Audio Engineering Society

Convention Paper

Presented at the 131st Convention
2011 October 20–23 New York, USA

Vibrato Detection Using Cross Correlation Between Temporal Energy and Fundamental Frequency

Henrik von Coler[1], Axel Roebel[2]

[1]*Audio Communication Group, Technical University Berlin*

[2]*STMS - IRCAM CNRS UPMC Paris, France*

Correspondence should be addressed to Henrik von Coler (von_coler@mailbox.tu-berlin.de)

ABSTRACT
In this work we present an approach for detecting quasi periodic frequency modulations (vibrato) in monophonic instrument recordings. Since a frequency modulation in physical instruments usually causes an amplitude modulation, our method is based on a block wise cross correlation between the extracted frequency- and amplitude modulation trajectories. These trajectories are obtained by removing the constant components. The resulting cross correlation curve shows significant positive peaks at vibrato regions and local minima at note boundaries. Our approach has the advantage of working without a previous note boundary detection and needs only a small look ahead. Furthermore no presumptions on vibrato parameters have to be made.

1. INTRODUCTION

Modulations are an essential part of musical performances. Musicians use them to create an individual and vivid interpretation and to achieve a richer timbre. In singing for example, such modulations are often used subconsciously yet usually synchronized to the musical content [1]. In solo instrument recordings we find, depending on the instrument and its excitation principle, different kinds of modulations. The terms regarding the modulations are not commonly used in literature. Some refer to vibrato as modulations in general including modulations of frequency, amplitude and spectral envelope. In this paper we differentiate between frequency modulations (FM), called vibrato, amplitude modulations (AM), called tremolo and the spectral envelope modulations (SEM). In physical instruments these three types usually appear together and related to each other. We will in this work focus on the relations between vibrato and tremolo. Vibrato is known to cause amplitude modulations by sweeping the excitation signal through the instruments resonances [2]. A way of describing this phenomenon is the use of the sinusoidal model. In [3], [4] and [5] the relations of instantaneous amplitude and instantaneous frequency have been investigated for the single partials inside a sinusoidal model. Here it is pointed out that in case of vibrato each partial varies its amplitude according to the section of the instruments frequency response it is located at. This leads to a modulation, depending on the slope of the relevant section. Thus each sinusoid has an individual change in amplitude induced by the vibrato what then leads to the overall amplitude modulation by summing up all partials modulations.

Vibrato is naturally used in all instruments which allow this kind of modulation. The singing voice, string, brass and wind instruments are regarded as being of interest in vibrato research in [6]. Synthesizers and guitars are also capable of playing a vibrato. Instruments like the piano and other percussion instruments do not allow frequency modulation techniques.

The analysis of vibrato in musical performances is of interest for multiple purposes. In signal transformations like pitch shifting or time stretching vibrato parameters have to be known [7], [8]. Vibrato information can also be used to evaluate the skill of musical performers like in [9] and [10]. Since vibrato features strongly depend on the instrument they can be used as an additional feature for instrument recognition, too. In [11] vibrato and tremolo features of single partials have been used to detect singing voice components.

The problem of extracting vibrato from monophonic instrument recordings has been addressed in [12], where several methods have been proposed. One general approach for vibrato extraction is to look for note boundaries first and then to analyze the parts in between. Hence, a significant lookahead is necessary. This makes a real time application impossible.

In this work we propose an algorithm for vibrato detection, based on the block wise cross correlation coefficient between the FM - and the AM trajectory. In combination with the proposed method for the frequency modulation trajectory extraction, a note boundary detection is not needed.

2. ALGORITHM DESCRIPTION

2.1. Modulation Trajectory Extraction

The first step in the presented vibrato detection is the extraction of the FM- and AM trajectories. The fundamental frequency trajectory and the short term energy trajectory of the whole signal have to be freed from their steady state component, leaving only the modulating components. Since the motivation of this project is also a possible integration into the phase vocoder, the f_0 analysis parameters are chosen to be applicable in that technology. A window size of 25 ms and a hop size of 10 ms are used. The calculation of the fundamental frequency trajectory is performed in SuperVP, yet it could be obtained with any other f_0 calculation method like YIN or SWIPE.

