BINAURAL PARTIAL TRACKING

Martin Raspaud and Gianpaolo Evangelista

Digital Media Group,
University of Linköping
Norrköping, Sweden
firstname.lastname@itn.liu.se

ABSTRACT

Partial tracking in sinusoidal models have been studied for over twenty years now, and have been enhanced, making it precise and useful to analyse noiseless harmonic sounds. However, such tools have always been used in a monophonic (single channel) context.

A method is thus proposed to adapt the partial tracking to the case of binaural signals. This gives a tool to perform spectral analysis of such signals, keeping relevant information from both left and right channels. Moreover, azimuth (position in the horizontal plane) information for each partial is gained using interaural cues, such as interaural time differences (ITDs) and interaural level differences (ILDs). The azimuth information can then be used as an attribute or as a constraint in the binaural partial tracking algorithm.

Finally, some classification results using the azimuth of partials are presented.

1. INTRODUCTION

Spectral models provide general representations of sound in which many audio effects can be performed in a very natural and musically expressive way. The analysis tool called partial tracking [1] has been widely studied and enhanced [2] over the years and can now be considered as rather robust in the case of noiseless harmonic sounds. To the best of our knowledge, partial tracking has always been applied to monophonic signals.

In this article, we present a new way to track partials in binaural contexts. Instead of tracking partials in a single signal, we perform the tracking in the left and right channels of binaural signals. This is done by tracking spectral peaks simultaneously in both left and right observation signals, while using the same base techniques as in the classical partial tracking algorithm. This gives 'stereo' partials, from which we can draw relevant data from either binaural channel.

In the meanwhile, techniques for binaural source localisation have been explored for a few years, showing promising results [3, 4]. Some of these techniques are based on level differences and phase delays between spectral peaks in the left and right channels of binaural recording, which we can obtain thanks to our stereo partials. It is then possible to obtain an accurate estimation of the azimuth (position on the horizontal plane) of each partial.

The azimuth can then be considered as a simple attribute of the partial, or can be used as a constraint for tracking the partial, in the same way as frequency is in classical algorithms. This opens the way to improved detection of overlapping partials, and thus to enhancements to the separation of such partials.

The azimuth of the partial is a very important cue for the purpose of Auditory Scene Analysis, since common direction of arrival is a relevant cue for source separation in the human brain [5]. Hence, we present some partial classification results based on this cue.

The work presented in this paper has been developped using binaural recordings, but it can be generalised to stereo recordings, as long as the azimuth estimation techniques can be adapted to stereophonic recordings.

In Section 2, we will present the classic partial tracking algorithm, followed by the presentation of our new algorithm for the tracking of partials in binaural signal. Section 3 will introduce the binaural based spacial cues, followed by the application of these cues to localisation of partials in the azimuth plane. In Section 4, we will show some classification results based on the azimuth of the partials. We will then conclude and present our future work.

2. SINUSOIDAL MODELLING

2.1. Model and Parameters

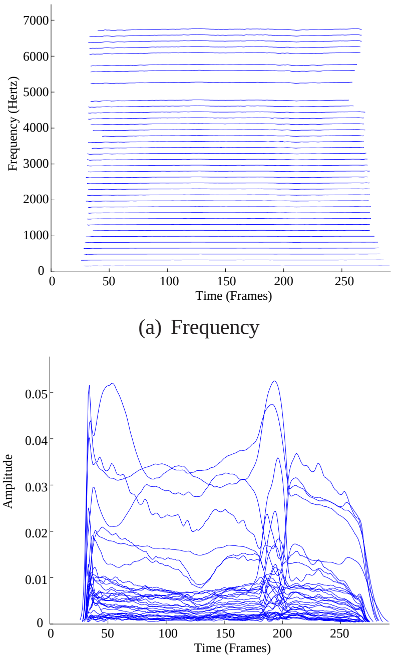
Additive synthesis is a spectrum modelling technique. It is rooted in Fourier's theorem, which states that any periodic function can be modelled as a sum of sinusoids at various amplitudes and harmonic frequencies. For stationary pseudo-periodic sounds, these amplitudes and frequencies continuously evolve slowly with time, controlling a set of pseudo-sinusoidal oscillators commonly called *partials*. This is the well-known McAulay-Quatieri representation [1]. The audio signal a can be calculated from the additive parameters using Equations (1) and (2), where P is the number of partials and the functions f_p, a_p, and ϕ_p are the instantaneous frequency, amplitude, and phase of the p-th partial, respectively. The P pairs (f_p, a_p) are the parameters of the additive model and represent points in the frequency-amplitude plane at time t. This representation is used in many analysis / synthesis programs such as Lemur [6], SMS [7], or InSpect [8].

