Phase-controlled sound transfer based on maximally-inconsistent spectrograms *

◎ Jonathan Le Roux (NTT CS Labs)

1 Introduction

It is generally considered that the magnitude part of an STFT spectrogram is a reliable cue to build an intuition of what a signal resynthesized from that spectrogram will sound like. Spectrogram reading and algorithms for sound reconstruction from magnitude spectrograms are striking illustrations of this idea, and one might be tempted to think that, whatever the phase combined with a given magnitude, the worse that could happen when resynthesizing a signal through inverse STFT is to obtain a noisy version of the sound that the intuition hints at.

We show in this paper that it is not the case: while results meeting the intuition can indeed be obtained by minimizing what we call the "inconsistency" of a spectrogram, we investigate in this paper what happens when inconsistency is maximized instead of minimized, and explain how the same magnitude spectrogram can lead, depending on the phase it is combined with, to extremely diverse resynthesized sounds, some of them very far from what one would expect.

2 Consistency and intuition

Let $(A_{m,n})_{0\leq m\leq M-1, 0\leq n\leq N-1} \in \mathbb{C}^{MN}$ denote an array of real non-negative numbers, where m will correspond to the frame index and n to the frequency band index. A classical task [1] is to estimate a real-valued time-domain signal x such that the magnitude of its STFT is closest to A in a least-squares sense, as this signal would thus sound close to what one would intuitively expect for a hypothetical sound with STFT magnitude given by A. Equivalently [2], this problem can also be formulated as that of estimating a phase ϕ such that $H = Ae^{j\phi}$ is "as consistent as possible", where we call "consistent spectrograms" the elements of $\mathbb{C}^{MN}$ which can be obtained as the STFT spectrogram of a time-domain signal. The lack of consistency, or inconsistency, of any array H is numerically characterized by the quantity $\mathcal{I}(H) = \|\mathcal{G}(H) - H\|$, i.e., the L^2 norm between H and the complex spectrogram of the signal resynthesized from H by inverse STFT, denoted as

$$\mathcal{G}(H) = \mathrm{STFT}(\mathrm{iSTFT}(H)).$$

Note that it will always be assumed that the synthesis window in the inverse STFT is equal to the STFT analysis window, up to the normalization required to obtain perfect reconstruction.

Obtaining a sound which corresponds to one's intuition given an STFT magnitude A thus corresponds to minimizing the inconsistency measure $\mathcal{I}(Ae^{j\phi})$ with respect to ϕ. A simple algorithm to perform this minimization was derived by Griffin and Lim [1], and consists in iteratively updating the phase estimate $\phi^{(k)}$ at step k by replacing it with the phase of the STFT of its inverse STFT, $\angle\mathcal{G}(Ae^{j\phi^{(k)}})$, while keeping A fixed. Fast approximations based on time-frequency domain computations were considered in [3].

3 Maximizing inconsistency

We are interested here in what happens when inconsistency is maximized. We shall find convenient to introduce the notation $\mathcal{F}(H) = H - \mathcal{G}(H)$. Under the above assumption for the analysis and synthesis windows, the operators $\mathcal{F}$ and $\mathcal{G}$ can be shown to be orthogonal projections on complementary subspaces of $\mathbb{C}^{MN}$, such that, for all H,

$$\|H\|^2 = \|\mathcal{G}(H)\|^2 + \|\mathcal{F}(H)\|^2. \tag{1}$$

Maximizing the inconsistency $\|\mathcal{F}(H)\|$ is thus equivalent to minimizing $\|\mathcal{G}(H)\|$. The extreme case will be that of spectrograms that can be obtained as $H = \mathcal{F}(S)$ for some S: they will indeed verify $H = \mathcal{F}(H)$ as $\mathcal{F}$ is a projection, and thus $\|\mathcal{G}(H)\| = 0$. These spectrograms will thus be resynthesized through inverse STFT as silence, as already noted in [2]. What happens is that the contributions of neighboring frames in such spectrograms perfectly cancel in the overlap-add procedure.

