COMPARISON OF INPAINTING METHODS FOR AUDIO RECONSTRUCTION

Daniele Ugo Leonzio, Simone Mariani

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[danieleugo.leonzio, simone5.mariani]@mail.polimi.it

ABSTRACT

Image inpainting methods have been much improved in the last few years. Their goal is to restore a corrupted image. In the audio field, some degradation/restoration problems can be faced with a similar approach for the same image related problems. In fact, an audio source can be considered an image just applying a time-frequency transform on it. In this paper we analyze and compare various inpainting methods on how they perform in restoring audio signals. From the obtained results we proved the effectiveness of the image inpainting on audio spectrogram. There is not a general solution to the problem, but there is a trade-off among the methods depending on the details of every specific problem.

1. INTRODUCTION

Image inpainting, as said in [1], refers to the process of filling-in missing data in a designated region of the visual input. The object of the process is to reconstruct missing parts or damaged image in such a way that the inpainted region cannot be detected by a causal observer. Typically, after the user selects the region to be restored, the inpainting algorithm automatically repairs the damaged area by means of image interpolation.
Our idea is to compare some inpainting methods on how they perform in reconstructing audio spectrograms, which can be treated as images.
We start from an audio file we convert it in time-frequency representation. Then we apply a mask on the spectrogram to select regions that need restoration or in which we want to filter out some audio parts. Now the masked spectrogram is treated as an image and we execute the selected inpainting method on it.
Finally we go back in time domain either using the original phase or approximating it with Griffin-Lim method. In our research we evaluated three different inpainting procedures ranging from mathematical models to deep convolutional neural networks.
The purpose of this study is to prove that image inpainting methods are useful also in the audio field both as restoration technique and as filtering way to remove undesired audio parts.
The paper is orgainzed as follows: section 2 describes theoretically how the method works, section 3 contains the details of our experimental setup, in section 4 we show the results, in section 5 there is a discussion on the obtained results and final conclusions.

2. AUDIO INPAINTING

The process flow is modeled on four stages pipeline as shown in Figure 1:

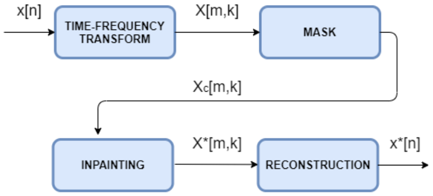


Figure 1: Pipeline flow

2.1. Time-Frequency Transform

In order to have more variability we adopted two types of time-frequency representations.

2.1.1. STFT

Given a signal x(t) we sample it to get its discrete version x[n], with t = nT = n/fs where fs is the sampling frequency. Then its Short Time Fourier Transform is:

$$\mathbf{STFT}\{x[n]\}(m,k) \equiv X(m,k) = \sum_{n=0}^{L-1} x[n]w[n-m]e^{-j2\pi kn/L}$$

In particular m indicates the temporal bins, k represents the frequency bins, w[n] is the analysis window with a length L.

2.1.2. MEL

The MEL spectrogram is a time-frequency representation where the frequency axis is based on the MEL scale. To obtain such spectrogram the starting point is a STFT and then the following transformation is applied to the frequency axis:

$$mel(f) = \begin{cases} f & \text{if } f \leq 1kHz \\ 2595 \cdot \log(1 + \frac{f}{700}) & \text{if } f > 1kHz \end{cases}$$

This is just one of the several ways in which the MEL scale can be approximated.

2.2. Masking

The mask is a binary element applied on the original image to select parts of it. Any type of mask can be implemented, depending on how the inpainting technique is used. The mask can, for example, recreate typical degradation patterns such as 'Salt Pepper Noise' or can be used to cancel the audio parts that we want to filter out.

2.3. Inpainting Method

2.3.1. Model Based

The model based algorithm is described in the papers [2] and [3]. In their work Damelin and Hoang described with numerical experiments how image inpainting can be done using harmonic and biharmonic functions. The problem to solve is:

$$\Delta^2 u = 0$$
$$\Delta u|_S = \Delta \mathrm{u}_0|_S$$
$$u|_S = \mathrm{u}_0|_S$$

Where the element u is a biharmonic function that solves this problem. The boundary data used for constructing biharmonic functions are the values of the laplacian and normal derivatives of the functions on the boundary. To solve the harmonic functions a finite difference approach has been adopted. Using the finite difference scheme the equation is reduced to a linear algebraic system of the form Au = b.
Chui and Mhaskar, in their study [3], intended to extend and generalize this approach to nonstationary smooth function. The idea was to extend previous works on image inpainting, from isotropic diffusion to anisotropic diffusion and from bi-harmonic extension to multi-level lagged anisotropic diffusion extension.

