IEEE Workshop on Applications of Signal Processing to Audio and Acoustics • October 2021

# Filtered noise shaping for time domain room impulse response estimation from reverberant speech

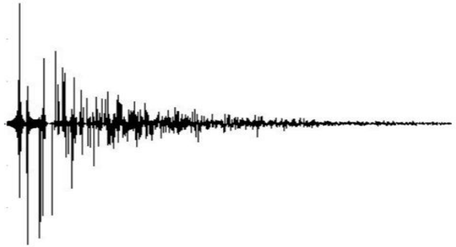

Christian J. Steinmetz
Centre for Digital Music, Queen Mary University of London
c.j.steinmetz@qmul.ac.uk

Paul Calamia
Facebook Reality Labs Research

Vamsi Krishna Ithapu
Facebook Reality Labs Research


Queen Mary University of London

FACEBOOK REALITY LABS

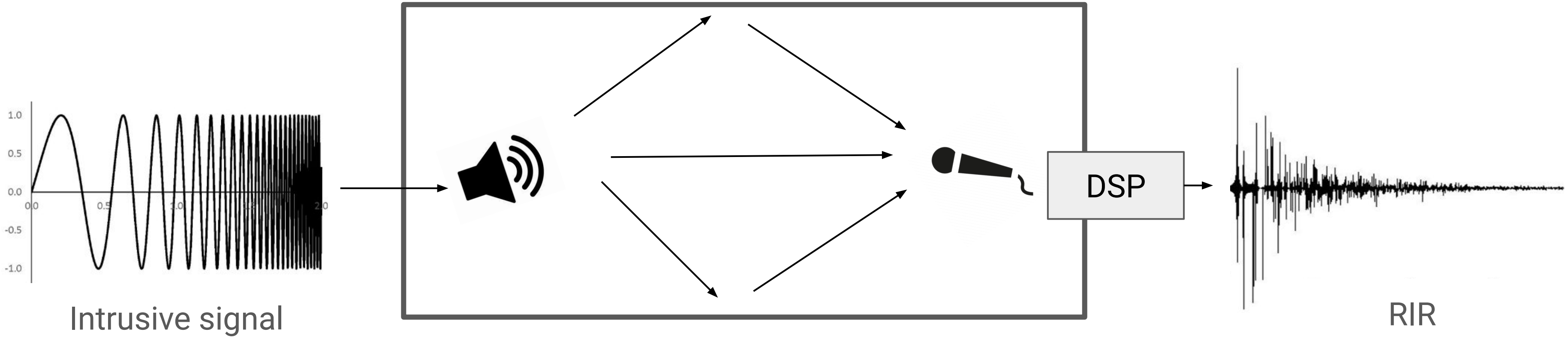


The room impulse response (RIR) has many applications

- Informing dereverberation and speech recognition algorithms
- Room acoustics analysis
- Virtual sound sources for VR/AR

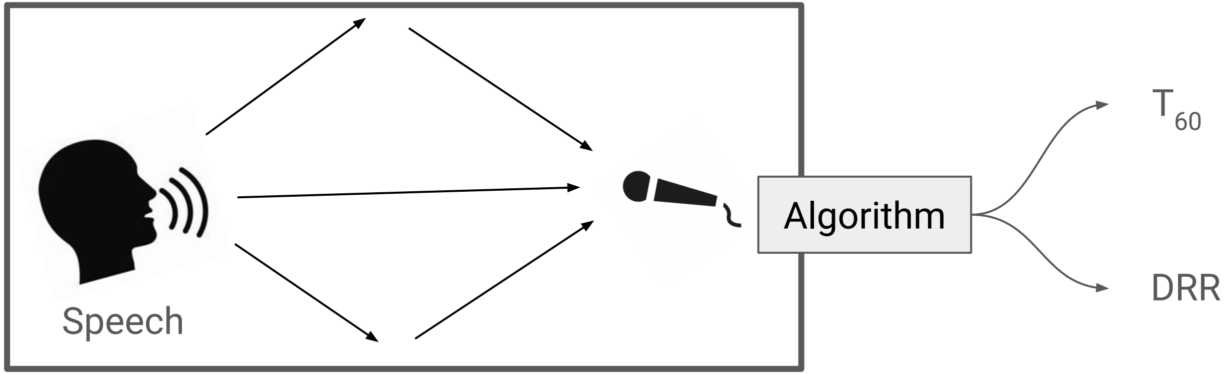
...but measuring the RIR can be difficult.

