

Abstract

Contributions

Recent **mel spectrogram inversion** models (vocoders) developed for speech achieve a high degree of realism on speech, but exhibit pitch instability when reconstructing sustained musical notes.

With respect to speech, music has longer sustained notes, may be polyphonic, and requires higher pitch precision.

We propose a new vocoder model that is specifically designed for music. Key to improving the **pitch stability** is the choice of a **shift-invariant target space** that consists of the magnitude spectrum and the phase gradient.

Our contributions include:

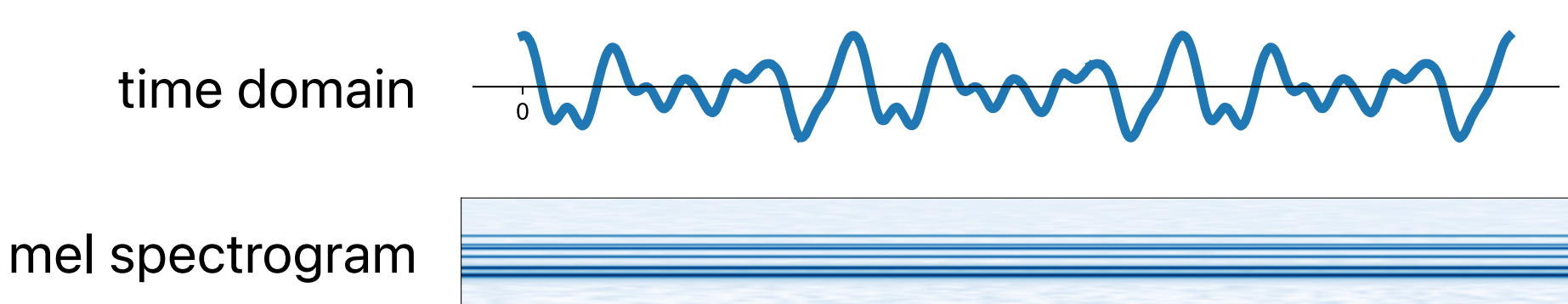
- a formulation of the mel spectrogram inversion task specifically designed for music, matching shift-invariant network and target, in order to improve the perceived stability of sustained notes
- a phase integration algorithm (see the paper)
- the harmonic error metric, which measures pitch stability on reconstructed audio including multiple sustained notes

Methods

Experiments

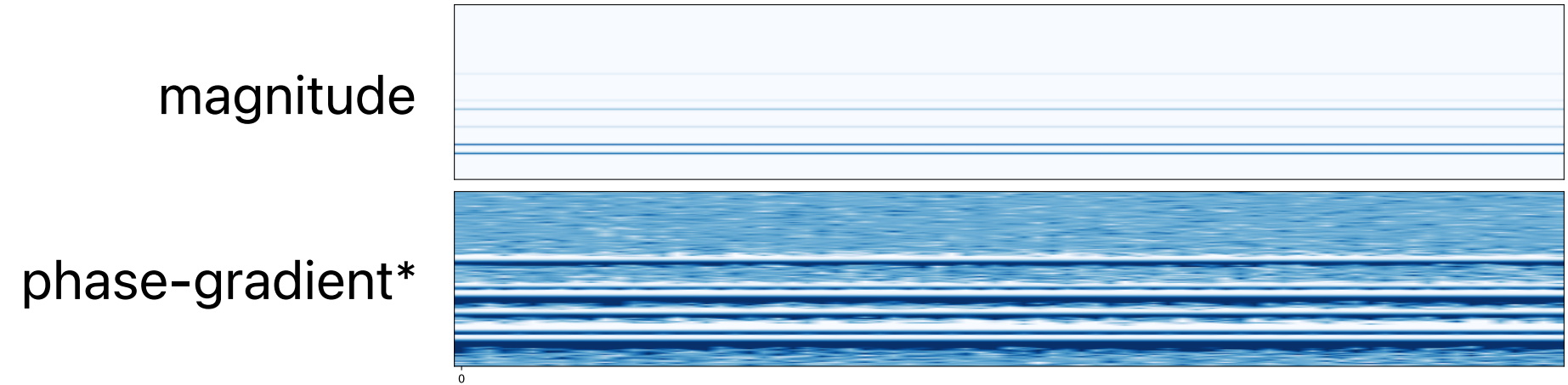
The stability of a sustained pitched note manifests in the time-domain audio signal as the steady repetition of a periodic waveform.

Periodic patterns are by definition not shift-invariant.



