SELF-SUPERVISED GENERATION OF SPATIAL AUDIO FOR 360° VIDEO

UC San Diego
Adobe

Pedro Morgado, Nuno Vasconcelos, Timothy Langlois and Oliver Wang

Motivation & Contributions

We introduce an approach to convert **mono audio** recorded by a 360° video camera into **spatial audio**, a representation of the distribution of sound over the full viewing sphere.

360° video provides viewers an **immersive viewing experience.** Spatial audio lets viewers turn their head, and have the audio follow the sound sources, rather than remaining fixed. However, spatial audio requires expensive equipment to record, and is rare in many 360 videos available for viewing today.

In order to close this gap, we introduce three main contributions:

- **Formalize the spatial audio generation problem for 360° video**
- **Design the first spatial audio generation procedure**
- **Collect two datasets and propose an evaluation protocol for future benchmark.**

Ambisonics

Ambisonics approximates the sound pressure field $f(\boldsymbol{\theta}, t)$ by it spherical harmonic decomposition at a single point in space.

$$f(\boldsymbol{\theta}, t) = \sum_{n=0}^{N} \sum_{m=-n}^{n} Y_n^m(\boldsymbol{\theta}) \phi_n^m(t)$$

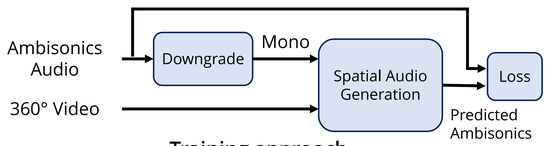
Ambisonic audio stores expansion coefficients $\phi_n^m(t)$, and can be **decoded in real time** to any set of speakers (e.g., headphones) to provide a realistic spatial audio experience.

First-order ambisonics (FOA) truncates the expansion at n=1. FOA coefficients are denoted $\phi_0^0 = \phi_w$, $\phi_1^{-1} = \phi_y$, $\phi_1^0 = \phi_z$ and , $\phi_1^1 = \phi_x$.

Spatial Audio Generation

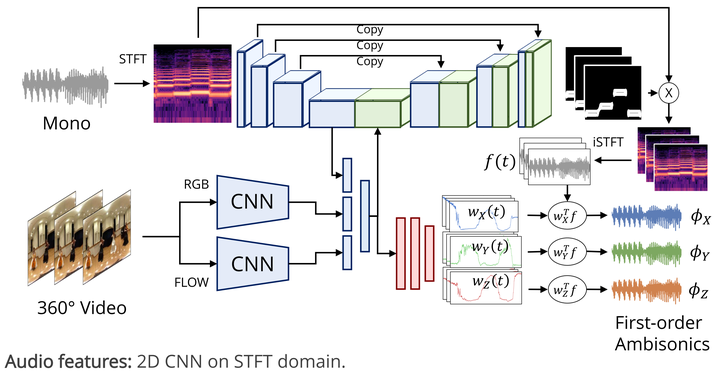
Given **non-spatial audio** $\boldsymbol{i(t)}$ (e.g., mono) and the corresponding **360° video** $\boldsymbol{v(t)}$, generate **first-order ambisonic channels** $\boldsymbol{\Phi(t)}$.

Self-supervision: To avoid collecting 360° videos with a pair of audio recordings for training, $i(t)$ and $\boldsymbol{\Phi}(t)$, we **downgrade** ambisonics audio into **mono**, and use the original as supervision.



