AUDIO RESTORATION FROM MULTIPLE COPIES

Pablo Sprechmann,[1] *Alex Bronstein*,[2], *Jean-Michel Morel*,[3] *and Guillermo Sapiro*[1]

[1] Duke University, USA;[2] Tel Aviv University, Israel; [3]ENS Cachan, France.

ABSTRACT

A method for removing impulse noise from audio signals by fusing multiple copies of the same recording is introduced in this paper. The proposed algorithm exploits the fact that while in general multiple copies of a given recording are available, all sharing the same master, most degradations in audio signals are record-dependent. Our method first seeks for the optimal non-rigid alignment of the signals that is robust to the presence of sparse outliers with arbitrary magnitude. Unlike previous approaches, we simultaneously find the optimal alignment of the signals and impulsive degradation. This is obtained via continuous dynamic time warping computed solving an Eikonal equation. We propose to use our approach in the derivative domain, reconstructing the signal by solving an inverse problem that resembles the Poisson image editing technique. The proposed framework is here illustrated and tested in the restoration of old gramophone recordings showing promising results; however, it can be used in other applications where different copies of the signal of interest are available and the degradations are copy-dependent.

Index Terms— Audio restoration, impulse noise removal, samples fusion, multi-signal alignment, dynamic time warping, Eikonal equation.

1. INTRODUCTION

Digital audio restoration has been widely studied in the literature for several decades, see [1, 2] for reviews. One of the most common types of degradation is impulse noise, that is, a localized distortion affecting the signal. Restoring this wide class of degradations arises naturally in many modern digital signal processing applications. In this work we will use the restoration of gramophone recordings as the illustrative example. In this particular case, the problem receives the name of *de-clicking* and *de-scratching*, and has been extensively studied in the literature.

Most of the work in audio restoration has been performed considering that only one copy of the signal is available [1, 2, 3], while this is not the case in many practical scenarios. The single-copy processing is motivated by several reasons. First, it is common practice in audio transfer and digitalization to select the best available record, and perform on it all the acquisition and signal restoration procedures [4]. Second, the quality of old recordings generally differs significantly, and many of the available copies are of considerably poorer quality than the best one. Finally, most available restoration systems can be operated in manual mode, allowing the user to search for the (time segment-dependent) parameters that produce the best results, and even modify any artifact introduced in the processing. Our goal is to obtain such high quality results by an automatic model free procedure, exploiting all the available recordings.

De-clicking methods normally start with click detection, where the objective is to find the distorted audio fragments. Classical approaches are based on outliers detectors, assuming a variety of models for the audio signal. Techniques based on autoregressive (AR) models have been demonstrated to be particularly successful [1]. In general, the results are moderately dependent on the selected parameters, which in turn depend on the often unknown level of degradation and the characteristics of the musical content.

Once the location of the distorted fragments has been determined, the remaining task consists of restoring the affected samples. Standard click removal algorithms use interpolation schemes which set the missing samples to some estimates of their true value based on the uncorrupted surrounding samples. Classical approaches typically use AR modeling or Bayesian estimation to recover the distorted samples [1, 5], while more recent methods employ sparse representations to model time-domain audio frames [6]. In [7], an elegant framework that fills in the missing samples by copying the statistical properties of the signal in the surrounding of the gap is presented. Note that model-based methods will always suffer from model and data inconsistency as well as parameter sensitivity, both for the detection and restoration steps.

In addition to computational burdens and model dependency, single signal methods present another important drawback: they are fundamentally restricted to working only with degradations well-localized in time. Due to the non-stationary nature of audio signals, meaningful model-based data reconstruction can only be achieved for relatively short-duration portions of audio signals. In [7] the authors explain that "For certain long lasting disturbances, e.g., those caused by really

P.S. and G.S. were supported by ONR, NSF, NGA, NSSEFF, and ARO. A.B. was supported by BSF. J-M.M. was supported by the Office of Naval Research under Grant N00014-97-1-0839 and by the European Research Council, advanced grant "Twelve labours".

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

large scratches, human intervention is necessary as it would be very dangerous to allow the system for detection and replacement of more than 100 "bad" samples in a row in an unsupervised, i.e., automatic, mode. Otherwise, it could potentially behave in an unpredictable manner."

