Conference: IEEE ICASSP 2021

Artificially Synthesising Data for Audio Classification and Segmentation to Improve Speech and Music Detection in Radio Broadcast

Satvik Venkatesh[1,*], David Moffat[1], Alexis Kirke[1], Gözel Shakeri[2], Stephen Brewster[2], Jörg Fachner[3], Helen Odell-Miller[3], Alex Street[3], Nicolas Farina[4], Sube Banerjee[5], Eduardo Reck Miranda[1]

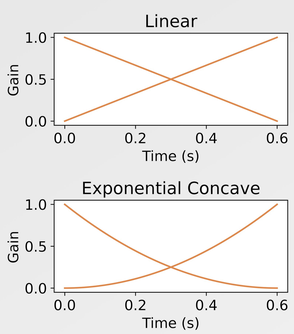
*satvik.venkatesh@plymouth.ac.uk

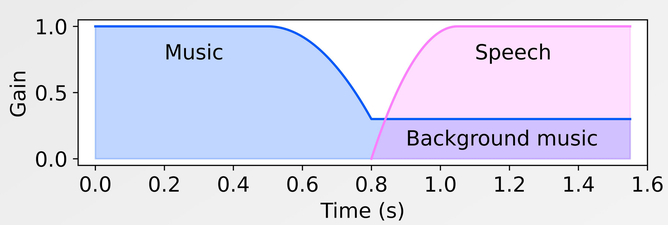
Introduction

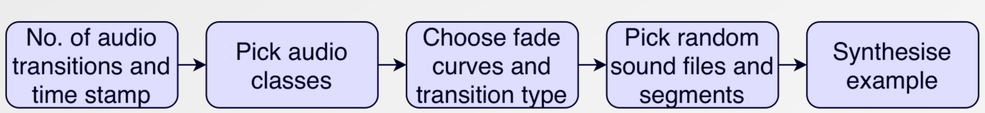
- Audio Segmentation divides an audio signal into homogeneous sections such as music and speech.
- Machine learning models are generally trained using proprietary audio, which cannot be shared. This hinders the reproducibility of research.

Training Set Synthesis

- We synthesise large-scale training sets for audio segmentation.
- Create realistic radio examples by incorporating different fade curves and audio ducking.

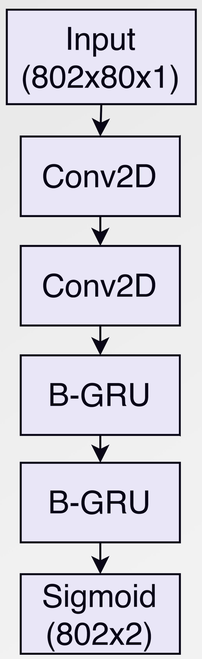






Deep Learning

- We designed a Convolutional Recurrent Neural Network (CRNN).
- We extract Mel Spectrograms as audio features and perform segmentation-by-classification.



Results

- In-house test set

Dataset	$F_{overall}$	F_{speech}	F_{music}
Without Synthesis	93.54	94.58	92.99
With Synthesis	**96.69**	**96.17**	**96.97**

- MIREX competition dataset

Algorithm	F_{music}	F_{speech}
Choi et al. [25]	49.36	77.18
Marolt [26]	38.99	91.15
Marolt [26]	54.78	90.9
Marolt [26]	31.24	90.86
Our model	**85.76**	**92.21**

Conclusion

- Used only synthetic data to train models.
- Obtained state-of-the-art performance for audio segmentation of music and speech.
- Significantly reduced the time and resources to label training sets for audio segmentation.

[1]Interdisciplinary Centre for Computer Music Research, University of Plymouth; [2]School of Computing Science, University of Glasgow; [3]Cambridge Institute of Music Therapy Research, Anglia Ruskin University, UK; [4]Centre for Dementia Studies, Brighton and Sussex Medical School, UK; [5]Faculty of Health, University of Plymouth, UK.

