Audio Engineering Society
Convention Paper
Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

Partial Tracking based on Future Trajectories Exploration

Mathieu Lagrange[1], Sylvain Marchand[2], Jean-Bernard Rault[1]

[1]*France Telecom R&D 4, rue du Clos Courtel, BP 59 F-35512 Cesson Sevigne cedex, France*

[2]*LaBRI, Universite Bordeaux 1, 351 Cours de la Liberation, F-33405 Talence cedex, France*

Correspondence should be addressed to Mathieu Lagrange (mathieu.lagrange@rd.francetelecom.fr)

ABSTRACT

This paper introduces a partial-tracking algorithm suitable for the sinusoidal modelling of polyphonic sounds. A new method, based on the backward exploration of possible extensions of the partials in future frames, is proposed to cope with the lack or corruption of spectral data. The allocation of spectral peaks to a partial is done by considering possible trajectories in future frames where frame hoping is allowed. A suitable transition probability that takes into account missing or rejected peaks is proposed. The trajectory that exhibits the highest probability is searched for and the corresponding peak for the current frame is chosen to extend the partial.

1. INTRODUCTION

Spectral sound models provide general representations for many applications such as compression, content extraction and transformation. Most of these models, such

as additive synthesis, are based on the Fourier analysis which has proven to be accurate under the condition of local stationarity.

Additive synthesis is the original spectrum modeling technique. It is rooted in Fourier's theorem, which states that any periodic function can be modeled 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*. The audio signal s can be calculated from the additive parameters using Equations 1 and 2, where n_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 n_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 SMS [1] or *InSpect* [2].

$$s(t) = \sum_{P=1}^{n_P} a_P(t)\ \cos(\phi_P(t)) \quad (1)$$

$$\phi_P(t) = \phi_P(0) + 2\pi \int_0^t f_P(u)\ du \quad (2)$$

Thanks to the enhanced Fourier transform based on the derivative of the signal [3], the precision of parameters of selected peaks in the spectrum is quite high. Since the short-term analysis uses a sliding time/frequency window, the resulting representation is discrete. For applications such as time scaling and pitch shifting of monophonic sources, this discrete representation is sufficient. But for many other applications, a continuous representation of the sinusoidal components of the sound is very useful. Low bit-rate audio coding (HILN and SSC) can be done using a sinusoidal model since the parameters controlling oscillators are slow-time varying and can be encoded very efficiently [4, 5]. Recently, this model has been also used for musical transcription and source separation [6]. The temporal integration of informations provided by the continuous representation of spectral components is used to interprete complex spectral data of a polyphonic sound mixture.

To analyse harmonic monophonic sounds, the size of the analysis window to be used can be adapted to an estimate of the pitch of the source. The frequency resolution is then sufficient to separate harmonics and the resulting time resolution is optimal, see Figure 1.1. The partial tracking can be done efficiently by linking a peak with the nearest frequency peak neighbor in the next frame as proposed by Mac Aulay and Quatieri in [7]. During the analysis of a polyphonic sound mixture, a good frequency resolution is required so that the size of the window should be fixed to an arbitrary high value, therefore breaking local stationarity condition. Intermodulations between sinusoidal components may lead to spurious peaks or some peaks may be missing, blurring the spectral representation, see Figures 1.2, 1.3. The task of tracking partials over time is then much more complex. Considering evolutions of sinusoidal components over several frames can be useful to avoid local disturbances, as in the Hidden Markov Model algorithm (HMM) proposed in [8].

An overview of these two tracking methods is first given, the Mac Aulay and Quatieri algorithm (MAQ) in Section 2 and the HMM algorithm in Section 2.3. Section 3 provide an overview of the tracking algorithm proposed in this article. The peak-to-peak distance is introduced in Section 4 and a statistical approach to extend this distance to allow frame hopping is proposed in Section 5. This new distance is then used to generate a set of trajectories in future frames that will be used as guides for existing partials, as described in Section 6. Results follow in Section 7.

2. MAQ ALGORITHM

The algorithm is based on the assumption that partials composing a voiced speech signal have stationary frequency evolutions. It is then proposed to consider frequency differences between peaks of immediate successive frames to form partials. A maximal frequency difference threshold Δ_f between successive peaks of a partial is set:

$$|f_i^k - f_j^{k+1}| < \Delta_f \quad (3)$$

where f_i^k is the frequency of the i^{th} peak of frame k.