In the case of gramophone recordings, the most disturbing degradations are record-dependent. Clicks and scratches appear due to surface imperfections in the record itself [4]. The largest portion of them are the consequence of aging of the record medium: wear, groove damage, and mishandling. Since these degradations occur at random locations for each record, a reasonable assumption is that different records are affected in different positions. The main contribution of this paper is to develop an efficient and simple de-clicking algorithm exploiting this observation as described in the sequel.[1]

2. MULTI-SIGNAL RESTORATION

The detection of the clicks and recovery of the missing or corrupted samples is performed fusing several copies of the same master recording. With this procedure we can accurately detect and restore the affected audio fragments while avoiding the difficult (and ambitious) problem of explicitly modeling the audio signals. Our main point is that, even if some of the available signals are of poorer quality than the best available one, they can still provide crucial information for detecting and restoring the distortions.

We begin by presenting in this section a general variational framework for robust multi-signal alignment. In this setting, the optimal alignment of all the signals and the impulse noise are obtained simultaneously. We model the misalignment via a continuous version of dynamic time warping that can be obtained by solving the Eikonal equation [8]. The use of the time warping allows us to depart from the widely used unrealistic assumption that the impulse noise does not distort the timing of the recording. The click artifacts are modeled as sparse outliers, following the spirit of the recent works in robust principal component analysis [9, 10] and robust image alignment [11].

It is common practice to high-pass filter the input signal, as a pre-emphasis filter, in order to enhance the presence of the impulses and obtain a more robust detection of the location of the distorted samples. In Section 3, we propose to perform the restoration in the derivative domain as a pre-emphasis. The restored signals are then reconstructed by solving an inverse problem that resembles the Poisson image editing technique proposed in [12]. To conclude, in Section 4 we evaluate the proposed approach with several experiments using real recordings.

To the best of our knowledge, [13] is the only work that considers the use of multiple records of the same signal for audio de-noising. In this patent, the author proposes a method for noise reduction in old recordings by averaging several different copies. The method is based on a heuristic sample-by-sample alignment, and does not consider the presence of outliers. All signals are independently aligned to a manually selected master copy in contrast to the collaborative approach proposed in this work. In addition, the inclusion of continuous DTW via Eikonal equations is novel for this application as well. Finally, our overall approach is to use the best part of each signal, as in manual systems, and not an average.

Despite being very different, the proposed method goes in the same direction as the approach recently used in [14, 15] for extracting the music and sound effects track of a movie from several a set of soundtracks of a movie in different languages.

2.1. Problem statement

Let us be given p signals $x_i(t)$, $i = 1, \ldots, p$; $t \in T$. The signals are misaligned and corrupted by click artifacts. We seek for the time warps $\gamma_i : T \to T$ and a common master signal $m(t)$ such that each $x_i(\gamma_i(t)) \approx m(t) + o_i(t)$. The outlier signals $o_i(t)$ represent the artifacts in each $x_i(t)$. This multi-signal restoration problem can be stated in a variational setting as the minimization of the functional

$$\begin{aligned}\mathcal{F}(m, o_1, \ldots, o_p, \gamma_1, \ldots, \gamma_p) = \\ = \frac{1}{2}\sum_{i=1}^{p}\int_T (x_i(\gamma_i(t)) - o_i(t) - m(t))^2 dt \\ +\lambda \sum_{i=1}^{p}\int_T |o_i(t)| dt, \qquad (1)\end{aligned}$$

The regularization term with the ℓ_1 norm on $o_i(t)$ promotes solutions with energy concentrated in small regions in time, which is characteristic of impulsive noise.

The optimization can be performed by alternating the minimization over m and the o_i's with fixed γ_i's, and over each γ_i while keeping the rest of the variables fixed. In what follows, we describe in details both minimization steps.

2.2. Continuous time warping

Let us fix all the variables of $\mathcal{F}$ except for one single γ_i for some index $1 \leq i \leq p$. To simplify notation, we will denote $z(t) = o_i(t) + m(t)$, $x(t) = x_i(t)$, and $\gamma(t) = \gamma_i(t)$. Let us further define the function

$$e(t, t') = \int (x(\tau - t) - z(\tau - t'))^2 d\tau \qquad (2)$$

describing the cost incurred by aligning a window of x centered at t to a window of z centered at t'. The minimization

[1] Randomness in the position of the degradation is not limited to this scenario, and can appear for example as a result of data loss in wireless transmission to multiple receivers, or distinct compression artifacts in multiple copies. The underlying ideas here introduced are applicable to these scenarios as well.

