Wave-U-Net:
A Multi-Scale Neural Network for End-to-End Audio Source Separation

19th International Society for Music Information Retrieval Conference (ISMIR 2018)

Authors: Daniel Stoller, Sebastian Ewert, Simon Dixon

Presenters: Court Zabel,
Brandon Pearl, &
Quinn Eggleston

APRIL 21, 2022

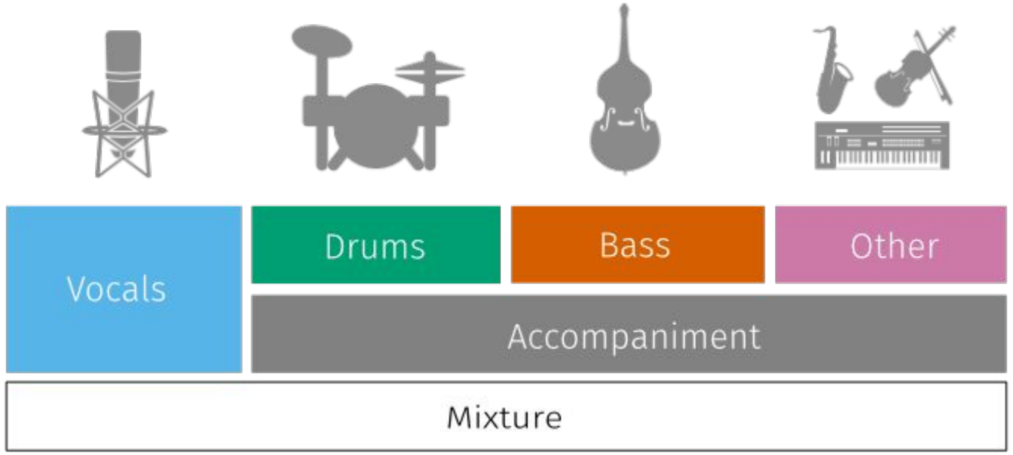
VIRGINIA TECH™

