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

Phase-controlled sound transfer based on maximally-inconsistent spectrograms

ルルー・ジョナトン　(NTT CS研)
Jonathan Le Roux

NTT

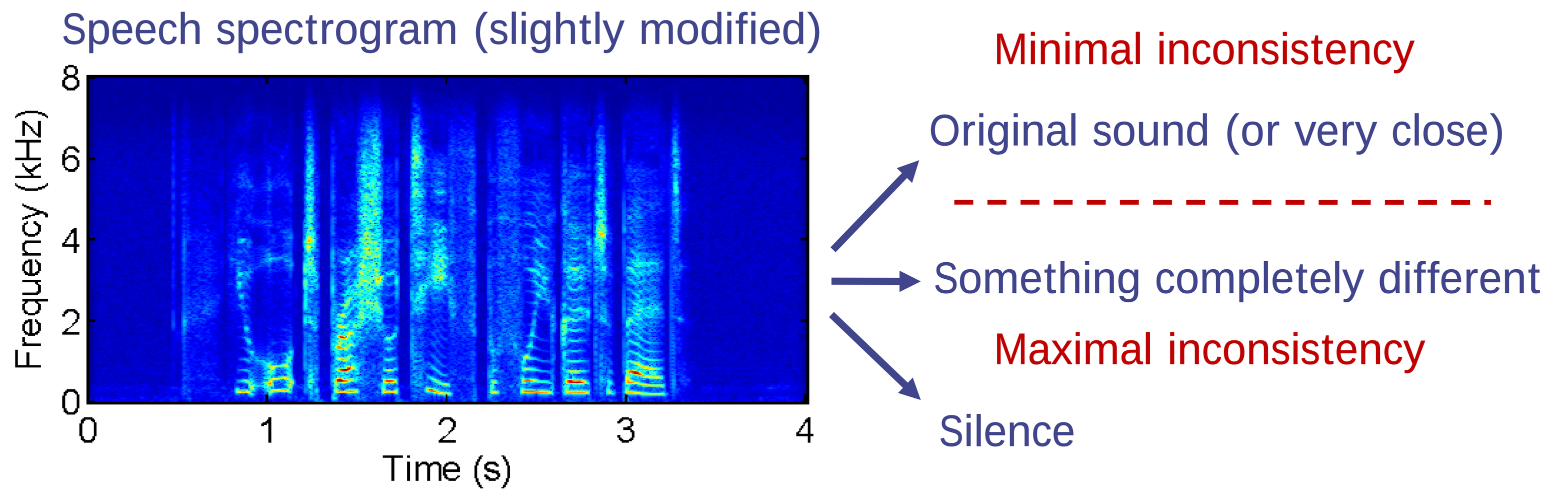
Abstract

- Magnitude of STFT spectrogram generally considered a reliable cue to build intuition on resynthesized signal
  - Spectrogram reading
  - Algorithms for sound reconstruction from magnitude only

→ At worst, bad choice of phase only leads to noisy reconstruction of what the intuition suggests?

Answer: Wrong!

- Intuition linked to spectrogram "inconsistency"
  - Results meet **intuition** for **minimal** inconsistency
  - What happens for **maximal** inconsistency?
- Same magnitude spectrogram can lead to extremely diverse resynthesized signals depending on phase



1. Consistency and intuition

Given an array M of real non-negative numbers $M_{\omega,k}$, what do we intuitively expect M to "sound like"?

