HYBRID AUDIO INPAINTING APPROACH WITH STRUCTURED SPARSE DECOMPOSITION AND SINUSOIDAL MODELING

Eto Sun and Philippe Depalle

Sound Processing and Control Laboratory, CIRMMT *
McGill University
Montréal, Canada
eto.sun@mail.mcgill.ca | philippe.depalle@mcgill.ca

ABSTRACT

This research presents a novel hybrid audio inpainting approach that considers the diversity of signals and enhances the reconstruction quality. Existing inpainting approaches have limitations, such as energy drop and poor reconstruction quality for non-stationary signals. Based on the fact that an audio signal can be considered as a mixture of three components: tonal, transients, and noise, the proposed approach divides the left and right reliable neighborhoods around the gap into these components using a structured sparse decomposition technique. The gap is reconstructed by extrapolating parameters estimated from the reliable neighborhoods of each component. Component-targeted methods are refined and employed to extrapolate the parameters based on their own acoustic characteristics. Experiments were conducted to evaluate the performance of the hybrid approach and compare it with other state-of-the-art inpainting approaches. The results show the hybrid approach achieves high-quality reconstruction and low computational complexity across various gap lengths and signal types, particularly for longer gaps and non-stationary signals.

1. INTRODUCTION

Audio inpainting involves the recovery of missing or degraded parts of an audio signal based on its reliable segments [1]. Suppose $\mathbf{y} \in \mathbb{R}^N$ be an audio signal with N samples, and the indices of its missing or degraded samples are known (referred to as unreliable samples). The recovered signal $\mathbf{s}$ should be the same as the original signal in the reliable part, in other words, it should belong to the following set $\Gamma_{\mathbf{y}}$:

$$\Gamma_{\mathbf{y}} = \{\mathbf{s} \in \mathbb{R}^N : \mathbf{M}_{\mathrm{R}}\mathbf{s} = \mathbf{M}_{\mathrm{R}}\mathbf{y}\} \tag{1}$$

where $\mathbf{M}_{\mathrm{R}} \in \mathbb{R}^{N\times N}$ is a square diagonal matrix whose k-th diagonal value is 1 if the k-th sample of the original signal is reliable, otherwise it is 0, which means $\mathbf{M}_{\mathrm{R}}\mathbf{y}$ contains all reliable samples. In this study, we focus on compact gaps that are well separated from one another, rather than random small gaps.

Various approaches proposed to address this issue will be presented in chronological order, each leveraging distinct assumptions about the underlying audio structure. The first approach relies on the assumption that the audio signal is relatively stationary around the unreliable region. This method analyzes the reliable neighborhood using autoregressive (AR) modeling and extrapolates the missing samples based on the estimated autoregressive coefficients [2, 3]. Another approach is based on the concept of sinusoidal modeling, which assumes that the audio signal can be decomposed into a sum of time-varying sinusoids (*partials*). By analyzing the behavior of these partials in the reliable neighborhood, this method predicts their trajectories within the unreliable segment [4, 5]. A third approach exploits sparsity, which assumes that signals can be efficiently represented using only a few significant coefficients within a transformed domain, typically the time-frequency domain. This approach aims to construct a signal that matches the original signal in the reliable parts and exhibits a sparse representation around the unreliable region in the chosen domain [1, 6, 7, 8]. Recently, a data-driven approach has proven effective in solving this task, especially for longer gaps [9, 10, 11]. This approach utilizes deep neural networks to find statistical audio structures (priors) from a large amount of training data and fill the gap based on these priors. In this research, we concentrate on the modeling approaches that rely solely on information from reliable samples around the gap for inpainting.

While the sparsity-based approach offers promising results in many scenarios, there are still challenges that degrade the reconstruction quality. One challenge arises when dealing with non-stationary signals, such as those containing fast time-varying components like modulations or inherently nonsparse elements like noise. Sparsity-based methods may encounter difficulties in accurately representing non-stationary signals around the gap, which could lead to the selection of inappropriate atoms and artifacts in the unreliable region (Figure 4b). The length of the gap is another challenge. For gaps longer than 50 milliseconds, sparsity-based methods often have energy drops in the unreliable region [7] because the partial trajectories are not maintained properly. These challenges also exist with other approaches.

To address these challenges, a structured representation of the sound can be incorporated as prior knowledge to guide the inpainting process. This structured approach, indirectly utilized in previous works on additive synthesis and autoregressive modeling, provides a way to combine these existing techniques within the sparse representation framework. A natural prior is to rely on the usual structure of sound signals, which considers them as a mixture of three components: *tonal*, *transients*, and *noise* [12]. The tonal part can be generalized as the slow-varying deterministic part, which is mostly stationary (or cyclo-stationary) in the longer term. It is usually made of time-varying sinusoids (also known as *partials*). The transients represent the fast-varying deterministic part, which consists of components that have a short duration, a wide spectral bandwidth, and are usually located at the beginning or end of a sustained sound. The noise refers to the stochastic part of

* Centre for Interdisciplinary Research in Music Media and Technology

the signal. This model provides a structured view of various audio signals and is widely used in the fields of additive synthesis [12, 13] and audio encoding [14]. However, it is rarely explored in the context of audio inpainting.

In order to separate these three components from the mixture, a technique known as *structured sparse decomposition* can be employed. This technique builds upon the idea of sparse decomposition, but incorporates prior knowledge from the signals into the decomposition process [15]. By leveraging the known characteristics of each component, structured sparse decomposition can effectively isolate them within the audio signal. This separation allows for the application of component-specific methods for audio inpainting.

In this research, we propose a hybrid audio inpainting approach to improve the perceived quality of the reconstruction while considering the diverse characteristics of audio signals. By combining the strengths of sparsity-based decomposition for efficient representation and tailored methods for each component, this hybrid approach aims to overcome the limitations of individual methods and achieve robust and high-quality reconstruction across a wider range of audio signals.