2.1. Basic Algorithm

The algorithm operates iteratively frame by frame and by increasing frequency. For a peak ρ_i^k of index i and frame k, we look for an unlinked peak ρ_j^{k+1} such that the frequency difference between those peaks is minimal. If the frequency difference is greater than Δ_f, the current partial is labeled as "dead". Else, ρ_j^{k+1} is selected.

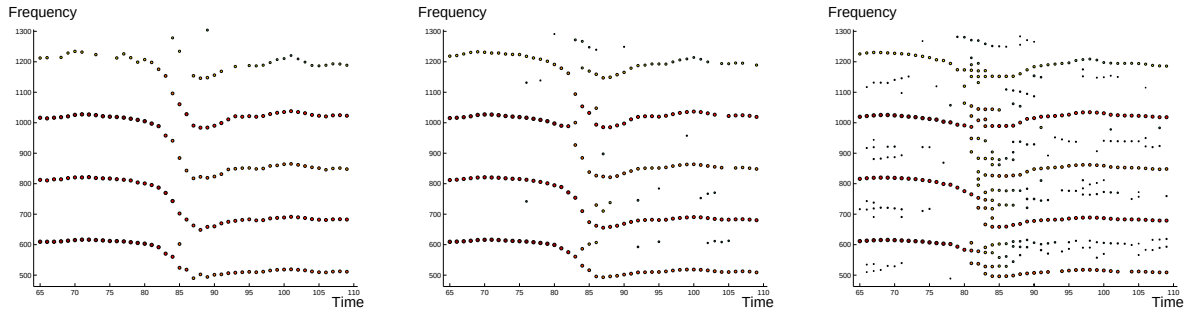

Fig. 1: Spectral peaks (spectrum local maxima) of a singing voice analyzed at a 512 samples frame rate using a window of 1024, 2048 and 4096 samples respectively. As the size of the window grow, the temporal smearing is getting more and more pronounced.

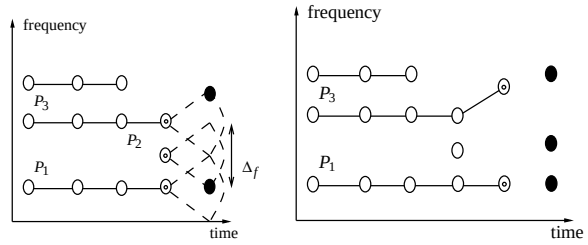

Fig. 2: *One step of the MAQ algorithm. White peaks cannot be candidate, whereas black peaks can. Peaks with dots are "tails" of partials.*

If this peak cannot be better linked with ρ_{i+1}^k, the partial having peak ρ_i^k is extended to peak ρ_j^{k+1}. If not, the current partial looks for another candidate in the next frame. If no alternative can be found (partial P_2 in Figure 2), the partial is also labeled "dead". After all reachable peaks of frame $k+1$ are linked, unlinked peaks of frame $k+1$ give rise to new partials.

2.2. "Zombie" extension

For various reasons such as decreasing amplitude, strong modulations or Fourier's transform bin corruption, the peak selection process can discard peaks [9]. This leads to missing peaks. To overcome this analysis drawback, it is proposed in [1] to add a "zombie" state to partials, so that if a partial cannot link to any peak in a frame, it can still look for a peak candidate in the next frames. If a peak can be found, the parameters of "zombie" peaks are interpolated. This extension is very useful to obtain a more concise set of partials.

2.3. HMM Algorithm

It is proposed in [8] to use the HMM formalism to track partials in musical sounds. This approach is very innovative because it tends to optimize a global tracking criterion over several frames. An HMM is composed of a set of states linked by transitions and a set of observation linked to states by observation probabilities. In this approach, a state is defined as the results of tracking between two immediately successive frames.

The transition probability between two peaks is computed jointly over time / frequency and time / amplitude planes considering the slope of parameters between the two last inserted peak. The transition probability between two states is then the product of transition probabilities between all linked peaks between the two considered frames. Hence, to find the "optimal" tracking, the sequence of states maximizing the product of transition probabilities between successive states over the considered frames is found by means of the Viterbi algorithm [10].