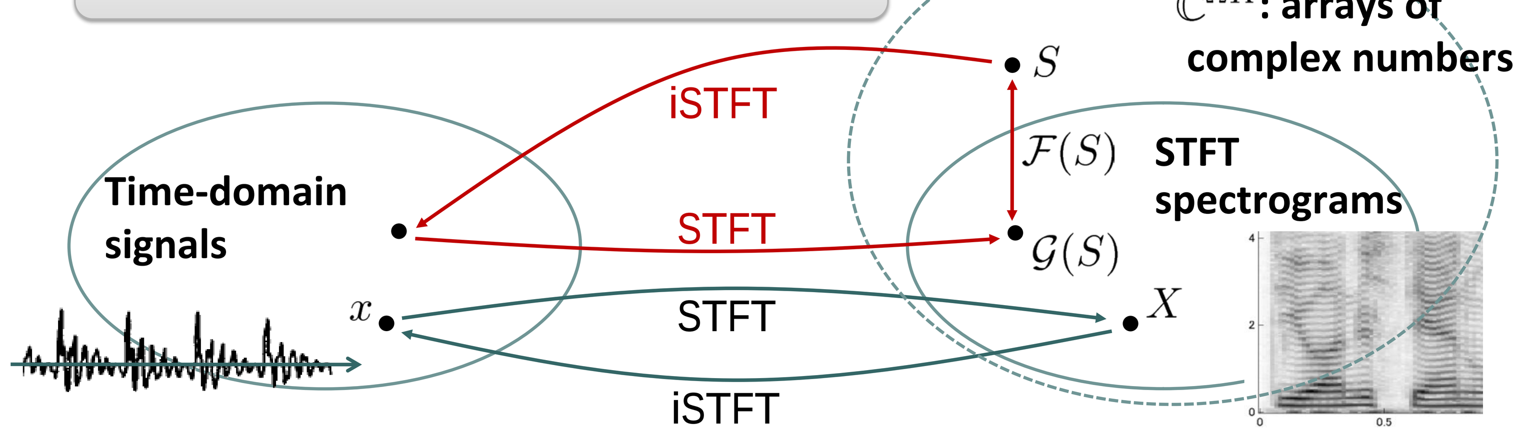
- Classical task:
  Estimate a time-domain signal whose magnitude spectrogram is closest to M in a least-squares sense

→ Reconstructed signal expected to sound close to intuition

- Equivalent formulation:
  Estimate phase ϕ such that $S = Me^{j\phi}$ is "as consistent as possible", i.e., as close as possible to the spectrogram of the sound resynthesized from itself, $\mathcal{G}(S) = \mathrm{STFT}(\mathrm{iSTFT}(S))$

Numerical criterion: "how inconsistent?"

$$\mathcal{I}(S) = \|\mathcal{F}(S)\|^2 = \|S - \mathcal{G}(S)\|^2$$

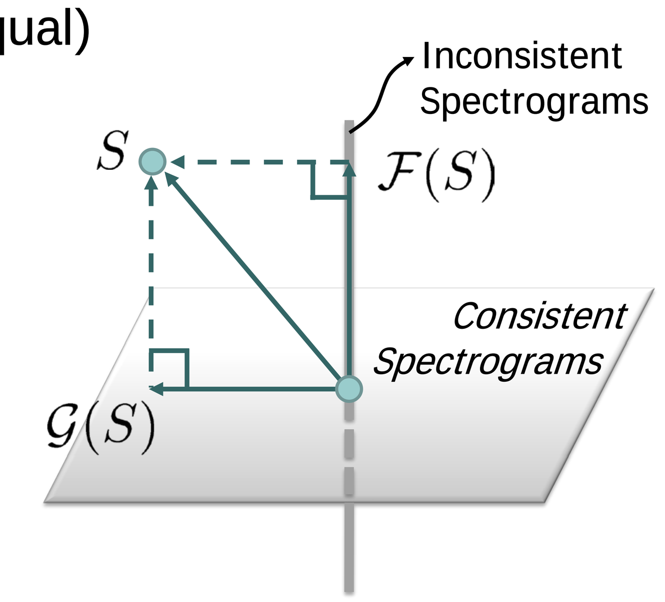


Intuition ⟷ minimum inconsistency

2. Maximizing inconsistency

- $\mathcal{G}$ orthogonal projection on consistent spectrograms
  (synthesis and analysis windows assumed equal)
- As $\mathcal{F} = \mathrm{Id} - \mathcal{G}$,

$$\|S\|^2 = \|\mathcal{G}(S)\|^2 + \|\mathcal{F}(S)\|^2$$



- Minimum inconsistency:
  $\mathcal{G}(S) = S$: STFT spectrograms
- Maximum inconsistency:
  $\mathcal{G}(S) = 0$: resynthesizes to silence
  - Trivial for rectangular windows, 50% or 75% overlap
  - Not trivial in general for other windows or overlap ratios

