Title: Audio inpainting

URL: https://en.wikipedia.org/wiki/Audio_inpainting

PageID: 74274244

Categories: Category:Deep learning, Category:Digital signal processing, Category:Machine learning

---

Audio inpainting (also known as audio interpolation) is an audio restoration task which deals with the reconstruction of missing or corrupted portions of a digital audio signal . Inpainting techniques are employed when parts of the audio have been lost due to various factors such as transmission errors, data corruption or errors during recording.

The goal of audio inpainting is to fill in the gaps (i.e., the missing portions) in the audio signal seamlessly, making the reconstructed portions indistinguishable from the original content and avoiding the introduction of audible distortions or alterations.

Many techniques have been proposed to solve the audio inpainting problem and this is usually achieved by analyzing the temporal and spectral information surrounding each missing portion of the considered audio signal.

Classic methods employ statistical models or digital signal processing algorithms to predict and synthesize the missing or damaged sections. Recent solutions, instead, take advantage of deep learning models, thanks to the growing trend of exploiting data-driven methods in the context of audio restoration.

Depending on the extent of the lost information, the inpainting task can be divided in three categories.

Short inpainting refers to the reconstruction of few milliseconds (approximately less than 10) of missing signal, that occurs in the case of short distortions such as clicks or clipping . In this case, the goal of the reconstruction is to recover the lost information exactly.

In long inpainting instead, with gaps in the order of hundreds of milliseconds or even seconds, this goal becomes unrealistic, since restoration techniques cannot rely on local information. Therefore, besides providing a coherent reconstruction, the algorithms need to generate new information that has to be semantically compatible with the surrounding context (i.e., the audio signal surrounding the gaps). The case of medium duration gaps lays between short and long inpainting.

It refers to the reconstruction of tens of millisecond of missing data, a scale where the non-stationary characteristic of audio already becomes important.

Definition

Consider a digital audio signal x {\displaystyle \mathbf {x} } . A corrupted version of x {\displaystyle \mathbf {x} } , which is the audio signal presenting missing gaps to be reconstructed, can be defined as x ~ = m ■ x {\displaystyle \mathbf {\tilde {x}} =\mathbf {m} \circ \mathbf {x} } , where m {\displaystyle \mathbf {m} } is a binary mask encoding the reliable or missing samples of x {\displaystyle \mathbf {x} } , and ■ {\displaystyle \circ } represents the element-wise product . Audio inpainting aims at finding x ^ {\displaystyle \mathbf {\hat {x}} } (i.e., the reconstruction), which is an estimation of x {\displaystyle \mathbf {x} } . This is an ill-posed inverse problem , which is characterized by a non-unique set of solutions. For this reason, similarly to the formulation used for the inpainting problem in other domains, the reconstructed audio signal can be found through an optimization problem that is formally expressed as

x ^ ∗ = argmin X ^ L (m ■ x ^ , x ~) + R (x ^) {\displaystyle \mathbf {\hat {x}} ^{*}={\underset {\hat {\mathbf {X} }}{\text{argmin}}}~L(\mathbf {m} \circ \mathbf {\hat {x}} ,\mathbf {\tilde {x}})+R(\mathbf {\hat {x}})} .

In particular, x ^ ∗ {\displaystyle \mathbf {\hat {x}} ^{*}} is the optimal reconstructed audio signal and L {\displaystyle L} is a distance measure term that computes the reconstruction accuracy between the corrupted audio signal and the estimated one. For example, this term can be expressed with a mean squared error or similar metrics .

Since L {\displaystyle L} is computed only on the reliable frames, there are many solutions that can minimize L (m ■ x ^ , x ~) {\displaystyle L(\mathbf {m} \circ \mathbf {\hat {x}} ,\mathbf {\tilde {x}})} . It is thus necessary to add a constraint to the minimization, in order to restrict the results only to the valid solutions. This is expressed through the regularization term R {\displaystyle R} that is computed on the reconstructed audio signal x ^ {\displaystyle \mathbf {\hat {x}} } . This term encodes some kind of a-priori information on the audio data. For example, R {\displaystyle R} can express assumptions on the stationarity of the signal, on the sparsity of its representation or can be learned from data.

Techniques

There exist various techniques to perform audio inpainting. These can vary significantly, influenced by factors such as the specific application requirements, the length of the gaps and the available data. In the literature, these techniques are broadly divided in model-based techniques (sometimes also referred as signal processing techniques) and data-driven techniques.