Note that for a rectangular analysis window and 50 % or 75 % overlap between frames, the STFT spectrogram of any sound can be trivially modified to resynthesize to silence while keeping the same magnitude: it suffices to add π to the phase of every other frame. This is however not true in general for other windows or overlap ratio, and we shall now derive an algorithm to maximize inconsistency with a given magnitude spectrogram A. In the same way as we can show that minimization of $\|\mathcal{F}(Ae^{j\phi})\|$ (i.e., maximization of $\|\mathcal{G}(Ae^{j\phi})\|$) can be performed through the updates $\phi^{(k+1)} \leftarrow \angle\mathcal{G}(Ae^{j\phi^{(k)}})$, we can show that maximization of $\|\mathcal{F}(Ae^{j\phi})\|$ can be performed through the updates

$$\phi^{(k+1)} \leftarrow \angle\mathcal{F}(Ae^{j\phi^{(k)}}). \tag{2}$$

Assuming the algorithm converged or was stopped after K iterations, we consider the complex spectrogram $\tilde{S} = \mathcal{F}(Ae^{j\phi_K})$. This spectrogram is very close to $Ae^{j\phi_K}$ by construction, and in particular its magnitude $\tilde{A}$ is very close to A. Furthermore, it also verifies $\mathcal{G}(\tilde{S}) = 0$. We thus built a complex spectrogram whose magnitude is very similar to A but which resynthesizes to silence through inverse STFT.

Let us look at the example of a speech signal s by a female speaker sampled at 16 kHz. The magnitude A of its complex STFT spectrogram $S = Ae^{j\phi}$, computed using a sine window with window length 512 and 75 % overlap between frames, is shown in Fig. 1 (all figures show the magnitude to the power 0.3 for better visibility). The above algorithm based on iterative STFT computations is initialized using approximate methods similar to those described in [3], and stopped after 200 iterations. It outputs a spectrogram $\tilde{A}e^{j\tilde{\phi}}$ which resynthesizes to silence and whose magnitude is very close to that of the original sound. The signal-to-distortion ratio (SDR) between these two magnitude spectrograms, which corresponds to the SDR between the original signal and the signal reconstructed from $\tilde{A}$ combined with the original phase, is +77 dB. Reconstructing the sound

* スペクトログラム矛盾性最大化と位相制御による音の転写、ルルー・ジョナトン (NTT)

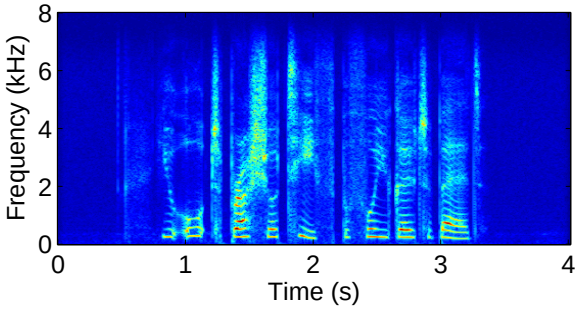

Fig. 1 *Magnitude spectrogram of a speech utterance.*

which "intuitively" corresponds to $\tilde{A}$, i.e., estimating the phase which minimizes inconsistency, leads as expected to a signal whose magnitude spectrogram is very close to that of the original sound, with an SDR of +31 dB. Depending on the phase information, the same magnitude spectrogram can thus lead to a very good reconstruction of a particular sound as well as silence. We can actually go further, as we now explain.

4 Phase-controlled sound transfer

Let $X = Be^{j\psi}$ be the complex STFT spectrogram of a (different) signal x with the same length as the signal s considered above. The STFT analysis conditions are also assumed to be identical. We consider the family of complex spectrograms

$$\tilde{S}_\lambda = \tilde{A}e^{j\angle(\tilde{S}+\lambda X)} \tag{3}$$

parameterized by $\lambda \geq 0$. Note that no information concerning X appears in the magnitude part of $\tilde{S}_\lambda$. The perceived quality of the sound reconstructed from $\tilde{S}_\lambda$ dramatically depends on the value of λ:

- For $\lambda = 0$, $\tilde{S}_0 = \tilde{S}$ resynthesizes to silence.
- For $\lambda \gg 1$, we get $\tilde{S}_\lambda \approx \tilde{A}e^{j\angle X} \approx Ae^{j\psi}$, i.e., $\tilde{S}_\lambda$ is close to the spectrogram formed by the magnitude of the first sound and the phase of the second sound. In general, this will lead to a resynthesized signal sounding like a noisy version of the first sound: the influence of the magnitude is greater as generally expected.
- For $0 < \lambda \ll 1$, we have

  $$\tilde{S}_\lambda = \tilde{S} + \lambda X + O(\lambda^2).$$

  To show this, we first rewrite $\tilde{S}_\lambda$ as

  $$\tilde{S}_\lambda = \frac{|\tilde{S}|}{|\tilde{S}+\lambda X|}(\tilde{S}+\lambda X).$$

  As $\tilde{S} = \mathcal{F}(S)$ and $X = \mathcal{G}(X)$ (as a consistent spectrogram) respectively lie in the images of $\mathcal{F}$ and $\mathcal{G}$, which are orthogonal projections on complementary subspaces, $\tilde{S}$ and X are orthogonal. In particular, the denominator $|\tilde{S}+\lambda X|$ will only lead to second and higher order terms in λ. As $\tilde{S}$ will perfectly cancel out in the resynthesis, the signal reconstructed from $\frac{1}{\lambda}\tilde{S}_\lambda$ will be close to x, the second signal. Note that the result holds strictly only in regions where $\tilde{A} > 0$.

Summarizing the above results, we see that for large contributions of the phase of x, the resynthesized signal will tend to sound like the sound s. On the other hand, theoretically, the smaller (while staying strictly positive) the contribution of the phase of the sound x on $\tilde{S}_\lambda$, the closer (up to rescaling) the resynthesized signal will sound to x.

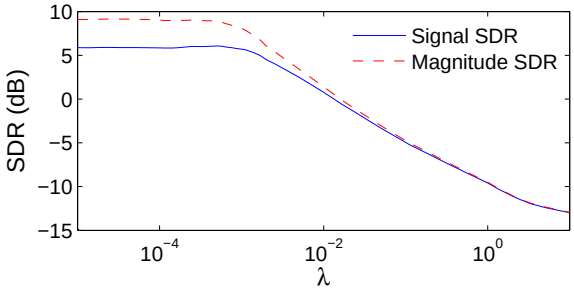

Fig. 2 *Evolution of the SDR w.r.t. λ.*

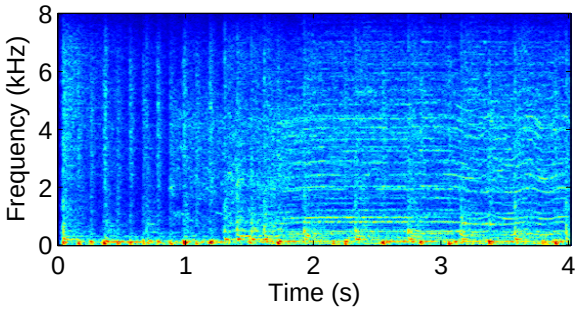

Fig. 3 *Magnitude spectrogram of the reconstructed rock music signal.*

Of course, in practice, dynamic range issues kick in for very small values of λ due to machine precision. As an example, we consider for the second sound x an excerpt of rock music. The evolution of the SDR of the reconstruction of x according to the value of λ is shown in Fig. 2, computed both on the signals themselves and on their magnitude spectrograms (i.e., assuming equal phase). For $\lambda = 3 \times 10^{-4}$, for example, the signal SDR is +6.0 dB, while the magnitude SDR is +9.0 dB. Perceptually, the reconstruction is very good. The magnitude spectrogram of the reconstructed signal is shown in Fig. 3. We stress again that the information about x in $\tilde{S}_\lambda$ only appears in the phase part. Reconstruction of the rock music is performed through subtle combinations of the speech magnitude, including the faint background noise before and after the utterance, controlled by phase.

5 Conclusion

We showed that, while inconsistency minimization with a given magnitude spectrogram leads to reconstruction of a sound corresponding to intuition, inconsistency maximization leads to resynthesized signals with very low energy. We devised a method to build highly-inconsistent spectrograms with a given magnitude and explained how a unique magnitude spectrogram can lead to resynthesized signals close to virtually any sound only through phase manipulation.

References

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

[2] J. Le Roux, N. Ono, and S. Sagayama, "Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction," in *Proc. SAPA*, Sep. 2008.

[3] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. DAFx-10*, Sep. 2010.