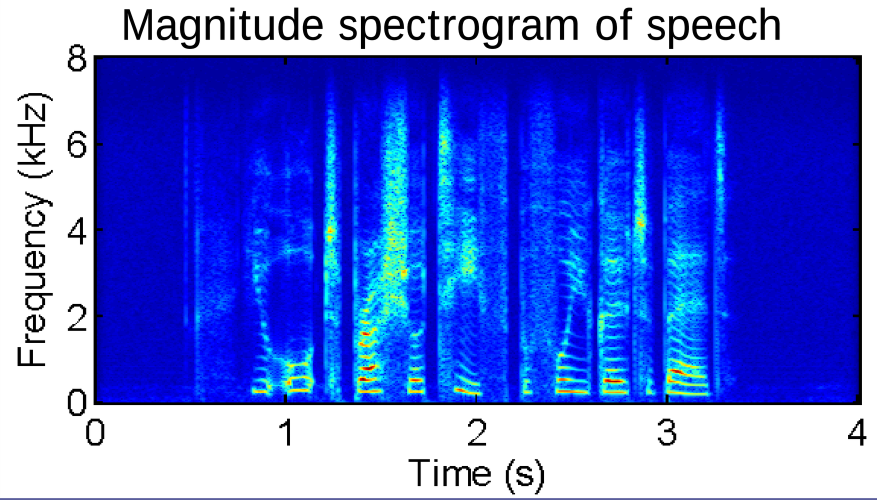
- Inconsistency maximization algorithm
  - Iterative STFT for minimization [1]:
    project on consistent spectrograms with $\mathcal{G}$, keep only the phase
  - Here: project on inconsistent spectrograms with $\mathcal{F}$

Inconsistency	Objective	Algorithm
Minimization	$\mathrm{argmin}_\psi \|\mathcal{F}(Me^{j\psi})\|^2$	$\psi^{(k+1)} \leftarrow \angle\mathcal{G}(Me^{j\psi^{(k)}})$
Maximization	$\mathrm{argmin}_\phi \|\mathcal{G}(Me^{j\phi})\|^2$	$\phi^{(k+1)} \leftarrow \angle\mathcal{F}(Me^{j\phi^{(k)}})$

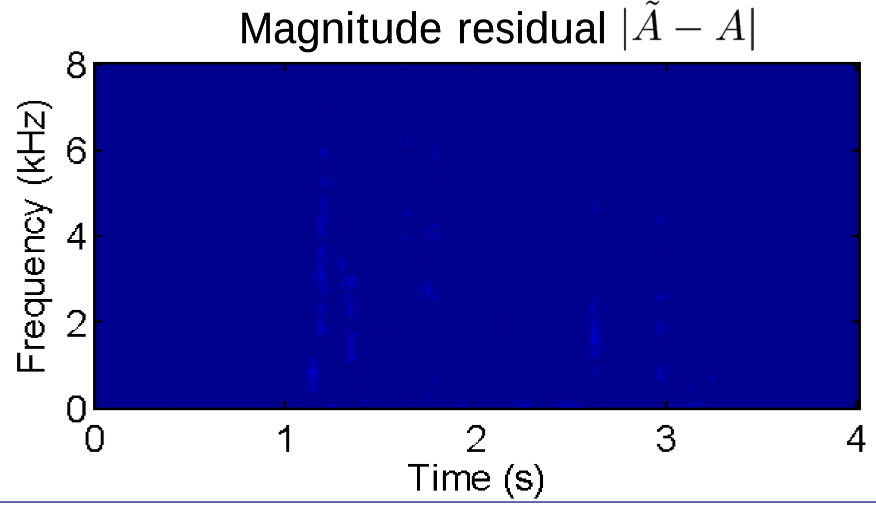
- Leads to $\tilde{S} = \mathcal{F}(Me^{j\phi_N})$
  - Very close to $Me^{j\phi_N}$: in particular, $|\tilde{S}|$ close to M
  - Verifies $\mathcal{G}(\tilde{S}) = 0$: resynthesizes to silence
- Fast approximations as in [3]