Training approach

Network Architecture

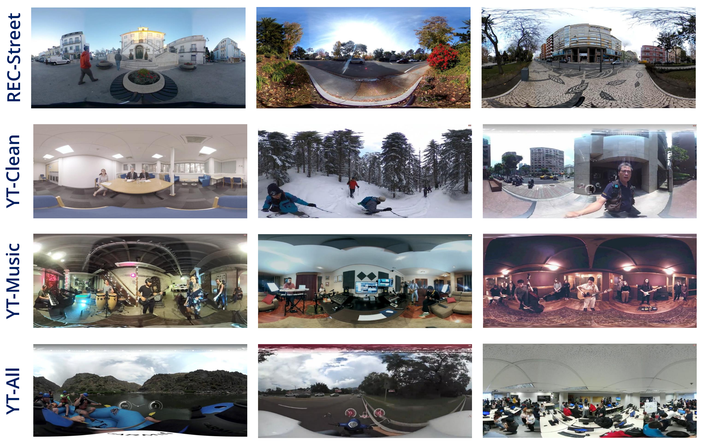


Audio features: 2D CNN on STFT domain.
Video features: ResNet18 on RGB and Flow streams.
Audio Separation: U-Net predicting separation masks for each source k, $a^k(\omega, t)$.
Source Localization: MLP predicting localization weights for each source k, $w^k(t)$.
Spatial Audio Generation: High-order ambisonic channels generated by

$$\widehat{\Phi}_c(t) = \sum_k w_c^k(t) \times iSTFT\left(a^k(\omega, t) \cdot STFT\big(i(t)\big)\right)$$

Training: Ambisonic audio $\Phi(t)$ is downgraded to mono $i(t)$. The model is learned by minimizing the reconstruction error in the STFT domain (l_2 distance).

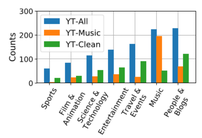
Spatial Audio Datasets



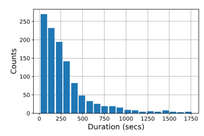
Dataset Stats

	REC Street	YT Clean	YT Music	YT All
# Videos	43	496	397	1146
# Hours	3.5	38	36	113
# Samples	123K	137k	128k	4M

Video Categories



Video Length



Evaluation

Quantitative evaluation:

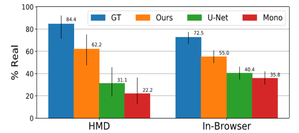
- **Frequency domain STFT:** Distance between STFTs averaged across channels.
- **Temporal domain ENV:** Distance between envelopes averaged across channels.
- **Localization EMD:** Earth Mover's Distance between the directional energy maps of spatial audio.

	REC-Street			YT-Clean			YT-Music			YT-All		
	STFT	ENV	EMD	STFT	ENV	EMD	STFT	ENV	EMD	STFT	ENV	EMD
Spatial Prior	0.187	0.958	0.492	1.394	2.063	1.478	4.652	4.355	3.479	2.691	3.394	2.246
U-Net Baseline	0.180	0.935	0.449	1.361	2.039	**1.403**	4.338	4.678	2.855	2.658	3.239	2.137
Ours-NoVideo	0.178	0.973	0.450	1.370	2.081	1.428	4.220	4.591	2.654	2.635	3.200	2.117
Ours-NoRGB	**0.158**	0.779	**0.425**	**1.339**	1.847	**1.405**	3.664	3.569	2.432	2.546	2.907	2.063
Ours-NoFlow	0.172	0.784	0.440	**1.349**	**1.778**	**1.402**	3.615	3.467	2.403	2.455	2.665	**2.023**
Ours-NoSep	**0.152**	0.790	**0.422**	1.381	**1.773**	1.415	3.627	3.602	2.447	**2.435**	2.694	2.050
Ours-Full	**0.158**	**0.767**	**0.419**	1.379	**1.776**	1.417	**3.524**	**3.366**	**2.350**	2.447	**2.649**	**2.019**

User study:

Participants shown a 360° video with predicted or ground-truth ambisonics and asked whether the audio matches the location of its sources (real) or not (fake). Study is conducted in two viewing environments:

- **In-browser 360° video platform:** 32 users recruited & 1120 clips watched.
- **Using head-mounted display (HMD):** 9 users recruited & 180 clips watched.



Conclusions

- Our architecture outperforms baselines by statistically significant margins in user studies.
- Low performance of NoVideo in quantitative evaluation indicates reliance on visual features.
- Spatial audio generation is still an open problem, which will benefit from enhancements in audio separation, visual grounding, multi-modal feature fusion, etc.

Predictions

Visit our website to watch 360 videos with spatial audio predictions on your phone.

Support

FCT Fundação para a Ciência e a Tecnologia
NSF

Website

https://pedro-morgado.github.io/spatialaudiogen/

Repository

https://github.com/pedro-morgado/spatialaudiogen/