The rest of this paper is organized as follows: Section 2 describes the proposed hybrid approach for audio inpainting and elaborates on each technique used in detail. Section 3 analyzes our hybrid approach and compares it with other state-of-the-art techniques through various experiments. Section 4 summarizes this paper, outlines the strengths and limitations of our approach, and addresses some possibilities for future research.

2. METHODOLOGY

2.1. Overview

The proposed hybrid inpainting approach views an audio signal as a mixture of three components. The input signal is first pre-processed (Sect. 2.2), then decomposed into tonal (Sect. 2.3), transient (Sect. 2.4), and noise components. Tonal and noise components are reconstructed independently by integrating and refining previous methods (Sect. 2.5 for tonal and Sect. 2.6 for noise). The final output is a combination of the reconstructed components with post-processing (Sect. 2.7). Figure 1 summarizes the overall process of our approach. Interested readers can find all the details of each proposed techniques at the accompanying website [16].

2.2. Pre-processing

Pre-processing entails shortening and aligning the input signal for downstream tasks. The minimum length of the shortened signal is determined based on the offset, window size, and time shift of the window [17]. In order to better estimate time-varying signals, we will set the length longer than the minimal support. The center of the gap is aligned in the midpoint of two adjacent Gabor windows, which refers to the "half" offset configuration in [7].

2.3. Estimation of tonal part

After the pre-processing, the next step is to decompose the signal into a deterministic part and a residual part based on structured sparse decomposition. In this case, our extracted tonal component will be a set of atoms that reflects the tonal structure of a sound. This technique builds upon the idea of sparse decomposition, which

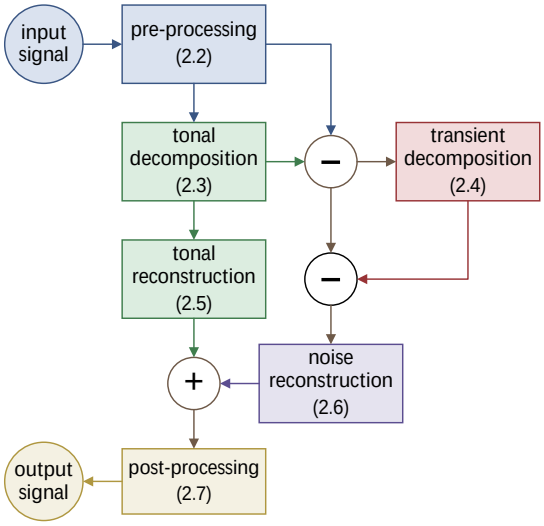


Figure 1: *The overview structure of the proposed hybrid approach. The number in parentheses represents the corresponding section.*

aims to represent or approximate an audio signal as a linear combination of simple waveforms (*atoms*) selected from a set of atoms (*dictionary*) [18]. That leads to the sparse approximation problem, which can be formalized as an optimization problem. In the context of audio inpainting, the problem can be written as:

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0 \;\text{ subject to }\; \|\mathbf{M}_\mathrm{R}\mathbf{y} - \mathbf{M}_\mathrm{R}\mathbf{\Phi}\mathbf{z}\|_2^2 \leq \epsilon \quad (2)$$

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{\Phi}^\mathrm{H}\mathbf{x}\|_0 \;\text{ subject to }\; \|\mathbf{M}_\mathrm{R}\mathbf{y} - \mathbf{M}_\mathrm{R}\mathbf{x}\|_2^2 \leq \epsilon \quad (3)$$

where $\mathbf{\Phi} : \mathbb{C}^P \to \mathbb{C}^N$ is the synthesis operator, $\mathbf{z} \in \mathbb{C}^P$ is a vector of atoms, $\mathbf{\Phi}^\mathrm{H} : \mathbb{C}^N \to \mathbb{C}^P$ is the conjugate transpose (Hermitian transpose) of the synthesis operator $\mathbf{\Phi}$ and is referred to as the analysis operator, signal $\mathbf{x} : \mathbb{R}^N$ is the cosparse representation of signal $\mathbf{y}$ [19], $\hat{\mathbf{x}}$ is the estimation of $\mathbf{x}$, and $\|\mathbf{x}\|_0$ is the ℓ_0 "norm" of $\mathbf{x}$. The first equation refers to the synthesis variant of the inpainting problem, and the second equation refers to the analysis variant.

Although finding the optimal solution to this non-convex problem is NP-hard [20], a suboptimal solution is usually built as an approximation based on available algorithms. One approach called relaxation is to replace the ℓ_0 "norm" with the ℓ_1-norm, which is a convex approximation of the ℓ_0 "norm" for sparse decomposition [18]. The relaxation approach can be expressed in the following unconstrained form:

$$\arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{M}_\mathrm{R}\mathbf{x} - \mathbf{M}_\mathrm{R}\mathbf{y}\|_2^2 + \lambda\|\mathbf{\Phi}^\mathrm{H}\mathbf{x}\|_1 \right\} \quad (4)$$

where $\|\mathbf{x}\|_1$ represents the ℓ_1-norm of $\mathbf{x}$, and λ is a parameter controlling the strength of the constraint.

A shrinkage operator needs to be defined to solve this problem, which is able to integrate structured information in the decomposition process. This technique, known as *social sparsity*, involves selecting atoms based on the coefficients within their respective neighborhoods $\mathcal{V}$ [21]. The neighborhood $\mathcal{V}(k)$ of an atom with index k is defined as a set of atoms that are near the atom k. The neighborhood can be of an arbitrary shape and can be weighted for more flexibility. In this research, we use the Persistent Empirical

Wiener (PEW) shrinkage operator [22]:

$$\mathcal{S}_\lambda(z_k) = z_k \cdot \max\left(1 - \frac{\lambda^2}{\|\mathbf{V}_k\mathbf{z}\|_2^2}, 0\right) \tag{5}$$

where $\mathbf{z}$ is a vector that contains all atoms, $\mathbf{V}_k$ is a diagonal matrix made of 0 and 1 such that $\mathbf{V}_k\mathbf{z}$ select all atoms in the neighborhood of atom z_k.

