LEVERAGING REPETITION TO DO AUDIO IMPUTATION

Ethan Manilow, Bryan Pardo

Northwestern University
Electrical Engineering and Computer Science
Evanston, IL, USA
ethanmanilow@u.northwestern.edu, pardo@northwestern.edu

ABSTRACT

In this work we propose an imputation method that leverages repeating structures in audio, which are a common element in music. This work is inspired by the REpeating Pattern Extraction Technique (REPET), which is a blind audio source separation algorithm designed to separate repeating "background" elements from non-repeating "foreground" elements. Here, as in REPET, we construct a model of the repeating structures by overlaying frames and calculating a median value for each time-frequency bin within the repeating period. Instead of using this model to do separation, we show how this median model can be used to impute missing time-frequency values. This method requires no pre-training and can impute in scenarios where missing or corrupt frames span the entire audio spectrum. Human evaluation results show that this method produces higher quality imputation than existing methods in signals with a high amount of repetition.

Index Terms— Audio imputation, repetition, PLCA, REPET

1. INTRODUCTION

Modern audio editing programs (e.g. Adobe Audition) let users directly edit and manipulate the time-frequency data (e.g. a spectrogram) of an audio recording. The tools offered to users are similar in nature to image editing tools: select a region, paint, erase, etc. Skilled editors can use these tools to remove unwanted sounds or artifacts from an audio recording. However, editing an audio file to remove an unwanted element may leave a blank spot in the spectrogram that sounds unnatural. This is analogous to editing out an element of a visual image, leaving a hole in the image that is immediately obvious. Repairing missing data in images is known as image inpainting. We seek to do inpainting in the audio domain.

Because of the analogy to image inpainting, a natural place to seek inspiration is visual image inpainting. Many automated image inpainting techniques depend on properties of the human visual system (specifically *isophotes*, or contours of equal luminance) [1, 2]. Thus, the set of assumptions made about image data and how to manipulate it do not hold when these ideas are applied to audio data. Also, not all manipulations that are suitable in the image domain can be appropriately applied to a spectrogram. For instance, copying the pattern of a brick wall to fill a blank spot higher up in the picture is perfectly valid, but an analogous operation on a spectrogram could produce an unusable result. Because the vertical axis corresponds to frequency on a spectrogram, moving a sound along the vertical axis fundamentally changes the nature of that sound.

This work sponsored by National Science Foundation Award 1420971.

Some solutions to imputing appropriate audio in the time domain have been proposed, such as repairing lost packets over VoIP protocols [3] or removing clipping (truncation of the amplitude of an audio signal) or distortion from a signal [4]. These solutions, often called audio inpainting, solve a different set of problems than those that arise from when a user is manually editing a spectrogram. Further, these solutions can only fix blank spots that are on the order of 10 ms, whereas the blank spot produced by a user edit of an audio file can span many seconds.

Matrix imputation is an area of study with applications in many fields, including audio. Smargdis et al [5] proposed a solution for filling in (imputing) missing parts of a magnitude spectrogram by using a Probabilistic Latent Component Analysis (PLCA) model trained on complete or partially complete time-frames. This model was then used to impute missing values in the spectrogram. This technique is closely related to Non-negative Matrix Factorization (NMF), which has been used for bandwidth expansion of audio [6].

While it is possible to adapt general methods of matrix imputation to infer missing time-frequency values in a damaged spectrogram, one must be careful in their application, because techniques such as NMF are blind to the overarching temporal structures inherent in many audio scenes (e.g. music). Blindness to temporal structure can, to an extent, be overcome by using a Non-negative Hidden Markov Model (NHMM) (e.g. Han et al [7]), but such a model requires that the current audio's temporal structure strongly mirror the temporal structure of prior training examples.

Non-negative Tensor Factorization (NTF) [8] is another approach that takes local temporal structure into account, but NTF will fail when trying to impute a repeating background sequence that is never fully isolated in the undamaged part of the file (i.e. the repeating pattern that would be useful for infill always has some foreground audio element overlapping it somewhere).

All of the aforementioned audio imputation methods–PLCA, NMF, NHMM, NTF–suffer when all or most of the amplitudes and phases across the spectrum at a particular time are damaged or missing. This can happen when someone edits out a wide-band noise (e.g. a snare drum hit). In this case, all frequencies have zero energy and general matrix-based imputation algorithms have no way to determine which dictionary elements are most applicable. Therefore, these methods will do nothing or produce very poor results. A model of the overarching temporal structure of the audio can overcome this problem, as it lets the algorithm infer the likely contents of a blank spot based on past repetitions of a pattern. General matrix imputation methods also share a second problem: phase. Current practice applies these approaches to magnitude spectrograms, which do not contain phase information. Once imputation is complete on the magnitude spectrogram, one must still estimate phase

for the imputed areas before the waveform can be reconstructed.