2.3.2. CNN Image Translation

Differently from previous one, this technique is based on 'Convolutional Neural Networks (CNN)'. The aim is to build a network able to learn a mapping from input to output images.
Among the various image to image translation networks we selected the 'Pix2Pix'. The release of the Pix2Pix software is associated with the paper [4], where the authors explain theoretically how the software works.
The network adopted in the Pix2Pix software is based on a Conditional GAN (Generative Adversarial Network) scheme.
A definition of GAN can be found in [5]: *"Generative adversarial networks are based on a game theoretic scenario in which the generator network must compete against an adversary. The generator network directly produces samples. Its adversary, the discriminator network, attempts to distinguish between samples drawn from the training data and samples drawn from the generator"*.
The Conditional GAN attach additional information to the two models, in order to generate samples of a given type. The objective of a conditional GAN can be expressed as:

$$\mathcal{L}_{cGAN}(G, D) =$$
$$E_{x,y}[log(D(x, y)] + E_{x,z}[log(1 - D(x, G(x, z))]$$

Where G tries to minimize this objective against an adversarial D that tries to maximize it.

2.3.3. Deep Prior Image

Also this approach is based on CNN. In their study [6], they show that a randomly-initialized neural network can be used as a handcrafted prior with excellent results in standard inverse problems such as denoising, superresolution, and inpainting.
Deep networks are applied to image generation by learning generator/decoder networks x = f(z) that map a random code vector z to an image x. This approach can be used to sample realistic images from a random distribution. The aim of that research is to investigate the prior implicitly captured by the choice of a particular generator network structure, before any of its parameters are learned. The inverse task they analyzed, can be expressed as energy minimization problems of the type

$$\mathbf{x}^* = \min_{\mathbf{x}} \mathbf{E}(\mathbf{x}; \mathbf{x_0}) + \mathbf{R}(\mathbf{x})$$

Where E(x; x_0) is a task-dependent data term, x_0 the noisy/low-resolution/occluded image, and R(x) a regularizer.
The corresponding data term in case of inpainting is given by

$$\mathrm{E}(\mathrm{x}; \mathrm{x}_0) = |(x - x_0) \odot m|^2$$

Where $\odot$ is Hadamard's product, m is the binary mask.

2.4. Reconstruction

2.4.1. Original Phase

In order to reconstruct with original phase, we assume to know the original signal at the reconstruction point. In case of a restoration of a pre-degradated signal, this approach could not be applied. Whereas this can be an effective way to go back to the time domain when the method is used to filter the original signal. To synthesize the time domain signal we applied the inverse STFT on the inpainted spectrogram.

2.4.2. Griffin-Lim

The Griffin-Lim algorithm as explained in [7], is a phase reconstruction method based on the redundancy of the short-time Fourier transform. It promotes the consistency of a spectrogram by iterating two projections, where a spectrogram is said to be consistent when its inter-bin dependency owing to the redundancy of STFT is retained. Griffin-Lim algorithm is based only on the consistency and does not take any prior knowledge about the target signal into account.

3. EXPERIMENTAL SETUP

3.1. Dataset

The data used in our experiment come from the 'Mobiphone' dataset. It is composed by voice recordings taken from twenty different mobile phones each one with twentyfour speakers. The mobile phones recordings are characterized by different sampling frequencies, so to have a standard value we set it to 16 kHz. In addition since the recordings have different durations we decided to cut them to their minimum, that is 20 seconds.

3.2. Technical Details

In order to treat audio as an image is essential to go into time-frequency domain, to do so we followed two alternative paths: STFT or log MEL spectrogram.
The STFT has been computed with a rectangular window 625 samples long, hop length equal to the window length to avoid overlap between consecutive windows. The number of FFT points has been set to 1023, so in the end we get a spectrogram with a size of 512x512 that fits the inpainting methods size requirements.
For the log MEL spectrogram computation we exploited the already implemented librosa function which expects as input a STFT. So we

Time-Frequency Transform	Reconstruction	MODEL BASED				PIX 2 PIX				DEEP PRIOR			
		SNR	PSNR	SSIM	PESQ	SNR	PSNR	SSIM	PESQ	SNR	PSNR	SSIM	PESQ
STFT	Original Phase	14.189	28.361	0.926	1.869	16.923	24.789	0.854	1.740	19.271	21.641	0.557	1.055
STFT	Griffin-Lim	15.013	27.439	0.923	1.374	17.138	24.800	0.854	1.331	20.310	21.752	0.561	1.056
log MEL	Original Phase	14.308	27.006	0.925	1.852	17.179	25.758	0.897	1.739	18.538	20.398	0.621	1.062
log MEL	Griffin-Lim	14.956	27.006	0.925	1.373	16.173	25.755	0.897	1.319	18.903	20.483	0.626	1.057

