Streams as Seams: Carving trajectories out of the time-frequency matrix

Giovanni Capizzi
Dep. of Math. and Computer Science
University of Palermo, Italy
capizzi94@gmail.com

Davide Rocchesso
Dep. of Math. and Computer Science
University of Palermo, Italy
davide.rocchesso@unipa.it

Stefano Baldan
Independent researcher
Berlin, Germany
singintime@gmail.com

ABSTRACT

A time-frequency representation of sound is commonly obtained through the Short-Time Fourier Transform. Identifying and extracting the prominent frequency components of the spectrogram is important for sinusoidal modeling and sound processing. Borrowing a known image processing technique, known as seam carving, we propose an algorithm to track and extract the sinusoidal components from the sound spectrogram. Experiments show how this technique is well suited for sound whose prominent frequency components vary both in amplitude and in frequency. Moreover, seam carving naturally produces some auditory continuity effects. We compare this algorithm with two other sine extraction techniques, based on peak detection on spectrogram frames. The seam carving skips this step and turns out to be applicable to a variety of sounds, although being more computationally expensive.

1. INTRODUCTION

For the analysis of audio signals, the most commonly-used representation is the intensity spectrogram, which is essentially a regular tessellation of the time-frequency plane, resulting from the magnitude of the Short-Time Fourier Transform (STFT). For impact sounds with stable frequency components it is of primary importance to identify the resonance frequencies and their respective decay rates, so to make re-synthesis and processing possible. More generally and more interestingly, in speech and audio signal processing it is important to extract – or to track – the prominent sinusoidal components from those signals where these components are time-varying, both in amplitude and in frequency. This opens the possibility of resynthesis with modification, as in time stretching or pitch transposition [1–3]. Most often, the trajectories in the time-frequency plane are drawn by tracking individual magnitude peaks of frames of the Discrete Fourier Transform (DFT), with no special regard to how frequency trajectories actually emerge in auditory perception, where they are typically segregated as streams [4].

In this contribution we propose a tracker that extracts the most prominent sinusoidal components (or partials) from a time-frequency spectrogram matrix, based on the *seam carving* algorithm [5], that was originally proposed for content-based image resizing. No preliminary computation of the DFT peaks is necessary and no heuristics on track creation, extinction, or continuation are used. To extract each trajectory (seam), the algorithm computes an energy matrix by dynamic programming in time proportional to the number of frames and to the number of frequency bins of the STFT, and then backtracks to extract the seam. We will show how this method would come useful both for impact sounds with frequency-stationary decaying modes, as well as for sounds with erratic resonances. Moreover, some continuity effects as found in auditory perception naturally emerge from seam carving as an analysis method.

Section 2 introduces the seam carving algorithm and its adaptation to audio spectrograms. Section 3 elaborates on the parameters of the carving algorithm, of STFT analysis, and on their impact on computational cost. Section 4 illustrates how the algorithm can be applied to extract relevant acoustic information from the extracted seams. Section 5 compares audio seam carving with two other techniques to track and extract the prominent sinusoidal components. Finally, section 6 shows how seam carving naturally reproduces some relevant perceptual contininuity effects.

2. CARVING SPECTROGRAMS AS IMAGES

An audio spectrogram, or at least the matrix containing its magnitude, can be treated as an image and manipulated by image-processing techniques. In the literature there are several examples of conversion of image-processing methods and algorithms to transform or synthesize audio [6]. Recent progress in generative neural networks for audio is largely due to transportation of image-based techniques to time-frequency representations such as the spectrogram [7].

Seam carving, used to resize an image by removing paths that are minimally relevant for the displayed content, is a popular image-processing algorithm that has found only limited use in audio [6, 8–10]. The image-processing algorithm is based on the construction of an energy matrix, and on the computation of a minimum-energy path connecting two opposite image edges. The energy function, that determines the energy matrix, can be tuned to the content to be removed, thus guiding the removal process towards those image areas that are the least important for the human eye.

The seam-carving algorithm was adapted by Tarrat-Masso to perform audio time scaling [6] with preservation of some important audio features such as tempo and sound attacks, that would be easily distorted if simply compressing or

stretching the signal in time. In his application, the seams are carved vertically from the spectrogram, connecting high to low frequencies. Barnwal et al. proposed an algorithm derived from seam carving to produce feature vectors that characterize the harmonic signature of human voice [9]. Their method assumes that the high-energy paths, as they are found as seams in the magnitude spectrogram, bring information that is relevant to classify speech sounds. The use of seam carving for sinusoidal component tracking was also proposed to estimate the speed from the captured sound of vehicles passing by [8]. All the three mentioned works were based on some spectrogram pre-processing: to preserve the sound features while scaling [6] or to improve seam extraction [8, 9].