$$a(t) = \sum_{p=1}^{P} a_p(t)\ \cos(\phi_p(t)) \tag{1}$$

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u)\, du \tag{2}$$

2.2. Mono Partial Tracking

This model requires an analysis method in order to extract the parameters of the partials from sounds which were usually recorded in the temporal model, that is audio signal amplitude as a function of time. The accuracy of the analysis method is extremely important since the perceived quality of the resulting spectral sounds



(a) Frequency

(b) Amplitude

Figure 1: *Frequency (top) and amplitude (bottom) trajectories of the partials of a note of alto saxophone, performed with a classical partial tracking algorithm.*

depends mainly on it. Moreover, the main interest of an accurate analysis method, providing precise parameters for the model, is to allow ever deeper musical transformations on sound by minimising deformations due to analysis artifacts.

The analysis method we use consists of two steps: spectral peaks are first extracted from the sound using a short-time spectral analysis, then these peaks are tracked from frame to frame in order to reconstruct the partials.

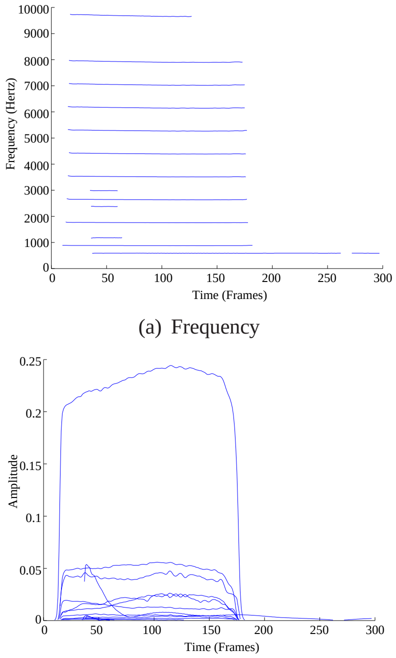
2.2.1. *Extraction of Spectral Peaks*

First, a short-time Fourier analysis produces a series of short-term spectra taken on successive temporal windows on the original signal. Information about the local maxima in magnitude (so-called peaks) is then extracted from these short-term spectra using the derivative algorithm proposed in [9], in order to provide the model with accurate spectral parameters (frequency, amplitude, and phase).

As for the practical side of this analysis, we used an analysis window of 2048 samples, moving by steps of $H = 512$ samples. These settings were chosen as a good compromise between time and frequency resolutions for our sound source separation objective. The test sound used for Figures 1(a) and 1(b) was a 16-bit, 44100-Hz mono recording of an alto saxophone playing at a fundamental frequency around 165 Hz with vibrato and tremolo. The length of each analysis window is thus about 50 ms and the resolution of the resulting Fourier spectrum is approximately 20 Hz.

2.2.2. *Tracking of Partials*

Since the short-time Fourier analysis delivers a short-time spectral representation of the analysed sound, we consider local maxima in the magnitude spectrum (so-called peaks, see above) to be the instantaneous representation of partials. We have then to link peaks of successive frames to recover the continuous evolution of the partials. For this purpose, we use the enhanced partial-tracking algorithm proposed in [2, 10]. This algorithm improves the classic McAulay-Quatieri algorithm [1] by using linear prediction in order to forecast, from their past, the future evolutions of the trajectories of the partials.



