Audio & Speech:
Audio Reconstruction from MFCC

DSP Lab 2021 autumn Audio and Speech week 2

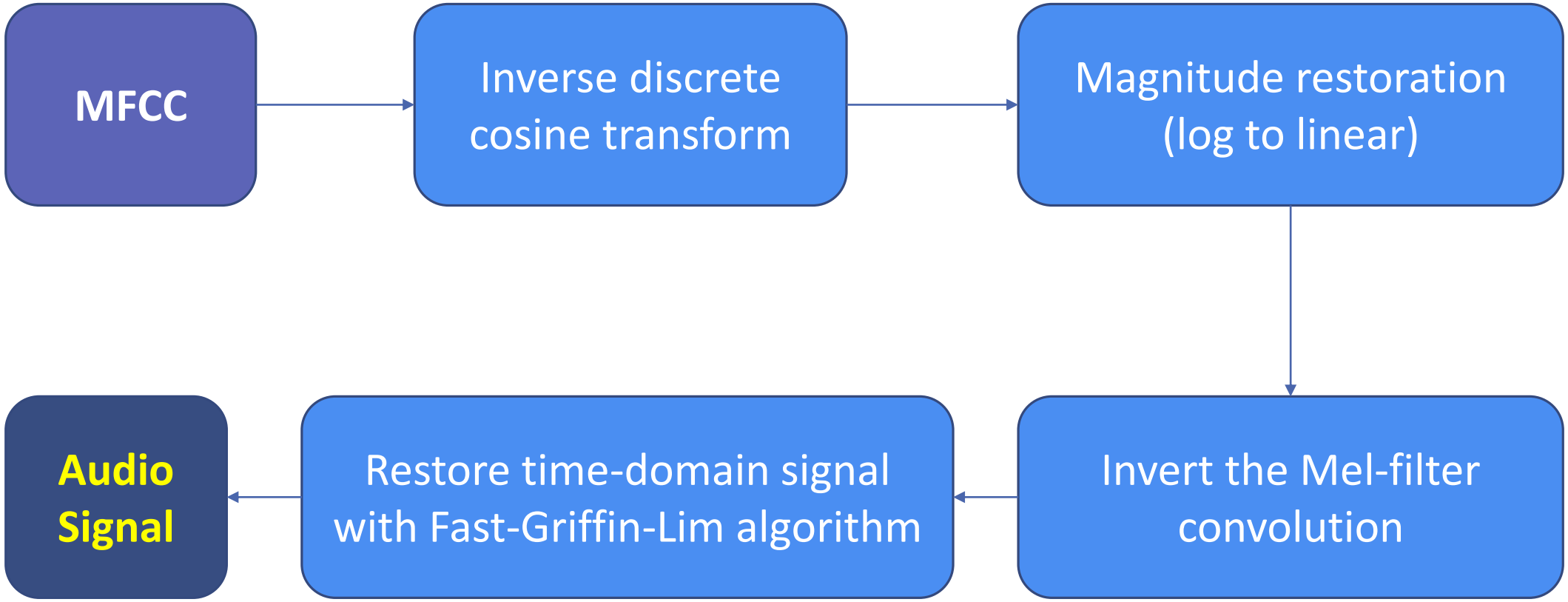
授課老師：劉奕汶教授, 李祈均教授

助教：陳靖杰, 陳舫慶, 邱信豪, 許暐彤

Outline:

- MFCC inversion overview.
- Elaboration on some components.
- Demo.
- Report requirements.

Audio reconstruction flow chart



Component(1): Filter inv-convolution

- Recall that the convolution was:

$$\boldsymbol{Features = STFT\{signal\} * Mel - Filter}$$

- Notice that for a linear transformation $\boldsymbol{Y = XA}$ with given $\boldsymbol{Y}$ and $\boldsymbol{A}$, the $\boldsymbol{X}$ that yields ordinary least square: $(\boldsymbol{Y - XA})^T(\boldsymbol{Y - XA})$ can be found with equation:

$$\boldsymbol{X = (A^T A)^{\dagger} A^T Y}$$

- Where $\boldsymbol{Y}$ is ***Features***, the ***Mel-Filter*** is $\boldsymbol{A}$, and we want ***STFT{signal}*** as $\boldsymbol{X}$.