Table 1: Time Results

Time-Frequency Transform	Reconstruction	MODEL BASED				PIX 2 PIX				DEEP PRIOR			
		SNR	PSNR	SSIM	PESQ	SNR	PSNR	SSIM	PESQ	SNR	PSNR	SSIM	PESQ
STFT	Original Phase	13.428	41.860	0.995	2.586	15.437	28.138	0.911	1.960	18.774	22.032	0.591	1.066
STFT	Griffin-Lim	16.463	41.860	0.995	1.634	16.007	28.138	0.911	1.375	19.931	22.053	0.594	1.056
log MEL	Original Phase	14.890	29.682	0.988	2.039	15.006	27.942	0.928	1.972	18.326	20.918	0.653	1.066
log MEL	Griffin-Lim	17.807	29.682	0.988	1.431	15.073	27.927	0.925	1.386	19.166	21.249	0.670	1.062

Table 2: Frequency Results

passed to the function a STFT computed with the same parameters explained before.

After that we converted the spectrogram magnitude in dB scale. Moreover, to handle spectrogram as an image we mapped the transform values into 256 levels, thanks to that we have grayscale images with values ranging from 0 to 255.

The next step is building the mask. The general idea is to create a mask that enlarges at every iteration of a loop. We investigated two different mask applications. The first one completely deletes rows, that is equivalent to cancel time instants. The second one acts on the columns, so on frequency removal. For both of them we implemented a 5 steps cycle, but due to a different impact between times and frequencies deletion, we adopted different increments for the two masks. The time mask increment is of 20 columns per iteration, instead the frequency mask increment is of 2 rows per iteration. Depending on the inpainting method, the region selected by the mask requires to be set to zeros or ones.

After applying the mask to the spectrogram image, we give it as input to one of the inpainting methods. The first one, the model based reconstruction, is implemented in the scikit image library through its 'Inpainting' function. In our code we used this function with default parameters.

The second one is the Pix2Pix net. Since it's a CNN and works on a loss function minimization, before applying the method it requires a training process. This process will strongly affect the performance of the reconstruction. In our setup, the training is performed over 5 epochs, on 456 signals (19 classes x 24 files/classes). For both generator and discriminator the method imposes an Adam optimizer with a learning rate of $2 * 10^{-4}$ and β_1 equal to 0.5. The loss function is a cross-entropy loss.

The last one, which is Deep Image Prior, allows to select different types of net inside its code. We choosed to work on the 'skip depth 6'. The other parameters are 400 epochs, Adam optimizer with a learning rate of 0.01 and MSE as loss function.

After the execution of the inpainting method, we get the restored image. Then we follow the inverse path to go back to the time domain signal. First, we map the image values to the original spectrogram value scale. Subsequently we reconstruct the audio signal either with original phase or with the Griffin-Lim. The original phase reconstruction needs to apply just the inverse STFT with the same parameters used in analysis. The Griffin-Lim method is implemented in Librosa and we used that function with default parameters.

4. RESULTS

To evaluate the performances of the inpainting methods we chose four metrics, two based on the reconstructed audio signal and the other two on the restored spectrogram image.

These metrics are:

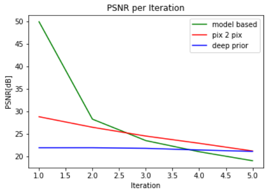
1. **SNR**
   Signal-to-noise ratio (SNR) is defined as the ratio of the power of a signal to the power of background noise.
   Often and also in our case this is measured in dB.
   The higher the value of SNR, the better will be the quality of the output signal.

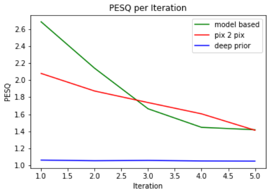
2. **PSNR**
   The PSNR calculates the PSNR ratio in decibels between two images. We often use this ratio as a measurement of quality between the original image and the resultant image. The higher the value of PSNR, the better will be the quality of the output image.

3. **SSIM**
   The Structural Similarity Index (SSIM) is a perceptual metric that quantifies image quality degradation caused by processing such as data compression or by losses in data transmission.
   SSIM actually measures the perceptual difference between two similar images.
   We used the function implemented in the Python 'Scikit Image' library which is based on the papers [8] and [9].
   SSIM values range from 0 to 1, where 0 means no similarity and 1 indicates perfect similarity.

4. **PESQ**
   Perceptual Evaluation of Speech Quality (PESQ) is a family of standards comprising a test methodology for automated assessment of the speech quality as experienced by a user of a telephony system.
   This metric was introduced by the work [10], in which the authors gave a theoretical description on this evaluation method.
   We added this metric in our system because our dataset is composed by mobile phone voice recordings, so it was useful to evaluate the perceptual speech quality.
   The PESQ values range from a minimum of -0.5 to a maximum of 4.5. The higher the value, the higher the audio quality.