Model-based techniques

Model-based techniques involve the exploitation of mathematical models or assumptions about the underlying structure of the audio signal. These models can be based on prior knowledge of the audio content or statistical properties observed in the data. By leveraging these models, missing or corrupted portions of the audio signal can be inferred or estimated.

An example of a model-based techniques are autoregressive models . These methods interpolate or extrapolate the missing samples based on the neighboring values, by using mathematical functions to approximate the missing data. In particular, in autoregressive models the missing samples are completed through linear prediction. The autoregressive coefficients necessary for this prediction are learned from the surrounding audio data, specifically from the data adjacent to each gap.

Some more recent techniques approach audio inpainting by representing audio signals as sparse linear combinations of a limited number of basis functions (as for example in the Short Time Fourier Transform). In this context, the aim is to find the sparse representation of the missing section of the signal that most accurately matches the surrounding, unaffected signal.

The aforementioned methods exhibit optimal performance when applied to filling in relatively short gaps, lasting only a few tens of milliseconds, and thus they can be included in the context of short inpainting. However, these signal-processing techniques tend to struggle when dealing with longer gaps. The reason behind this limitation lies in the violation of the stationarity condition, as the signal often undergoes significant changes after the gap, making it substantially different from the signal preceding the gap.

As a way to overcome these limitations, some approaches add strong assumptions also about the fundamental structure of the gap itself, exploiting sinusoidal modeling or similarity graphs to perform inpainting of longer missing portions of audio signals.

Data-driven techniques

Data-driven techniques rely on the analysis and exploitation of the available audio data. These techniques often employ deep learning algorithms that learn patterns and relationships directly from the provided data. They involve training models on large datasets of audio examples, allowing them to capture the statistical regularities present in the audio signals. Once trained, these models can be used to generate missing portions of the audio signal based on the learned representations, without being restricted by stationarity assumptions. Data-driven techniques also offer the advantage of adaptability and flexibility, as they can learn from diverse audio datasets and potentially handle complex inpainting scenarios.

As of today, such techniques constitute the state-of-the-art of audio inpainting, being able to reconstruct gaps of hundreds of milliseconds or even seconds. These performances are made possible by the use of generative models that have the capability to generate novel content to fill in the missing portions. For example, generative adversarial networks , which are the state-of-the-art of generative models in many areas, rely on two competing neural networks trained simultaneously in a two-player minmax game : the generator produces new data from samples of a random variable, the discriminator attempts to distinguish between generated and real data. During the training, the generator's objective is to fool the discriminator, while the discriminator attempts to learn to better classify real and fake data.

In GAN-based inpainting methods the generator acts as a context encoder and produces a plausible completion for the gap only given the available information surrounding it. The discriminator is used to train the generator and tests the consistency of the produced inpainted audio.

Recently, also diffusion models have established themselves as the state-of-the-art of generative models in many fields, often beating even GAN-based solutions. For this reason they have also been used to solve the audio inpainting problem, obtaining valid results. These models generate new data instances by inverting the diffusion process, where data samples are progressively transformed into Gaussian noise.

One drawback of generative models is that they typically need a huge amount of training data . This is necessary to make the network generalize well and make it able to produce coherent audio information, that also presents some kind of structural complexity. Nonetheless, some works demonstrated that, capturing the essence of an audio signal is also possible using only a few tens of seconds from a single training sample. This is done by overfitting a generative neural network to a single training audio signal. In this way, researchers were able to perform audio inpainting without exploiting large datasets.

Applications

Audio inpainting finds applications in a wide range of fields, including audio restoration and audio forensics among the others. In these fields, audio inpainting can be used to eliminate noise, glitches, or undesired distortions from an audio recording, thus enhancing its quality and intelligibility. It can also be employed to recover deteriorated old recordings that have been affected by local modifications or have missing audio samples due to scratches on CDs .

Audio inpainting is also closely related to packet loss concealment (PLC). In the PLC problem, it is necessary to compensate the loss of audio packets in communication networks. While both problems aim at filling missing gaps in an audio signal, PLC has more computation time restrictions and only the packets preceding a gap are considered to be reliable (the process is said to be causal).

See also

Audio forensics

Audio restoration

Image inpainting

Packet loss concealment

References