- Example: sound a, speech by female speaker, $S_a = Ae^{j\phi_a}$
  - Above algorithm leads to $\tilde{S}_a = \tilde{A}e^{j\tilde{\phi}_a}$
    - Magnitude close to A: +77dB SDR between $\tilde{A}$ and A
    - $\tilde{A}e^{j\tilde{\phi}_a}$ resynthesizes to silence
  - ! Estimate minimally-inconsistent phase ψ for $\tilde{A}$
    - $\tilde{A}e^{j\psi}$ resynthesizes to speech with +31dB magnitude SDR

Magnitude spectrogram of speech



Magnitude residual $|\tilde{A} - A|$



3. Phase-controlled sound transfer

- Another sound b, with complex spectrogram $S_b = Be^{j\phi_b}$

→ Consider the family of spectrograms:

$$\tilde{S}^{(\lambda)} = \tilde{A}e^{j\angle(\tilde{S}_a + \lambda S_b)}$$

No information on b in the magnitude!

$\lambda = 0$: $\tilde{S}^{(0)} = \tilde{S}_a$ → Silence

$\lambda \gg 1$: $\tilde{S}^{(\lambda)} \approx \tilde{A}e^{j\angle S_b} \approx Ae^{j\phi_b}$

→ Noisy version of sound a

$0 < \lambda \ll 1$: $\tilde{S}^{(\lambda)} = \tilde{S}_a + \lambda S_b + O(\lambda^2)$

$\Rightarrow \mathrm{iSTFT}(\tilde{S}^{(\lambda)}) = \lambda\,\mathrm{iSTFT}(S_b) + O(\lambda^2)$

→ Scaled-down version of sound b

Proof of the $0 < \lambda \ll 1$ case:

$$\tilde{S}^{(\lambda)} = \frac{|\tilde{S}_a|}{|\tilde{S}_a + \lambda S_b|}(\tilde{S}_a + \lambda S_b)$$

$$\tilde{S}_a = \mathcal{F}(S_a) \perp \mathcal{G}(S_b) = S_b$$

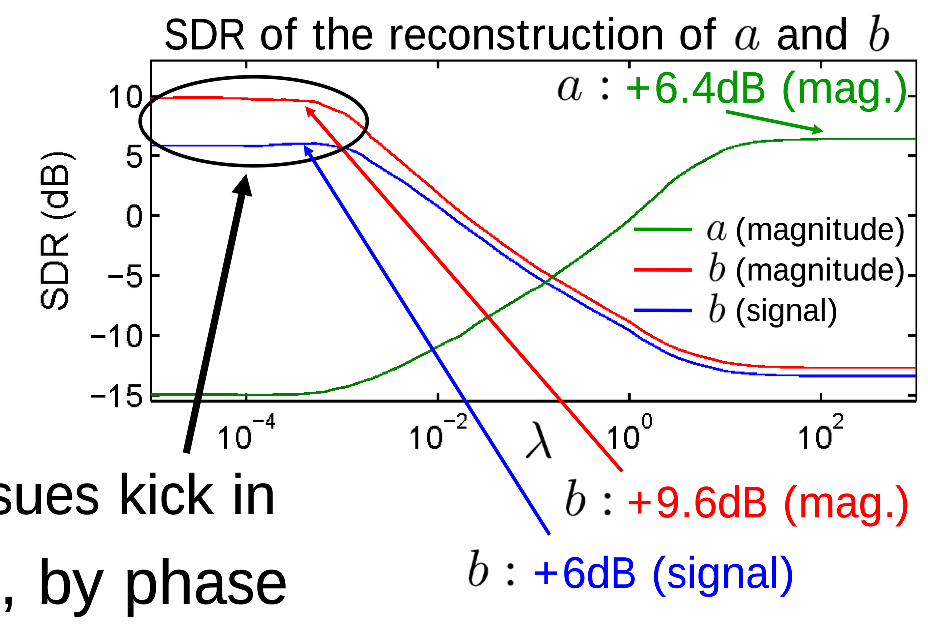
$$\hookrightarrow |\tilde{S}_a + \lambda S_b|^2 = |\tilde{S}_a|^2 + \lambda^2|S_b|^2$$

$$\Rightarrow \frac{|\tilde{S}_a|}{|\tilde{S}_a + \lambda S_b|} = 1 + O(\lambda^2)$$