(a) Frequency

(b) Amplitude

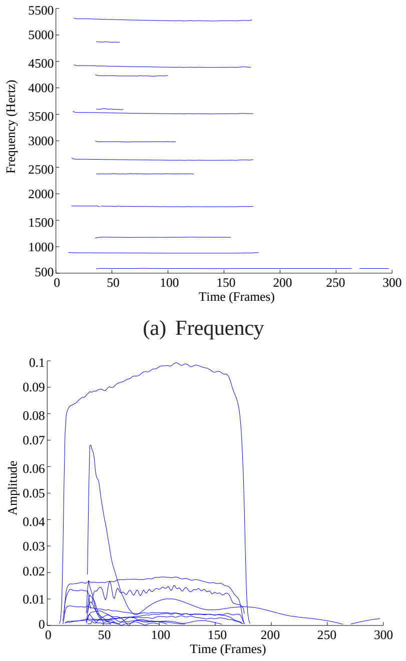
Figure 2: *Partials (frequency (top) and amplitude (bottom) trajectories) of a mix of a piano note on the right and a clarinet note on the left (starting first), left channel. Here the clarinet partials are clearly seen, while the piano partials are barely present.*

As for the practical side of this analysis, the maximal frequency difference between two successive frames for each partial was set to $\Delta = 1\%$ of the current frequency. Partials whose amplitude are always below -60 dBFs or whose length is smaller than 0.2 s are considered as noise, since we are interested only in reliable – long and strong – partials.

An example of the result of mono partial tracking is shown in Figures 1(a) (frequency trajectories of the partials) and 1(b) (amplitude trajectories of the partials).

2.3. Binaural Partial Tracking

In the context of binaural recordings, a recording provides two observation signals of the same sonic environment. In the case of a mix of several instruments playing together, these two observations might give different information about the same instrument. For example, Figures 2(a), 2(b), 3(a), and 3(b) show the frequencies and amplitudes of the partials found in the left and right channels of a binaural mix of a clarinet note (placed 40°on the left, starting first) and a piano note (placed 40°on the right). As we can see, the partials of the piano are logically stronger on the right channel, while the clarinet partials are stronger on the left channel.



(a) Frequency

(b) Amplitude

Figure 3: *Partials (frequency (top) and amplitude (bottom) trajectories) of a mix of a piano note on the right and a clarinet note on the left (starting first), right channel. Here the piano partials are clearly seen and strong, while the clarinet partials are only available up to about 5500 Hz.*

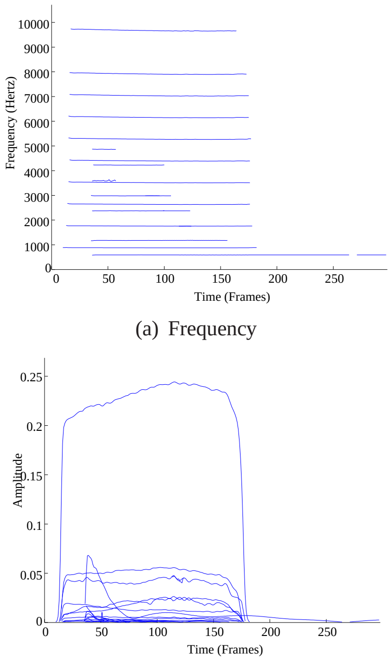
The objective then is to gather the most interesting information that each channel has to offer. In our previous example, that would be to keep the clarinet partials from the left channel and the piano partials from the right channel, while estimating the direction of arrival of each partial using the previously presented spacial cues.

2.4. Tracking Partials in Two Observation Signals

In order to realize meaningful partial tracking for binaural recordings, partial tracking has to be performed in parallel in the two observation signals, so what we obtain is a set of 'stereo' partials, containing information of amplitude, frequency, and phase from both left and right channels.

This can be done by enhancing the classical partial tracking algorithm to handle stereo. The first step is to use stereo spectral peaks, which contain for each spectral peak the frequency, amplitude and phase information of both channels. Next, matching of the spectral peaks to stereo tracks is done using the frequency, amplitude and phase from the loudest channel of the peak. Hence, a peak having a higher amplitude from the left channel information will use this information in the partial-to-peak matching phase. Finally, the matched peaks are added to the corresponding partials. These 'stereo' partials hold the information of both channels, but it is only the information from the channel with highest mean amplitude that is used when needed.

Using these enhancements, we then obtain the most relevant information from each channel. Figures 4(a) and 4(b) show the frequency and amplitude information gathered this way. Comparing these to the Figures 2(a), 2(b), 3(a), and 3(b), we can see that both the clarinet and the piano partials are shown to their full length and strength using the stereo partial tracking.