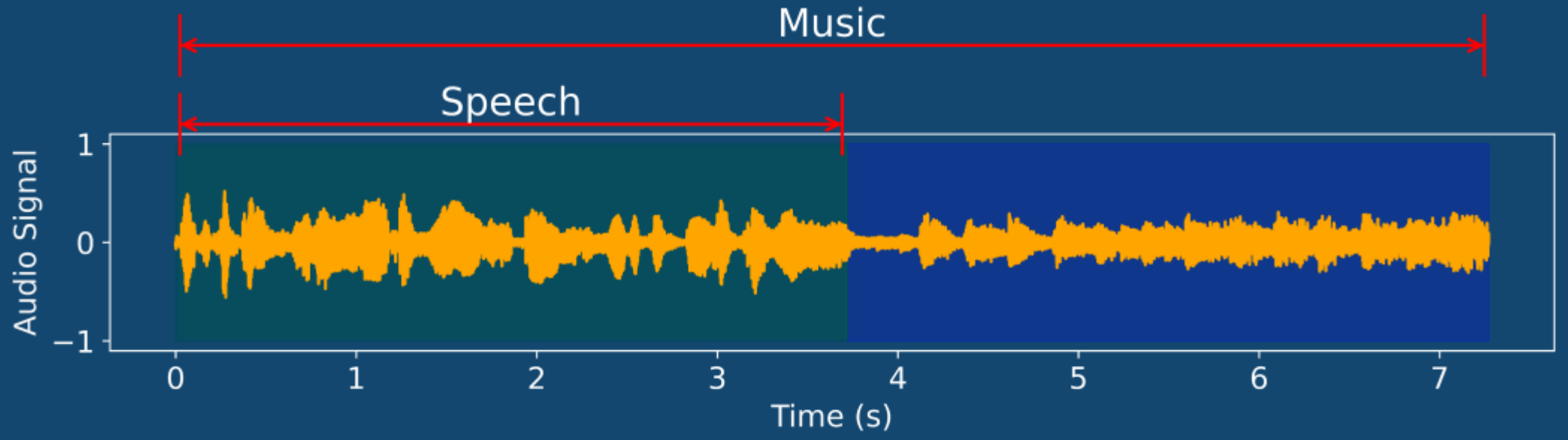
Research supported by Engineering and Physical Sciences Research Council (EPSRC) grants EP/S026991/1, EP/S027491/1, EP/S027203/1, and EP/S026959/1.

Conference: IEEE ICASSP 2021

Artificially Synthesising Data for Audio Classification and Segmentation to Improve Speech and Music Detection in Radio Broadcast

Satvik Venkatesh, David Moffat, Alexis Kirke, Gözel Shakeri, Stephen Brewster, Jörg Fachner, Helen Odell-Miller, Alex Street, Nicolas Farina, Sube Banerjee, Eduardo Reck Miranda