Instead of using a constant 2D kernel as the neighborhood weights for the 2D convolution to calculate the coefficient sum of each atom's neighborhood in the time-frequency plane [22, 23], we define the neighborhood with two median filters. The median filter is a non-linear spatial filter that sets the coefficient based on the median value among the defined neighbors [24]. In order to better separate the tonal component from the mixture, it is desirable to both suppress the sparsity in the time direction and promote the sparsity in the frequency direction. Therefore, we propose to jointly determine the time-frequency (TF) neighborhood $\mathcal{V}_{tf}$ using neighbors in both the time and frequency directions, as illustrated in Figure 2. The TF neighborhood weight w_k of atom z_k can be formulated as:

$$w_k = \mathrm{med}\{\mathbf{V}_k^t\mathbf{z}\} - \gamma\,\mathrm{med}\{\mathbf{V}_k^f\mathbf{z}\} \tag{6}$$

where $\mathbf{V}_k^t\mathbf{z}$ selects the atoms in the time neighborhood $\mathcal{V}_t(k)$, $\mathbf{V}_k^f\mathbf{z}$ selects the atoms in the frequency neighborhood $\mathcal{V}_f(k)$, γ is a parameter controlling the sparsity along frequency, and $\mathrm{med}\{\mathbf{x}\}$ is the median of $\mathbf{x}$.

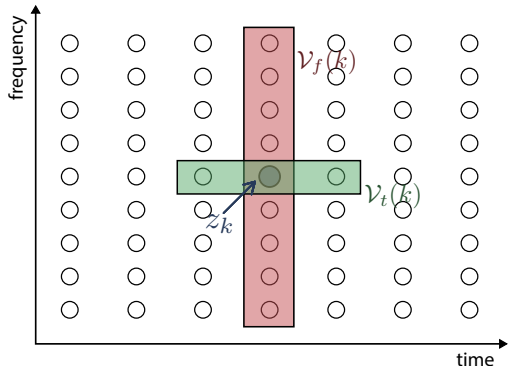


Figure 2: *The neighborhood configuration for tonal decomposition. The dark solid circle represents the center atom z_k, the green horizontal area represents the time neighborhood $\mathcal{V}_t(k)$, and the red vertical area represents the frequency neighborhood $\mathcal{V}_f(k)$.*

The Loris-Verhoeven (LV) algorithm [23] to solve the problem in Eq. (4) is employed. The sparsity parameter λ is automatically tuned based on the time-varying spectral flatness of the signal. The procedure is described on the accompanying website [16].

2.4. Estimation of transient part

After obtaining the residual signal without most of the deterministic part, the next step is to further decompose it into a transient and a stochastic component. The same structured sparse decomposition method as in Section 2.3 is applied, except that the TF neighborhood $\mathcal{V}_{tf}$ is simply the frequency neighborhood $\mathcal{V}_f$. The decomposition result is the transient part of the residual signal. The residual of the decomposition is considered to be the stochastic part, which will be analyzed and reconstructed in a subsequent process. Although the transient part is not synthesized in this research, it is still valuable to consider it as it improves the analysis of the stochastic component.

2.5. Reconstruction of tonal part

The tonal component resulting from the sparse decomposition mainly results from a superposition of partials, which are temporarily evolving sinusoids. Therefore, techniques for analyzing and re-synthesizing partials can be applied to them (from sets of atoms originating from the non-gap regions) and to reconstruct the tonal part in the gap region. In practice this is achieved by performing the following steps: first, signals from the right and left sides of the gap are analyzed to extract partials, which are further processed to be more reliable and consistent; second, the corresponding partials on both sides of the gap are matched and predicted over the gap; finally, the resulting partials are synthesized to obtain the reconstructed signal $\hat{\mathbf{y}}_{\mathrm{tonal}}$.

2.5.1. Partial tracking

Partials are extracted by a technique called partial tracking that aims to build partial trajectories by linking the spectral peaks across frames according to their parameters. For this, we rely on Neri and Depalle's method [25], which treats partial tracking as a combinatorial optimization problem to obtain the optimal connections between peaks by minimizing connection costs. We refine this method by specifically converting partial frequencies from Hz to equivalent rectangular bandwidth (ERB) scale, and by adding an extra term in the cost function that apply a constrain on the frequency derivative differences between spectral peaks.

Let's now describe the assignment problem as presented in [25]. Suppose that the first set of peaks S_1 contains N_1 elements, and the second set of peaks S_2 contains N_2 elements. The assignment problem can be formalized as follows:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} C_{ij}X_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{N_1} X_{ij} = 1 \qquad j = 1,\ldots,N_2 \\ & \sum_{j=1}^{N_2} X_{ij} = 1 \qquad i = 1,\ldots,N_1 \end{aligned} \tag{7}$$

where C_{ij} is the cost of assigning element i in set S_1 to element j in set S_2, X_{ij} is a binary variable indicating the assignment, which is set to 1 if element i is assigned to element j and 0 otherwise. The optimal solution of the assignment problem can be obtained by the Hungarian algorithm [26].