As the number of peaks in frames grows, the number of possible combinations between states can be very important, so it seems difficult to consider also transitions between peaks of non adjacent frames. However, frame hopping capability is of great importance to avoid a spurious peak at a given frame so as to to link with a good one in the next frames.

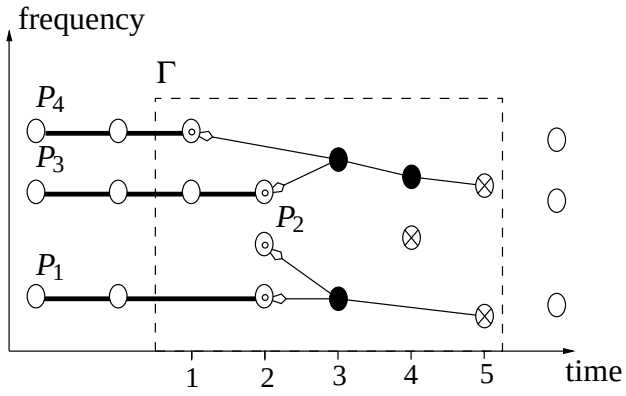

Fig. 3: *Backward incremental generation of optimal trajectories inside* Γ. *Note that optimal trajectories can share peaks. Unfilled peaks are inactive, double circles indicate the last inserted peak in a partial (the tail), Partials are plotted with bold lines, optimal trajectories with thin arrows.*

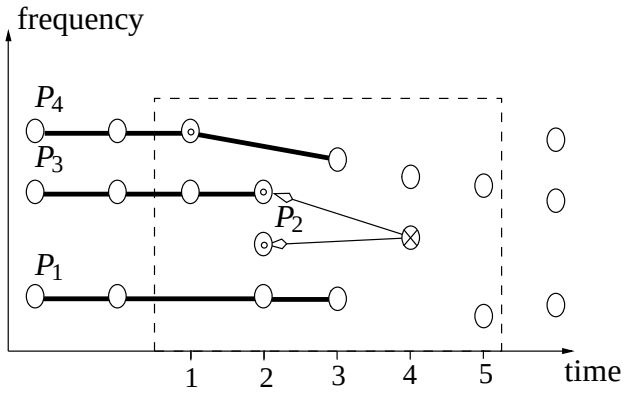

Fig. 4: *Second round of a tracking step.*

3. OVERVIEW OF THE PROPOSED METHOD

The analysis of possible evolutions of partials over numerous frames is a solution to cope with local spectral artifacts. Let be a set of partials tracked until frame t, some of these partials may be extended using some peaks of frame $t+1$. This choice should be made according to possible evolutions in future frames. The principle of the proposed method is to build backward short trajectories going through peaks of a restricted number of future frames called Γ according to a given transition probability distance. These trajectories finally link with the tail of a partial, indicating the optimal extension for this partial, see Figure 3.

A peak-to-peak distance is computed using spectral informations of the extracted peaks as detailed in Section 4. This distance cannot be used in the case of frame hopping since it does not consider frame distance between the two peaks: two peaks having similar spectral properties may be linked even if several frames separate them. To address this problem, we propose in Section 5 a statistical approach which, given a spectral distance and a frame distance between two peaks, computes a transition probability between those peaks.

Trajectories in future frames are computed with an algorithm described in Section 6.1 that takes advantage of this transition probability. Partials then select the trajectory that ends at their tail, see Figure 3. Partials having the best trajectory (P_1 and P_4 in Figure 4) lock their trajectory and link with the first peak of the trajectory if this peak is in frame $t+1$, a peak having interpolated parameters is considered otherwise. The peaks used in this trajectory are not considered any more during this tracking step. Trajectories that go through remaining peaks are then computed and assigned until no extension can be done for existing partials, see Figure 4. If a partial has not extended itself like partial P_2, it is considered as dead and removed from the tracking process. All unlinked peaks of frame $t+1$ give rise to new partials.