# Measured room impulse response




- Uses an intrusive test signal
- Require low noise floor in the environment
- High fidelity transducer and microphone

# Blind estimation of room characteristics



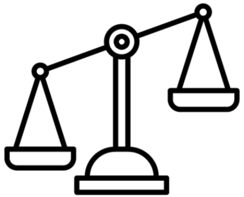

- Uses an unobtrusive test signal
- More robust to external noise
- Use consumer grade microphones

...but T60 and DRR alone do not fully characterize the room.

# Recent deep learning approaches

**a) Estimate parameters of artificial reverberators**

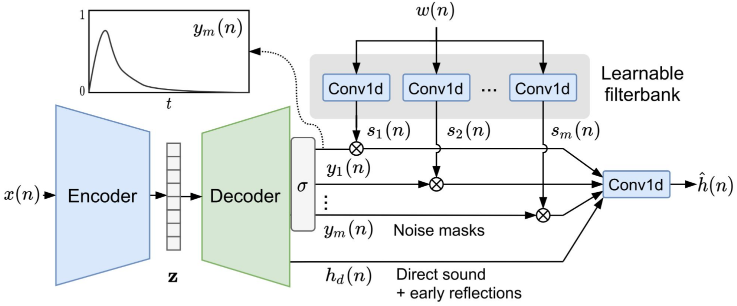
- May not generalize to real rooms
- Highly dependant on quality of artificial reverberation algo.



**b) End-to-end neural network processes audio signals**

- Requires significant compute
- Potential to add artifacts

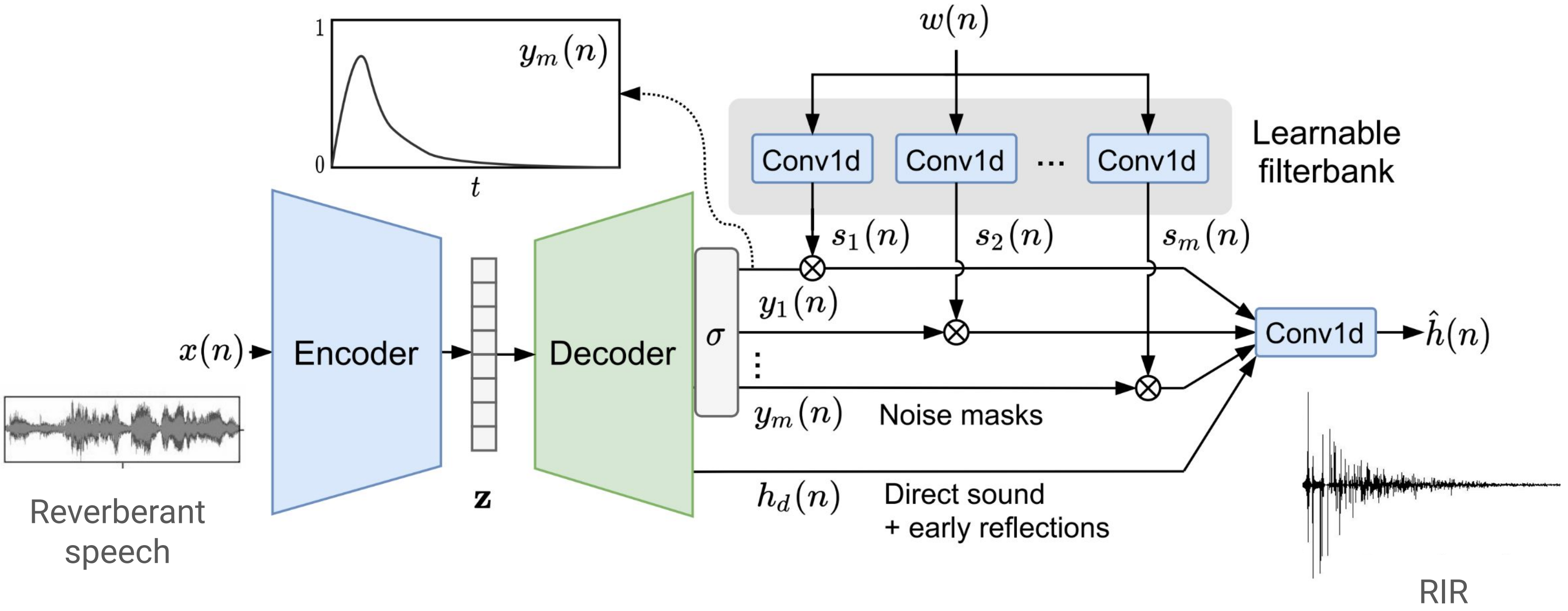
*Balance these approaches by estimating the RIR directly and perform convolution*