The spectral peak parameters are estimated using the distribution derivative method (DDM) [27]. Continuity constraints between the midpoints of consecutive frames are introduced by incorporating the frequency, amplitude, and frequency derivative differences between the peaks at the midpoint of the frames in the cost function. However, a fixed frequency threshold in Hz may make tracking partials in high frequencies difficult, since the frequency variation is greater at high frequencies than at low frequencies. Therefore, we use the ERB scale frequency difference instead. The frequency (in ERB scale), amplitude, and frequency derivative differences

between peak i in frame $k-1$ and j in frame k are defined as:

$$\Delta f_{ij}^{[k]} = \mathrm{ERB}(f_i^{[k-1]}[H/2]) - \mathrm{ERB}(f_j^{[k]}[-H/2]) \tag{8}$$

$$\Delta a_{ij}^{[k]} = a_i^{[k-1]}[H/2] - a_j^{[k]}[-H/2] \tag{9}$$

$$\Delta \beta_{ij}^{[k]} = \frac{f_i'^{[k-1]}[H/2]}{\mathrm{ERB}\left(f_i^{[k-1]}[H/2]\right)} - \frac{f_j'^{[k]}[-H/2]}{\mathrm{ERB}\left(f_j^{[k]}[-H/2]\right)} \tag{10}$$

where H is the hop size, $f_i^{[k]}[n]$ and $a_i^{[k]}[n]$ are the instantaneous frequency and log-amplitude of partial i in frame k, respectively. The definition $\mathrm{ERB}(f) = \frac{1000}{24.7\times 4.37}\ln(\frac{4.37f}{1000}+1)$ is from [28].

These constraints lead to two types of assignments: *useful* assignments and *spurious* assignments. Useful assignments are those that satisfy the continuity constraints, while spurious assignments are those that do not satisfy them and are thus ignored in the partial tracking process.

The cost of a useful assignment from peak i in frame $k-1$ to peak j in frame k is defined as:

$$C_{ij}^{\text{useful}[k]} = 1 - \exp\left(-\frac{\Delta f_{ij}^{[k]^2}}{2\sigma_f^2} - \frac{\Delta a_{ij}^{[k]^2}}{2\sigma_a^2} - \frac{\Delta \beta_{ij}^{[k]^2}}{2\sigma_\beta^2}\right). \tag{11}$$

The parameters σ_f^2, σ_a^2, and σ_β^2 are the variances of the frequency, amplitude, and normalized frequency derivative distributions, respectively, which are computed as:

$$\sigma_\chi^2 = \frac{\zeta_\chi^2}{2\ln(\delta^{\text{track}}-2) - 2\ln(\delta^{\text{track}}-1)} \tag{12}$$

where χ is a placeholder that represents f, a, and β, ζ_f, ζ_a, and ζ_β are predefined thresholds that control the range of the frequency and amplitude matching, respectively. δ^{track} is the parameter that controls the trade-off between useful and spurious assignments. The cost of a spurious assignment is defined as:

$$C_{ij}^{\text{spurious}} = 1 - (1-\delta^{\text{track}})C_{ij}^{\text{useful}}. \tag{13}$$

To obtain both useful and spurious assignments using the Hungarian algorithm, the cost matrix can be defined as:

$$C_{ij} = \min\{C_{ij}^{\text{useful}}, C_{ij}^{\text{spurious}}\}. \tag{14}$$

Consequently, assignments $X_{ij} = 1$ with $C_{ij} = C_{ij}^{\text{useful}}$ are considered as useful assignments, while those with $C_{ij} = C_{ij}^{\text{spurious}}$ are categorized as spurious assignments. If a useful assignment is not connected to any previous trajectories, this assignment is considered as a born partial. If a previous trajectory does not correspond to any useful assignments in the current slice, the partial is considered as dead. The Hungarian algorithm is used to obtain the optimal assignment matrix by providing the cost matrix $\mathbf{C}$ [25].

2.5.2. *Partial modeling and prediction*

A general model and prediction method is proposed in this section, which will be extensively used and applied to the subsequent processing of partials.

Observations of sound signals indicate that partials may exhibit both trends and periodicities, or only one of them. For instance, a violin can play portamento and vibrato simultaneously, which not only involves a gradual increase or decrease in frequency on a macro scale but also introduces periodic fluctuations in frequency on a micro scale. Therefore, the long-term trend and short-term periodicity should both be taken into account in the model, which leads to a two-step analysis.

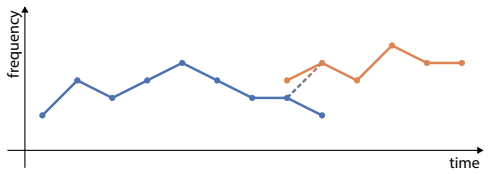
In order to predict the trend component, a linear regression of the frequency (or amplitude) of the partial is performed by calculating the coefficient of determination (R^2). If R^2 is greater than a threshold, the trend exists, and the model is used to predict its value. If it's less than or equal to the threshold, the trend is non-existent, and the trend component is set to 0. A large R^2 indicates the periodicity component is non-existent.

Burg's autoregressive model is employed to predict the periodicity component [29]. The synthesized partial is then obtained by adding the trend prediction to the periodicity prediction. This allows both long-term and short-term variations in partial frequency and amplitude to be captured and rendered.

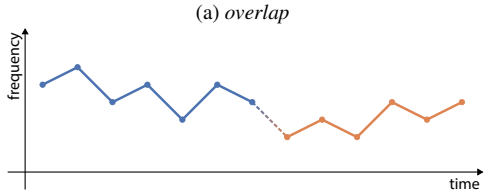
2.5.3. *Partial reconnection*

Ideally, a partial should correspond to an actual part of the sound. In practice, however, the "partials" that we analyze from the algorithms tend to be shorter and more fragmented, which is a distortion compared to the actual partials. These extracted "partials" reduce the accuracy of the prediction, since they carry very limited information. To address this issue, we propose a method to reconnect these fragmented partials based on their frequency and amplitude continuity.

The proposed method can be applied to the partials that fall into the following two scenarios. The first scenario is when the two partials overlap in time by a small amount. The second scenario is when the two partials do not overlap in time, but are close together, which means that there is a small gap between their end and start points. Figure 3 illustrates these two cases accordingly.