For our purposes the fundamental frequency trajectory f_0 will be regarded as a composition of three parts

$$f_0 = f_{step} + f_{cor} + f_{mod} \tag{1}$$

with the leading component f_{step}, which contains the discrete value note information, the slow varying correction component f_{cor} which features glissandi and correction movements and finally the quasi periodic modulating components f_{mod}.

Extracting f_{mod} from the other components is a bandpass filtering problem. The high frequency parts of the note transitions and the low frequency parts of the notes mean values have to be eliminated to isolate the modulation trajectory which ranges from about 4 to 12 Hz, with a preferred frequency of 6 Hz. Applying normal bandpass filters with the chosen hop size for f_0 calculation causes errors, since the note transitions in performances lead to leaps in the fundamental frequency trajectory what causes overshoots when filtering with IIR filters.

As mentioned previously, most algorithms for vibrato detection and extraction of vibrato parameters regard single notes only or work with a prior note boundary detection. Approaches of separating the modulation trajectory from the fundamental frequency trajectory like this have been described in [7], [13] and more recently in [8]. In that case the mean value is extracted within each note to obtain f_{step}. The instantaneous vibrato amplitude can then

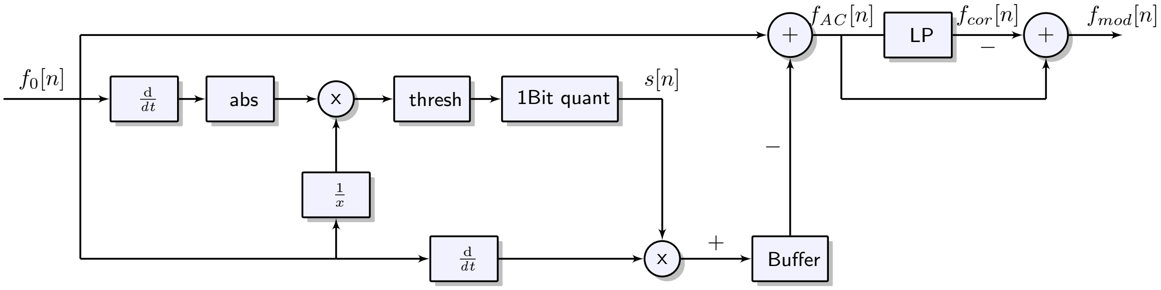


Fig. 1: Flow chart of the SOM - f_0 decomposition algorithm

be calculated as the deviation from the center (note) frequency. In our approach we extract the modulating trajectory without this prior analysis to increase the real time capability.

In the presented method a non linear bandpass filter which can be described as a derivative limiter is used. A block diagram of the principle entitled *Slope Overload Memory* (SOM) Filter is shown in Figure 1. Whenever the absolute relative derivative of the f_0-trajectory exceeds a threshold, the switch signal $s[n]$ which is usually 0 will be set to 1 and the corresponding f_0 derivative value will be added to an accumulating buffer. The value of this buffer, which is the sum of all f_0 derivative values which exceed the threshold, is then subtracted from all following f_0 values. The result of this procedure is the sum of f_{cor} and f_{mod} which is named f_{AC}. It contains all information with exception of the discrete note frequency. After a simple low pass filter with a cut off frequency at 4 Hz and a subsequent subtraction, the modulation trajectory f_{mod} is isolated. The first sample of the input signal always has to be set to 0, otherwise the output signal can be biased. In Figure 2 a resulting trajectory is shown. It can clearly be seen here that not all leaps in the f_0 trajectory are smoothed out. Yet for the parts between the note boundaries the clean sinusoidal modulations are left. Figure 3 shows both the spectra of the f_0 trajectory and the output of the SOM filter. It can clearly be seen here that the strong constant component in the f_0 trajectory is eliminated. In the resulting f_{mod} trajectory a significant peak around the reasonable modulation frequency of approximately 10 Hz is left.