# **FiNS**: **Fi**ltered **N**oise **S**haping network

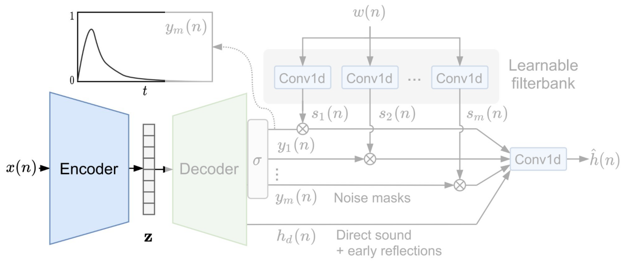
## reconstructs RIRs from reverberant speech

- Analyze reverberant speech to estimate time domain RIR
- Model RIR as sum of decaying filtered noise signals
- Operate at 48 kHz for use in high fidelity audio processing
- Outperforms DL based approaches in listening test

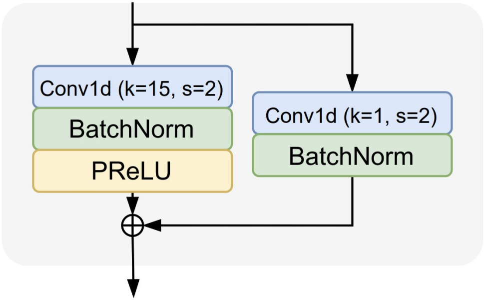


$y_m(n)$
1
0
$t$
$w(n)$
Learnable filterbank
Conv1d
Conv1d
Conv1d
$s_1(n)$
$s_2(n)$
$s_m(n)$
$x(n)$
Encoder
$\mathbf{z}$
Decoder
$\sigma$
$y_1(n)$
$y_m(n)$
Noise masks
$h_d(n)$
Direct sound
+ early reflections
Conv1d
$\hat{h}(n)$
Reverberant speech
RIR

# Encoder



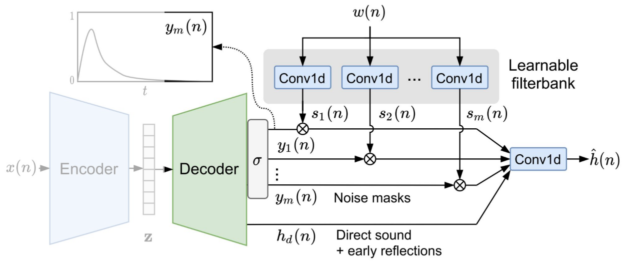

- Time domain encoder
- 13 layers of Conv1d residual blocks
- Strided convolutions downsample signal
- Produces 128 dim embedding
- Receptive field ~2.4 seconds @ 48 kHz