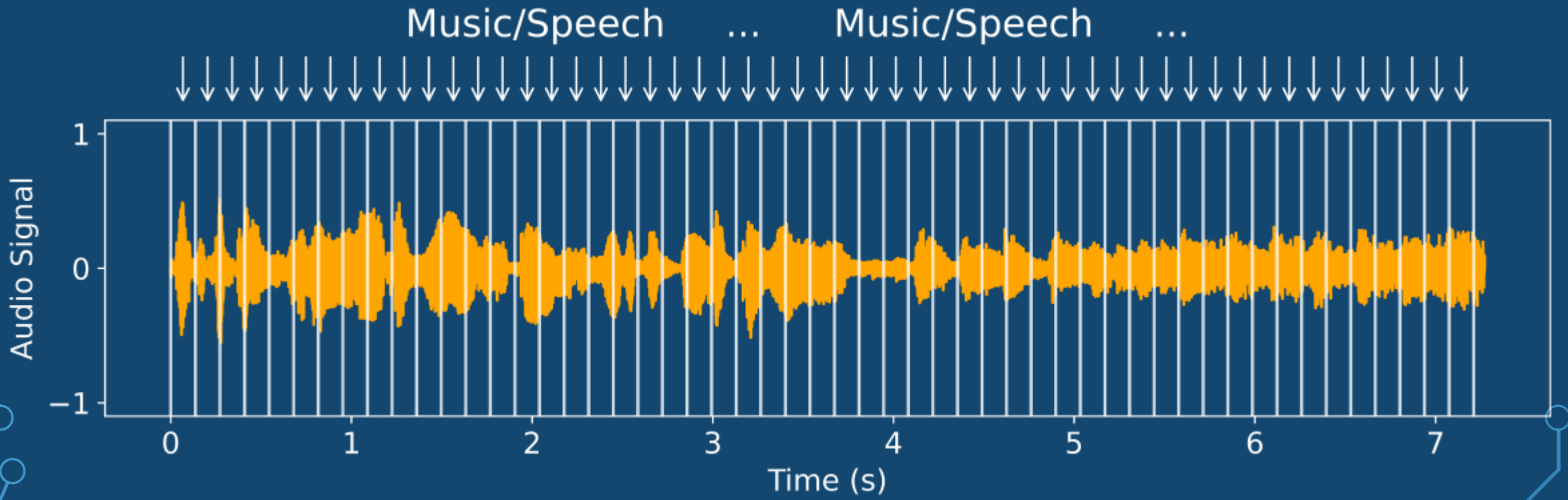
Audio Segmentation



0.0, 3.7, Speech

0.0, 7.2, Music

Why Audio Segmentation?



Segmentation by classification

Challenges

- Labelled radio recordings cannot be shared.
- Open datasets
  - MuSpeak dataset (MIREX, 2018)

Proposed Study

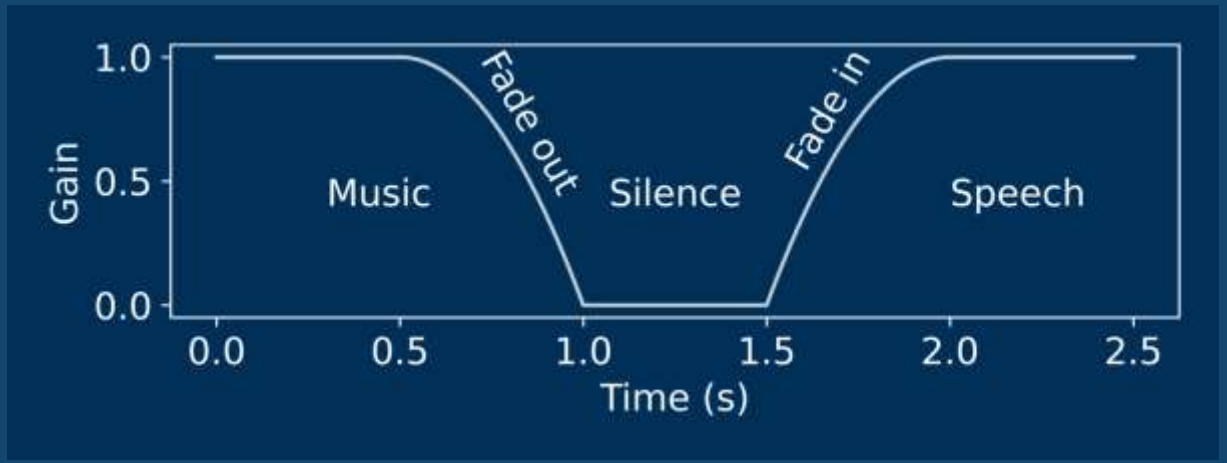
- Use open datasets with separate files of music and speech
- Artificially synthesise radio-like examples
- Fade curves and audio ducking