Here, we propose to impute missing or corrupted portions of an audio signal with repeating elements by leveraging these repeating patterns to guide imputation. We adapt the source separation algorithm REPET (REpeating Pattern Extraction Technique) [9] for this purpose. The resulting technique can impute values in any part of a spectrogram, including places where the missing audio extends through all frequencies and can last for a full second or more. The exact length one can impute is dependent on the length of the repeating pattern used to guide the imputation. This approach does not require prior training on similar signals and preserves overarching temporal structures in the audio. Further, this approach automatically imputes both the phase and amplitude of a complex spectrogram, allowing reconstruction of the waveform without need for additional post-processing to estimate phase.

2. THE MEDIAN MODEL

The crux of natural sounding imputation is creating a believable model of what the audio should sound like. One common situation where infill will be needed is where there is a repeating pattern in the audio. This occurs frequently in music, although it is not the only situation (e.g. frogs croaking, engine noise, footsteps on pavement). If it is possible to create a model of repeating structures in audio, then it is possible to impute missing sounds from an audio signal using knowledge of the repeating structures.

2.1. Creating a Median Model

To create a model of repetition for a given audio signal, we look to the REpeating Pattern Extraction Technique, or REPET [9]. REPET is a blind audio source separation algorithm that is designed to separate out a non-repeating foreground from a repeating background (see [9] for the full details of REPET). To do imputation, we stop short of creating a mask for separation and instead use the modeled repeating structures REPET creates to fill in missing values in the audio. Here we outline the process for creating the median model.

To create a model of repetition in the audio, we estimate the repeating period, p, of the signal from the beat spectrum. The beat spectrum, $b(l)$, represents the self-similarity of a signal as a function of time lag, l. Given an audio signal x, we compute its short-time Fourier transform (STFT), S. We create a magnitude spectrogram, $V = |S|$, by taking the absolute value of the STFT and then element-wise squaring it to create the power spectrogram, V^2.

We compute the autocorrelation over time for each frequency channel of V^2 to obtain the matrix of autocorrelations A, as in Eq. 1. The overall self-similarity, b (Eq. 2), of x is then obtained by taking the mean over the frequency axis of A.

$$A(i,l) = \frac{1}{m-l+1} \sum_{j}^{m-l+1} V^2(i,j)V^2(i,j-l+1) \quad (1)$$

$$b(l) = \frac{1}{n}\sum_{i}^{n} A(i,l) \quad \text{then } b(l) = \frac{b(l)}{b(1)} \quad (2)$$

for $i = 1...n$ where $n =$ number of frequency channels
for $l = 1...m$ where $m =$ number of time frames

From the beat spectrum, b, we estimate the repeating period of the signal, p by using a peak finder to select a local maximum value within a specified lag range of the beat spectrum [9].

Once a period p is determined, the STFT of the signal is partitioned into r windows (W) of length p that are "overlayed" on one another. From the stack of overlayed windows we calculate a median value for each time (τ), frequency (ω) point to create the median model, M of length p:

$$M(\omega,\tau) = \underset{\text{for } k=1...r}{\text{median}^*}\{W_k(\omega,\tau)\} \quad (3)$$

Calculating a median is undefined on complex values, so at each time-frequency point of the overlayed frames, we ignore the imaginary part in the calculation of the median. Once the median point is determined from those overlayed frames, we insert the corresponding complex-valued STFT time-frequency point into the model. In this way, the resultant median model is complex valued. We denote this operation with an asterisk (*) in Eq. 3. While this technique produced sufficient results, a more principled approach would involve using just the magnitude for the median calculation and inserting both the magnitude *and* phase in to the median model.

Building the median model from the complex STFT keeps the association of the magnitude and phase of each time-frequency point. This is important because we impute missing values using the median model and it lets us automatically impute both phase and magnitude in a reasonable way. Other approaches to imputation (e.g. NMF, PLCA) use models built from the magnitude spectrogram and therefore have no phase information. This introduces a problem when reconstructing the waveform from the imputed data: one must still estimate phase in some way.

2.2. Imputing with a Median Model

Once we have calculated the median model, M with length p, it is possible to impute arbitrary missing time-frequency values from an input signal represented as an STFT, S. For a missing value time-frequency value at a specific position $S(\omega,\tau)$, we simply insert the value of the model at $M(\omega, \tau \bmod p)$.