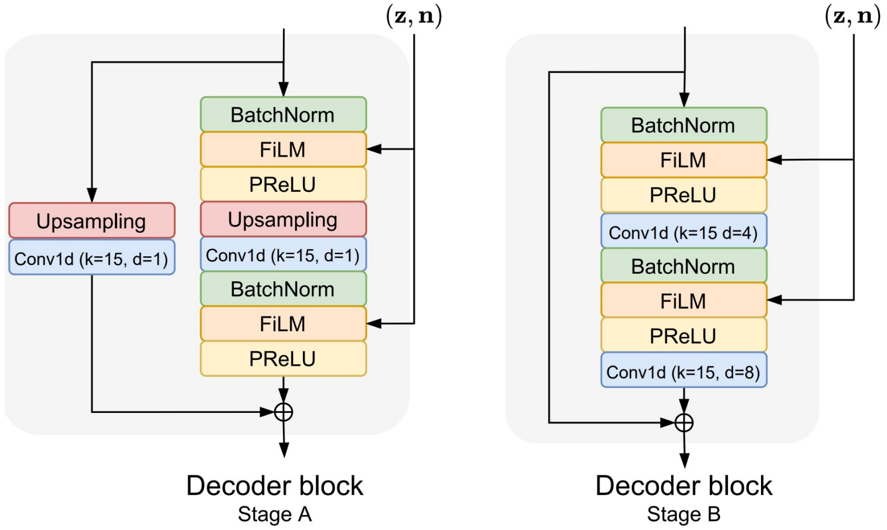



Encoder block

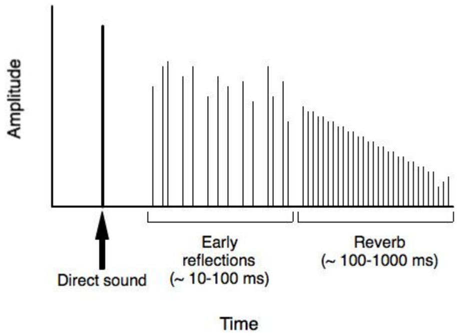
# Decoder

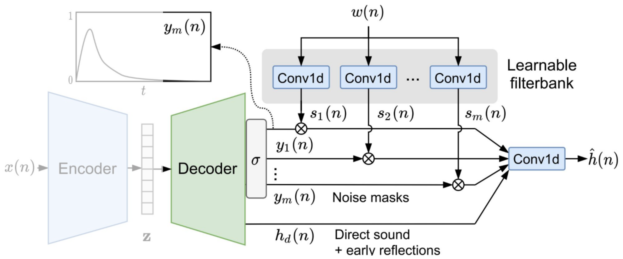
- Upsample latent (z) to produce RIR
- Design based on decoder of GAN-TTS
- Use feature-wise linear modulation (FiLM) to inject latent and noise at each block