Data Repository

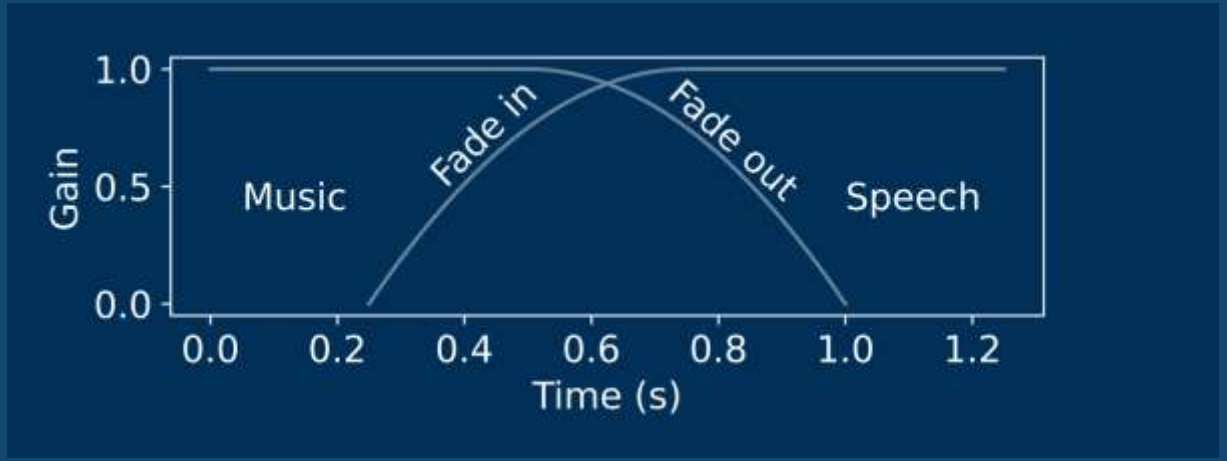
- MUSAN corpus
- GTZAN genre recognition
- Singing Voice Audio Dataset
- LibriSpeech corpus

Types of Transitions

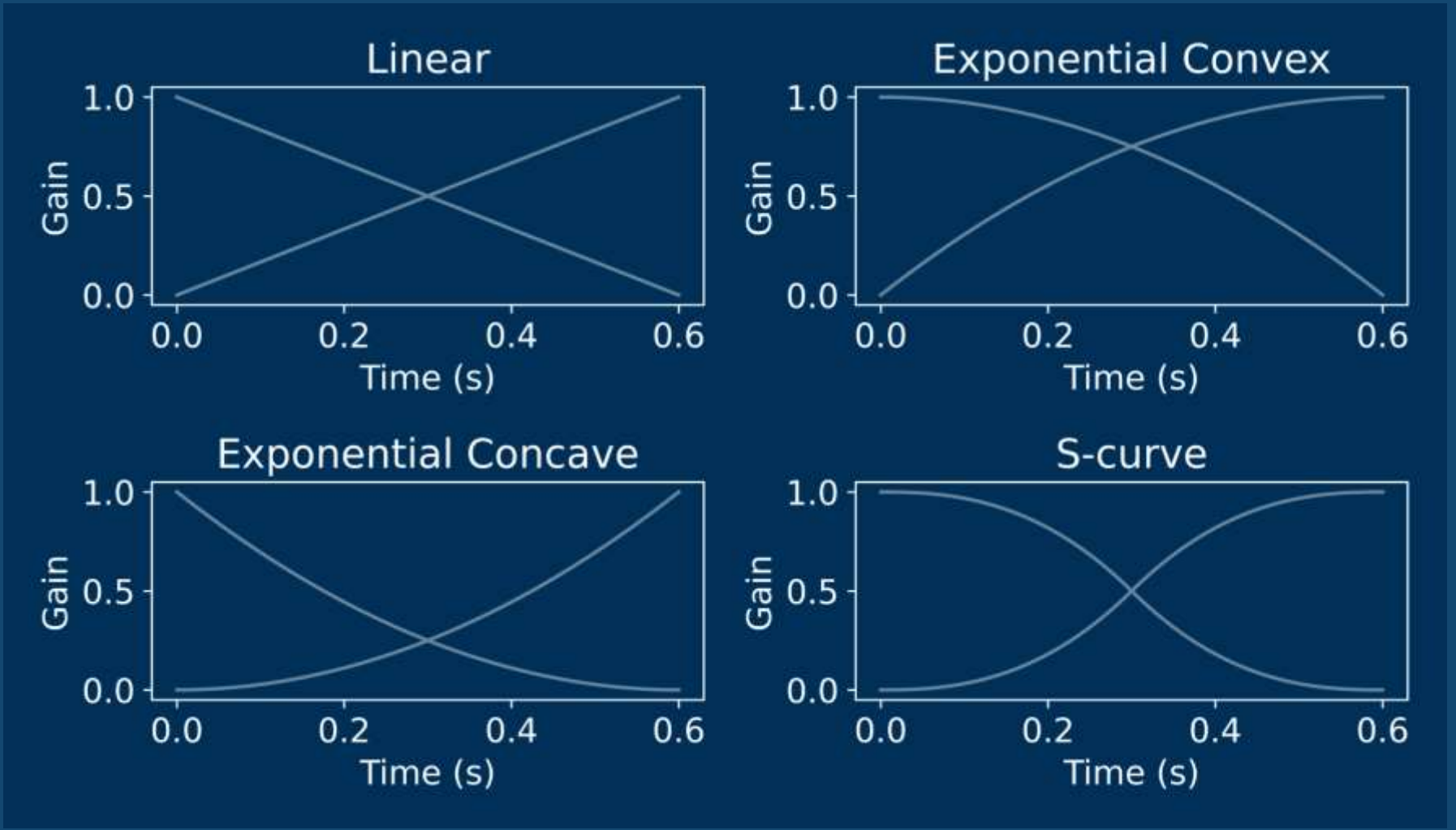
(1) Normal fade:



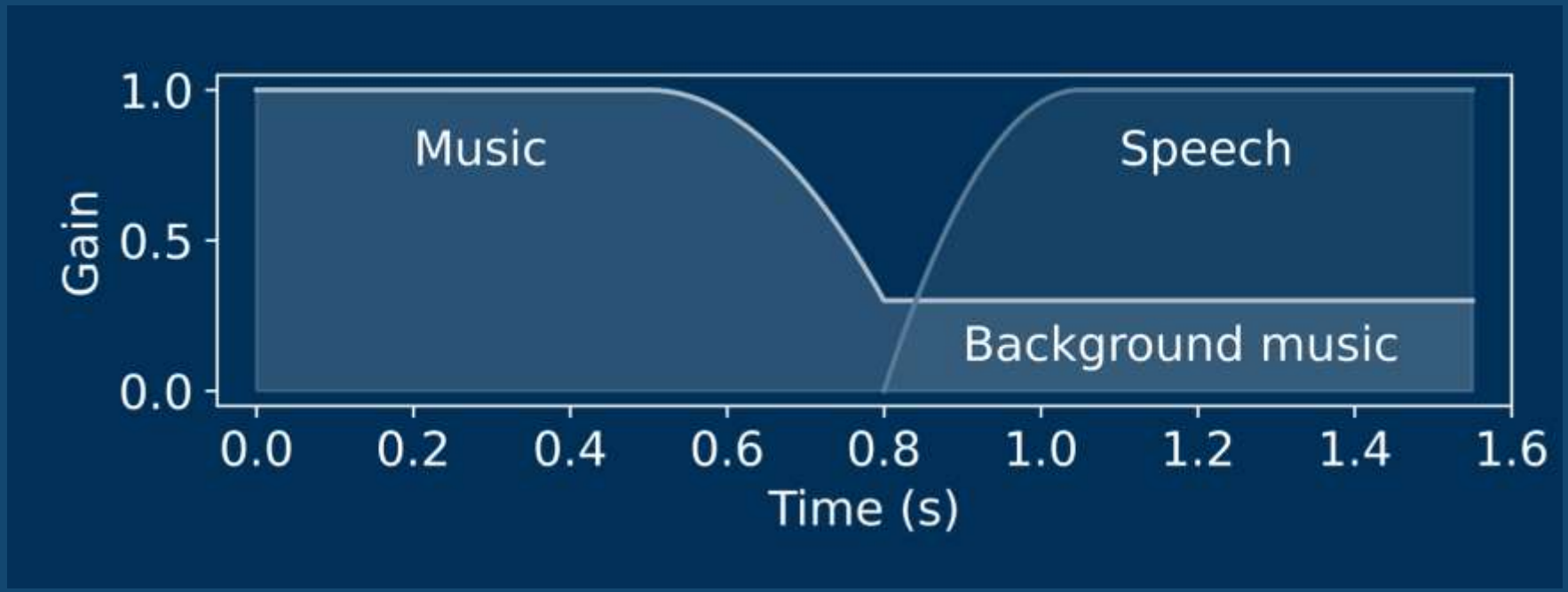
(2) Cross-fade:



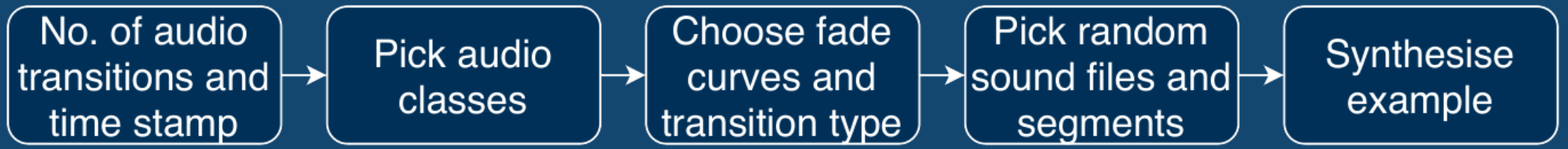
Fade Curves


Linear
Exponential Convex
Exponential Concave
S-curve
Gain
1.0
0.5
0.0
0.0
0.2
0.4
0.6
Time (s)

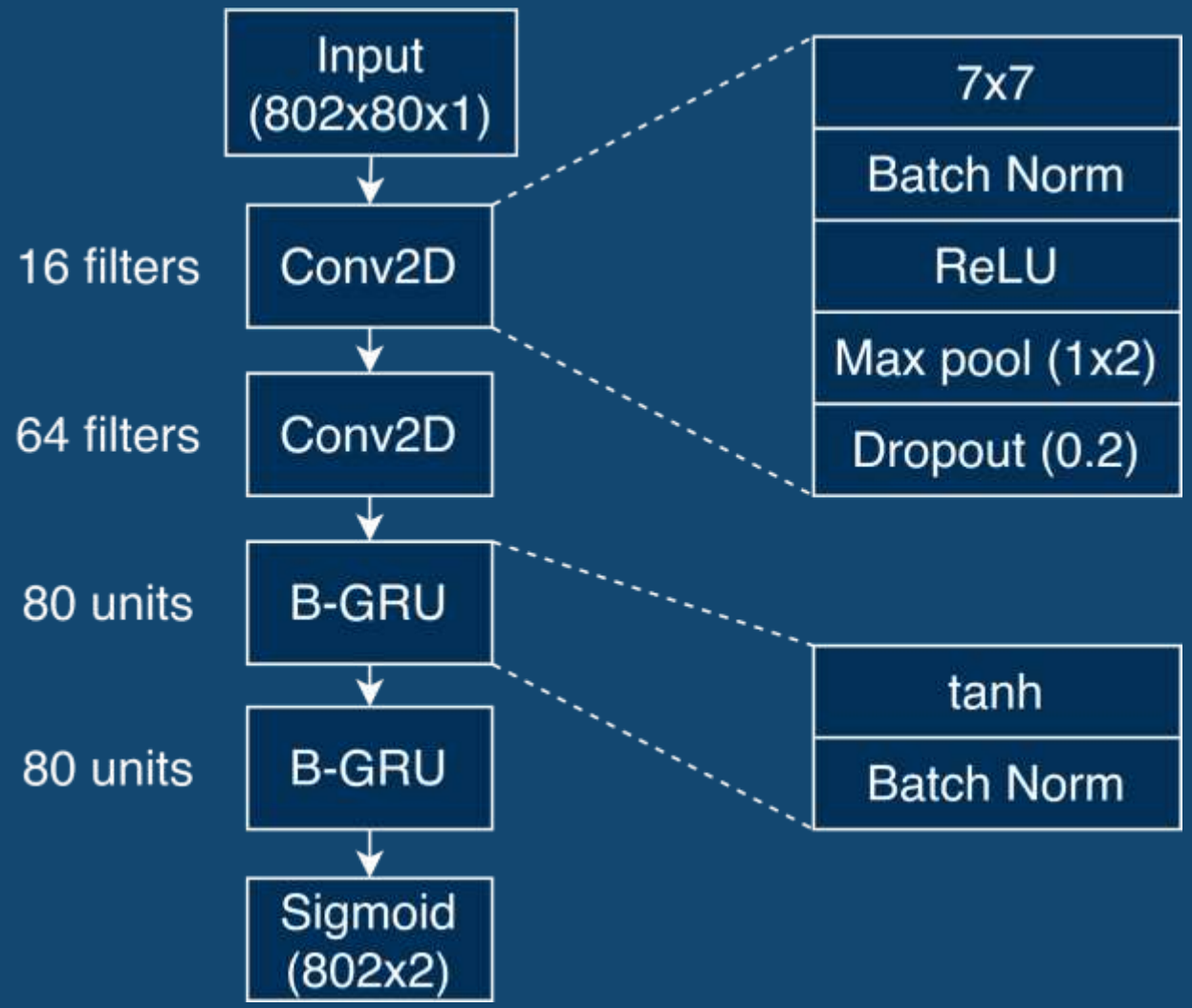
Background Music



Overview


No. of audio transitions and time stamp
Pick audio classes
Choose fade curves and transition type
Pick random sound files and segments
Synthesise example

Architecture



Validation and Test Sets

- MuSpeak dataset: approx. 5 hours
- BBC Radio Devon: 9 hours

Training datasets

1. Dataset-only files (d-OF)
2. Dataset-only files and background music (d-OFB)
3. Dataset-no normalisation (d-NN)
4. Dataset-data synthesis (d-DS)

Results

Dataset	$F_{overall}$	F_{speech}	F_{music}
d-OF	93.54	94.58	92.99
d-OFB	93.68	94.95	92.99
d-NN	95.33	**96.44**	94.73
d-DS	**96.69**	96.17	**96.97**

Results on MIREX dataset

Algorithm	F_{music}	F_{speech}
Choi et al. [25]	49.36	77.18
Marolt [26]	38.99	91.15
Marolt [26]	54.78	90.9
Marolt [26]	31.24	90.86
Our model	**85.76**	**92.21**

Example detection

- https://github.com/satvik-venkatesh/audio-seg-data-synth

```
0.0, 19.01, speech
16.37, 50.84, music
36.1, 38.31, speech
46.8, 87.71, speech
74.74, 87.71, music
88.77, 191.92, speech
192.49, 237.0, speech
237.64, 265.79, speech
265.72, 273.35, music
268.87, 318.72, speech
276.23, 305.33, music
306.16, 567.3, music
323.63, 341.5, speech
560.53, 645.04, speech
```

```
644.6, 857.02, music
856.06, 860.53, speech
858.28, 1062.05, music
1054.39, 1088.3, speech
1089.07, 1161.2, speech
1142.3, 1351.85, music
1349.16, 1352.69, speech
1353.59, 1610.56, music
1607.25, 1667.96, speech
1663.71, 1925.69, music
1670.01, 1676.72, speech
1906.89, 1935.66, speech
1927.22, 1935.96, music
1936.61, 1945.1, speech
```

...

Conclusion

- Used only synthetic data to train models.
- Obtained state-of-the-art performance for audio segmentation of music and speech.
- Significantly reduced the time and resources to label training sets for audio segmentation.

Thank you!

Email: satvik.venkatesh@plymouth.ac.uk

Twitter: @SatvikVenkatesh