Component(2): **Fast Griffin-Lim**

$P_{C_1}(c) = STFT\{\, invSTFT\{\, c \,\}\,\}$

$P_{C_2}(c) = s \cdot e^{i\angle c}$

α_n: $step\ size$

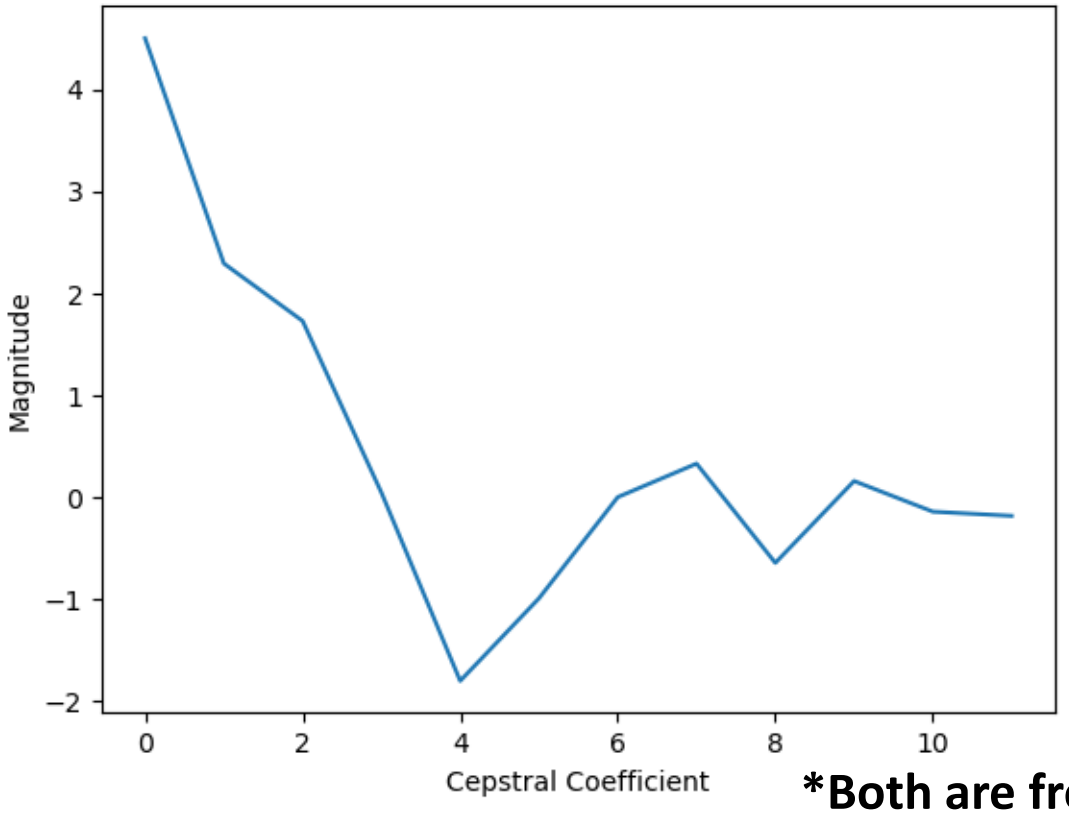
$G^{\dagger}$: $invSTFT\{\}$

Fix the initial phase $\angle c_0$
Initialize $c_0 = s \cdot e^{i\angle c_0}$, $t_0 = P_{C_2}\left(P_{C_1}(c_0)\right)$
Iterate for $n = 1, 2, \ldots$
$t_n = P_{C_1}\left(P_{C_2}(c_{n-1})\right)$
$c_n = t_n + \alpha_n(t_n - t_{n-1})$
Update α_n
Until convergence
$x^* = \mathbf{G}^{\dagger} c_n$