(a) Frequency

(b) Amplitude

Figure 4: *Frequency (top) and amplitude (bottom) trajectories of the piano-clarinet mix, using the stereo partial tracking. Both notes have been tracked and relevant information from both channels have been retained.*

Next we will show how we can make use of the second part of the stereo information to track the position of the partial in space.

3. BINAURAL SOURCE LOCALISATION

3.1. Spatial Cues

Binaural recordings of sound provide two different observations of the sonic environment[1]. As presented for example by Viste and Evangelista [3], we can define binaural cues that will give us some indication of the location of the content of the environment. We will use two such cues here, namely the ILD and ITD. These two cues are based on the sliding short-time Fourier transform (STFT) of the two observations.

The ILD (in dB) at the n-th frame is defined as follows:

$$\Delta L_n(\omega) = 20 \log_{10} \left| \frac{S_n^r(\omega)}{S_n^l(\omega)} \right| \tag{3}$$

where ω is the frequency and S_n^r and S_n^l respectively are the STFTs of the right and left channel of the binaural signal s. ΔL_n is thus simply the ratio in dB of the amplitudes of the right and left STFTs, i.e. the difference of the amplitudes in dB of the right and left STFTs.

[1] We make the assumption that the signal of a given source is present in both observations of the sonic environment. Without this assumption, the spacial cues cannot be computed correctly.

Also based on the right and left spectra of the n-th frame, we define the ITD (in seconds) as:

$$\Delta T_{n,p}(\omega) = \frac{1}{\omega}\left(\angle\frac{S_n^r(\omega)}{S_n^l(\omega)} + 2\pi p\right) \tag{4}$$

with p as the phase unwrapping factor. The use of this factor is made necessary by the fact that the angle of the spectra ratio is computed modulo 2π. This thus makes the phase become ambiguous above a certain frequency, which is dependent on the size and shape of the head mainly, and is averaged to 1500 Hz.

3.2. Estimation of Azimuth

In order to estimate the azimuths, two methods were proposed in [3] : looking up in a reference table, or using a model. In order to be as generic as possible, in this paper we will take the model based approach, since the lookup table implies knowledge of the subject's head-related transfer function (HRTF). This model allows for a simpler computation of azimuths, but at the cost of decreased accuracy.

As a basis for the estimation of the parameters of this model, we use the CIPIC HRTF Database [11].

3.2.1. Interaural Time Differences

The model we use is the following:

$$\Delta T_s(\theta,\omega) = \beta_s(\omega) r \frac{\sin\theta + \theta}{c} \tag{5}$$

where r is the "head radius", and c is the wave propagation speed (344 m/s). We make use of the frequency-dependent scaling factor $\beta_s(\omega)$. This scaling factor is first estimated individually for each subject, and then is averaged to be used in this generic model.

3.2.2. Interaural Level Differences

Based on a study of the HRTFs in the CIPIC database [11], Viste and Evangelista [3] propose the following model:

$$\Delta L_s(\theta,\omega) = \alpha_s(\omega)\sin\theta \tag{6}$$

with frequency dependent scaling factor $\alpha_s(\omega)$. Here again, the scaling factor used is an average over all the subjects.

3.2.3. Computation of the Azimuth

In order to retrieve the azimuth from the spectra using this method, we have to inverse equations 6 and 5 such that:

$$\theta_{L,n}(\omega) = \arcsin\frac{\Delta L_n(\omega)}{\alpha(\omega)} \tag{7}$$

$$\theta_{T,n,p}(\omega) = g^{-1}\left(\frac{c}{r\beta(\omega)}\Delta T_{n,p}(\omega)\right) \tag{8}$$

where $\Delta L_n(\omega)$ and $\Delta T_{n,p}(\omega)$ are defined respectively in equations 3 and 4, and g^{-1} is the inverse function of $g(\theta) = \sin\theta + \theta$. This function cannot be inverted algebraically. However, using a Chebyshev series, we can compute a polynomial approximation $\tilde{g}$ of g over the interval of interest, then inverse it:

$$\tilde{g}^{-1}(x) = \frac{x}{2} + \frac{x^3}{96} + \frac{x^5}{1280} \tag{9}$$

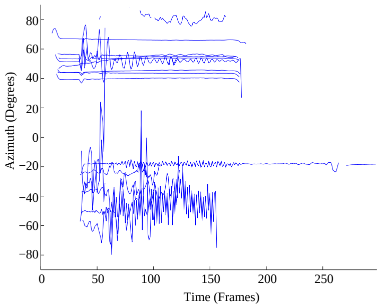


Figure 5: *Azimuth trajectories of the partials based on level differences. Mix of a clarinet, 40° on the left, starting first, and a piano note, 40° on the right.*

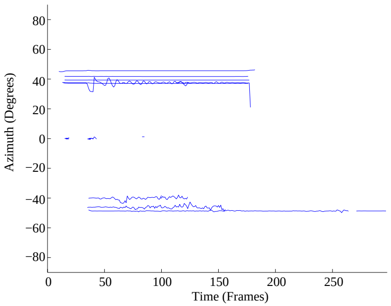


Figure 6: *Azimuth trajectories of the partials based on phase delays. Mix of a clarinet, 40° on the left, starting first, and a piano note, 40° on the right. The azimuth trajectories are less dispersed thanks to the enhanced precision of the joint azimuth estimation method.*

In practice we use this approximation in Equation (8).

Using this model, the estimation thus becomes continuous along the azimuth axis.

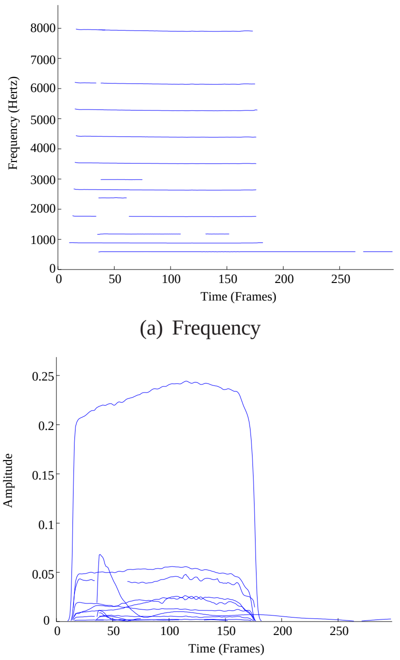
3.3. Tracking the Azimuth

In this article, we consider that sound sources are placed spatially on the horizontal plane, and that they come from the front of the subject (-90° to 90°).

As shown before, spatial cues can be estimated using the information from both left and right channels. During the stereo partial tracking, we gather information from each channel on both amplitude and phase, which are needed in order to compute the interaural spatial cues, ILD and ITD.

Hence, at the same time as the partial tracking is taking place, at each frame we compute the ILD and ITD. These cues are then in turn used to compute the azimuth of each stereo spectral peak, as shown in Equations (7) and (8).

The azimuth we obtain from these cues are however not ideal. Indeed, the azimuth computed from the ILD, based on the amplitude ration, is noisy. On the other hand, the ITD, based on the phase delay, gives a more precise estimation, but it is ambiguous at a wavelength smaller than the diameter of the head. However, as shown in [3], it is possible to obtain a more precise and non-ambiguous estimation of the azimuth using a joint estimation process: the noisy azimuth based on the ILD is used to disambiguate the more precise azimuth based on the ITD. An example of this is given in Figures 5 and 6. In the first figure, the azimuth trajecto-



(a) Frequency

(b) Amplitude

Figure 7: *Using azimuth as a constraint for peak matching in the binaural partial tracking algorithm. Resulting frequency (top) and amplitude (bottom) trajectories.*

ries of some of the partials of the mix, computed using the ILD, is shown. We can recognise there the two notes, vaguely grouped in two sets, one on the left, starting earlier, which is the clarinet note, and a second set on the right, which is the piano note. These trajectories are noisy, and hardly usable. On the second figure however, using the joint azimuth estimation, we obtain much more precise trajectories.