4. PEAK-TO-PEAK DISTANCE

It is considered in the MAQ algorithm that the frequency of partials should be constant over time. Yet, the evolutions of partials in frequency and amplitude may varies - due to musical modulations - but not chaotically. For example, there is a strong correlation between the evolution of the amplitude and the frequency of a natural vibrato, as can be seen in Figure 5. Note that these evolutions are out of phase, suggesting that when the frequency grow up, the amplitude fall down to preserve some kind of energy conservation. Therefore, we propose the use of a peak-to-peak distance between ρ_i^k and ρ_j^{k+1} that consider this property :

$$d^2(\rho_i,\rho_j) = (f_i - f_j)^2\,(a_i - a_j)^2 \qquad (4)$$

Where f_i and a_i are the frequency and the amplitude of

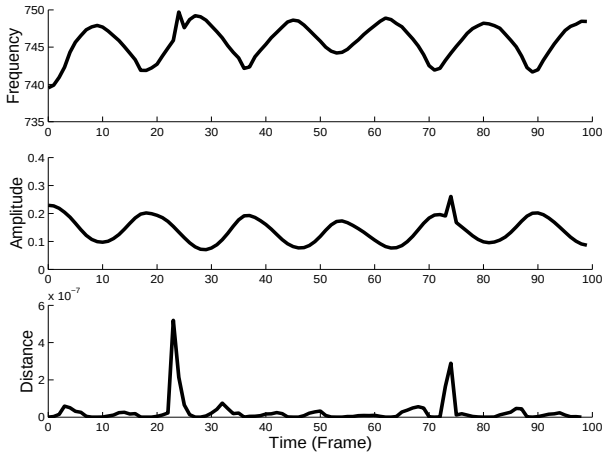

Fig. 5: *Frequency (top), amplitude (middle) and the peak-to-peak distance (bottom) for the first harmonic of a vibrato saxophone tone. Note that frequency and amplitude are out of phase. Values of frequency or amplitude are artificially corrupted at frame 25 or 75 respectively.*

the peak ρ_i.

Although this example can not be a proof and this property would require in depth acoustical studies, this distance has proven successful to identify peak successor in a given frequency neighborhood and to avoid spurious peaks of wrong frequency or amplitude. Figure 5.3 illustrate the influence of noisy frequency or amplitude values on the proposed distance.

5. TRANSITION PROBABILITY

We propose to weigh a transition between ρ_i^k and ρ_j^{k+1} by a Gaussian function of variance σ applied to a given peak-to-peak distance between the two peaks:

$$p_1(\rho_i^k, \rho_j^{k+1}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(d^2(\rho_i^k, \rho_j^{k+1})/2\sigma^2)} \qquad (5)$$

The variance σ is a parameter that is the variance of the distance between peaks of a "well-tracked" partial. σ can be estimated by considering distance variance between successive peaks within a partial over a large set of already tracked partials.

A *n*-transition is a direct transition (without intermediate peaks) from a peak ρ_i^k of frame k to a peak ρ_j^{k+n} belonging to frame $k+n$. Let us denote by $p_n(a,b)$ the probability of a *n*-transition from a peak a to a peak b. In the following, the other peaks are described by their distance relative to a.

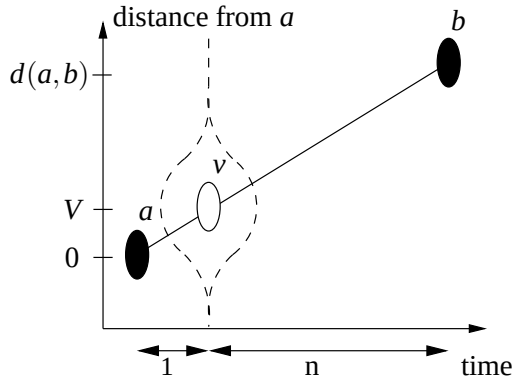

Fig. 6: *Extension principle of the probability measure. Real peaks (a, b) are black, the virtual peak v in white. Its position follow a Gaussian (dashed line) and has a maximum occurrence probability at V.*

For a 2-transition, we consider that we take a step through a virtual peak v of unknown characteristics. It has its ideal position V when $d(a,v) = d(v,b) = 1/2\,d(a,b)$. Its position is considered distributed with a Gaussian probability law $g(\Delta)$ centered at the ideal position V having variance σ. The probability of a 2-transition is then expressed by:

$$p_2(a,b) = \int_{-\infty}^{\infty} g(\Delta)\, p_1(a,v)\, p_1(v,b)\, d\Delta \qquad (6)$$

Where Δ is the distance between the position of v and its ideal position V. p_2 is the integral over all possible position of v of the probabilities of transition (a, v, b), weighted by the probability of finding the virtual peak at this position, that is at distance Δ of its ideal position V.