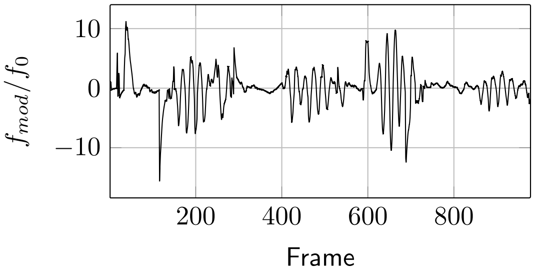


Fig. 2: Extracted modulation trajectory f_{mod} for a violin solo, relative to the note center frequency

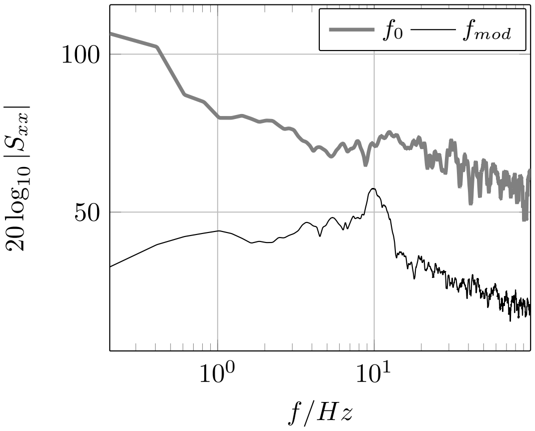
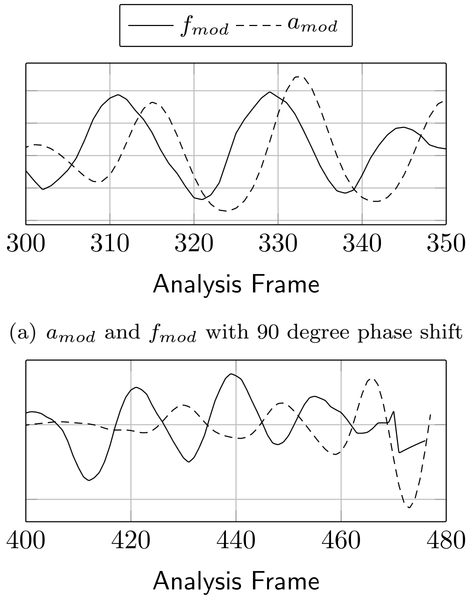


Fig. 3: Power spectral density of the fundamental frequency trajectory f_0 and the modulation trajectory f_{mod} computed with SOM filter for a violin solo

The basic parameter of the slope overload memory filter is the threshold for the maximum allowed relative deviation. The threshold values depend on the hop size L_{hop} of the f_0 calculation and on the content to be analyzed. Applied threshold values vary from 0.01 for violin to 0.04 for singing voice recordings. The latter usually contain large vibrato ranges and hence need a higher threshold in order not to smoothen out the vibrato.



(a) a_{mod} and f_{mod} with 90 degree phase shift

(b) a_{mod} and f_{mod} with 180 degree phase shift

Fig. 4: Amplitude modulation trajectories (AM) and frequency modulation trajectories (FM), both normalized, of subsequent notes from a violin recording

The extraction of the amplitude modulation trajectory a_{mod} is less complicated for the analyzed passages, since there are no leaps in the temporal energy with the parameters used. The temporal energy trajectory is also obtained with a hop size of 10 ms. Hence a simple bandpass filter is sufficient for extracting the amplitude modulation trajectory.
The obtained trajectories f_{mod} and a_{mod} are correlated in case of vibrato, however they do not have a fixed phase relation. Figure 4 shows how phase relations between amplitude- and frequency modulation trajectories can vary between two adjacent notes of the same violin recording. In 4(a) the two trajectories are shifted by about 90°, whereas in 4(b) the phase is shifted by about 180°. The phase between these signals could only be predicted if the frequency response of the resonating body of the instrument was known. For our purpose we assume that the phase between the two signals is arbitrary and hence not predictable.