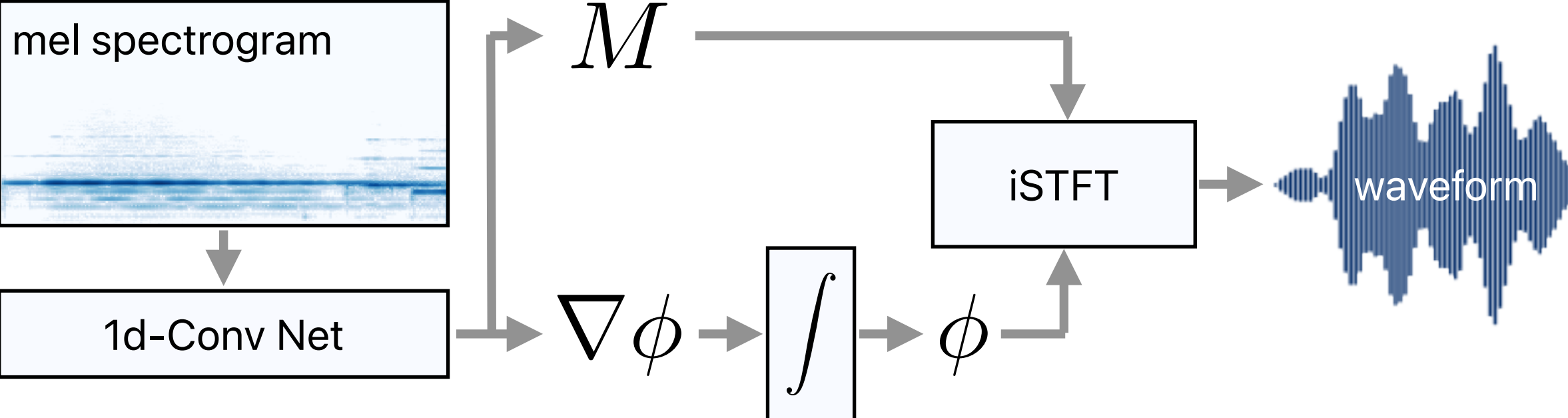
End-to-end models are required to learn all possible **shifts** of periodic patterns, a space that increases exponentially for polyphonic music, and how to **sequence** these patterns.

The proposed formulation uses, as the target of a convolutional network, a time-frequency representation that is shift-invariant for sustained notes.

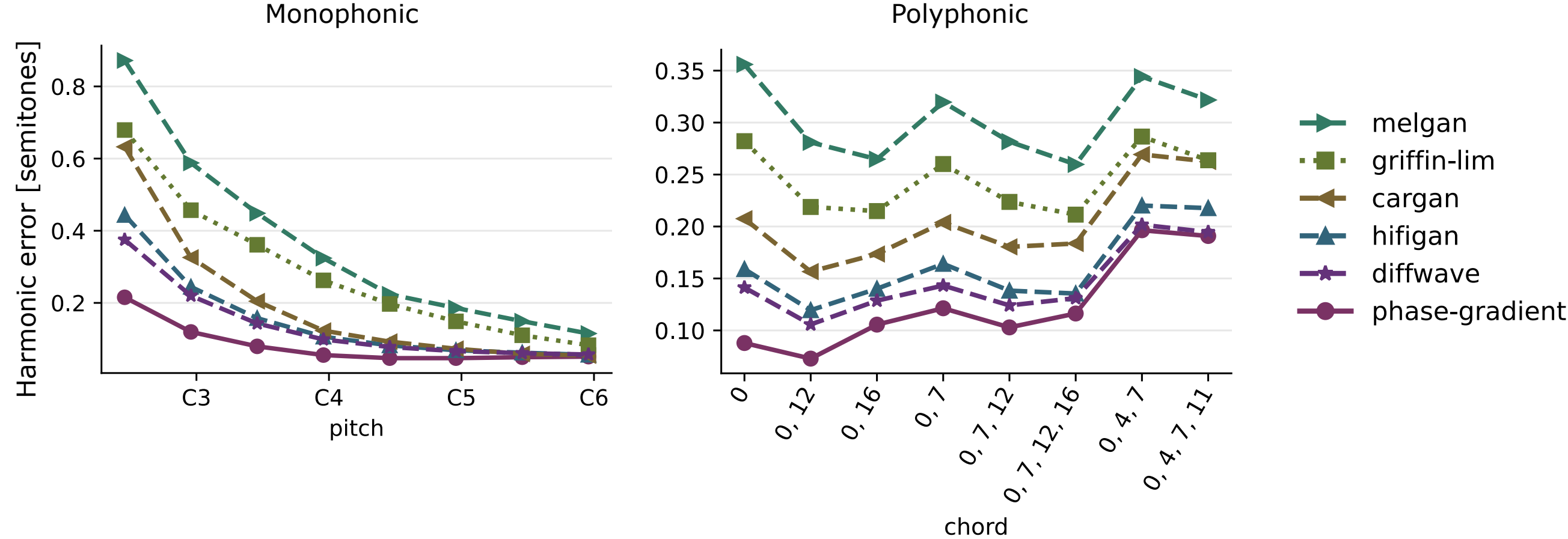


\* The figure shows only the frequency bin offset component, related to the partial derivative of phase along time (see the paper for details)

The proposed model for mel spectrogram inversion:



**Pitch stability:** we synthesized a dataset of one second-long notes and chords from midi using different sounds, and measured stability by comparing the original and reconstructed audio with the harmonic error metric.