It has to be noted however that we show in Figures 5 and 6 only the partials with a mean frequency lower than 6000 Hz. This is due to the constraints of the ILD model. Indeed, above that frequency the $\alpha(\omega)$ parameter is varying greatly from subject to subject, making the average $\alpha(\omega)$ parameter inaccurate, and thus creating very noisy azimuth trajectories. Hence, only the partials with a frequency lower than 6000 Hz are considered reliable as far as the azimuths are concerned.

3.4. Azimuth as a Tracking Constraint

Until now, we have only considered the azimuth of the partials as an attribute of the stereo partials that help locate the partials in space. However, let us consider the case of overlapping harmonics of notes from two distinct sources (in space). In this case, in the binaural partial tracking system we proposed, only the continuity in frequency is the criterion for the tracking of a partial: a spectral peak is considered to be matching a partial only if it is within a 10 Hz range of the predicted spectral peak. Hence, a partial tracking will not be discontinued when an overlap occurs. This makes it at least difficult to detect overlapping at least, and even more difficult to separate overlapping partial.

However, in the light of the azimuth estimation we presented above, we can now put on constraints not only on frequency, but also on azimuth during the peak matching phase. This leads to interesting results for overlapping partials. Figures 7(a), 7(b), and 8 show our piano and clarinet example one more time, except that this time from one frame to the next, each partial is not moving in azimuth more than one degree. We can see on the figures that adding this constraint degrades the partial tracking, but sometimes in an interesting way. For example, we notice that the fourth partial from the bottom on Figure 7(a) (just below 2000 Hz) is missing a part at the onset of the piano note. This is of course not a coincidence, since it comes from the fact that the azimuth of the partial if strongly modified due to the overlapping of this clarinet partial with one of the piano's. Hence, this could be used as a tool for overlapping detection and thus also for disambiguation (such as presented in [12] for example). One could also use this azimuth constraint technique to tracks moving sources.

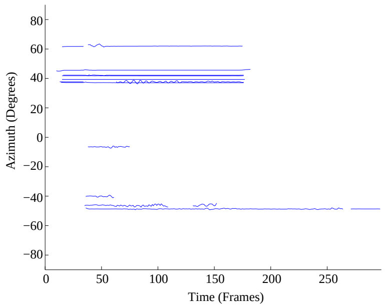


Figure 8: *Using azimuth as a constraint for peak matching in the binaural partial tracking algorithm. Resulting azimuth trajectories.*

Overlap	# Cases	Correct in %	Error in %
0	76	93.3 (12.9)	0.24 (1.50)
1	71	92.6 (11.8)	4.43 (10.28)
2	28	94.9 (9.0)	7.63 (10.74)
3	20	94.3 (10.4)	14.21 (20.48)
4	9	91.5 (18.8)	12.63 (18.31)
more than 5	6	99.2 (1.3)	30.20 (32.52)
1.2238	**210**	**93.5 (12.0)**	**5.36 (13.03)**

Table 1: Results of a simple classification of partials, using their azimuth as classification criterion. The numbers in parentheses are the standard deviation of the result.

4. CLASSIFICATION

A logical application of binaural partial tracking is the classification of the partials using the spatial cues. In order to test the results of such a classification, we have set an experiment to classify about 200 mixed pairs of sounds (results given in Table 1).

For each case of the experiment, we take two mono sounds and analyse each of them using the regular mono partial tracking algorithm. This gives us the reference partial set of each sound.

Then, we perform binaural mixing of the two sounds at angles -30°and 30°. We analyse the mix, and gather the stereo partials in two sets according to their azimuth: one set that is closer to -30°and the other set that is closer to 30°.

We then compare the new found sets to the reference sets. A partial of an experimental set is consider to match a partial from a reference set if the frequencies are within a 10 Hz range over the length of the common partials length. If there is a match, the common length is then added to the score of the matching of the experimental set to the reference set. Eventually, the scores are divided by the total length of all the partials of the experimental set, so that the results are given in percentage. Another way of putting it is that we measure the pourcentage of the partials (of a given test set) corresponding respectively to the correct and erraneous reference partial sets in order to get the correctness and error percentage.

The sounds used for this experiment are sounds from the Iowa database [13]. These are harmonic sounds recorded in a low-noise environment. We have used sounds of Bassoon, Cello, Clarinet, Flute, Oboe, Piano, Saxophone, Trombone and Trumpet, starting at the same time in order to have a maximum time overlap. Notes range from E3 to C6. The considered partials are below 6000 Hz in order to avoid errors due to the ITD and ILD unreliability in the model.