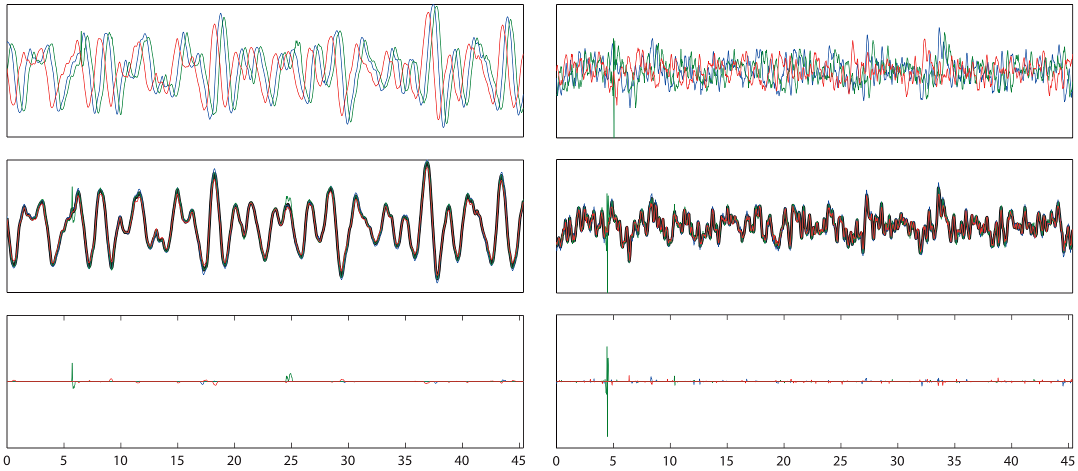

Fig. 1. Algorithm described on Section 2 applied to the original signals (left column); and to first-order derivatives (right column). First row: three 45 ms fragments of unaligned audio signals (or derivatives) with clearly audible click artifacts. Second row: the aligned signals and the reconstructed master signal m (bold black). Third row: the estimated outliers.

of $\mathcal{F}$ with respect to γ_i can now be stated as minimizing the path integral

$$\min_{\gamma} \int_{\gamma} e = \min_{\gamma} \int_{T} e(\gamma(t), t) dt. \tag{3}$$

Similar problems are encountered in optics, where the minimum path γ represents the path of a light beam propagating in an inhomogeneous medium with the refractive index e. Using Maxwell equations, it has been shown that the light propagation is governed by the non-linear first order partial differential equation of the form $\|\nabla E\|_2 = e$. This equation is usually known as the *Eikonal equation* and its solution E as the *eikonal* [8]. The characteristics of the Eikonal equation are the light propagation paths,

$$\dot{\gamma}(t) = \frac{\nabla E(\gamma(t), t)}{e(\gamma(t), t)}. \tag{4}$$

While the Eikonal equation does not have continuously differentiable solutions in the classical sense, existence and uniqueness of the so-called *viscosity* solutions has been established [16]. Using this interpretation, finding the globally optimal time warp γ_i of x_i can be reduced to integrating the viscosity solution of the Eikonal equation with the refractive index described by the corresponding e. We discretize the Eikonal equation on a Cartesian grid describing the narrow band $\gamma(t) \in [t-h, t+h]$. The value of h can be obtained by roughly aligning small frames of the signals x_i. We use the multi-stencil fast marching method [17] to solve the discretized Eikonal equation, and a Runge-Kutta method to integrate the characteristic equation for the timewarp γ_i.

2.3. Robust estimation of the master signal

Let us now fix the o_i's and the time warp transformations γ_i, denoting by $y_i(t) = x_i(\gamma_i(t))$. The minimization of $\mathcal{F}$ now reduces to

$$\min_{m} \sum_{i=1}^{p} \int_{T} (y_i(t) - o_i(t) - m(t))^2 dt,$$

for which the Euler-Lagrange equation gives the minimizer as

$$m(t) \quad = \quad \frac{1}{p} \sum_{i=1}^{p} (y_i(t) - o_i(t)). \tag{5}$$

Fixing m, all the time warp transformations and all the outlier terms except for one o_i for some i, we obtain

$$\min_{o_i} \frac{1}{2} \int_{T} (y_i(t) - o_i(t) - m(t))^2 dt + \lambda \int_{T} |o_i(t)| dt,$$

for which the minimizer is given by

$$o_i(t) \quad = \quad \sigma_{\lambda} \left(y_i(t) - m(t) \right). \tag{6}$$

Here $\sigma_{\lambda}(x) = \operatorname{sign}(x) \max\{|x| - \lambda, 0\}$ denotes soft thresholding (shrinkage). Alternating (5) and (6) with the initialization $o_i \equiv 0$ yields the robust estimate of the master signal m and the outliers o_i.