$$p_2(a,b) = \frac{1}{\sigma 2\pi} \int_{-\infty}^{\infty} g(\Delta)\, e^{-\frac{(V+\Delta)^2}{2\sigma^2}}\, e^{-\frac{(V-\Delta)^2}{2\sigma^2}}\, d\Delta \qquad (7)$$

Recursively, we define the probability of a $n+1$-transition by:

$$p_{n+1}(a,b) = \int_{-\infty}^{\infty} g(\Delta).p_1(a,v).p_n(v,b)\, d\Delta \qquad (8)$$

Where v has its ideal position V defined by :

$$d(a,v) = \frac{1}{n+1} d(a,b) \quad (9)$$

$$d(v,b) = \frac{n}{n+1} d(a,b) \quad (10)$$

And $p_1(a,v).p_n(v,b)$ is the probability of a $(n+1)$-transition from a to b with taking a step with a peak v (see Figure 6). By solving the integral, we find:

$$p_n(a,b) = \frac{1}{\sqrt{\frac{2(2n)}{n-1}} K_{n-1} (\sigma\sqrt{\pi})^n} e^{-(d^2(a,b)/2n\sigma^2)} \quad (11)$$

Where K_n is recursively defined by:

$$K_1 = \sqrt{2} \quad (12)$$

$$K_n = K_{n-1} \sqrt{2 \frac{2n-1}{n-1}} \quad (13)$$

By simplifying, we find:

$$p_n(a,b) = \frac{1}{\sqrt{2n \binom{2n-1}{n}} (\sigma\sqrt{\pi})^n} e^{-(d^2(a,b)/2n\sigma^2)} \quad (14)$$

Note that this transition probability depend on the σ parameter. Some properties on p_n can be computed to be able to choose a meaningful σ considering the chosen peak-to-peak distance and a given hop-size.

$$p_n(a,c) > p_1(a,b)\, p_{n-1}(b,c) \quad \text{if } d(a,b) > \sigma \sqrt{\frac{ln(\frac{n-1}{2n-1})}{n}} \quad (15)$$

$$p_n(a,b) > p_1(a,c)\, p_{n-1}(c,b) \quad \text{if } d(a-b) > \sigma \sqrt{ln(\frac{2n-1}{n-1} n(1-n))} \quad (16)$$

Where $d(a,c) = 0$ and $d(a,b) = d(b,c)$. Now that we have defined the transition probability between peaks of non adjacent frames, we will show how this probability is used to build optimal trajectories in a restricted set of future frames.

6. TRAJECTORIES IN FUTURE FRAMES

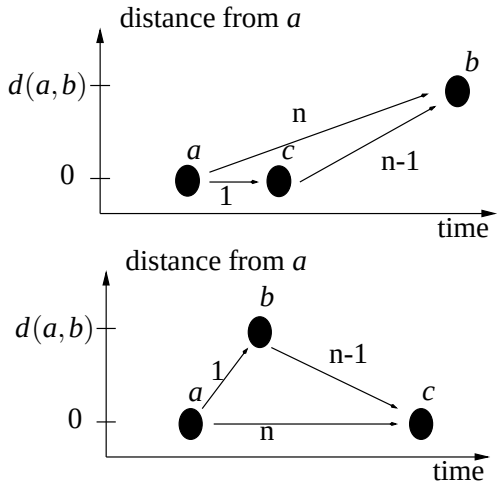

Fig. 7: *Representation of the two inequality 15 and 16 useful for the estimation of σ parameter.*

An optimal trajectory from ρ_k^i to ρ_{k+n}^j is a set of successive transitions starting from ρ_k^i and leading to ρ_{k+n}^j, maximizing the associated weight W, defined as the product of transition probabilities of transitions between peaks of the trajectory.

6.1. Generation of Trajectories

The set of transitions in considered future frames can be seen as a probability trellis similar to those used by the Viterbi algorithm [10] for HMM training. The efficiency of the Viterbi algorithm is based on the property that there is only one optimal trajectory from ρ_k^i to ρ_{k+n}^j. Unfortunately, since we allow transitions of size greater than 1, this property is not directly verified. Indeed, $p_{n+m}(a,c)$ can be equal to $p_n(a,b)\, p_m(b,c)$. In practice, we always choose the trajectory having the greatest number of peaks, since this trajectory is more relevant, based on extracted peaks.