2.3. Advantages and Limitations

One major advantage that the proposed method has is that the size and shape of the damaged section is less of a concern than in previous methods, such as PLCA. PLCA imputation will fail if the majority or all of the frequencies in the STFT are damaged or have missing values, which occurs when one edits out a wide band signal (e.g., a snare or cymbal). The proposed method has no such requirement. The only requirement is that there are at least three repetitions of the repeating structure to make a median model.

Another major advantage that this method has is its speed and simplicity. It is very computationally efficient to create the median model and impute from it. The bottleneck for the proposed method is calculating the Fourier transforms when building the autocorrelation matrix, which is $O(n\log(n))$, whereas PLCA/NMF-based methods are solvable in polynomial time [10].

Furthermore, the median model can be created from the complex-valued spectrogram, meaning that imputation does not require another step to reconstruct the phase of the signal. PLCA imputation is applied to a non-negative magnitude spectrogram, thus there is additional work required to reconstruct the phase (often using some variant of the Griffin-Lin Algorithm [11]), further adding to its overhead.

Because this method is closely related to the original REPET formulation [9], imputations from the proposed method will also

remove any non-repeating elements in the signal, similar to how REPET removes a singing voice from a repetitive background.

Finally, effective use of the proposed method is limited to signals with a regularly repeating structure. As the background repetition becomes less regular, the quality of the imputation will degrade.

2.4. Practical Concerns

We have determined a number of best-practices that makes imputation sound more natural when using the proposed method.

First, because the creation of the median model is predicated on a regular repeating period p, it is crucial that p is calculated correctly. As noted above, p is estimated in the time-frequency domain, but if the true repeating period (in the time domain) is not close to a multiple of the hop size used when calculating the STFT, the proposed method will not produce good results. For instance, if the "true" repeating period is 102,912 samples (about 2.33 seconds at 44.1 kHz) and the hop size is 1024 samples, the repeating period falls exactly halfway between hops (102,912 mod 1024 = 512). In this case, imputation would lead to audible artifacts. But if we had selected a hop size of 512 samples, the model would be able to capture the "true" repeating period and the artifacts would be less discernible or non-existent. In practice, this means the best technique to mitigate this issue is to use smaller hop sizes, but this obviously comes at a computational cost.

Additionally, when imputing values at the edge of the damaged part of the STFT, i.e., imputing values where $S(\omega, \tau \pm 1)$ are still in tact, it is useful to "cross-fade" values to avoid discontinuities when S is transformed back into the time domain.

Finally, the median model M typically has less overall amplitude than the original signal due to the mathematics of taking the median. By definition, taking the median value of each time-frequency point over each repetition does not select values that are loud or soft, but aggregating the median of each bin results in a quieter model. To counteract this, it is wise to boost the amplitude of the median model. The amount of increase required will depend on the signal; a good rule is to match the signal's RMS between the imputed part and its immediate surroundings.

3. EVALUATION

We compared two versions of the proposed method against two versions of PLCA imputation [5]. For PLCA, we evaluated one version where the ground-truth phase was copied to the imputed magnitude spectrogram, and another where the true phase was not provided to the system and was reconstructed from the imputed magnitude spectrogram using the Griffin-Lin Algorithm [11]. In all cases, PLCA used 120 components for 300 iterations on an STFT with a fixed window size of 1024 samples and fixed hop size of 512 samples.

For the proposed method, we used two variants of the window and hop size in the STFT provided as input to the system. The first variant used the same window and hop sizes as PLCA, 1024 and 512 samples, respectively, so as to provide a direct comparison to PLCA at their best window size and hop settings. The second variant used a window size of 256 samples and a hop size of 128 samples. This was chosen to maximize perceived audio quality when applying REPET-based imputation, as described in Section 2.4. These values were selected on a prior set of audio.

We did not do any "cross-fading" at the edge of imputation, nor did we do RMS matching (as discussed in section 2.4). We did however scale every value in the learned median model by 1.5 to boost the volume. The value 1.5 was determined on a prior set of data and applied to all imputations of the test data. Note that our system was never provided ground-truth phase and learns phase from the input audio.

For each audio file, all methods were only given the damaged audio file using the fixed meta-parameters described previously. There was no pre-training on any other audio. Quality of the results was evaluated by a set of human evaluators.

3.1. Dataset

Our imputation method is based on the REPET source separation algorithm. REPET has been shown to work well on many periodic patterns (e.g. live musicians vamping on a repeating chorus), even when the periodicity is not exact. For a full analysis of how REPET degrades as the audio becomes less periodic we refer the reader to the original journal article [9]. For this work we created an illustrative set of audio examples that show effectiveness of the proposed method in scenarios with regular repetition. The use case we envision is one where a large (e.g. 1 second long) blank spot was created, as might have been produced by a manual edit of an STFT.