(a) *overlap*



(b) *no-overlap*

Figure 3: *Two scenarios of potential partial connection. The solid lines represent extracted partials, the solid points represent data points (per frame), and the dashed lines with gradient color represent potential partial connections.*

The costs of connection are then calculated for each pair of partials. Suppose the long partial is denoted by p_i and the other partial is denoted by p_j. We define the cost for connecting p_j to p_i as:

$$C_{\hat{p}_i \leftarrow p_j}^{\text{connect}} = \delta^{\text{connect}} \frac{\bar{d}_f(\hat{p}_i, p_j)}{\zeta_f^{\text{connect}}} + (1-\delta^{\text{connect}}) \frac{\bar{d}_a(\hat{p}_i, p_j)}{\zeta_a^{\text{connect}}} \tag{15}$$

where ζ_f^{connect} and ζ_a^{connect} are the thresholds for frequency and amplitude, $0 \leq \delta^{\text{connect}} \leq 1$ is a parameter that controls the influence of the two metrics on the cost. $\bar{d}_f(\hat{p}_i, p_j)$ and $\bar{d}_a(\hat{p}_i, p_j)$ represent the normalized Euclidean distances in frequency (ERB scale) or amplitude (dB scale) between two partials p_i and p_j in the range of p_j, which requires the frequency and amplitude of p_i to be mostly extrapolated using the partial model of Section 2.5.2. We use the definition of normalized Euclidean distance from [5]:

$$\bar{d}_f(\hat{p}_i, p_j) = \frac{\| \operatorname{ERB}(\hat{\mathbf{f}}_i) - \operatorname{ERB}(\mathbf{f}_j) \|_2 \Big/ \sqrt{N_{p_j}}}{1 + \sigma\{\operatorname{ERB}(\hat{\mathbf{f}}_i)\} + \sigma\{\operatorname{ERB}(\mathbf{f}_j)\}} \tag{16}$$

$$\bar{d}_a(\hat{p}_i, p_j) = \frac{\|\hat{\mathbf{a}}_i - \mathbf{a}_j\|_2 \big/ \sqrt{N_{p_j}}}{1 + \sigma\{\hat{\mathbf{a}}_i\} + \sigma\{\mathbf{a}_j\}} \tag{17}$$

where $\hat{p}_i$ refers to the prediction of partial p_i, $\hat{\mathbf{f}}_i$ and $\hat{\mathbf{a}}_i$ are the (predicted) frequency and amplitude of partial p_i in the range of partial p_j, N_{p_j} is the length (in frames) of partial p_j, and $\sigma\{\mathbf{x}\}$ is the standard deviation of $\mathbf{x}$.

The merged partial with minimal connection cost (C^{connect}) is determined only if the minimal cost is less than 1. If all costs are greater than 1, the long partial is unable to connect to any other partials. The selected partial is then merged to the long partial. If it overlaps, crossfading is used in the overlapping area to smooth the transition. If it does not overlap, it is simply concatenated to the long one. Once merged, the shorter partial is removed from the list. The process is repeated until all valid partials are processed. The reconnection method can reduce the artificially high number of partials, and results in a more accurate and consistent representation of the signal.

2.5.4. Partial matching

The next step is to determine which partial near the gap's left boundary should be connected to which partial near the gap's right boundary in order to form a merged partial. To achieve this, a method for matching two partials before and after the gap is proposed.

First, all partials with enough length (more than a threshold $l_{\min}^{\text{match}}$) that are near the gap are selected as candidates for matching. Then, all candidate partials around the gap region are extrapolated in the gap region using the prediction method described in Section 2.5.2. The prediction in amplitude extrapolation does not use amplitude parameters from semi-reliable frames[1].

Next, the normalized Euclidean distances (defined in Section 2.5.3) between the left and right predictions for each pair of candidate partials are calculated. A cost matrix based on the normalized Euclidean distances is constructed as follows, similar to the cost matrix for partial tracking:

$$C_{ij} = \min\{C_{ij}^{\text{match}}, C_{ij}^{\text{mismatch}}\} \tag{18}$$

and

$$C_{ij}^{\text{match}} = 1 - \exp\left(-\frac{\left(\bar{d}_a(\hat{p}_i, \hat{p}_j)\right)^2}{2\sigma_a^2} - \frac{\left(\bar{d}_f(\hat{p}_i, \hat{p}_j)\right)^2}{2\sigma_f^2} \right) \tag{19}$$

$$C_{ij}^{\text{mismatch}} = 1 - (1 - \delta^{\text{match}}) C_{ij}^{\text{match}}. \tag{20}$$

Finally, the Hungarian algorithm in [25] is employed to determine the optimal matching. This matching indicates which partials should be connected across the gap.

[1] A semi-reliable frame is defined as a frame in which the portion of the signal being analyzed contains unreliable samples.

2.5.5. Partial extrapolation

After matching the partials near the gap, further extrapolation of these partials is required for inpainting. All partials involved in the partial matching process are further partitioned into three groups: *matched* partials, unmatched *born* partials, and unmatched *dead* partials.

The proposed general partial prediction method is employed for extrapolating all frequency trajectories. However, for the amplitude extrapolation, different strategies are applied to the three groups of partials. The matched partials are interpolated using our partial prediction method. The unmatched born partials are further separated into two types based on their slope of the trend line calculated from the general partial prediction method. Interested readers can find the details process and figures about the strategies at the accompanying website [16].

The phase is reconstructed in the same way as the phase interpolation method in [5], which is based on the method of [30] and further spreads the phase error over the whole gap. The signal $\hat{\mathbf{y}}_{\text{tonal}}$ with all partials is reconstructed using the synthesis method in [30].

2.6. Reconstruction of the noise part

To analyze the stochastic part in order to reconstruct the noise from the residual, the region close to the boundary of the gap is set as unreliable. Because the residual resulting from the sparse decomposition of the tonal part may not be accurate enough near the boundary, and may leak some energy of the tonal part to the residual signal.

We use Burg's AR model to estimate the power spectral density (PSD) of the left and right reliable neighborhoods that consist of noise [31]. Two noise signals that have the same PSDs as the left and right reliable neighborhoods' are generated by filtering a normalized Gaussian noise with linear prediction coefficients calculated from Burg's method. Then, a cosine window is used to crossfade these two noise signals to obtain a smooth transition.