The use of seam carving on the STFT matrix for multi-trace frequency tracking has been recently proposed by Zhu et al. [10], and shown to outperform probabilistic models in both accuracy and speed, especially for noisy signals. Our approach is similar, as it extracts traces by dynamic programming and spectral compensation, and was independently developed as part of the master's thesis of the first author [11]. The method by Zhu et al. is computationally more expensive, as it admits arbitrarily steep frequency trajectories, while at the same time introducing a regularization term that penalizes ample frequency deviation. They demonstrated the validity of the method to separate a heart-pulse signal from a motion signal in a mixture. Our method, instead, sets the maximum frequency slope as a parameter, and does not use any regularization terms. We present this technique in the context of audio signal processing, showing its usefulness for sinusoidal modeling synthesis and transformation, as well as its relation with perceptual auditory streaming.

2.1 Seam carving for image resizing

The seam-carving algorithm for image resizing was proposed by Avidan and Shamir in 2007 [5], as an improvement over conventional resizing that is blind to the image content, thus introducing distortions of the relevant portrayed objects [12]. Conversely, seam carving is content aware, as it preserves important image features, such as the object proportions. The algorithm operates by iteratively removing seams, which are paths of adjacent pixels traversing the image, vertically or horizontally. Each seam is formed and removed by minimizing an energy function that weighs the importance of each pixel in terms of local variation. By playing with the energy function, different operators can be implemented according to the desired image manipulation, for example for the removal of selected objects or to change the aspect ratio for image retargeting.

On images, the seam-carving algorithm is usually formulated as dynamic programming, and it is linear in the number of pixels. Alternatively, the optimal seam may be extracted as the shortest path in a graph connecting the adjacent pixels.

Consider an image I of size $h \times w$, and a cost (or energy) function E representing the importance of each pixel as referred to the scene content. The objective is that of removing a given number of optimal (or minimum cost) seams from the image. The algorithm is iteratively applied, each iteration requiring the energy matrix computation in order to detect and remove the optimal path of pixels. If proceeding top to bottom, the seam removal implies a shift of all pixels on the left of the removed seam.

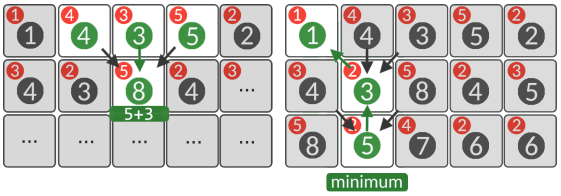


Figure 1: Example computation of the energy matrix C and identification of the minimal seam. Numbers in corners represent intensity values. Numbers in the middle of cells represent energy accumulated along paths.

Formally, a vertical seam is defined as

$$s^x = \{s_i^x\}_{i=0}^{h-1} = \{(i, x(i))\}_{i=0}^{h-1}, \quad s.t.\ \forall i, |x(i) - x(i-1)| \leq 1. \quad (1)$$

For each row i, the value $x(i)$ is the column position of the seam pixel. A vertical seam is a connected path of pixels, from top to bottom of the image, containing one pixel per row. Similarly, a horizontal seam can be defined.

Given a vertical seam s and the image I, the image pixels belonging to the seam are

$$I_s = \{I(s_i)\}_{i=0}^{h-1} = \{I(i, x(i))\}_{i=0}^{h-1}. \quad (2)$$

The cost of the seam is defined as its cumulated pixel energy $E(s) = E(I_s) = \sum_{i=0}^{h-1} e(I(s_i))$ and, therefore, the minimal cost seam s^* to be found is

$$s^* = \arg\min_s E(s) = \arg\min_s \sum_{i=0}^{h-1} e(I(s_i)). \quad (3)$$

The minimal seam s^* can be found by dynamic programming. The image is scanned from its second to its last row to define the matrix C containing the minimal costs for all possible seams. For each pixel in position (i, j), with $1 \leq i \leq h-1$ and $0 \leq j \leq w-1$, we have

$$C(i,j) = e(i,j) + \min(C(i-1, j-1), C(i-1, j), C(i-1, j+1)). \quad (4)$$

At the end of this first phase, the minimal seam is indicated by the smallest element of the last row of C. The second phase is a backtracking process starting from such element, which removes the identified seam and shifts the pixels as required. The whole process is exemplified in figure 1.