3. DERIVATIVE DOMAIN RESTORATION

Sharp discontinuities of small magnitude in a signal translate into low-energy distortions in terms of the ℓ_2 norm but create severe audible artifacts. This type of click artifacts are not well captured by the model presented in Section 2.1 since they can be absorbed by the data fitting term. In the derivative domain, however, these type of distortions are amplified and can be much better captured by the outlier signal. The derivative acts as a pre-emphasis filter, which is common practice in audio processing. We propose to apply the

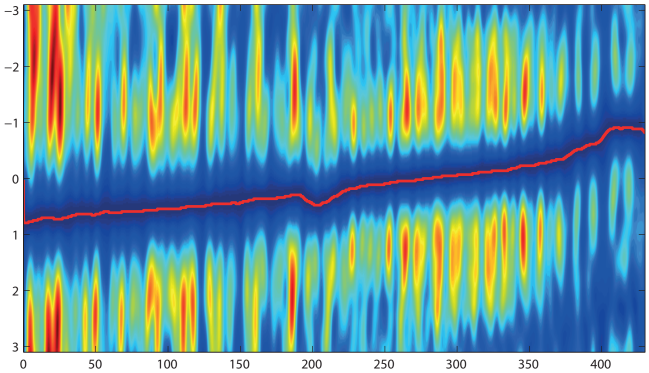

Fig. 2. $e(\gamma(t), t)$ plotted as the function of $\tau = \gamma(t) - t$ (vertical axis), and t (horizontal axis) in ms. The optimal time warp is depicted in bold red. w was set to an 8 ms Hamming window. Time axes are in 44.1 KHz samples.

framework described in Section 2 to the derivative of the signals. Specifically, we assume that the signals verify the model $x_i'(\gamma_i(t)) \approx m'(t) + o_i'(t)$. The algorithm then finds the set of m', o_i''s and γ_i's that minimizes $\mathcal{F}$.

Note that when running the robust joint alignment in the derivative domain a post-processing step needs to be applied in order to obtain the restored signals. We propose to use the support of the outlier signal for finding the location of the impulse noise and then use the estimated master $m'(t)$ to guide the interpolation of the missing samples in the signal domain. For simplicity, we consider the case in which we have an isolated interval of missing components, say T_G, for a j-th signal in the set, x_j. More complex patterns of missing data can be handled in a similar way.

We propose to recover the signal x_j using a guided interpolation technique that resembles the Poisson image editing methodology [12], consisting of solving,

$$\min_x \int_{T_G} (x' - \hat{x}_j')^2 \text{ s.t. } x|_\Omega = x_j \tag{7}$$

where Ω represents the extreme points of the missing interval.

Problem (7) can be discretized as follows. Let $\mathbf{x} \in \mathbb{R}^n$ be the signal to be reconstructed, assuming that the interval T contains n equally spaced samples. With a slight abuse of notation, we also use $T_G = [i_0, \ldots, i_0 + m]$ to denote the m samples corresponding to the continuous interval T_G. Then we propose to solve,

$$\min_{\mathbf{x} \in \mathbb{R}^m} ||\mathbf{D}\mathbf{x} - \hat{\mathbf{x}}_j'||_{\mathrm{F}}^2 \text{ s.t. } \begin{cases} \mathbf{x}[1] = \mathbf{x}[i_0] \\ \mathbf{x}[m] = \mathbf{x}[i_0 + m], \end{cases} \tag{8}$$

where $||.||_{\mathrm{F}}$ represents the Frobenius norm, and $\mathbf{D} \in \mathbb{R}^{m \times (m-2)}$ is a linear operator computing the derivatives via finite difference in the interior points of T_G. Problem (8) is a quadratic program with simple linear equality constraints. The solution of (8) needs to satisfy the following linear equations,

$$\mathbf{D}^{\mathrm{T}}\mathbf{D}\mathbf{x} = \mathbf{D}^{\mathrm{T}}\hat{\mathbf{x}}_j', \quad \mathbf{x}[1] = \mathbf{x}[i_0] \quad \mathbf{x}[m] = \mathbf{x}[i_0 + m],$$

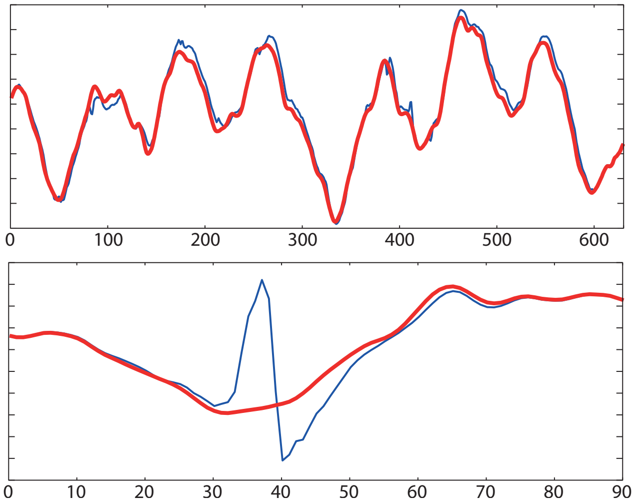

Fig. 3. Examples of reconstruction (red) of affected intervals using the derivative editing technique. The original signal is shown in blue. Time axes are in 44.1 KHz samples.

which can be solved efficiently using iterative methods.