2.7. Post-processing

The complete reconstructed signal $\hat{\mathbf{y}}_{\text{rec}}$ is obtained by superimposing the tonal signal $\hat{\mathbf{y}}_{\text{tonal}}$ and the noise signal $\hat{\mathbf{y}}_{\text{noise}}$ together. In order to keep the reliable part of the original signal unchanged, only the gap region will be replaced by the reconstructed signal, with a short crossfade at the boundaries of the gap to suppress potential discontinuities.

3. EXPERIMENTS AND RESULTS

In the following two experiments, we compare the reconstruction quality of our hybrid approach (referred to as Hybrid) with four state-of-the-art inpainting methods: the analysis variant of SPAIN method (A-SPAIN) [6], the weighted Chambolle-Pock method (w-CP) [7], the iteratively reweighted Chambolle-Pock method (re-CP) [7], and the frame-wise Janssen method (Janssen) [2]. The first three methods are based on sparse decomposition, and the last method is based on AR modeling. All sparsity-based methods use the half offset configuration. The window sizes (from 2800 to 8400 samples) and hop sizes (1/4 of the window sizes) are determined based on the length of the gaps. Other parameters are set as the same values in [7]. For the Janssen method, we set the number of iterations to 20. For all signals used in the experiments, the sampling rate is 44.1 kHz.

Two metrics that are used to evaluate the quality of the reconstructed signals for the proposed hybrid inpainting approach and other methods. The first metric is the signal-to-noise ratio (SNR), which is defined as:

$$\mathrm{SNR}(\mathbf{y}, \hat{\mathbf{y}}_{\mathrm{rec}}) = 10 \log_{10} \frac{\|\mathbf{y}\|_2^2}{\|\mathbf{y} - \hat{\mathbf{y}}_{\mathrm{rec}}\|_2^2} \tag{21}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}}_{\mathrm{rec}}$ represent the original signal (without gaps) and the reconstructed signal, respectively. A higher SNR value indicates a better reconstruction of the audio signal. In the following experiment, the SNR is computed only in the gap region.

The second metric is a perceptual metric called objective difference grade (ODG), which measures the perceptual similarity between the original and reconstructed signals [32]. The ODG corresponds to the subjective difference grade obtained from subjective listening tests, which ranges from 0 to –4 and can be interpreted as imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying, respectively. We use the PEMO-Q method to calculate the ODG, which has an advanced auditory model based on a modulation filterbank and demonstrates high prediction accuracy [33]. In addition, we record the elapsed time for each inpainting method to produce results.

In the first experiment, we demonstrate the adaptability and flexibility of our hybrid approach on handling two non-stationary signals. The first example is a synthesized sound with quadratic chirps, random exponentially damped sinusoids, and added noise (with a noise level of –20 dB), and the second is a soprano recording with vibrato from the Sound Quality Assessment Material (SQAM) dataset [34]. The gap length is 50 milliseconds for both signals. The reconstruction results using the proposed hybrid approach, A-SPAIN, and Janssen for the two test signals are presented in Figures 4 and 5, respectively. The figures for other methods are at the accompanying website [16].

As for the synthesized signal, w-CP (SNR = –6.38 dB, ODG = –2.12) and re-CP (SNR = –7.08 dB, ODG = –2.37) fail to reconstruct the tonal (chirp) part, and the re-CP method discards the noise in the gap. At the same time, A-SPAIN (Figure 4b) (SNR = –1.18 dB, ODG = –1.50) and Janssen (Figure 4c) (SNR = –1.31 dB, ODG = –2.11) cannot adapt to the variations of frequencies in the gap due to their stationary assumptions, leading to frequency jumps and the "freezing" of noise. However, the hybrid approach (Figure 4a) (SNR = 4.23 dB, ODG = –0.76) successfully captured the features from the reliable neighborhoods and accurately predicted both tonal and noise components for this signal.

As for the soprano signal with vibrato, w-CP (SNR = –3.37 dB, ODG = –1.28) and re-CP (SNR = –2.39 dB, ODG = –2.35) failed to inpaint the tonal component with modulation. Meanwhile, A-SPAIN (Figure 5b) (SNR = 0.46 dB, ODG = –1.04) and Janssen (Figure 5c) (SNR = 2.67 dB, ODG = –1.00) fail to connect the correct partials. The hybrid approach (Figure 5a) (SNR = –3.85 dB, ODG = –2.30) shows the most similar trajectories as the original audio signal with the correctly inpainted partials with modulation and captures some of the noise.

In the second experiment, we compare these methods quantitatively using six recordings with different characteristics from the SQAM dataset. Each signal has 8 gaps at random positions.

The evaluations of different audio inpainting methods under three metrics are shown in Figure 6. In terms of SNR, when the gap length is lower than 50 ms, the hybrid approach does not have the same good SNR as other methods. However, it starts to outperform other methods (except re-CP) for gaps longer than 50 ms. The results in terms of ODG is similar, except that the cutoff is at 100 ms. The re-CP method achieve the best SNR for gaps longer than 50 ms, but it has worst ODG for gaps longer than 25 ms. Furthermore, as for the running time, the sparsity-based methods (A-SPAIN, w-CP, re-CP) and AR-based model (Janssen) have an increasing runtime when the gap length grows, and Janssen's runtime increases exponentially. Our hybrid approach, unlike other compared methods, are insensitive with gap length.