2.2 Carving sinusoidal components from audio

Seam carving can be used to extract the parameters characterizing the most relevant time-varying sinusoidal components of sound. The magnitude spectrogram mX generated by the STFT analysis provides the image where the seams are to be found:

$$mX(i,j),\ 0 \leq i \leq h-1\ e\ 0 \leq j \leq w-1. \quad (5)$$

As opposed to visual image carving, however, in audio seam carving we seek for maximal, rather than minimal, seams, that represent the most relevant sinusoidal components of the signal. Equation 4 is, therefore, replaced by

$$C(i,j) = mX(i,j) + \max(C(i-1,j-1), C(i-1,j), C(i-1,j+1)). \quad (6)$$

The spectrogram image is visualized so that sounds proceed along time, left to right. The i index indicates the i-th time slice, and the j index represents the j-th frequency bin of the DFT. Called t the number of modal resonances to be extracted, algorithm 1 proceeds on the computed magnitude spectrogram through t iterations.

Algorithm 1 Extraction of audio seams from spectrogram

Inputs: number of seams, magnitude spectrogram
for each seam **do**
  for all (i,j) in mX with $1 \leq i \leq h-1$ and $0 \leq j \leq w-1$ **do**
    compute the cumulative energy matrix C according to equation 6
  find the maximum element in the last row of C and backtrack to find the maximal seam;
  extract the parameters (frequency tracks, magnitude peaks) of the seam and remove it from mX;

In a `python` implementation [1], the cumulative energy matrix C is called `distTo`, as every element can be considered as a distance from the seam beginning. The matrix `upLink` is also simultaneously produced, which contains the directions to backtrack along the optimal seam: -1 : for `distTo[i-1][j-1]`; 0 : for `distTo[i-1][j]`; 1 : for `distTo[i-1][j+1]`.

Figure 2 (left) shows the spectrogram of a bell sound [2] as well as the seams found after 40 iterations. Figure 2 (right) shows the cumulative energy matrix that is computed to find the first seam. Many of the found seams can be visually clustered in bold paths. This can be both due to resonances that are close to each other or to deficiencies in the seam removal process. In fact, the simple removal of a path of pixels may be adequate for images, but it does not take the specific properties of spectrograms into account.

Two other examples of carving, for sounds [3] whose sinusoidal component frequencies are clearly time-varying, are reported in figure 3.

Of course, results will change according to the parameters of the STFT (DFT size N, window size M, hop size H, window type), to the kind of signal, and to the strength of overlapping noise. In any case, we can say that the seam cancellation is often not sufficient to eliminate the energy band of the partial. Some different strategies have been attempted to pre-process the spectrogram before carving. In particular, we tried smoothing the spectra in frequency and eliminating all components that lie below the smoothed profile. We also tried classic image processing filters such as Sobel. These pre-processing procedures, however, did not produce significantly better carved trajectories and, for the sake of simplicity, they were not used any further. All the examples presented in this paper use a Blackman-Harris window, and the effect of window type on seam carving and cancellation is not discussed.

[1] `https://github.com/GiovanniCapizzi/SeamCarvingAudio`
[2] `https://freesound.org/people/cheira/sounds/430511/`
[3] `https://freesound.org/people/josepharaoh99/sounds/368175/` and `https://freesound.org/people/qubodup/sounds/331381/`

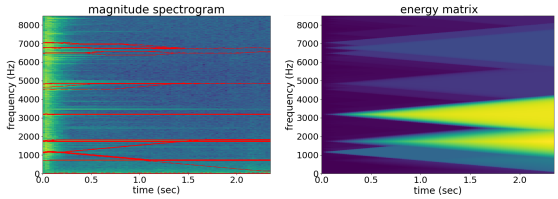


Figure 2: (left) Forty seams extracted from the spectrogram of a bell sound; (right) The cumulative energy matrix that is computed to find the first seam. DFT computed on N = 4096 points with a Blackman-Harris window of M = 4096 points, hop size H = 1024

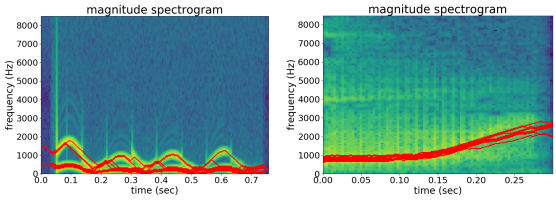


Figure 3: Carving eight seams from sounds whose sinusoidal component frequencies vary in time: (left) N = 1024, M = 511, H = 127; (right) N = 1024, M = 611, H = 152

3. TUNING THE CARVING

3.1 Extending the neighborhood

One of the limitations of the proposed carving algorithm is evident when there are rapidly varying frequency components, as in figure 3. In order to induce the algorithm to search beyond diagonal trajectories, we need to enlarge the neighborhood that is explored during the computation of the cumulative energy matrix. The modifications to the original carving algorithm are straightforward, and a new parameter `hwidth` can be introduced in the code and set to a small positive integer. This parameter represents the integer half number of pixels to include in the neighborhood during the matrix computation. The algorithm, being able to find its way through larger neighborhoods, behaves as depicted in figure 4 for `hwidth = 2` (i.e., five-cell neighborhood). This parameter acts as a limit in the steepness of frequency trajectories.