Two parameters are taken into account: n_f, the number of future frames considered and n_m, the maximal size of transition allowed. We have the constraint: $n_f \geq 2n_m - 1$ (in the following, we consider $n_f = 5$ and $n_m = 3$).

We give an initial weight for each unlinked peak in Γ (the set of considered frames), equal to the weight of a trajectory from this peak to a virtual peak of same spectral property in the next frame after Γ. Optimal trajectories are then generated backwards. For each peak in a given

frame, we look for an optimal trajectory starting from this peak. If this trajectory has a weight greater than the initial weight, the trajectory is selected. At frame 4, in Figure 3, the lower peak is too far from peaks in frame 5 so no trajectory is selected. On the other hand, for the upper peak, the weight of the optimal trajectory is greater than initial one, so the trajectory is selected.

6.2. Comparison of Trajectories

As can be seen on Figure 3, optimal trajectories can share peaks. To favor partials having stable evolutions, the system must allocate the optimal trajectories having the greatest weights first, so that competing trajectories having lower weights are discarded. Trajectories can have different lengths or be shifted, we can therefore only compare them on the weight associated to their common part, defined as a trajectory portion starting from the largest starting frame index of the two trajectories we want to compare to the smallest ending frame index of the two trajectories. If the common part is the same or empty, we favor the longest trajectory.

7. RESULTS

The aim of the presented method is to be able to obtain partials that are reliable concerning the frequency and amplitude evolutions. To illustrate this, we compare the MAQ method and the proposed method on their ability to extract the third harmonic of a violin vibrato. For both method, the maximal frequency difference Δ_f defined in 2 is set to 100 Hz, the number of "zombie" states is 3. Concerning the proposed method, 12 future frames are considered. For both methods, to remove small and low amplitude partials, we use a rejection criterion. If a partial has its mean amplitude multiplied by the number of peak in the partial below .02, the partial is not plotted.

On strong vibrato, spurious peaks appear due to the strong modulations (see Figure 8.1). Processed by the MAQ method, these peaks give birth to new partials that are closer in frequency to future peaks than "older" partials (see Figure 8.2) whereas, by considering trajectories in future frames having lowest energy distortion, we are able to track the harmonic and represent it with an unique partial (see Figure 8.3). An overall view is given by Figure 9. A reliable set of partials provided by a tracking module even in case of musical modulations is of great importance to possible applications of sinusoidal models. In case of source identification, indexing or separation, the interpretation task is much easier. Concerning coding applications where the number of sinusoids is a critical parameter for coding efficiency, the quality gain is appreciable as presented in the listening tests results in Figure 10. A positive diff score indicate that the new method performs better than the MAQ method and vice and versa. This subjective MUSHRA test performed at France Telecom R&D by five audio coding experts compare the MAQ method and the proposed method using the sinusoidal output of a SSC encoder. Therefore, the tracking module use a hop size of 360 samples at 44100 kHz and up to 60 sinusoids may be encoded at a time. Some sequences of the Mpeg set plus violin solo, piccolo solo and a Tracy Chapman song are used. The grading scale ranging from 0 to 100. Excepting voice sequences, the subjective quality is significantly improved.

Taking account of future trajectories induce a non negligible complexity overhead. The proposed method process roughly 3 to 20 times slower than the MAQ method depending on the number of peaks per frames. The violin vibrato – 120 frames long – plotted on Figure 9 was processed in 5 seconds by our C++ implementation on a 1gHz processor.

8. CONCLUSION

In this article, we have presented a new tracking method designed to extract reliable partials from sounds mixture. By considering several future frames, the partials can choose the link that leads to an "optimal" evolution in the future. This new approach is helpful to extract coherent and concise informations on the evolution of sinusoidal components.