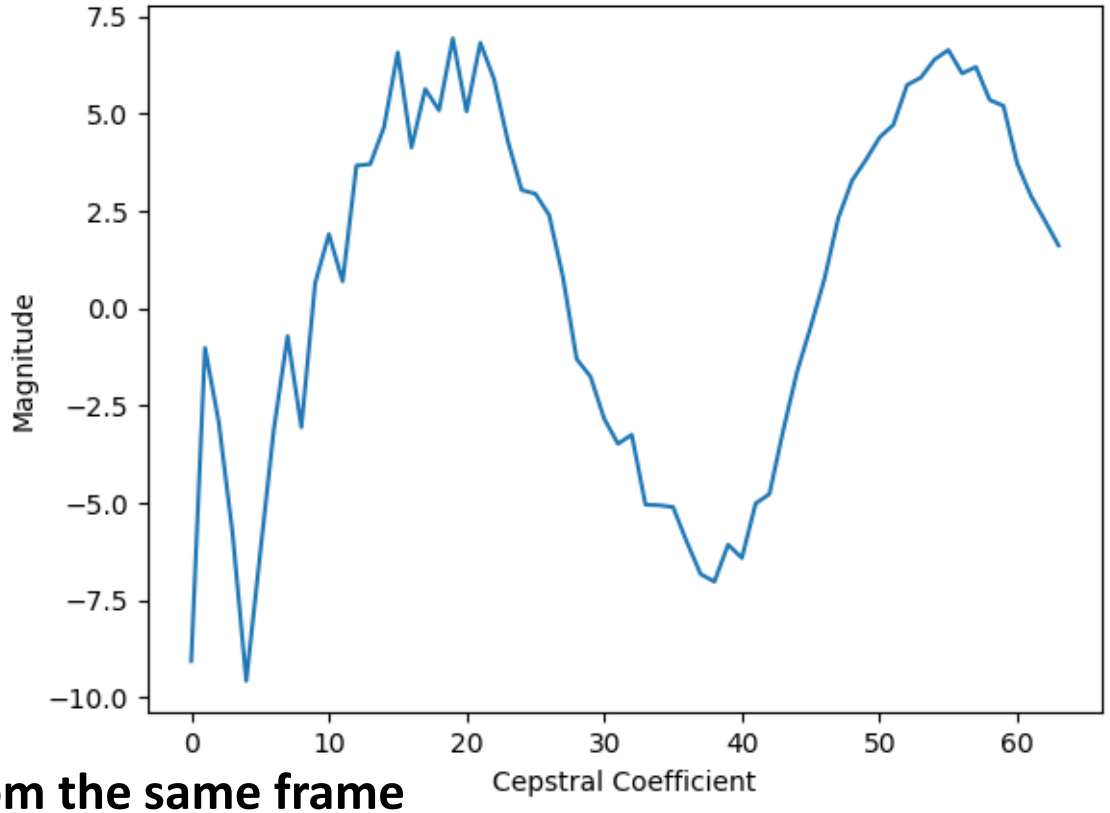
Source: Perraudin, Nathanaël & Balazs, Peter & Søndergaard, Peter. (2013). 'A Fast Griffin–Lim Algorithm.' IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 1-4. 10.1109/WASPAA.2013.6701851.