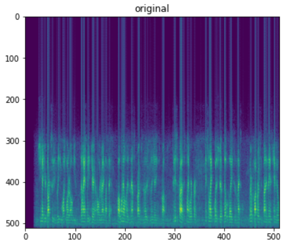
As shown in Table 1 and Table 2, the best results are obtained with the Model Based method with the following combination: STFT and Original Phase reconstruction. So, we selected this combination to plot a comparison between the three methods, considering the time axis degradation:

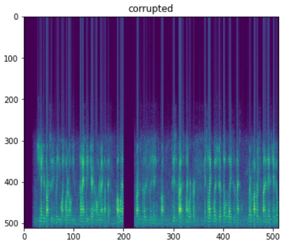


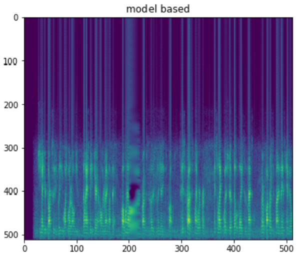


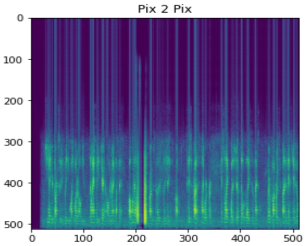
From this graphs, we notice that the three inpainting methods have different slopes. The model based has the fastest decrease, whereas the deep prior is almost constant. This suggests that the model based is very good for the restoration of small portions of image, instead the other two methods perform better for bigger losses.

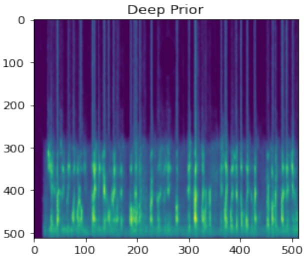
Furthermore, we show you how the methods act on the spectrogram images. We focused on the following spectrogram, applying a degradation of 20 columns (so on time axis):











5. DISCUSSION AND CONCLUSIONS

The results we presented in the previous section demonstrate that there is no unique solution to solve the problem in all the conditions.

As we expected, the results obtained in case of original phase reconstruction are, from the perceptive point of view (PESQ), better than the Griffin-Lim ones. This because Griffin-Lim gives just an approximation of the original phase.

In general the results obtained with model based method have the best scores because this method restores just the region selected by the mask.

The other two methods, instead, reconstruct completely the whole image.

Nevertheless the Pix2Pix method has similar values with respect to the Model Based method, especially in the case of a log Mel time-frequency transform.

Deep Prior method seems to have the worst performances, but we realized that this depends on the number of iterations selected in order to minimize the loss function. Potentially, this could be the best method also because it has a slowly decreasing slope as shown in the PSNR and PESQ plots. The problem of this technique is that to achieve high scores needs a high number of iterations which means lot of time spent in computations. So depending on the size of the selected region, we can choose the best method to use. In case of small regions we suggest to use the Model Based because it's faster and has good results. If the size of the region is large, the best solution is the Deep Prior trained with a high number of iterations, which can be very slow but the final results will be particularly nice.

To conclude, this paper proves that the image inpainting techniques can be applied also to audio spectrograms achieving satisfactory results.

In the future this work could be improved following this suggestion. Since the CNN based methods generate a completely new image, the improvement could be to evaluate the performance not on the net ouptut image but to take as final output an image composed by the original masked image with the mask selected region taken from the generated one. In this way we get a fairer comparison between the Model Based and the CNN based methods.

If you want to go deeper into our project, all the material is available at [11].

6. REFERENCES

[1] B. Furht, Ed., *Image Inpainting*. Boston, MA: Springer US, 2006, pp. 315–316. [Online]. Available: https://doi.org/10.1007/0-387-30038-4_98

[2] S. B. Damelin, N. S. Hoang, "On surface completion and image inpainting by biharmonic functions: Numerical aspects," *International Journal of Mathematics and Mathematical Sciences*, vol. 2018, 2018.

[3] Charles K. Chui, H. N. Mhaskar, "Mra contextual-recovery extension of smooth functions on manifolds," *Applied and Computational Harmonic Analysis*, vol. 28, pp. 104–113, Jan 2010.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros, "Image-to-image translation with conditional adversarial networks," 2017, cVPR2017.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[6] Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky, "Deep Image Prior," 2018, CVPR2018.

[7] D. W. Griffin, J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. ASSP*, vol. 32 no.2, p. 236–243, Apr 1984.

[8] Zhou Wang, A.C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *Signal Processing Magazine*.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*.

[10] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 32 no.2. IEEE, May 2001.

[11] "Audio inpainting github repository." [Online]. Available: https://github.com/mindofsimon/PC_L13_Audio_Inpainting