9. REFERENCES

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

[2] S. Marchand and R. Strandh, "InSpect and ReSpect: Spectral Modeling, Analysis and Real-Time Synthesis Software Tools for Researchers and Composers," in *Proceedings of the International*

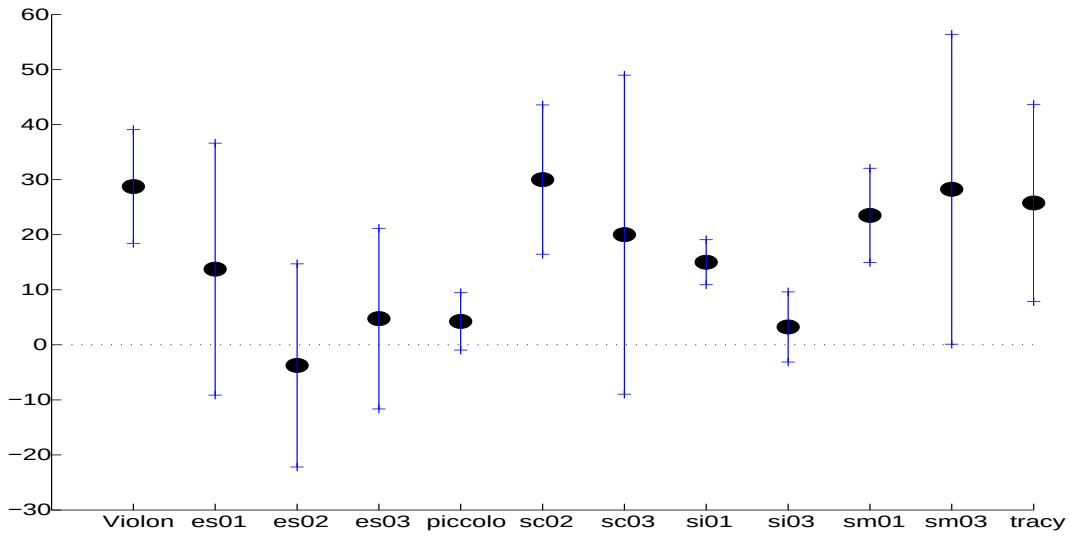

Fig. 10: *Results of subjective test with grading scale ranging from 0 to 100, comparing the proposed method versus the MAQ method. A positive diff score indicate that the new method performs better than the MAQ method and vice and versa. Some sequences of the Mpeg set plus violin solo, piccolo solo and a Tracy Chapman song are used. Mean scores are plotted with circles and confidence interval with thin line. Excepting voice sequences, the subjective quality is significantly improved.*

Computer Music Conference (ICMC). Beijing, China: International Computer Music Association (ICMA), October 1999, pp. 341–344.

[3] M. Desainte-Catherine and S. 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.

[4] H. Purnhagen and N. Meine, "HILN - The MPEG-4 Parametric Audio Coding Tools," in *IEEE International Symposium on Circuits and Systems (ISCAS 2000)*, vol. 3, May 2000, pp. 201–204.

[5] B. den Brinker, E. Schuijers, and W. Oomen, "Parametric Coding for High-Quality Audio," in *112th Convention of the Audio Engineering Society*. Audio Engineering Society (AES), May 2002.

[6] A. K. Tuomas Virtanen, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, April 2000, pp. 765–768.

[7] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[8] P. Depalle, G. Garcia, and X. Rodet, "Tracking of Partials for Additive Sound Synthesis using Idden Markov Model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, April 1993, pp. 225–228.

[9] M. Lagrange, S. Marchand, and J.-B. Rault, "Sinusoidal Parameter Extraction and Component Selection in a Non Stationary Model," in *Proceedings of the Digital Audio Effects (DAFx) Conference*. University of the Federal Armed Forces - Hamburg, Germany, September 2002, pp. 59–64.

[10] J. G. David Forney, "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, March 1973.