Harmonic error: for each (frame, note, harmonic index), the distance of the frequency of the closest spectral peak, between the reconstructed and the original audio.

**Overall reconstruction quality:** we computed the Frechét Audio Distance (FAD) on several datasets and ran a listening test.

	Frechét Audio Distance			Harmonic Error		Listening Test	Inference Speed	Capacity
	FAD (↓)			H <sub>err</sub> (↓)		MOS (↑)	RTF (↑)	#Params
	Ambient	NSynth	N+C	Notes	Chords	NSynth		
griffin-lim [17]	10.59	6.16	6.79	0.28	0.24	1.42 ± 0.11	7.14	0
melgan [1]	2.07	2.80	2.84	0.36	0.30	1.58 ± 0.15	179.90	4M
cargan [5]	6.47	8.31	8.45	0.21	0.21	1.74 ± 0.12	4.36	25M
hifigan [2]	<b>0.85</b>	1.31	<b>1.81</b>	0.16	0.17	2.89 ± 0.12	68.12	13M
diffwave [8]	2.62	6.84	1.89	0.14	0.15	3.19 ± 0.12	0.32	7M
phase-gradient	1.26	<b>1.26</b>	1.86	<b>0.09</b>	<b>0.14</b>	<b>3.73</b> ± 0.13	3.58	28M
oracle	0.51	0.33	0.00	0.00	0.00	4.34 ± 0.15	—	—

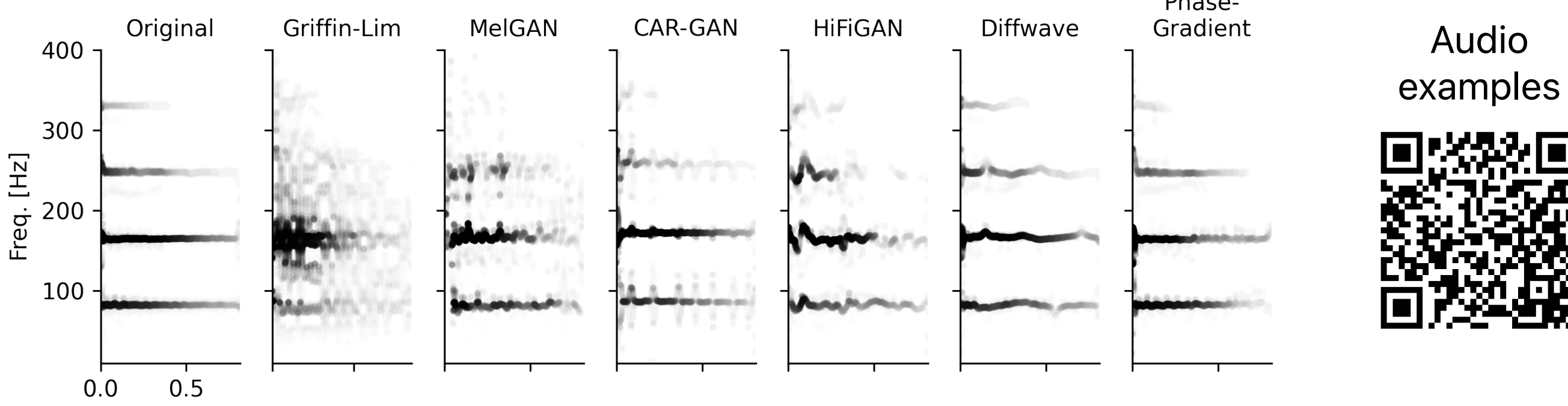
Table 1: Reconstruction results

Example

Conclusions

Reconstructions of an E2 nylon guitar note.

The figure highlights the instability on the fundamental and the harmonics caused by the lack of temporal phase coherence.



We have proposed a new mel spectrogram inversion model designed for music, using a frequency-domain target representation that is time shift invariant for harmonic signal components.

The proposed model is able to reconstruct single notes and chords more precisely than existing models, while still being competitive on generic music loop reconstruction.

Future directions include investigating log-frequency target representations, and improving the reconstruction of percussive components.