4. EXPERIMENTS

We tested our framework in a variety of real examples. The records were played on a commercial turntable with a spinning speed of 45 RPM. The acquisition was done with Audacity[2] software with 16 bits and a sampling rate of 44.1 KHz. The gain was adjusted to prevent saturation at all times. Signals were stop-band filtered to eliminate the 50 Hz hum due to the power transmission interference. Fixed filter differences were compensated by equalizing the spectral components to match the median across signals. This seems sufficient when signals are acquired with the same device. In more general settings equalization schemes as in [18] could be required.

Figure 1 shows the results obtained by running the algorithm described in Section 2 on both the signal (left) and derivative (right) domains. In this fragment all three signals are affected by clearly audible and visible clicks. One can see that outlier signals capture the click artifacts. One can see that in the derivative domain, the artifacts are amplified enabling the detection of clearly audible yet almost invisible artifacts.

The algorithm is capable of precisely aligning long segments of audio, since DTW can capture the complex patterns of the misalignment between signals. Note that this would be very difficult to model with a parametrized family of transformations. In Figure 2 we show an example of an optimal time warp obtained aligning a signal to the estimated mean.

Finally, we show the reconstruction obtained using the guided interpolation described in Section 3. Figure 3 show the interpolation of an affected signal. Note that the interval contains several hundred samples, which would be very challenging for a model based approach.

[2] http://audacity.sourceforge.net

5. REFERENCES

[1] S. H. Godsill and P. J. W. Rayner, *Digital Audio Restoration: A Statistical Model Based Approach*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1st edition, 1998.

[2] P. A. A. Esquef, *Handbook of Signal Processing in Acoustics. Part VI Audio Engineering*, pp. 773–784. Chapter 40, Springer, New York, 2008.

[3] T. Kasparis and J. Lane, "Adaptive scratch noise filtering," *IEEE Transactions on Consumer Electronics*, vol. 39, no. 4, pp. 917 – 922, 1993.

[4] *The AHRC Research Centre for the History and Analysis of Recorded Music*, http://www.charm.rhul.ac.uk/.

[5] A. J. E. M. Janssen, R. N. J. Veldhuis, and L. B. 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.

[6] 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*, 2011.

[7] M. Niediwiecki and K. Cisowski, "Smart copying-a new approach to reconstruction of audio signals," *IEEE Transactions on Signal Processing*, vol. 49, no. 10, pp. 2272 – 2282, Oct. 2002.

[8] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*, Cambridge University Press, 1999.

[9] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, May 2011.

[10] G. Mateos and G. B. Giannakis, "Robust PCA as bilinear decomposition with outlier-sparsity regularization," *IEEE Trans. on Signal Process.*, vol. 60, no. 10, pp. 5176–5190, 2012.

[11] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *CVPR*, 2010, pp. 763–770.

[12] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM Trans. on Graphics*. ACM, 2003, vol. 22, pp. 313–318.

[13] R. I. Webster, "Method and apparatus for reducing noise using a plurality of recording copies," *US Patent 5740146*, Apr 1998.

[14] J.-J. Burred and P. Leveau, "Geometric multichannel common signal separation with application to music and effects extraction from film soundtracks," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 201–204.

[15] P. Leveau, S. Maller, J.-J. Burred, and X. Jaureguiberry, "Convolutive common audio signal extraction," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*. IEEE, 2011, pp. 165–168.

[16] M. G. Crandall, L. C. Evans, and P. L. Lions, "Some properties of viscosity solutions of Hamilton-Jacobi equations," *Trans. Amer. Math. Soc*, vol. 282, no. 2, 1984.

[17] M. S. Hassouna and A. A. Farag, "Multistencils fast marching methods: A highly accurate solution to the eikonal equation on cartesian domains," *PAMI*, vol. 29, no. 9, pp. 1563–1574, 2007.

[18] A. Liutkus and P. Leveau, "Separation of Music+Effects sound track from several international versions of the same movie," *128th Convention of the Audio Engineering Society, London, UK*, May 2010.