Adversarial Generation of Time-Frequency Features with application in audio synthesis

Speaker: Andrés Marafioti
Co-Authors: Nathanaël Perraudin, Nicki Holighaus, Piotr Majdak

Acoustics Research Institute, Vienna
Austrian Academy of Sciences

International Conference on Machine Learning
Long Beach, California, June 11th, 2019
Time to time-frequency

**Time-domain signal**

**STFT Magnitude**
Time-frequency to time
Is it consistent?

Real STFT Magnitude

Fake STFT Magnitude
Applied to GANs

Real waveform $\in \mathbb{R}^{B \times T}$

TF transform

TF coefficients real $\in \mathbb{C}^{b \times M_2 \times T/a}$

Drop phase
Log transform

Fake waveform $\in \mathbb{R}^{B \times T}$

Inverse TF transform

TF coefficients fake $\in \mathbb{C}^{b \times M_2 \times T/a}$

Heap integration
Phase deriv. est.

TF representation real $\in \mathbb{R}^{b \times M_2 \times T/a \times c}$

GAN

Discriminator

Real/Fake $\in \mathbb{R}^{b \times 2}$

TF representation fake $\in \mathbb{R}^{b \times M_2 \times T/a \times c}$

Generator

Latent variable $\in \mathbb{R}^{b \times d}$

Consistency

Convergence Measure

Stop Training

Marafioti (ARI)  Adversarial Generation of TF Features  ARI 5 / 6
We trained on a dataset of spoken English digits [0-9].
We evaluated our results with perceptual tests.
Audio examples and implementations are available at tifgan.github.io

<table>
<thead>
<tr>
<th></th>
<th>WaveGAN digits</th>
<th>TiFGAN-M digits</th>
</tr>
</thead>
<tbody>
<tr>
<td>vs TiFGAN</td>
<td>vs WaveGAN</td>
<td></td>
</tr>
<tr>
<td>Real</td>
<td>86%</td>
<td>94%</td>
</tr>
<tr>
<td>TiFGAN</td>
<td>–</td>
<td>75%</td>
</tr>
<tr>
<td>WaveGAN</td>
<td>25%</td>
<td>–</td>
</tr>
</tbody>
</table>

Thank you for your attention!
Supported by the Austrian Science Fund (FWF; MERLIN, I 3067-N30).