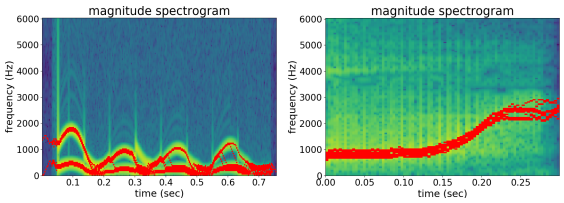
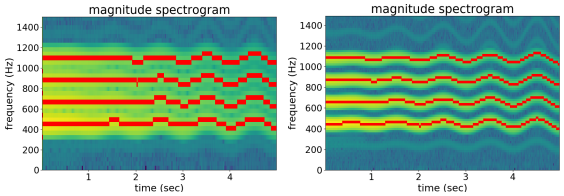
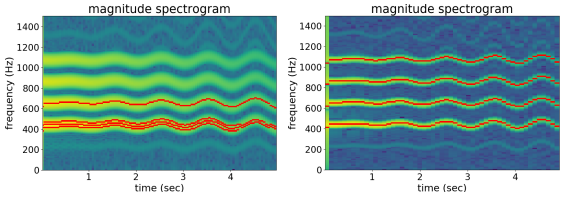


Figure 4: Carving eight seams as in figure 3, with `hwidth = 2`, corresponding to a 5-element search neighborhood



(a) 1024, 1024, 1024 (b) 2048, 2048, 512



(c) 2048, 4096, 2048 (d) 4096, 4096, 4096

Figure 5: Carving four seams for different STFT parameters. Each subcaption reports the values of parameters M, N, H.

3.2 Choosing the STFT parameters

To see the effect of different STFT parameters on the quality of carving we consider a synthetic sound made of frequency-modulated and amplitude-decaying sinusoidal components. Figure 5 shows some combinations of M, N, and H. It is evident how a small window, while producing low frequency resolution, produces consistent traces. Choosing N larger than M (zero padding) often leads to poor tracking, so it is preferred to keep N equal to M.

The interaction between the window size M and the hop size H is illustrated in figure 6. The parameters of figure 6c ($M = 4096$, $H = 2048$) and those of figure 6d ($M = 2048$, $H = 512$) afford a complete tracing of the three components of the inharmonic chirp. The analysis of figure 6c is more computationally demanding, but it has higher resolution and affords a higher quality resynthesis from the extracted seams.

3.3 Cost of spectrogram carving

If we consider a spectrogram with m frames, each computed on n bins, and a neighborhood of k elements, the algorithm to carve a seam:

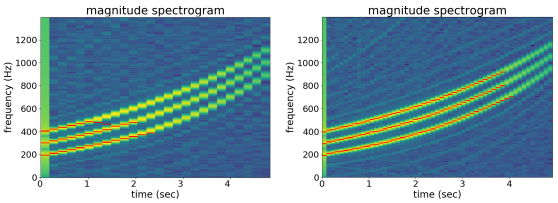
- zeroes the elements of the first row (temporal column in the rotated spectrum) of the distance matrix;
- for each element of each subsequent row, performs at most $k - 1$ comparisons among neighboring elements of the previous row, searching for the maximum.

By protecting the frame boundaries with very large values it is possible to have the same number of comparisons for all bins in a frame, so that the total number of comparisons is simply calculated as

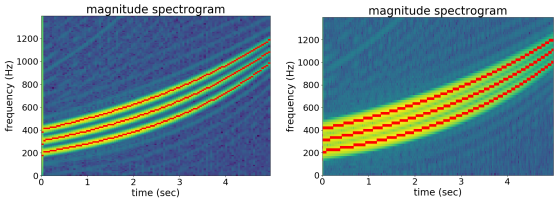
$$n_c = n(k-1)(m-1). \tag{7}$$

The operation of backtracking and seam canceling takes time proportional to the m frames, and is therefore superseded by n_c.

Figure 7 shows the measured experimental computation time required by the extraction of a seam from the chirp of figure 6, for different values of window size and relative window overlap. Different choices of M lead to very similar execution times, just because shorter windows are computed proportionally more often, for a given percent overlap.