Table 1 gives the results of our experiment. The first column shows the number of overlapping partials between the two reference sets, and the second column shows the number of cases it occurred in. The third column gives the percentage of correctness, that is the amount of partials that are matched from the experimental set to the correct reference set. In the fourth column, the error in percentage is given, that is the amount of partials that are matched from the experimental set to the wrong reference set. In both the third and fourth column, the standard deviation of the result is shown in parentheses. The final line shows the results over all the tests.

Table 1 shows that this simple classification procedure leads to very good results. The correctness is above 90% in all situations, while the error stays quite low until four partials are overlapping. We can see that, logically, the error raises as the number of overlapping partials increases. Indeed, the number of partials that can be matched in both sets increases with the number of overlapping partials, resulting in erroneous matches. In the case of more than 5 partials overlapping, we can see that the result is hardly meaningful since the error reaches more than 30%.

The classification experiments we show here are quite simple. However, the obtained results are sufficiently promising to conclude that spatial cues have a positive impact over classification and separation of sounds.

5. CONCLUSION

We have presented a method for coherent tracking of partials in a binaural context, along with a method to estimate the localisation of the partial in the horizontal plane. The method allows the detection of overlapping partials, and the classification of partials. The next steps will be towards the use of these techniques as a gathering tool for partials, while enhancing the estimation of azimuths for higher frequency partials.

6. REFERENCES

[1] Robert J. McAulay and Thomas F. Quatieri, “Speech Analysis/Synthesis Based on a Sinusoidal Representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[2] Mathieu Lagrange, Sylvain Marchand, Martin Raspaud, and Jean-Bernard Rault, “Enhanced Partial Tracking Using Linear Prediction,” in *Proceedings of the 6th Int. Conference on Digital Audio Effects (DAFx,04)*. "Queen Mary University of London", September 2003, pp. 141–146.

[3] Harald Viste and Gianpaolo Evangelista, “Binaural source localization,” in *Proceedings of the 7th Int. Conference on Digital Audio Effects (DAFx,04)*, Naples, Italy, October 2004, "Federico II University of Naples", pp. 145–150.

[4] Harald Viste, *Binaural Source Localization and Separation Techniques*, Ph.D. thesis, "École Polytechnique Fédérale de Lausanne (EPFL)", Switzerland, 2004.

[5] Albert S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, 1990.

[6] Kelly Fitz and Lippold Haken, “Sinusoidal Modeling and Manipulation Using Lemur,” *Computer Music Journal*, vol. 20, no. 4, pp. 44–59, Winter 1996.

[7] Xavier Serra, *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122, Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997.

[8] Sylvain Marchand and Robert Strandh, “InSpect and ReSpect: Spectral Modeling, Analysis and Real-Time Synthesis Software Tools for Researchers and Composers,” in *Proceedings of the International Computer Music Conference (ICMC)*, Beijing, China, October 1999, International Computer Music Association (ICMA), pp. 341–344.

[9] Myriam Desainte-Catherine and Sylvain Marchand, “High Precision Fourier Analysis of Sounds Using Signal Derivatives,” *Journal of the Audio Engineering Society*, vol. 48, no. 7/8, pp. 654–667, July/August 2000.

[10] Mathieu Lagrange, Sylvain Marchand, and Jean-Bernard Rault, “Long Interpolation of Audio Signals Using Linear Prediction in Sinusoidal Modeling,” *Journal of the Audio Engineering Society*, vol. 53, no. 10, pp. 891–905, October 2005.

[11] V.R. Algazi, Richard O. Duda, D.M. Thompson, and C. Avendano, “"the cipic hrtf database",” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, New York, USA, October 2001, pp. 99–102.

[12] H. Viste and G. Evangelista, “Separation of Harmonic Instruments with Overlapping Partials in Multi-Channels Mixtures,” in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-03)*, New Paltz, NY, Oct. 2003, pp. 25–28.

[13] “The Iowa Music Instrument Samples,” Online. `http://theremin.music.uiowa.edu`.