Audio Mixing & Separation

Audio & Music Source Separation

ISMIR

International Society for
Music Information Retrieval



Vocals

Drums

Bass

Other

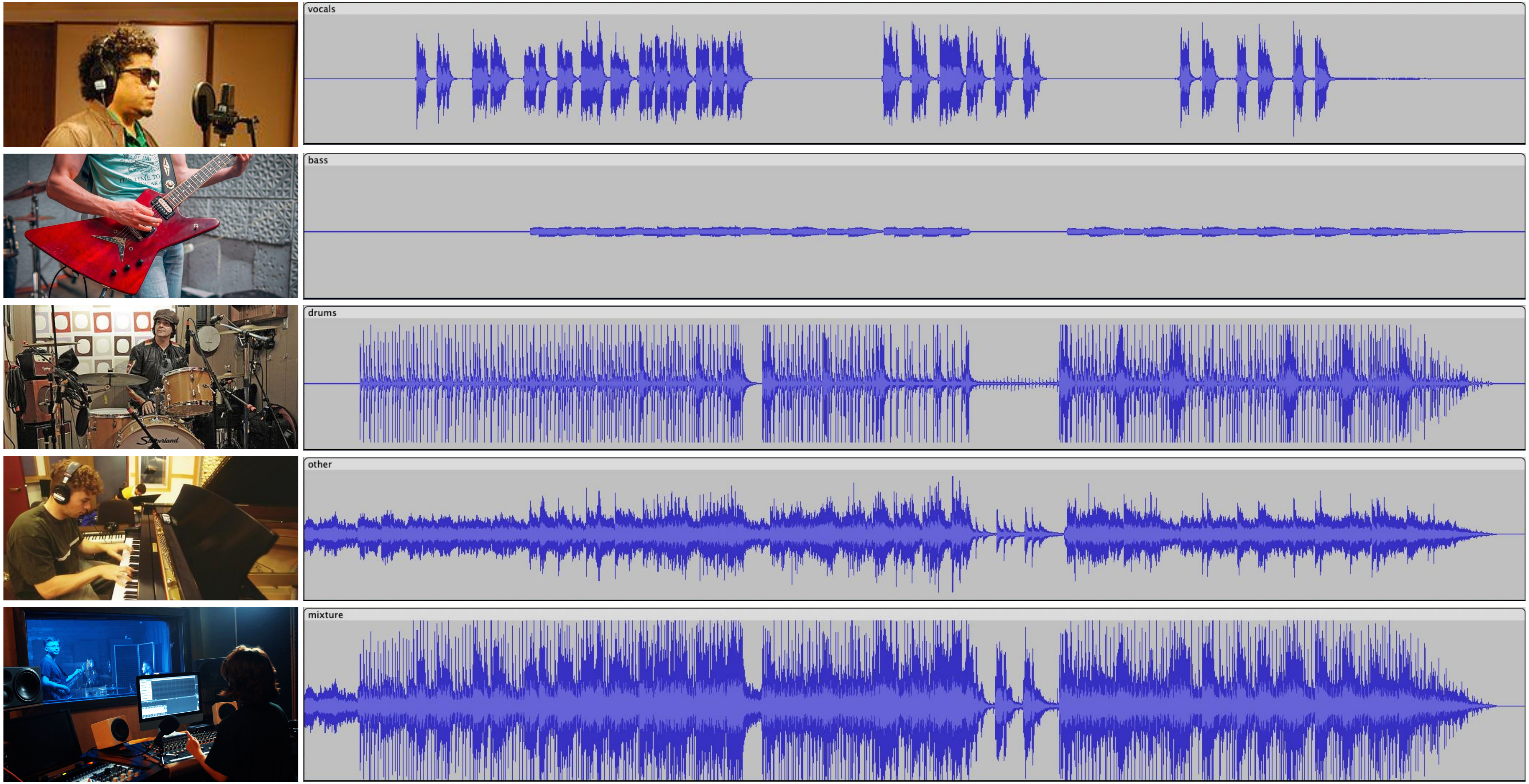
Accompaniment

Mixture

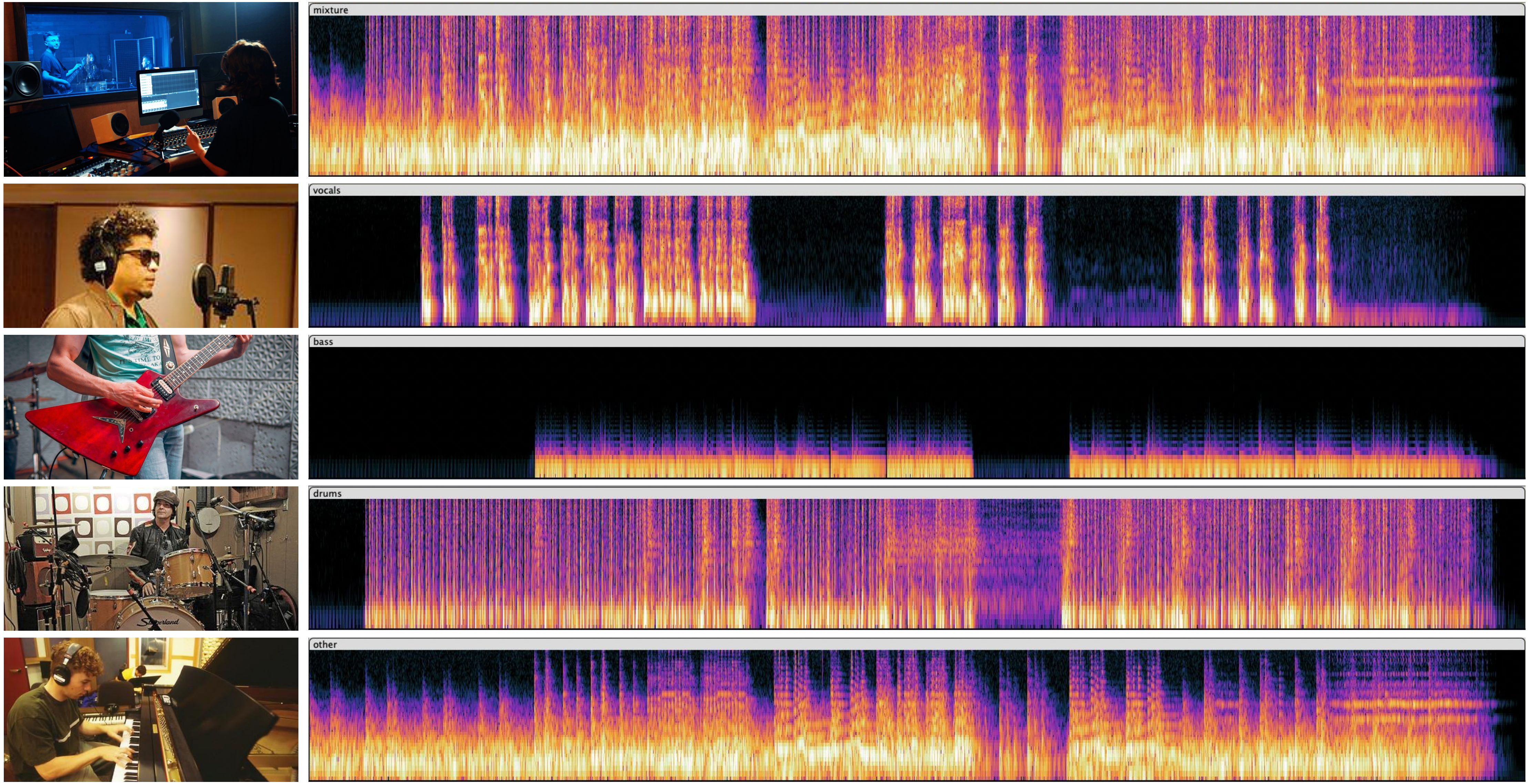
- Task of separating mixed audio into its source components:
  - Vocals, base, drums, &c