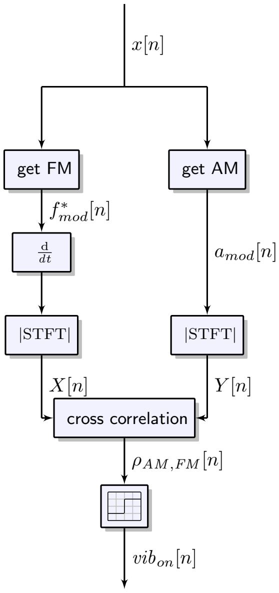


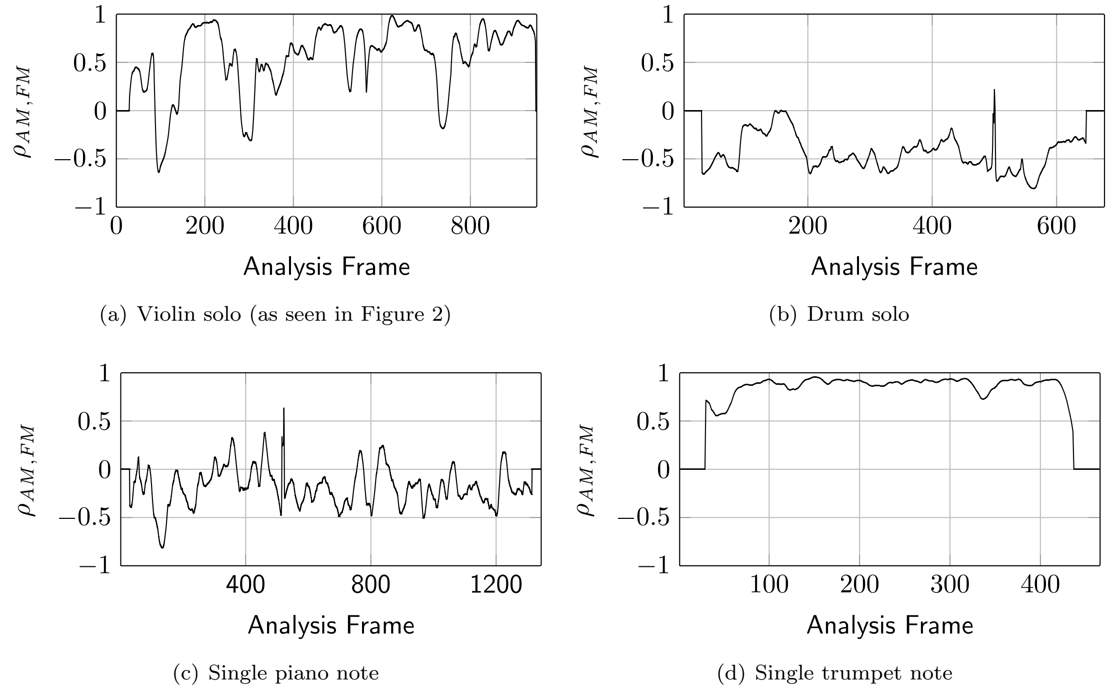
Fig. 5: Flow chart of the vibrato detection algorithm

2.2. Vibrato Detection

After the modulation trajectories have been extracted, the vibrato detection is performed. As we have mentioned before, it is widely known that a vibrato in musical instruments causes amplitude modulations. Moderate vibrato strengths of about ±35 musical cent can lead to amplitude modulations in the partials from 3 to 15 dB [14]. This relation will be used for the vibrato presence detection in our approach. Whenever the frequency modulation trajectory and the amplitude modulation trajectory are similar to each other, regarding their amplitude spectra, a vibrato is likely to be played. The amplitude spectra are used here, since the phase relations are ignored like this.
A block diagram of the vibrato detection algorithm is shown in Figure 5. After the relative modulation trajectory

$$f^*_{mod}[n] = f_{mod}[n]/f_{step}[n] \quad (2)$$

has been differentiated, an absolute STFT is calculated for both trajectories with a triangular window of $L_{\text{win}}60$ samples length and a hop size L_{hop} of 1



(a) Violin solo (as seen in Figure 2)

(b) Drum solo

(c) Single piano note

(d) Single trumpet note

Fig. 6: AM/FM cross correlation coefficient trajectories for different signals

sample. Since both trajectories have a sampling rate of 10 ms this leads to a look ahead of 300 ms. For each step the cross correlation coefficient $\rho_{AM,FM}[n]$ between the resulting signals $X[n]$ and $Y[n]$ is calculated, using Pearson's method:

$$\rho_{AM,FM}[n] = \frac{\mathrm{Cov}(X[n], Y[n])}{\sqrt{\mathrm{Var}(X[n])\mathrm{Var}(Y[n])}} \tag{3}$$