We made three base recordings that were then damaged in some way. The first recording contained a solo acoustic guitar strumming chords, the repeating period was 4 seconds and was looped 4 times for a 16 second signal. The second recording had the same repeating guitar strumming as the first, but also had a male vocalist singing a non-repeating melody and the acoustic guitar background was repeated 6 times for a 24 second audio file. In this example, the singer coughed. This creates a situation where no time-frame is exactly like any other, although there is still an underlying periodic structure that can be inferred. The third recording had an electric bass guitar and a drum kit (kick, snare, and shaker) playing for one bar lasting 3.4 seconds, and was looped 5 times for a 17 second audio file. This example is more timbrally complex than the singer and guitar. All recordings were mono, 16-bit, PCM sampled at 44.1kHz.

We damaged each of the three recordings in two ways; the first where energy at all frequencies was missing, as in a user removing an extraneous snare hit in an audio editor, leaving a complete blank spot. The second where a particular (wide) frequency range was set to zero volume. In damaging particular frequency ranges, we aimed to examine three different scenarios. For the solo acoustic guitar, we removed 130.8-1046.5 Hz (C3-C6) because those frequencies are the most common ones for fundamental frequencies to occur in music. For the acoustic guitar with singing, we removed a muffled cough by the singer, leaving blank frequencies between 0-2000 Hz (the frequencies required to remove the cough completely). For the drum and bass, we removed 27.5-1760 Hz (A0-A6).

This resulted in a total of 6 damaged audio recordings. Damage in each recording lasted between 1-1.5 seconds. We applied each imputation method (two variants of PLCA and two of the proposed method) to each recording, as described previously. This resulted in 3 original recordings, 6 damaged signals, and 24 damaged-then-imputed signals, for a total of 33 recordings.

3.2. Perceptual Evaluation

MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [12] is a protocol for subjectively assessing audio quality, usually done in a lab setting by experts. A recent study by Cartwright et al.

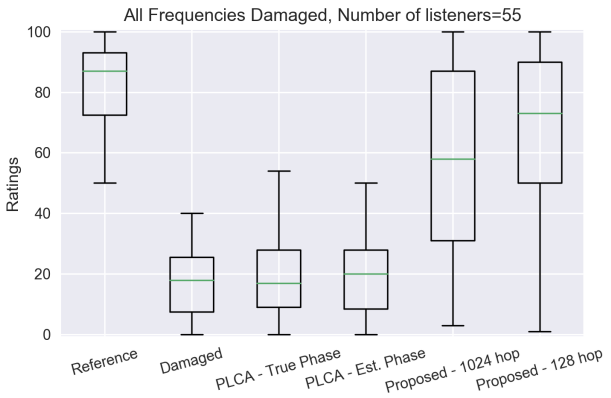

Figure 1: Perceptual evaluation results from 55 participants listening to 3 signals where all frequency bins are damaged. Lines inside the boxes indicate medians. "PLCA - True Phase" used the ground truth phase and "PLCA - Est. Phase" had reconstructed phase. "Proposed - 1024 hop" and "Proposed - 128 hop" are the proposed method with an STFT hop size of 1024 and 128 samples.

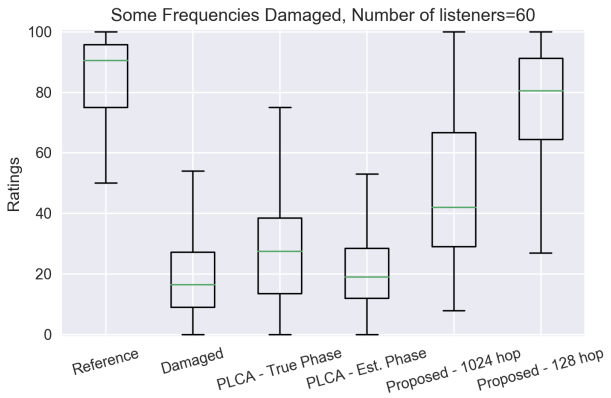

Figure 2: Perceptual evaluation results from 60 participants listening to 3 signals where some frequency bandwidth was damaged (see Sec 3.1 for details). Lines inside boxes indicate medians. The labels are the same as in Fig. 1

[13] showed that the qualitative results of a MUSHRA-like[1] evaluation of audio with medium levels of degradation could be performed much more cheaply and quickly by collecting evaluations from many listeners over the web, while maintaining high agreement with a lab-based MUSHRA study.