(a) M=8192, Overlap=0% (b) M=8192, Overlap=50%



(c) M=4096, Overlap=50% (d) M=2048, Overlap=25%

Figure 6: Carving three seams for different values of window and hop size

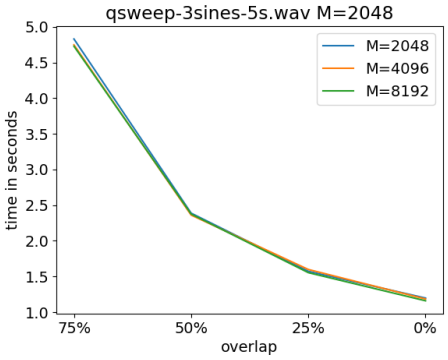
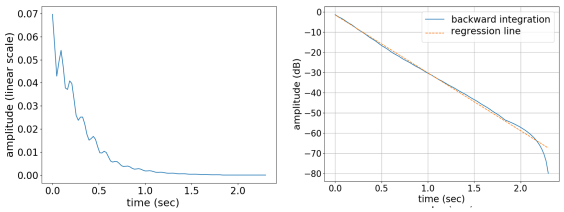
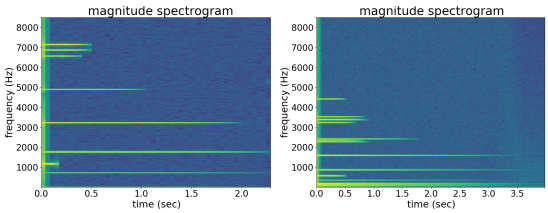


Figure 7: Empirical effect of window and hop size on seam carving time for figure 6. The reported times are computed as the average extraction time on three seams. The computation time of the STFT is not included.



(a) Extracted magnitude of a seam (b) Linear regression of the backward-integrated decay

Figure 8: Measuring the decay time of a sound partial



(a) Re-synthesis from extracted parameters (b) Resizing the bell

Figure 9: Extraction and synthesis of a bell sound. In 9b a variation of the original sound is produced.

4. APPLICATIONS OF SEAM CARVING

4.1 Parameterizing resonances from impact sounds

Impact sounds are most often characterized by exponentially-decaying partials, each at a given frequency. The identification of such components can be done by seam carving where, for each extracted seam, the following parameters are computed: (i) peak magnitude; (ii) decay time; (iii) frequency.

Once a seam has been found, the decay time can be computed by Schroeder backward integration [13] and by linear regression. The instant when the regression line has decreased below a predefined threshold (typically, 60dB) is used to compute the decay time. Figure 8a shows a typical amplitude decay of a detected seam. The decay time is computed by backward integration and linear regression, illustrated in figure 8b. For accurate frequency estimation and for better removal of the seam, parabolic interpolation and spectral resynthesis are used [2, 3].

In figure 9 an example of extraction and synthesis is provided. The audio files are available in the example folder of the code repository [1] . The original sound is used to extract the main spectral content of the bell, that becomes ready for sound manipulation. For example, adding new components and scaling them down in frequency, the bell is effectively resized and made to sound bigger.

The identification of exponentially-decaying sinusoidal components is often required to inform sound models of contact sounds for interaction in everyday virtual environments [14]. In these application contexts, sometimes the constraint of exponential decay is released and the whole amplitude envelope is retained [15]. Given the ubiquity of everyday sound modeling for interactive virtual and augmented environments, it would be interesting to exploit frequency constancy to speed up the carving process, as proposed in section 5.

4.2 Following erratic resonances

As opposed to other identification methods, such as the Matrix Pencil [16], that assume the signal as made of decaying exponentials, with seam carving we can extract sinusoidal components that vary both in amplitude and in frequency. Therefore, seam carving can be used within the analysis/synthesis framework based on spectral processing, which makes a large range of high-quality audio effects and sound transformations possible [2]. Among these, we just mention pitch transposition, time stretching, timbre modification and sound morphing. Moreover, from the observation that seam carving is based on a definition of energy and the construction of an energy matrix, we expect it to naturally mimic those auditory phenomena that are energy based. In particular, we expect that some kind of perceptual inertia, as found in auditory continuity effects, is reproduced while carving frequency trajectories. The examples of section 6 are aimed at verifying this expectation.

5. SEAM CARVING VS. PEAK TRACKING

The sinusoidal model for the analysis and resynthesis of sound belongs to the classic literature of signal processing [1, 3, 17, 18]. Such model relies on tracking the sinusoidal components, as they are previously detected on single spectrogram frames. Peak detection with parabolic interpolation is found in the Spectral Modeling Synthesis (SMS) tools [4] [2]. With such methods spurious peaks are often found as a result of rapid spectral changes or as sidelobes of the transform window.

McAulay and Quatieri proposed a method for tracking sinusoidal components in speech, based on the concepts of birth and death of partials [1], with frame-to-frame peak matching. Tracks are declared as dead or as born when a match is not possible within a frequency interval Δ. This method was also used by Smith and Serra in the more general context of analysis/resynthesis of musical, possibly inharmonic, sounds [19]. The method was later extended by introducing the concept of frequency guides, whose behavior may be set according to the characteristics of the sound being analyzed [2]. The complexity of the McAulay–Quatieri matching operation between two frames turns out to be quadratic in the number of peaks per frame, in the worst case.