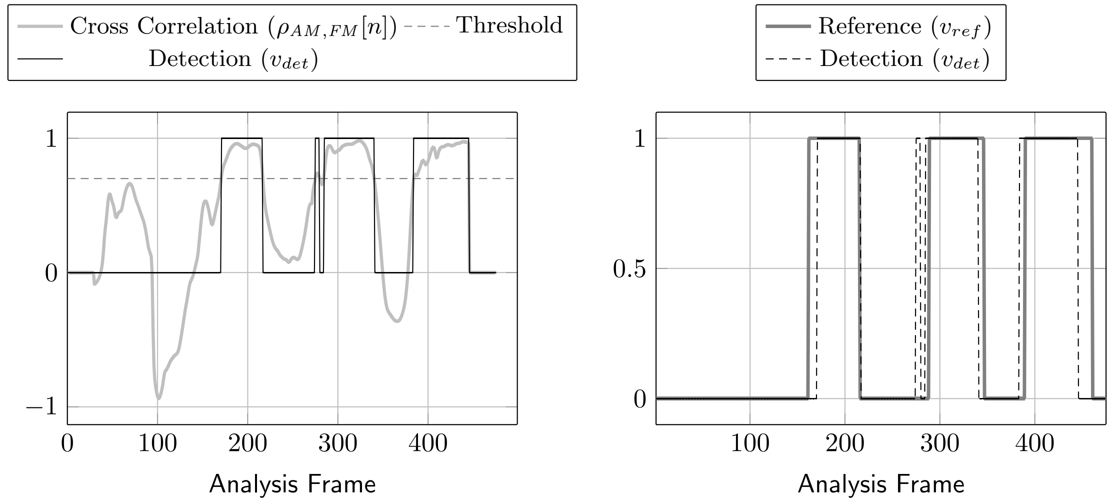
with $X[n]$ and $Y[n]$:

$$X[n] = \left|\mathcal{F}\left\{ f^{*'}_{mod}\left[\left(n - \tfrac{L_{win}}{2}\right)\dots\left(n + \tfrac{L_{win}}{2} - 1\right)\right]\right\}\right| \tag{4}$$

$$Y[n] = \left|\mathcal{F}\left\{ a_{mod}\left[\left(n - \tfrac{L_{win}}{2}\right)\dots\left(n + \tfrac{L_{win}}{2} - 1\right)\right]\right\}\right| \tag{5}$$

Figure 6 shows results of the AM/FM cross correlation coefficient trajectory $\rho_{AM,FM}[n]$ for different input signals. For the violin solo in Figure 6(a) long segments above a possible threshold of 0.5 can be seen. These regions are likely to contain vibrato. Significant local minima in this plot indicate note transitions. In case of the drum solo in Figure 6(b) no significant regions with a high cross correlation coefficient value can be seen since no vibrato is possible here. In fact this signal is polyphonic and does not always have a fundamental frequency. The results for a single piano note in Figure 6(c) also show no high cross correlation coefficient values. The piano also is a typical "no vibrato instrument", since the strings can note be modulated after excitation. Results for the single trumpet note in Figure 6(d) show a high cross correlation coefficient value throughout the whole sustain phase of the note which has been played with a strong vibrato.

The binary (on/off) vibrato detection vector v_{det} is then obtained by applying a threshold to the trajectory, as shown in Figure 7(a). All values below this threshold are set to 0 (no vibrato) and all values above the threshold are set to 1 (vibrato on). Recent experiments showed that a threshold of 0.35 to 0.6 leads to useful results. As results of the following evaluation section show, thresholds can be different for individual instrument groups.



(a) Vibrato detection by application of a threshold to AM/FM cross correlation

(b) Evaluation example for a violin passage (90% F-measure)

Fig. 7: Vibrato detection by applying a threshold (a) and example of comparison between detection results and reference (b)

3. EVALUATION

The vibrato detection algorithm is evaluated by comparing the results of the algorithm to manually created label files, applying the F-measure as used in [15] and [11] for similar tasks. As test set, 28 solo instrument passages with a length of 2 sec to 12 sec are used. The set is made of four instrument groups and features seven violin passages, seven woodwind passages, seven brass passages and seven singing voice passages. The musical content varies from classic to jazz performances. The material is mostly dry and without remarkable effects. Recordings within a group are taken from different individual instruments.

For the labeling process the audio software Audacity has been used with one audio track containing the raw wave file for the aural information, another audio track featuring a representation of the f_0 trajectory to give a visual help in labeling and a text track to place the labels in. In the label files the starting point and ending point of each vibrato segment was marked. Labeling was done by two persons independently.

