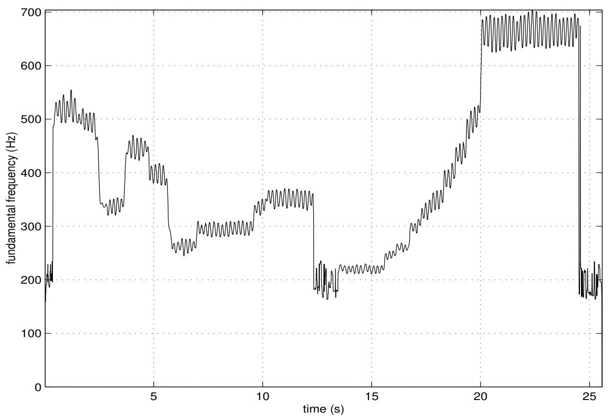

Figure 5: f_0 for the singing voice excerpt

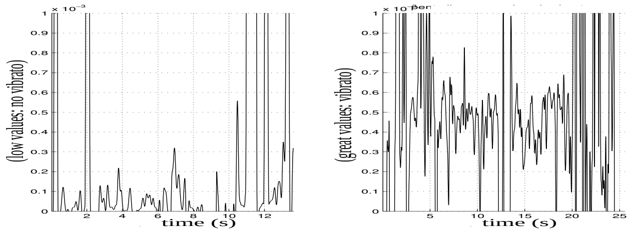

Figure 6: Temporal evolution of α for the flute and for the singing voice excerpts

signal. Then, the vibrato frequency can not be precisely extracted by a short-time Fourier spectral analysis, since the number of periods per segment of stationary signal (i.e. per note) is less than required [4 ... 5] for a sufficient frequency resolution.

The idea of the method is to double the number of samples by extrapolating the f_0 signal with a predictive model. 0.3 second segments are considered. The sampling frequency of the f_0 trajectory is 100 Hz, then each segment is composed of $N = 30$ samples. The N samples are 0-centered. An AR modelling is performed. A backward and a forward prediction are estimated. The length of the two predictions is equal to $N/2$ samples. Then, finally, we obtain $2N = 60$ samples. A FFT is computed on this new segment and the vibrato frequency is extracted by a precise localisation of the maximum of the magnitude spectrum. Experimental results for the flute excerpt (figures 7 and 8) and for the singing voice excerpt (figures 9 and 10) prove the noticeable improvments provided by the auto-regressive estimation and extrapolation. The precision and the stability of the results are improved. For the flute excerpt, it can be noticed that the low vibrato (A_v is small) present on the longest note (from 4 to 6 second) is detected.

2.4 Method based on the "analytic signal"

We use this method directly on the temporal evolution of the fundamental frequency f_0, although it is normally used directly on sound signals (see [Hes83]). The first step of the original method is to obtain an approximation of the fundamental component $s(i) \simeq \cos(2\pi f_0 i/f_e + \phi)$

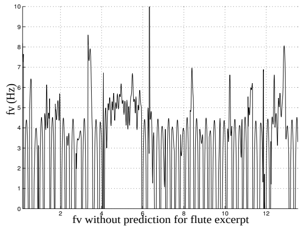

Figure 7: f_v found for the flute excerpt. No prediction is done.

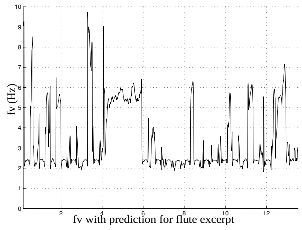

Figure 8: f_v found for the flute excerpt. Prediction is done.

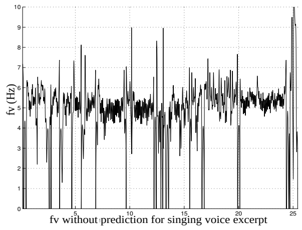

Figure 9: f_v found for the singing voice excerpt. No prediction is done.

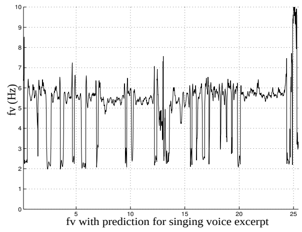

Figure 10: f_v found for the singing voice excerpt. Prediction is done.

by using a low-pass filter (in $\frac{1}{f^K}$) on the 0-centered version of the original signal. Then an analytic version $S(i) \simeq \exp(2j\pi f_0 i/f_e + j\phi) = \exp(j\phi_i)$ is obtained using an Hilbert filtering. The instantaneous fundamental frequency is deduced as $f_0 \simeq f_e/(2\pi)\arg(S(i+1)S^*(i))$, where $*$ stands for the conjugation. Finally, the instantaneous frequency is Kay filtered (see [Kay88]):

$$\bar{f}_0(i) = \frac{1}{2\pi}\sum_{n=i}^{i+N-2}\frac{1.5N}{N^2-1}\left(1-\left[\frac{n-(N/2-1)^2}{N/2}\right]^2\right)f_0(n)$$

As we use this method on the f_0 trajectory itself, harmonics which are naturally low if not absent can be neglected. Then we only need to band-pass filter the signal inside [3 ... 11] Hz. $f_v \simeq f_e/(2\pi)\arg(S(i+1)S^*(i))$ is obtained by the system presented on figure 11. The length of the three impulse responses is $T = 0.35s$, i.e. $N1 = N2 = N3 = 35$ samples.