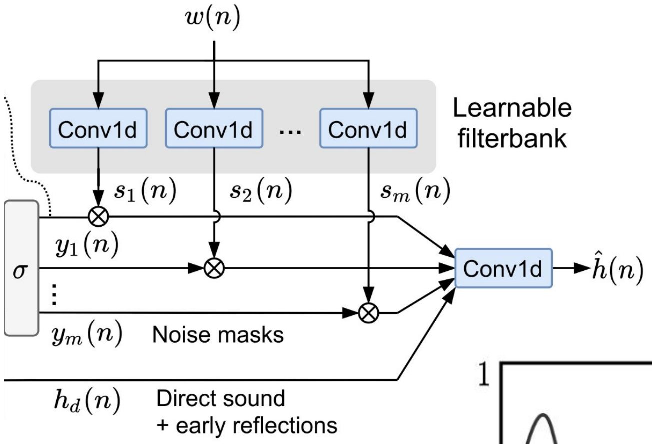


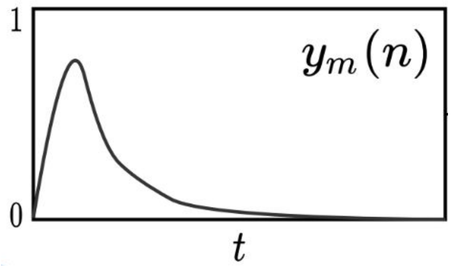



# Decoder (Noise shaping)



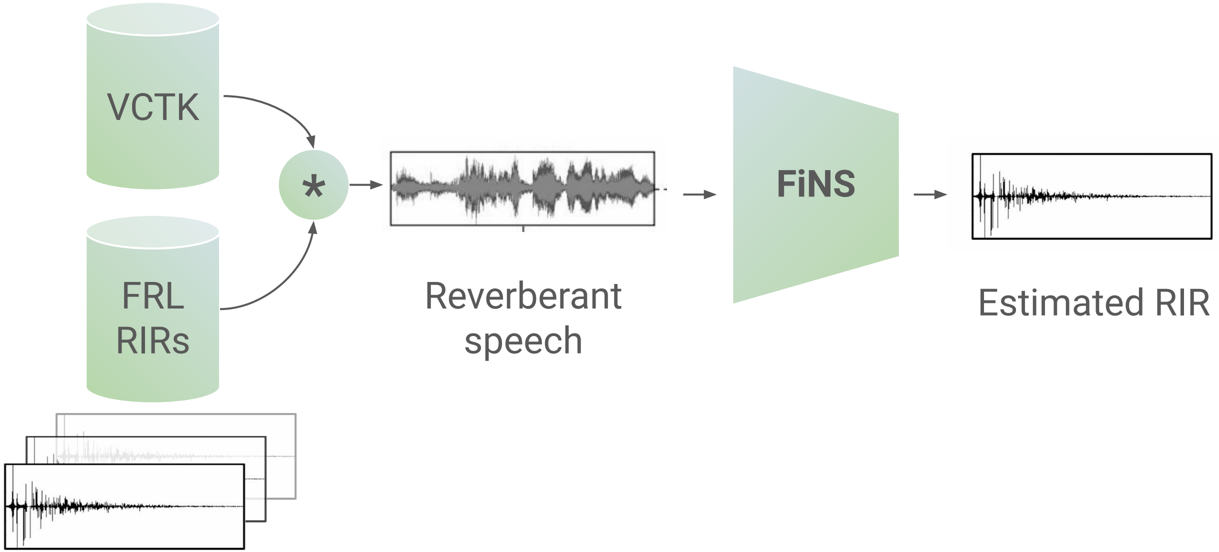




- Model the RIR in two parts:
    - **Late reverberation** generated with a sum of filtered noise signals
    - **Direct and early parts** estimated directly in the time domain

*Enable generative model without adversarial training*

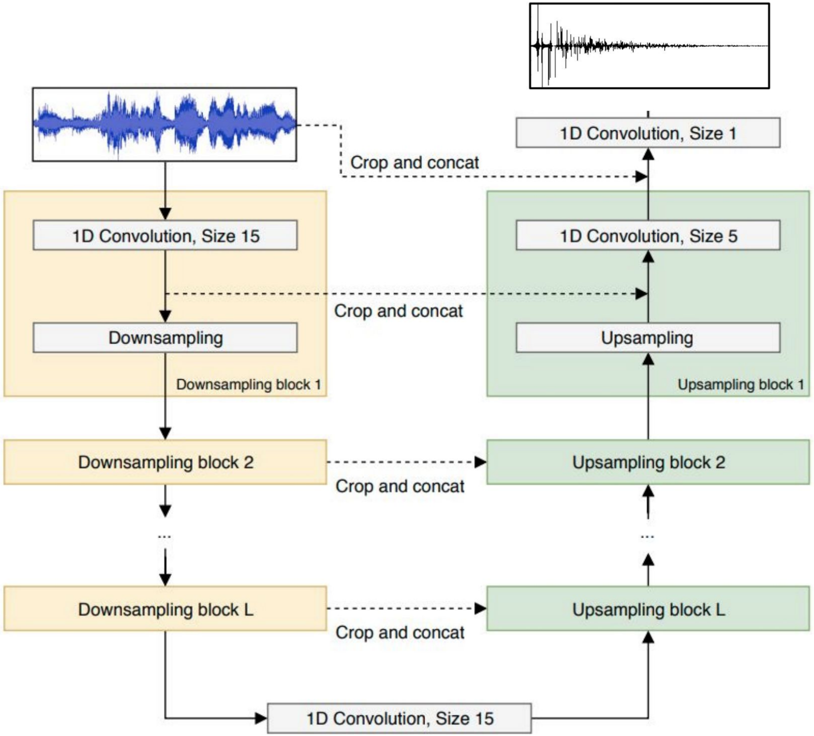
# Data generation



# Baselines

## Wave-U-Net

- Adapt model for source separation for estimation of RIR
- No inductive bias for the task of estimating RIRs
- Train using MRSTFT loss