Another method, closer in spirit to seam carving, has been proposed by the third author and implemented as a `modal_tracker` object for the Sound Design Toolkit [20]. The algorithm has been specifically tailored for the analysis of impact sounds, and for their re-synthesis through a physically-informed resonator made of exponentially-decaying sinusoidal oscillators, each controllable in ampli-

[4] A page on SMS tools is `https://www.upf.edu/web/mtg/sms-tools`, and `python` source code is available at `https://github.com/MTG/sms-tools`. The SMS tools have been used in the proposed implementation of audio seam carving.

tude, frequency and decay time. The `modal_tracker` algorithm proceeds in three steps:

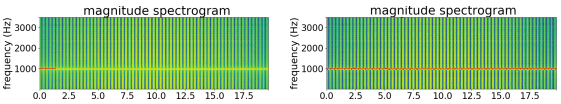
1. create a matrix of spectrogram peaks;
2. create a matrix of cumulated sums;
3. extract the partials one by one.

The peak-picking phase proceeds by selecting local maxima in each spectrogram frame, namely the frequency bins having the largest magnitude in a given neighborhood. In the `modal_tracker` implementation, a bin is considered to be a peak if its magnitude is greater than the one of the two bins above and the two bins below it. Only the peak magnitudes are retained, while all the other values are discarded.

In the second step, the cumulated sums are computed in a way which is similar, though slightly different, to seam carving: Sums are retained only where peak values in the original spectrogram exist, and values are cumulated only until there is a connecting path between peaks in the previous frame and peaks in the current frame. The width of the search window used to determine if a path is connected could be left as a free parameter of the algorithm, as it happens for seam carving, but in the `modal_tracker` implementation it is rigidly set fixed to three cells, or radius = 1. Only the maximum cumulated sum in the previous frame is added to the peak in the successive frame, following the same principle of seam carving. The additional constraints though, namely the peak picking phase and most importantly the interruption of the cumulated sum if no peaks are found in the neighborhood of a summation path, make the algorithm behave in a very different way: Instead of having "cones" of energy covering the whole length of the spectrogram, as shown in figure 2, the cumulated sum matrix of the `modal_tracker` algorithm shows well defined tracks, exactly one pixel wide, which can begin and end anywhere across the whole breadth of spectrogram. The combined setup of a five-cell peak-picking window and a three-cell peak tracking window avoids the forking or joining of different summation tracks, because local maxima are at least three bins apart from each other and therefore the presence of at most one peak in a three-cell search window is guaranteed. This is a desirable property for the particular use case of resynthesizing sounds using a fixed-size oscillator bank.

The final step consists in extracting the cumulated partials. The absolute maximum in the summation matrix should correspond to the end of the most prominent partial. By performing backtracking, as already described for seam carving, information about the partial is retrieved and then removed from the summation matrix, in order to be able to find other prominent components through consecutive iterations of this extraction phase.

If we call n the number of frequency bins in a single spectrogram frame, m the number of frames in the spectrogram, and p the number of partials to extract, the whole algorithm has complexity $O(nmp)$. In particular the peak picking phase has complexity $O(n)$, the energy summation has complexity $O(nm)$, and the backtracking phase has complexity $O(mp)$ if the positions of the p energy summation maxima are memoized during phase 2, and therefore



(a) `modal_tracker` (b) seam carving

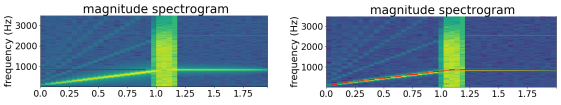
Figure 10: Extraction of a sinusoidal component interrupted by amplitude-varying noise bursts. The `modal_tracker` is about six times faster, but it fails to go through the bursts.

don't need to be searched in the energy summation matrix.

The `modal_tracker` requires, similarly to peak-tracking methods, the preliminary extraction of peaks from single DFT frames. It also relies on slow variation of partial frequencies, and is therefore mostly suitable for impact sounds. Similarly to seam carving, however, it computes a matrix of cumulated energies. Therefore, this method can be seen as intermediate between peak tracking and seam carving, and it turns out to be very efficient although not as general as the two other methods.