To evaluate the perceptual quality of the proposed method against our benchmarks, we employed a MUSHRA-like web-based evaluation campaign using the Crowdsourced Audio Quality Evaluation framework (CAQE) [13]. CAQE is a web framework to set up a website for gathering audio quality evaluation results. We used Amazon Mechanical Turk (AMT) to recruit and pay participants.

The hearing screening ensured that participants were able to hear the test on a system that had an adequate frequency response and were able to follow the directions correctly. The subjects were asked to adjust the volume of a 1000Hz sine wave to a comfortable level and then listen to and count a predetermined number of pitches with frequencies between 55 Hz and 10 kHz in an 8s audio clip. Participants were allowed two chances to answer correctly.

After passing the hearing screening, each participant was asked to rate a set of 6 audio examples based on a 0 to 100 sliding scales that represented "overall quality" in terms of how similar the recording was to the original undamaged audio (reference). The 6 audio examples used were the original recording (reference and high-quality anchor), the damaged audio recording (low-quality anchor), and the 4 recordings produced by applying an imputation algorithm to the damaged audio (as in Section 3.1). The whole task of listening to all 6 audio examples is considered a single "condition." Because there were 6 audio examples in each condition, we presented truncated versions of the audio signals to the participants. The original signals ranged between 16-24 seconds and were highly repetitive, so we truncated the signals to include two repetitions of the loop, one undamaged and one damaged-then-imputed.

We collected at least 21 trials for each condition. The mean number of conditions per trial was 24.3 with a max of 29. Participants were paid \$0.80 for the first trial and \$0.50 for each subsequent trial. Only participants with 1000 prior AMT assignments and 97% approval rating were allowed to participate. We eliminated responses from participants who responded to a post-evaluation survey that they heard external noise or were not listening on headphones. Additionally, we eliminated responses where the reference signal was rated lower than the damaged signal. After these eliminations we had 115 trials from 99 participants, with at minimum of 14 trials per condition, a mean of 19.1 and max of 22.

[1] As Cartwright, et. al. noted in [13], some specifications of MUSHRA are infeasible over the web, so we refer to these tests as "MUSHRA-like".

4. RESULTS

Aggregate results from the perceptual evaluation are shown in Figures 1 and 2. Fig. 1 shows results of the three signals where the damage spanned the entire frequency range, and Fig. 2 shows results of the three signals where damage only affected some of the frequencies (see Sec 3.1). PLCA imputed nothing in the examples where all frequency values were missing (Fig. 1) and participants rated the PLCA examples as close to the damaged signal. Conversely, in these cases, the proposed method *was* able to impute missing values. In the case where some frequency values were missing (Fig. 2) PLCA created audibly noticeable artifacts in the signal, especially on the "drums and bass" example where PLCA failed and imputed values that made the signal sound highly degraded. Though output from the proposed method was not without artifacts (especially with hop size of 1024), the artifacts were minimal and overall participants rated output from the proposed method as having equal or higher quality than PLCA imputation.

5. CONCLUSION

We have proposed a simple method for modeling repeating background structures for imputing missing time-frequency values in audio with repeating structure. It can do imputation in situations that existing matrix-imputation methods cannot, e.g., when damage occurs across the entire frequency spectrum. This method performs better than existing methods on crowd-based perceptual evaluation tasks in cases where there is a repeating structure in music.

6. REFERENCES

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 417–424.

[2] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on image processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[3] Y. Bahat, Y. Y. Schechner, and M. Elad, "Self-content-based audio inpainting," *Signal Processing*, vol. 111, pp. 61–72, 2015.

[4] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Missing data imputation for spectral audio signals," in *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*. IEEE, 2009, pp. 1–6.

[6] D. L. Sun and R. Mazumder, "Non-negative matrix completion for bandwidth extension: A convex optimization approach," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*. IEEE, 2013, pp. 1–6.

[7] J. Han, G. J. Mysore, and B. Pardo, "Audio imputation using the non-negative hidden markov model," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 347–355.

[8] U. Şimşekli, Y. K. Yılmaz, and A. T. Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5369–5372.

[9] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 73–84, 2013.

[10] S. Arora, R. Ge, R. Kannan, and A. Moitra, "Computing a nonnegative matrix factorization–provably," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012, pp. 145–162.

[11] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[12] I. Recommendation, "1534-1: Method for the subjective assessment of intermediate quality level of coding systems," *International Telecommunication Union*, 2003.

[13] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *41st IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.