The results for the flute excerpt and for the singing voice excerpt are shown on figures 12 and 13.

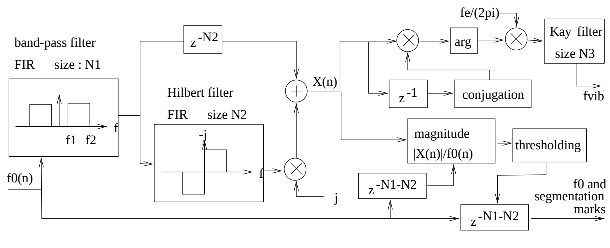

Figure 11: Synoptic of the method based on the analytic signal

2.5 Method based on "Minima – maxima detection"

Local maxima of the f_0 trajectory are detected and precisely pinpointed by interpolation. All the temporal dis-

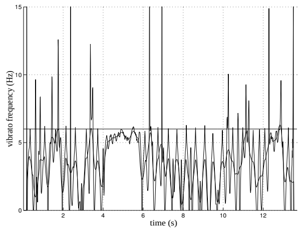

Figure 12: f_v found for the flute excerpt using the "analytic signal" method)

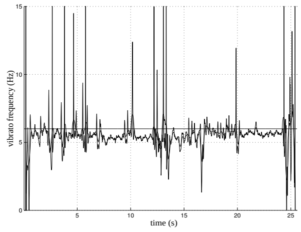

Figure 13: f_v found for the singing voice excerpt using the "analytic signal" method

tances between two successive local maxima are calculated. Then variance of these distances is computed as well as the number of distances comprised between $0.15\,s$ and $0.25\,s$ (which correspond to a $4\,Hz$ to $6.7\,Hz$ vibrato). The same processing is performed for the local minima. In the presence of a vibrato, both variances have low values and number of distances reach high values. Moreover, a large value for: $M_{\text{freq}} = \text{mean}\left(\frac{maxinterp - mininterp}{(maxinterp + mininterp)/2}\right)$ shows that the vibrato is significant. More details concerning the method can be found in [RRS+99]. Some results obtained for the flute and for the singing voice excerpts described above are given in the table 1.

	Vmax	Vmin	Pmax	Pmin	Mfreq
song	0.0010	0.0012	85 %	84 %	0.087
flute	0.0090	0.0100	50 %	46 %	0.032

Table 1: Vibrato detection results

3 Vibrato parameters estimation and vibrato extraction and modification

"Spectrum modelling", "AR prediction", "analytic signal" and "minima-maxima detection" methods provide estimations of the vibrato parameters. Then it is possible to remove or modify the existing vibrato of a sound. The vibrato is extracted from the f_0 trajectory using a by-product of the "minima – maxima detection" method: $\frac{maxinterp+mininterp}{2}$ is a good estimation of what would be the f_0 trajectory with no vibrato. Then the new vibrato can be added to this new f_0 trajectory and a sound signal can be synthetized using this processed f_0 evolution. Results using an IRCAM software called DIPHONE on a singing voice excerpt can be found at:

`http://www.ircam.fr/equipes/analyse-synthese/rossigno/`

4 Conclusion

In this paper, some efficient vibrato detection techniques have been presented. Very different approaches of the problem have been considered. The results are comparable or complementary each other. Other methods can be found in the literature (see for example [Lar89], [HB98]).

Vibrato estimation can be very useful for segmenting musical sounds into notes. Indeed, segmentation using the derivative of f_0 trajectory may be fail, as the frequency deviation generated by a vibrato can be greater than the f_0 deviation between two successive notes. Then removing the vibrato from the f_0 trajectory by using the results of the previous stages of the analysis helps to improve the segmentation into notes.

As a conclusion, we can also notice that analysing the tremolo is rather different: The "spectrum modelling" method has to be adapted. The "spectral envelopes distortion" method cannot be used as there is no typical distortion of the spectral envelope due to tremolo. But the methods based on f_0 trajectory analysis can be applied directly on power trajectory for tremolo detection and extraction.

Acknowledgements

This work is supported by CNET/RENNES (Centre National d'Études des Télécommunications), FRANCE.

References

[HB98] Perfecto Herrera and Jordi Bonada, *Vibrato extraction and parameterization in the spectral modeling synthesis framework*, Workshop on Digital Audio Effects (DAFx98), 1998, pp. 107 – 110.

[Hes83] Wolfgang Hess, *Pitch determination of speech signals*, Springer-Verlag, 1983.

[Kay88] S. Kay, *Statistically/computationally efficient frequency estimation*, 1988, pp. 2292 – 2295.

[Lar89] Jean Laroche, *étude d'un système d'analyse et de synthèse utilisant la méthode de prony - application aux instruments de musique du type percussif*, Ph.D. thesis, École Nationale Supérieure des Télécommunications, Octobre 1989.

[RRS+99] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette, and P. Depalle, *Automatic characterisation of musical signals: feature extraction and temporal segmentation*, Journal of New Music Research, 1999.

[Ser89] Xavier Serra, *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition*, Ph.D. thesis, Stanford University, October 1989.

[Ven90] Dominique Ventre, *Communications analogiques*, Collection pédagogique de Télécommunications, Éllipses, 1990.