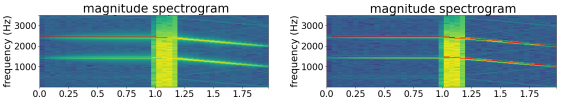
6. STREAMS AS SEAMS

For the comparison of algorithms for the identification and tracking of sinusoidal components, it is interesting to refer to auditory continuity illusions [4] in auditory scene analysis. Among the many examples of continuity illusions studied in the literature, we consider the perceived continuity of frequency glides and stationary frequency components, when they are interrupted by noise bursts [21]. Figures 10, 11 and 12 show the spectrograms of three such cases of continuity illusion. In all cases, the `modal_tracker` is faster but fails to go through the noise bursts.



(a) `modal_tracker` (b) seam carving

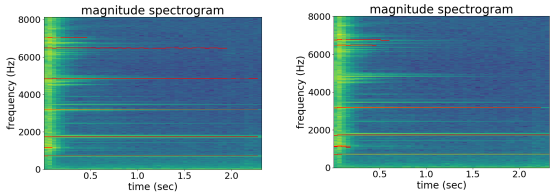
Figure 11: Glide to stationary sine transition interrupted by a noise burst. The `modal_tracker` fails to on the gliding start.



(a) `modal_tracker` (b) seam carving

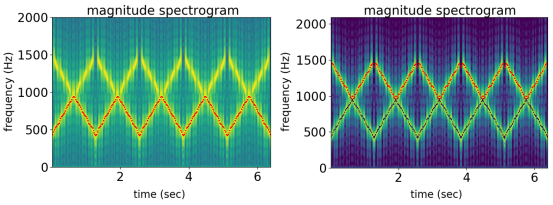
Figure 12: Stationary sine to glide transition interrupted by a noise burst. The `modal_tracker` fails at the noise interruption.

If the purpose is that of identifying the most prominent partials of a sound of impact, where the frequency of sinusoidal components is stationary, both `modal_tracker` and seam carving behave well, and the former is to be



(a) `modal_tracker` (b) seam carving

Figure 13: Extraction of seven partials from a bell sound. `modal_tracker` is 280 times faster.



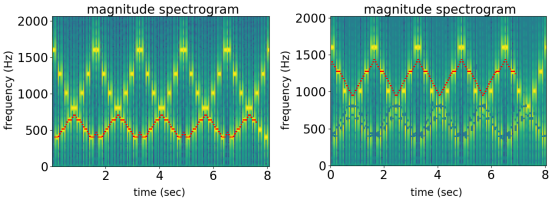
(a) First seam (b) Second seam

Figure 14: Extracting two seams from crossing glides

preferred as significantly faster. However, seam carving proves to be robust to the presence of noise in the spectrogram. Figure 13 shows the extraction of seven partials from a recording of bell sound. Despite the differences in the chosen partials, the resynthesis sounds convincing in both cases.

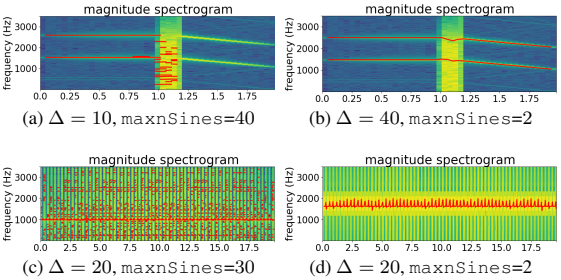
Another interesting case from the literature of auditory scene analysis is the perceptual segregation of crossing glides. Figure 14 shows how seam carving reproduces the results of experimental perceptual segregation of continuous crossing glides [22]. Figure 15 shows an example of stepped crossing glides. Although a spurious seam is extracted after the first, due to rapid frequency changes, the first and third extracted seams correspond to perceptually segregated streams.

Properly parameterized sinusoidal tracking, as implemented in the SMS tools, behaves well on both glides and stationary sounds, as shown in figure 16. In general, we found that such tracking is more prone to the production of spu-



(a) First seam (b) Third seam

Figure 15: Extraction of three seams from crossing stepped glides. The second seam is not shown, as largely attributable to the residual energy left from the removal of the first.



(a) $\Delta = 10$, `maxnSines`=40 (b) $\Delta = 40$, `maxnSines`=2

(c) $\Delta = 20$, `maxnSines`=30 (d) $\Delta = 20$, `maxnSines`=2

Figure 16: Sinusoidal tracking with the SMS tools on the sounds of figures 12 and 10, respectively.

rious tracks, due to the stochastic component of the spectrum. Figure 16c shows that, as the number of admitted tracks `maxnSines` is lowered and Δ is properly adjusted, the tracking improves. On the other hand, figure 16d shows oscillations due to noise. With seam carving such oscillations are much smaller, thanks to the energy function and to the limited neighborhood.

7. CONCLUSIONS

The most popular techniques for tracking sinusoidal components from audio rely on peak picking on spectrogram frames, and on matching and continuation algorithms across frames. As an alternative borrowed from image processing, we propose seam carving: a technique that skips the peak-picking stage and looks for paths of maximal accumulated energy in the spectrogram matrix.