One problem that arises is the correct labeling of the files. Since vibrato is usually appearing with a fade in and fade out it is not always clear where to mark the boundaries. Note transitions are another difficulty when labeling the vibrato regions. The interrelation between vibrato and note transitions has been addressed in [1]. In some cases the vibrato seems not to be discontinued when a note transition appears. That means it continues with the same phase and frequency. In other transitions the vibrato makes leaps in phase yet it is not discontinued. We decided to keep all note transitions as non vibrato segments.

Based on the label files, vectors v_{ref} are created which contain a 0 for non vibrato segments and a 1 for vibrato segments. These vectors have the same sampling rate as the detection results v_{det}. Figure 7(b) shows a comparison of a detection result and the corresponding reference vector. The evaluation is then performed using the F-measure:

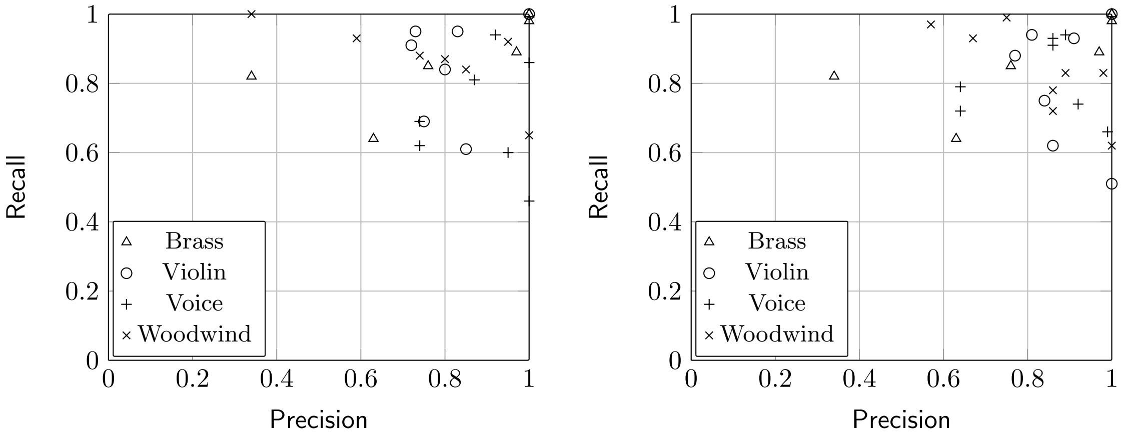
$$F = \frac{2PR}{P+R} \tag{6}$$

with the precision:

$$P = \frac{\textit{Number of vibrato frames detected correctly}}{\textit{Total number of frames containing vibrato}} \tag{7}$$

and the recall:

$$R = \frac{\textit{Number of non-vibrato frames detected correctly}}{\textit{Total number of frames containing no vibrato}} \tag{8}$$



(a) Results for common threshold (0.45)

(b) Results for individual group thresholds

Fig. 8: Test results in precision and recall

The algorithm was evaluated in two separate tests. In the first test all groups were evaluated with the same parameters. A threshold of 0.02 was chosen for the SOM filter and the cross correlation threshold was set to 0.45. This lead to a mean F-measure of 0.8000 with a mean precision of 0.8307 and a mean recall of 0.8107. The results are presented in Figure 8(a) as a scatter plot. Woodwind and Voice each show one significant outlier. The results at $(1,1)$ are samples which contain no vibrato at all (Violin, Woodwind). In these cases the algorithm shows no detection errors. The F-measure formula therefor had to be modified, since in a case of no vibrato frames it would lead to a devision of 0 by 0.

In a second test individual cross correlation thresholds were chosen for each instrument group. The values for the threshold, as well as the results, can be seen in Table 1. In Figure 8(b) the results are visualized in a scatter plot. Violin and Brass show slightly better performances than Woodwind and Voice. Over all groups a mean F-measure of 0.8207 could be reached. The algorithm also showed reasonable results for analog and digital synthesizers like a Roland Juno (F-measure = 0.9387) or a Yamaha DX7 (F-measure = 0.8535). In case of the Juno this is expected, since the voltage controlled filter represents a resonant body with resonant frequencies. For the DX7 no useful results were expected. It is not obvious whether this can be explained with the FM algorithm or the analog amplifier. Experiments with pure frequency modulated sines showed no correlation between AM and FM at all.