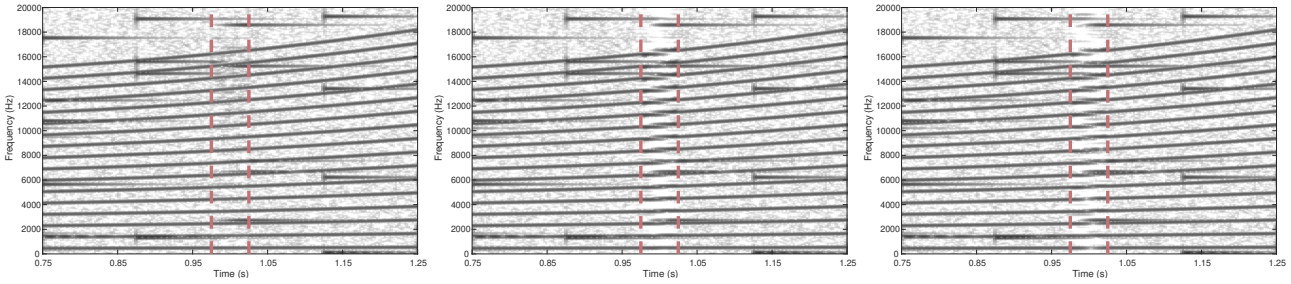
From the results of first experiment, both sparsity-based and AR-based methods failed in reconstructing incorrect partial trajectories, but for different reasons. For the sparsity-based methods (A-SPAIN, w-CP, and re-CP), although a long window might reduce the energy loss in the gap, time-varying partials will be analyzed as smoothed stationary atoms, resulting in a reconstruction that looks like a jump from the left reliable part to the right reliable part with a crossfade rather than a continuity within the gap region. The AR-based method (Janssen) benefits from an increased window size, thus better capturing the underlying temporal relationships within the signal, such as modulations. However, it is difficult to predict the long-term trend of partial parameters using an AR model.

The degradation of the reconstruction quality of our hybrid approach may be attributed to the following reasons. First, it is difficult to spread the phase error when synthesizing the partials when gap is short, which leads to a more pronounced discontinuity and lowers the ODG. Since we did not use the information within the semi-reliable frames to extrapolate the partial parameters, the phase of the reconstructed signal in the gap region may vary from the original signal, resulting in a reduced SNR, even though they are similar in terms of perception. Moreover, since the preset parameters are used to inpaint all types of signals, the partial matching method sometimes mismatch two partials that represent different notes together. Fine-tuning parameters for partial tracking and matching methods based on signal type and gap length can improve the accuracy of these methods. In future research, it is possible to incorporate a harmonicity constraint in order to enhance robustness in this situation.

The proposed hybrid approach includes a large number of parameters that need to be flexibly adjusted to different types of signals to obtain better reconstruction quality. A detailed description and explanation of the parameters as well as experimental results are available at the accompanying website with parameters' information, audio excerpts, supplemental figures, and MATLAB implementation [16].

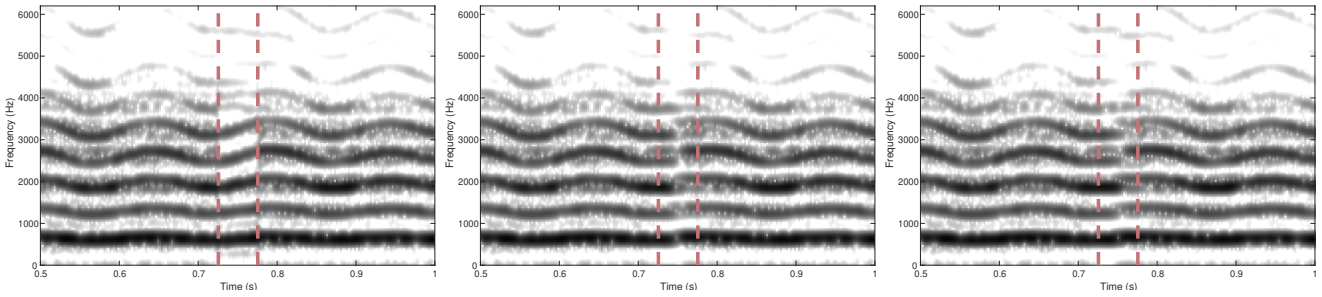
4. CONCLUSION

This paper proposes a hybrid audio inpainting approach that takes into account the diversity of audio signals. This approach solves the inpainting problem in a structured way as it decomposes the signal into tonal, transient, and noise components and reconstructs them separately using refined component-targeted methods with various controlling parameters for fine-tuning the behaviors of the methods. Results show that the proposed approach is flexible and adaptive with various lengths of gaps, especially for signals with medium gaps (50–150 ms) and non-stationary components. Furthermore, our hybrid approach scarcely increases the running time as the gap length grows. Future work may reconstruct the transient component in the gap. Moreover, other audio degradations, such as clipping and bandlimiting, may be reconstructed with the three-layer structured audio processing approach.



(a) *Hybrid* (b) *A-SPAIN* (c) *Janssen*

Figure 4: *Comparison of reconstruction of audio inpainting methods for synthesized chirps and exponentially damped sinusoids with added noise. The area between the two red dashed lines represents the unreliable region.*



(a) *Hybrid* (b) *A-SPAIN* (c) *Janssen*

Figure 5: *Comparison of reconstruction of audio inpainting methods for the soprano recording with vibrato. The area between the two red dashed lines represents the unreliable region.*

5. REFERENCES

[1] Amir Adler, Valentin Emiya, Maria G. Jafari, Michael Elad, Rémi Gribonval, and Mark D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[2] Augustus Janssen, Raymond Veldhuis, and Lodewijk Vries, "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330, 1986.

[3] Walter Etter, "Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters," *IEEE Transactions on Signal Processing*, vol. 44, no. 5, pp. 1124–1135, 1996.

[4] Robert C. Maher, "A method for extrapolation of missing digital audio data," in *Proceedings of the Audio Engineering Society 95th Convention*, New York, NY, USA, 1993.

[5] 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, 2005.

[6] Ondřej Mokrý, Pavel Záviška, Pavel Rajmic, and Vítězslav Veselý, "Introducing SPAIN (SParse audio INpainter)," in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019, pp. 1–5.

[7] Ondřej Mokrý and Pavel Rajmic, "Audio inpainting: Revisited and reweighted," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2906–2918, 2020.