Demo(1): **Effect of num Mel-Filter banks**

12 banks



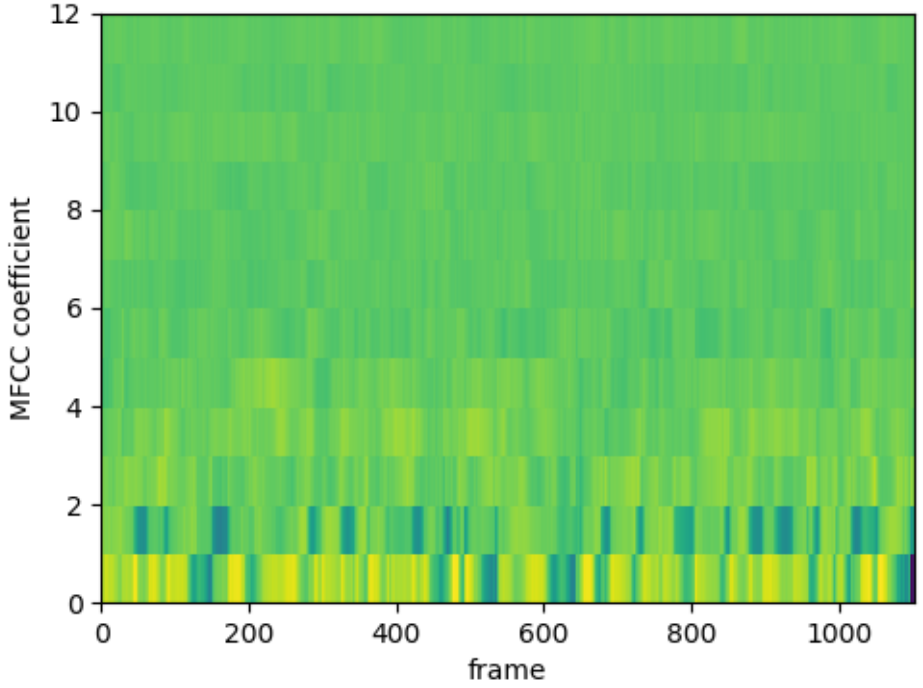
64 banks



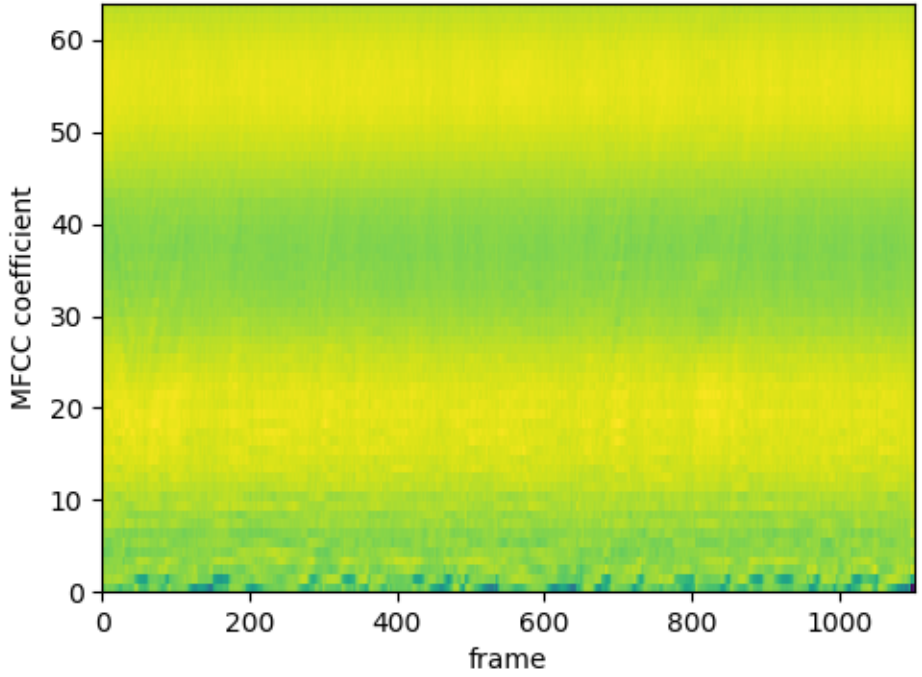
***Both are from the same frame**

Demo(2): **Effect of num Mel-Filter banks**

12 banks MFCC

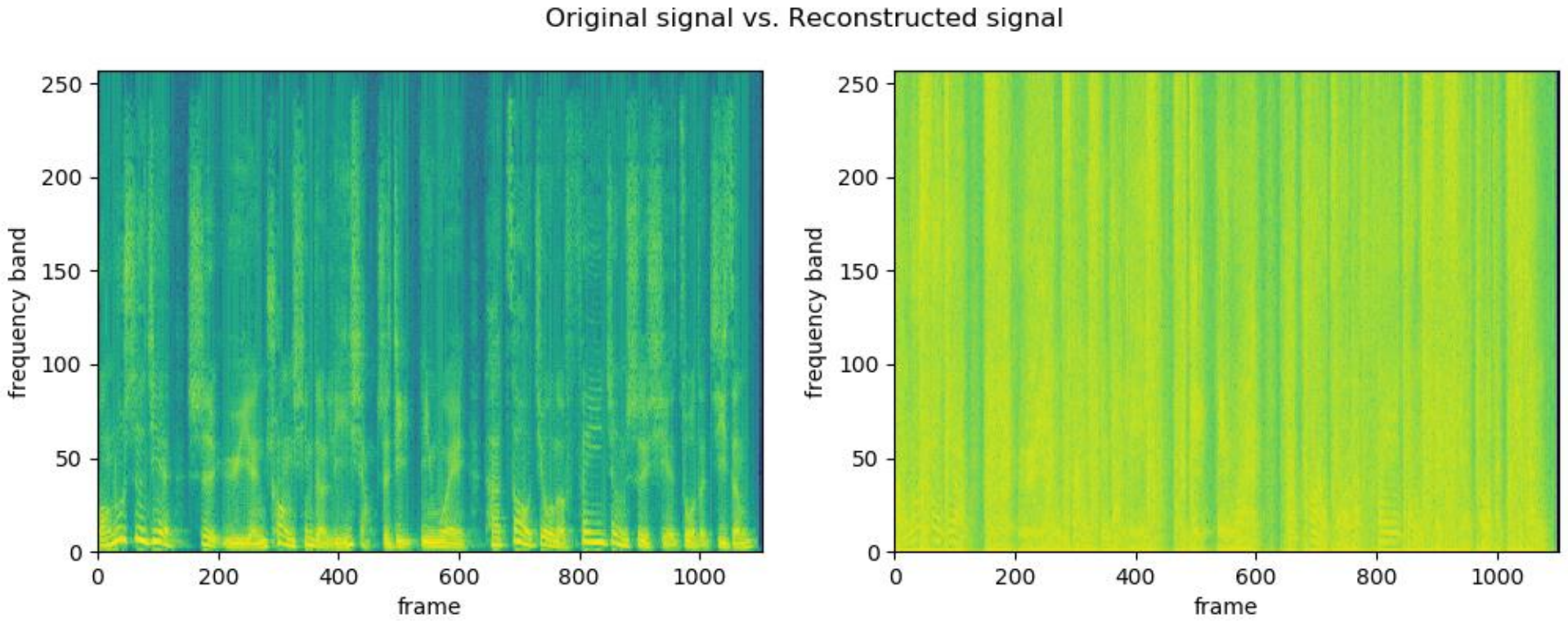


64 banks MFCC



Demo(3): **Ori vs Reconstructed**

64 banks, magnitude is log-scaled



Demo(4): **Reconstructed Audio**

- 64 banks pre-emphasized: 
- 64 banks NOT pre-emphasized: 
- 12 banks pre-emphasized: 
- 12 banks NOT pre-emphasized: 
- Original: 

Report questions:

1. **Question 1:** What are the artifacts and distortions in the reconstructed audio? Suggest what the causes of these degradations are. (i.e. which sections of the MFCC extraction process are not invertible?)
2. **Question 2:** Experiment with different frame length, step length, and number of fbanks; discuss what effects each of them has in the reconstruction process.
3. **(Bonus 1)** Aside from setting optimal parameters, what can be added in the reconstruction algorithm to improve the end quality? Implement your proposal and present some experiments of it.
4. **(Bonus 2)** We did not perform dimension reduction/reconstruction in the DCT/inv-DCT sections. Modify those parts such that we have a complete algorithm that performs compression/decompression. Discuss how this influences the reconstruction quality.