fnet

Kuan Yu
kuanyu@uni-potsdam.de

May 25, 2018

outline

complex output

adversarial network

attention

evaluation

stft

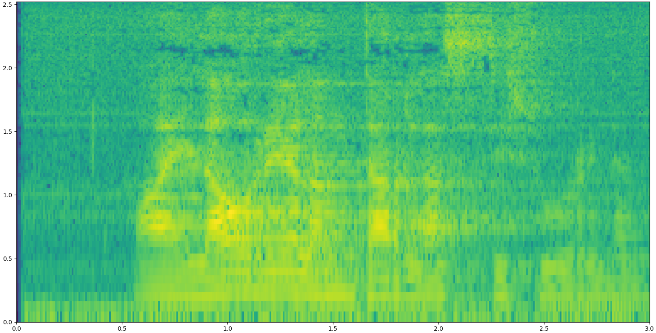
$$fs = 16,000$$
$$x : [-1, 1)^{3fs}$$

```python
from scipy.signal import stft
f, t, s = stft(x, fs)
```

$$s : \mathbb{C}^{129,376}$$

mel

```
plt.pcolormesh(t, np.log1p(f/700), np.log(np.abs(s)))
```



istft

```
from scipy.signal import istft
t2, x2 = istft(s, fs)
assert np.allclose(x, x2)
```

frames vs samples

- predicting frames takes much fewer steps
- an individual sample has no interpretable meaning
- a model predicting samples has to model much more complicated dependencies across a much longer time

vocoder

- most of the models we've seen has a trainable vocoder (wavenet, samplernn)
- to reconstruct the samples from frames
- which is unnecessary when we have complex-valued frames

complex network for speech

- 2016 Drude el al. "inappropriate for speech enhancement"
- 2016 Hu et al. "initial investigation"
- 2017 Fu el al. "complex spectrogram enhancement"
- 2018 Nakashika el al. "complex-valued rbm"

objective

- output expected complex-valued frames

	min	max	mean
`s.real`	-0.08	0.10	0.00
`s.imag`	-0.14	0.12	0.00
`s.abs`	6.65^{-09}	0.14	0.17^{-02}

- how to define the loss?

adversary

$$
\begin{aligned}
\text{frames} \quad & s: & \mathbb{C}^{f,t} \\
\text{generator} \quad & g: & ? \to \mathbb{C}^{f,t} \\
\text{discriminator} \quad & d: & \mathbb{C}^{f,t} \to \{0,1\}
\end{aligned}
$$

- zero-sum game $\arg\min_{\mathrm{g}} \max_{\mathrm{d}} v(g,d)$
- payoff $v(g,d) = \mathbb{E}_{s\sim p_{data}} \log d(s) + \mathbb{E}_{s\sim p_{model}} \log(1-d(s))$

attenttion

- lots of attention

problem

- how to evaluate

baby steps

- not to explode
- to drop the loss
- to output more than noises