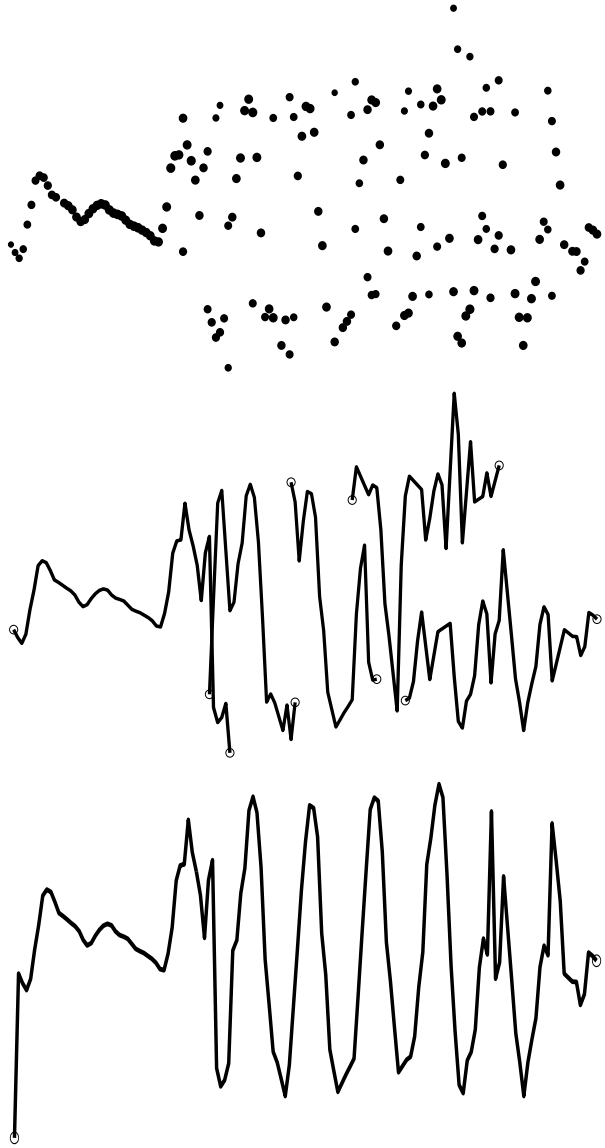

Fig. 8: *Tracking of the third harmonic of a violin vibrato plotted on the time / frequency plane. The peak view is on top, at the middle, partials extracted by the MAQ algorithm are plotted, and at the bottom is the partial as extracted by the proposed method.*

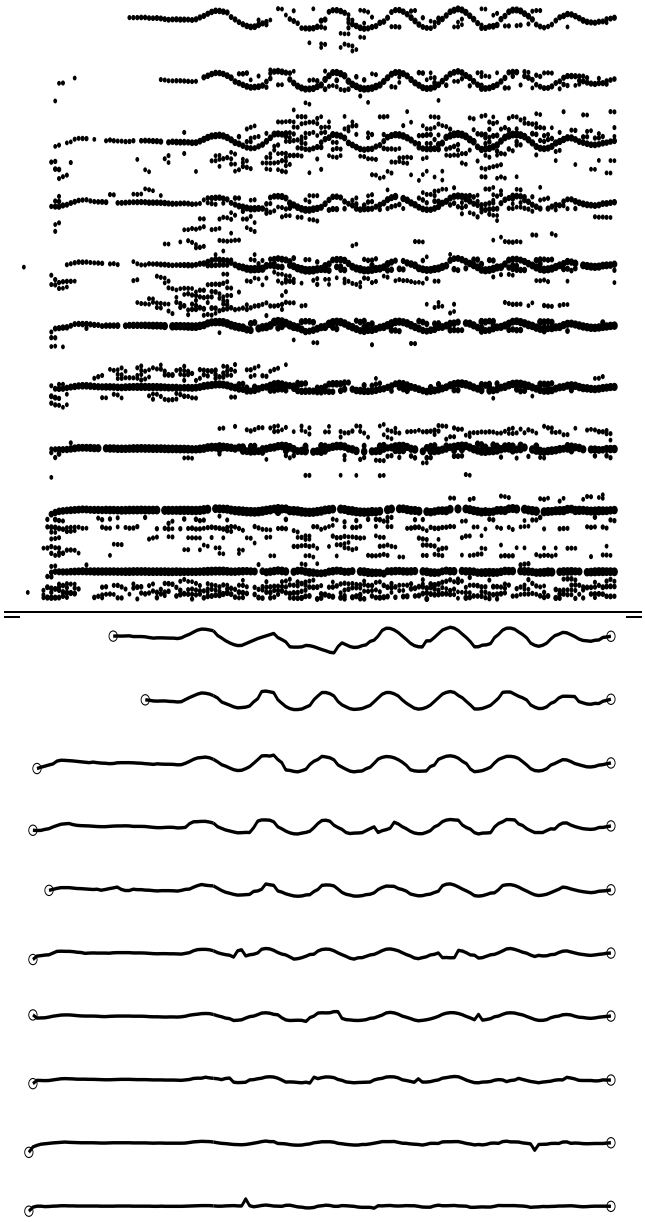

Fig. 9: *Tracking of a violin vibrato using proposed method and plotted on the time / frequency plane.*