Table 1: Mean evaluation results for instrument groups with individual cross correlation threshold

Group	Threshold	F-Measure	Precision	Recall
Violin	0.45	0.8286	0.8843	0.8043
Woodwind	0.35	0.8087	0.8225	0.8337
Brass	0.5	0.8343	0.8143	0.8829
Voice	0.6	0.8114	0.8286	0.8129

4. DISCUSSION

It could be shown that the FM/AM cross correlation coefficient is a valid indicator for vibrato presence. Results of the evaluation look promising and the presented approach might be used for several music information retrieval tasks. The relation between AM and FM however depends on the individual instrument and the musical content. For a fixed set of parameters this could mean that the outcome of the cross correlation coefficient contains information about the instrument. Evaluating further statistical analysis on the cross correlation trajectory might lead to significant differences between different instruments and interpreters.

If a reliable vibrato detection is desired, the parameters will have to be adjusted according to the instrument. Up to this point some parameters have been tuned for the single instrument to achieve better results. To improve the algorithm an automatic parameter adjustment of the SOM filter threshold and the cross correlation threshold is necessary. The algorithm has not been tested using different values for the hop size and frame size in analysis. Robustness might be increased with a different set of parameters.

Future work will include the attempt towards a more robust vibrato detection, adjusting important parameters according to the input signals features as well as tests for using the cross correlation coefficient results in instrument recognition processes. In terms of vibrato detection the presented attempt will be combined with other features to ensure highest reliability, since the cross correlation trajectory itself might not lead to a robust solution.

Finally, an application of the presented approach to the partials amplitude modulation seems helpful because different partial amplitude modulations can compensate each other so that the overall AM is less meaningful. Appropriate combination of the results obtained from individual partial AM can help here and lead to better results.

5. REFERENCES

[1] Robert C. Maher. Control of synthesized vibrato during portamento musical pitch transitions. *J. Audio Eng. Soc*, 56(1/2):18–27, 2008.

[2] Perfecto Herrera and Jordi Bonada. Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proceedings of the Digital Audio Effects Workshop (DAFX98)*, 1998.

[3] I. Arroabarren, M. Zivanovic, , and A. Carlosena. Analysis and synthesis of vibrato in lyric singers. *Proc. 11th European Signal Processing Conference*, 2002.

[4] I Arroabarren, M Zivanovic, X Rodet, and A Carlosena. Instantaneous frequency and amplitude of vibrato in singing voice. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 5, pages 537–540, 06/04/2003 2003.

[5] Ixone Arroabarren, Xavier Rodet, and Alfonso Carlosena. On the measurement of the instantaneous frequency and amplitude of partials in vocal vibrato. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 2006.

[6] Vincent Verfaille, Catherine Guastavino, and Philippe Depalle. Perceptual evaluation of vibrato models. *Proceedings of the Conference on Interdisciplinary Musicology (CIM05)*, 2005.

[7] D Arfib and N Delprat. Selective transformations of sounds using time-frequency representations: An application to the vibrato modification. *AES Convention:104*, 1998.

[8] Axel Roebel, Simon Maller, and Javier Contreras. Transforming vibrato extend in monophonic sounds. *Proc. of the 14 th Int. Conference on Digital Audio Effects*, 2011.

[9] Tomoyasu Nakano. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. *INTERSPEECH*, 2006.

[10] Helen F. Mitchell and Dianna T. Kenny. Change in vibrato rate and extent during tertiary training in classical singing students. *Journal of Voice*, 24(4):427–434, 2010.

[11] L. Regnier and G. Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 1685–1688, Washington, DC, USA, 2009. IEEE Computer Society.

[12] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette. Vibrato: detection, estimation, extraction, modifcation. *Digital Audio Effects Workshop*, 1999.

[13] Hee-Suk Pang and Doe-Hyun Yoon. Automatic detection of vibrato in monophonic music. *Pattern Recognition*, 38(7):1135–1138, 2005.

[14] Jürgen Meyer. Musikalische Akustik. In Stefan Weinzierl, editor, *Handbuch der Audiotechnik*, VDI-Buch, pages 123–180. Springer Berlin Heidelberg, 2008.

[15] H Lukashevich, M Gruhne, and C Dittmar. Effective singing voice detection in popular music using arma filtering. *Audio*, pages 75–75, 2007.