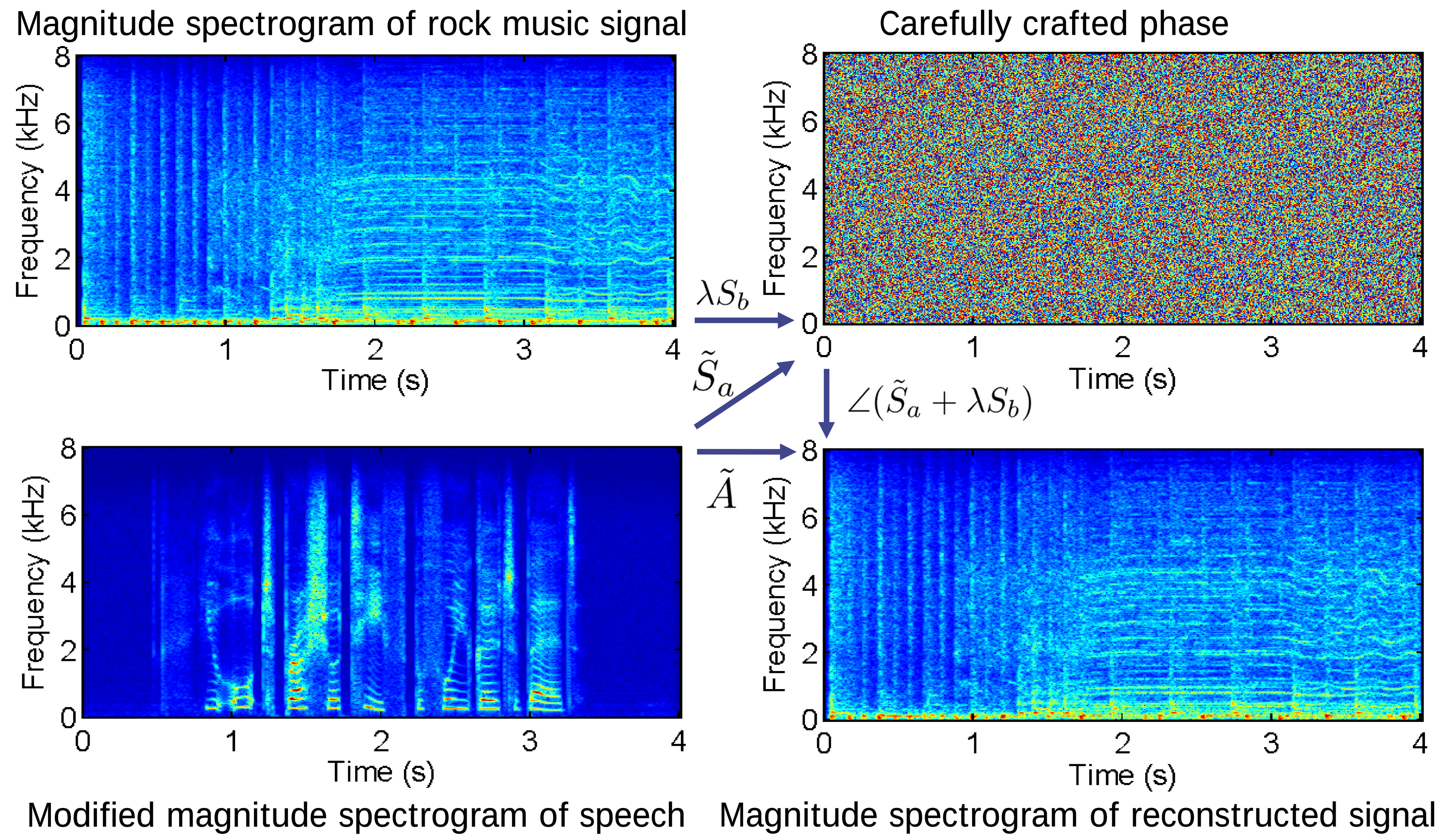
(for $\tilde{A} > 0$)

- Surprising relation:

Contribution of b to the phase	$\mathrm{iSTFT}(\tilde{S}^{(\lambda)})$ sounds like
Strong	a
Weak	b



- Example: from speech to rock, by phase



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.