Seam carving is quite general, as it behaves well for a large variety of sounds, and is still practically efficient when implemented via memoization and dynamic programming. Some parameter calibration is still needed, as with the other techniques, to achieve the best separation of sines from noise and to minimize spurious tracks, given a sound mixture. However, seam carving shows good robustness to noise, to the point that it recreates some auditory continuity effects, by naturally sticking to trajectories that are indeed submerged in noise bands. These extensions of trajectories appear to the ears of the listener and are commonly explained by gestalt principles of proximity or good continuation [23], but are not easily captured by techniques based on peak picking.

There are ample opportunities to investigate the behavior of audio seam carving further, by systematically varying all aspects of time-frequency analysis (e.g., the window type), by extending the repertoire of examples, and by considering other noteworthy perceptual effects. Horizontal (sine) beam extraction can be combined with vertical (transient) seam extraction, and the beams themselves may become the object of analysis, comparison, warping, and morphing in audio signal processing.

8. REFERENCES

[1] R. McAulay and T. 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] J. Bonada, X. Serra, X. Amatriain, and A. Loscos, "Spectral processing," in *DAFX: Digital Audio Effects*, U. Zölzer, Ed. John Wiley & Sons, Ltd, 2011, pp. 393–445.

[3] J. O. Smith III, *Spectral Audio Signal Processing*. W3K Publishing, 2011. [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/

[4] R. M. Warren, *Auditory Perception: An Analysis and Synthesis*, 3rd ed. Cambridge University Press, 2008.

[5] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *ACM Transactions on graphics (TOG)*, vol. 26, no. 3. ACM, 2007, p. 10.

[6] J. M. Tarrat-Masso, "Adaptation of the seam carving technique for improving audio time-scaling," *Master's thesis, Pompeu Fabra University*, 2008.

[7] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," *CoRR*, vol. abs/1706.09559, 2017. [Online]. Available: http://arxiv.org/abs/1706.09559

[8] S. Barnwal, R. Barnwal, R. Hegde, R. Singh, and B. Raj, "Doppler based speed estimation of vehicles using passive sensor," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 2013, pp. 1–4.

[9] S. Barnwal, K. Sahni, R. Singh, and B. Raj, "Spectrographic seam patterns for discriminative word spotting," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4725–4728.

[10] Q. Zhu, M. Chen, C. Wong, and M. Wu, "Adaptive multi-trace carving based on dynamic programming," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, Oct 2018, pp. 1716–1720.

[11] G. Capizzi, "Identificazione ed inseguimento di componenti sinusoidali da matrici di analisi tempo-frequenza di segnali audio," Master's thesis, Department of Mathematics and Computer Science, University of Palermo, Italy, 2019.

[12] A. Shamir and S. Avidan, "Seam carving for media retargeting," *Communications of the ACM*, vol. 52, no. 1, pp. 77–85, 2009.

[13] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 03 1965.

[14] J. Traer, M. Cusumano, and J. H. McDermott, "A perceptually inspired generative model of rigid-body contact sounds," in *Proc. International Conference on Digital Audio Effects*, Birmingham, UK, 2019.

[15] D. B. Lloyd, N. Raghuvanshi, and N. K. Govindaraju, "Sound synthesis for impact sounds in video games," in *Symposium on Interactive 3D Graphics and Games*, ser. I3D '11. New York, NY, USA: ACM, 2011, pp. 55–62. [Online]. Available: http://doi.acm.org/10.1145/1944745.1944755

[16] J. Laroche, "The use of the matrix pencil method for the spectrum analysis of musical signals," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 1958–1965, 1993. [Online]. Available: https://doi.org/10.1121/1.407519

[17] P. Hedelin, "A tone oriented voice excited vocoder," in *ICASSP'81. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6. IEEE, 1981, pp. 205–208.

[18] L. Almeida and F. Silva, "Variable-frequency synthesis: An improved harmonic coding scheme," in *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 9. IEEE, 1984, pp. 437–440.

[19] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proceedings of the International Computer Music Conference*, Urbana Champaign, IL, 1987, pp. 290–297.

[20] S. Baldan, S. Delle Monache, and D. Rocchesso, "The sound design toolkit," *SoftwareX*, vol. 6, pp. 255–260, 2017.

[21] V. Ciocca and A. S. Bregman, "Perceived continuity of gliding and steady-state tones through interrupting noise," *Perception & Psychophysics*, vol. 42, no. 5, pp. 476–484, 1987.

[22] Y. Tougas and A. S. Bregman, "Crossing of auditory streams." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 6, p. 788, 1985.

[23] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT press, 1990.