[8] Tomoro Tanaka, Kohei Yatabe, and Yasuhiro Oikawa, "Phase-aware audio inpainting based on instantaneous frequency," in *Proceedings of 2021 APSIPA Annual Summit and Conference*, Tokyo, Japan, 2021, pp. 254–258.

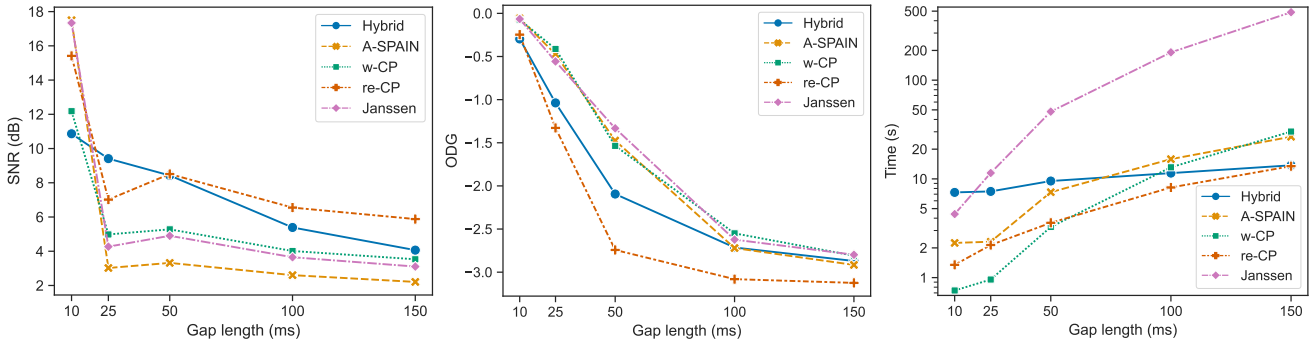
[9] Andrés Marafioti, Nicki Holighaus, Piotr Majdak, and Nathanaël Perraudin, "Audio inpainting of music by means of neural networks," in *Proceedings of the Audio Engineering Society 146th Convention*, Dublin, Ireland, 2019.

[10] Eloi Moliner and Vesa Välimäki, "Diffusion-based audio inpainting," *Journal of the Audio Engineering Society*, vol. 72, no. 3, pp. 100–113, 2024.

[11] Federico Miotello, Mirco Pezzoli, Luca Comanducci, Fabio Antonacci, and Augusto Sarti, "Deep Prior-Based Audio Inpainting Using Multi-Resolution Harmonic Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 113–123, 2024.

[12] Tony S. Verma and Teresa H. Y. Meng, "Extending spectral modeling synthesis with transient modeling synthesis," *Computer Music Journal*, vol. 24, no. 2, pp. 47–59, 2000.

[13] Charturong Tantibundhit, J. Robert Boston, Ching-Chung Li, John D. Durrant, Susan Shaiman, Kristie Kovacyk, and Amro El-Jaroudi, "Speech enhancement using transient speech components," in *Proceedings of the 2006 IEEE International Conference on Acoustics Speed and Signal Processing (ICASSP)*, Toulouse, France, 2006, vol. 1, pp. 833–836, IEEE.



(a) *SNR* (b) *ODG* (c) *Time*

Figure 6: *Comparison of audio inpainting methods under different gap lengths in terms of SNR (lower are better), ODG (higher are better), and runtime. The runtime of Janssen exceeds the boundary at a gap length of 25 ms and keeps growing as the gap length increases.*

[14] Laurent Daudet and Bruno Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, 2002.

[15] Kai Siedenburg and Monika Dörfler, "Structured sparsity for audio signals," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*, Paris, France, 2011, pp. 23–26.

[16] Eto Sun, "Hybrid Audio Inpainting Approach," `https://etosphere.github.io/hybrid-audio-inpainting-approach/`, 2024.

[17] Pavel Rajmic, Hana Bartlová, Zdeněk Průša, and Nicki Holighaus, "Acceleration of audio inpainting by support restriction," in *Proceedings of the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Brno, Czech Republic, 2015, pp. 325–329, IEEE.

[18] Corey Kereliuk and Philippe Depalle, "Sparse atomic modeling of audio: A review," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx)*, Paris, France, 2011, pp. 81–92.

[19] Sangnam Nam, Mark E. Davies, Michael Elad, and Rémi Gribonval, "The cosparse analysis model and algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[20] Balas K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[21] Matthieu Kowalski, Kai Siedenburg, and Monika Dörfler, "Social sparsity! Neighborhood systems enrich structured shrinkage operators," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2498–2511, 2013.

[22] Kai Siedenburg, Matthieu Kowalski, and Monika Dörfler, "Audio declipping with social sparsity," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 1577–1581.

[23] Pavel Záviška and Pavel Rajmic, "Audio declipping with (weighted) analysis social sparsity," in *Proceedings of the 45th International Conference on Telecommunications and Signal Processing (TSP)*, Prague, Czech Republic, 2022, pp. 407–412.

[24] Derry FitzGerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, 2010.

[25] Julian Neri and Philippe Depalle, "Fast partial tracking of audio with real-time capability through linear programming," in *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx)*, Aveiro, Portugal, 2018, pp. 326–333.

[26] Harold W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[27] Michaël Betser, "Sinusoidal polynomial parameter estimation using the distribution derivative," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4633–4645, 2009.

[28] Brian R Glasberg and Brian C.J Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[29] Steven M. Kay, "Autoregressive spectral estimation: Methods," in *Modern Spectral Estimation: Theory and Application*, Prentice-Hall Signal Processing Series, pp. 217–270. Prentice Hall, Englewood Cliffs, NJ, 1988.

[30] 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.

[31] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, New York, NY, USA, 1996.

[32] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation BS.1387-0, Geneva, Switzerland, 1998.

[33] Rainer Huber and Birger Kollmeier, "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[34] European Broadcasting Union, "Sound quality assessment material recordings for subjective tests," https://tech.ebu.ch/publications/sqamcd, 2008.