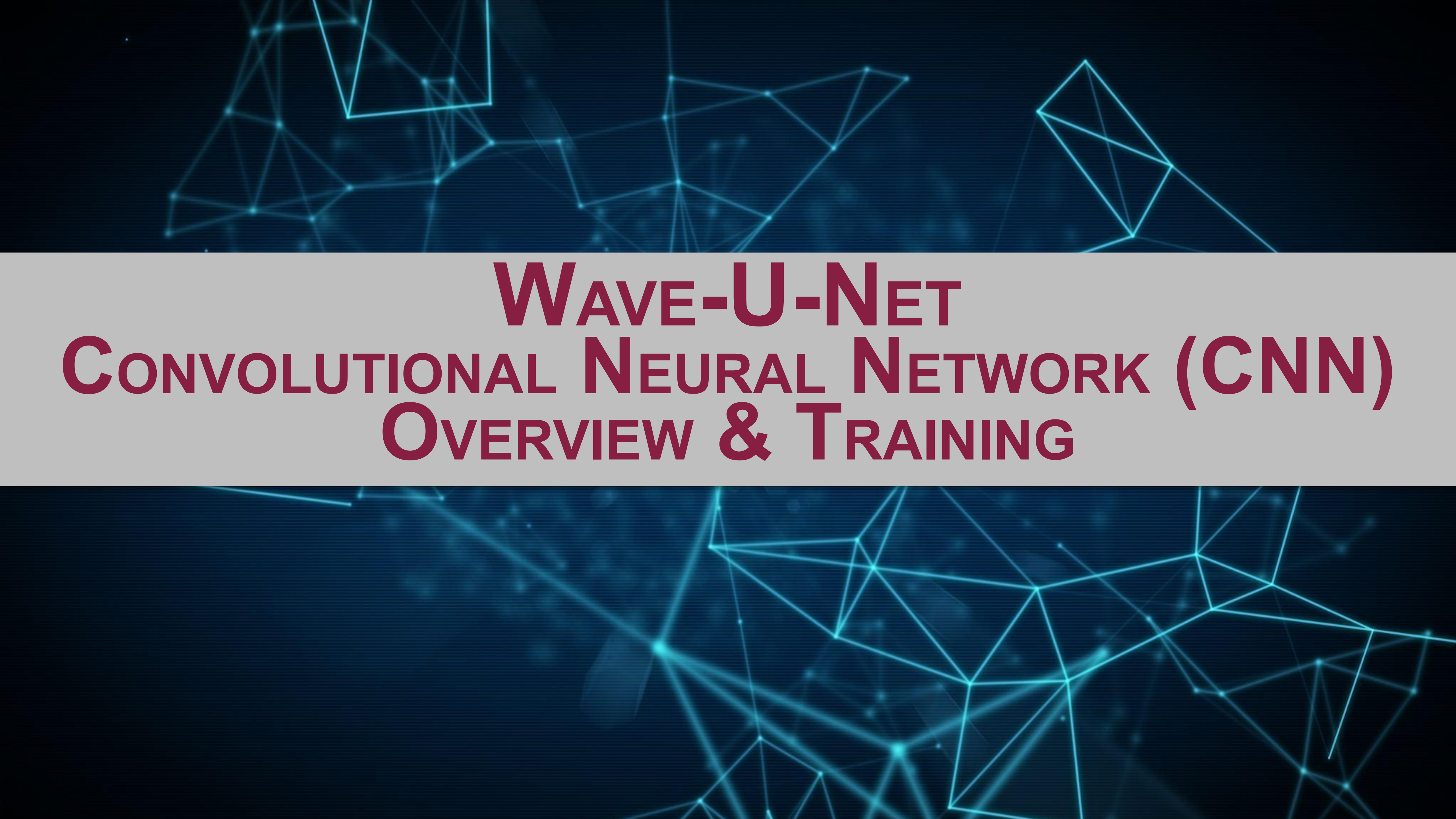
- Applications can include:
  - Performance extraction
  - Vocal Enhancement
  - Post-production, remixing, & 3D up-mixing
  - Hearing aids

Audio Mixing



Audio Separation Challenges

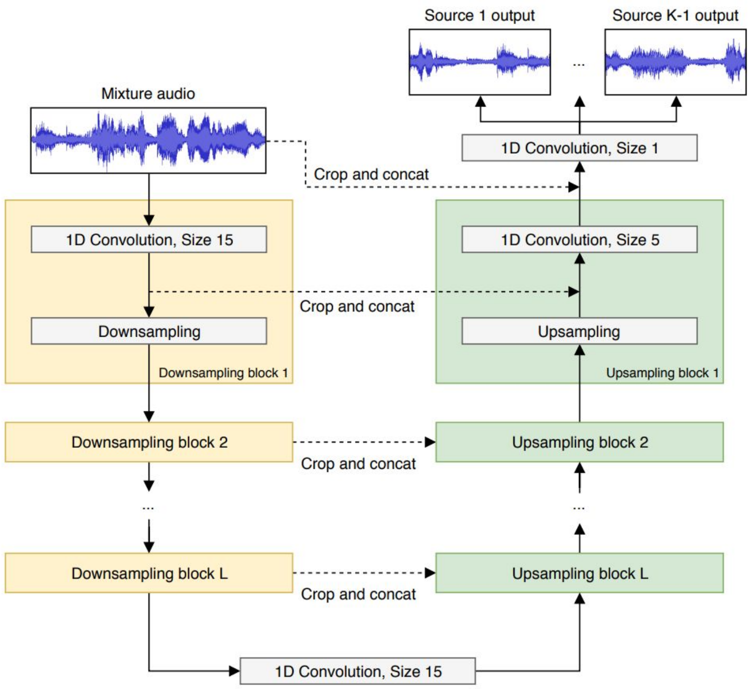




Wave-U-Net
Convolutional Neural Network (CNN)
Overview & Training

Wave-U-Net – Architecture Review

- Downsampling block structure
  - Convolution layer
  - Downsampling layer
- Upsampling block structure
  - Convolution layer
  - Upsampling layer
- Comprehension of many temporal contexts
- Improves upon artifacting issues in previous networks



From https://arxiv.org/abs/1806.03185

Wave-U-Net – Training attempts

- We wished to train our own Wave-U-Net model to compare results with pre-trained separation models
- Encountered challenges
  - Dependency issues
  - Exploding gradients
  - Insufficient VRAM
  - Slow CPU train times
- Unable to train a model fully
- Multiple attempts to resolve issue
  - Reduced learning rate
  - Increased batch size
  - Increased number of upsample/downsample blocks
  - Could not run identical architecture to paper

Wave-U-Net - Conclusions

- Unable to train model due to hardware limitations
- Not all bad news - repository provides pre-trained models
- Experiments in next section
- Will make further attempts to train before final report due date
  - Amazon S3?

Audio Separation Results

Audio Separation - Results

- Successfully achieved multi-instrument separation, but was not perfect
  - Did face challenges separating trumpet audio in from human voice in some music
  - Faced challenges separating a group singing in harmony from other instrumentals
- Song in Training Set
  - Full 🔊
  - Drums 🔊
  - Vocals 🔊
  - Bass 🔊
- Song Outside of Training Set
  - Full 🔊
  - Drums 🔊
  - Vocals 🔊
  - Bass 🔊

Lessons Learned

```
peration == "MIRROR_X":
irror_mod.use_x = True
irror_mod.use_y = False
irror_mod.use_z = False
_operation == "MIRROR_Y":
irror_mod.use_x = False
irror_mod.use_y = True
irror_mod.use_z = False
_operation == "MIRROR_Z":
irror_mod.use
```

```
ntext.scene.objects.active
("Selected" + str(modifier
irror_ob.select = 0
= bpy.context.selected_obj
ata.objects[one.name].se
print("please select exactly
--- OPERATOR CLASSES ----
types.Operator):
X mirror to the selected
ject.mirror_mirror_x"
rror X"
```

Caption if desired

Source: TBD