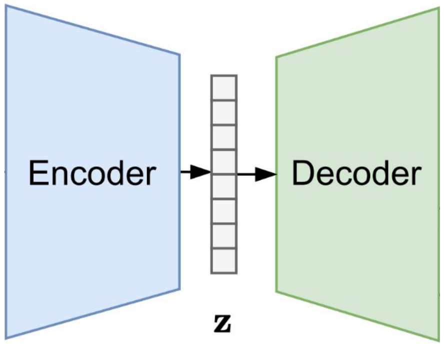


# Baselines

## FiNS Direct (D)

*Does filtered noise shaping aid in RIR estimation?*

- Use same encoder and decoder, except the decoder directly estimates the time domain RIR
- Conceptually similar to Wave-U-Net except without skip connections

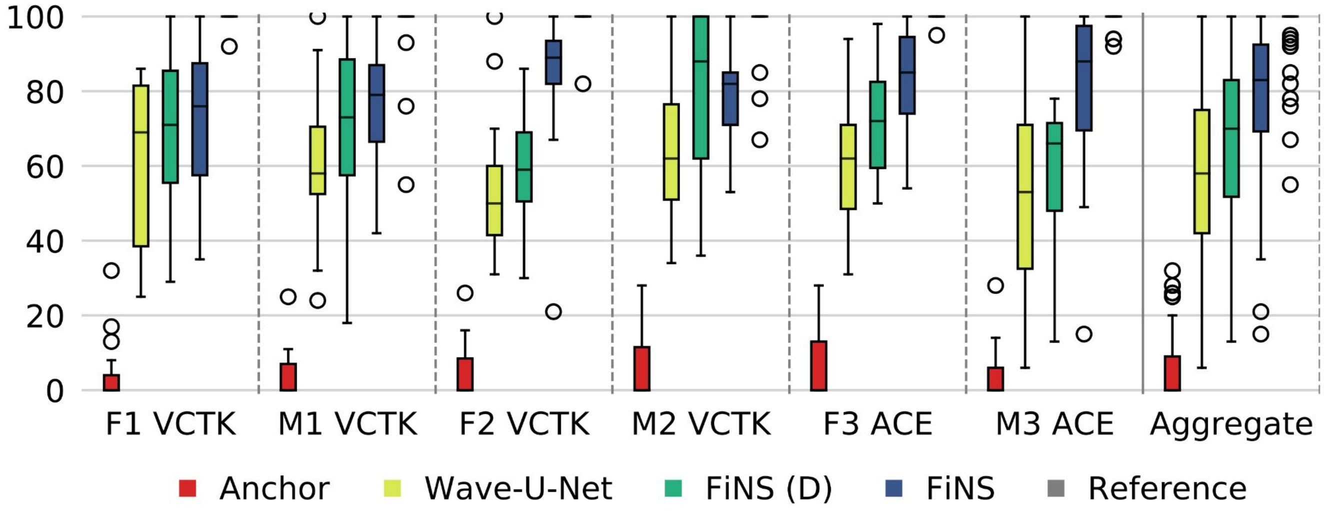


# Objective results

| RIR | Speech | Model | $\mathcal{L}_{\text{STFT}} \downarrow$ | $T_{60}$ | | | DRR | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Bias $\lvert \downarrow \rvert$ | MSE (s) $\downarrow$ | $\rho \uparrow$ | Bias $\lvert \downarrow \rvert$ | MSE (dB) $\downarrow$ | $\rho \uparrow$ |
| FRL | VCTK | Wave-U-Net | 1.127 | -0.016 | 0.005 | 0.480 | -0.25 | 4.19 | 0.736 |
| | | FiNS (D) | **1.064** | -0.001 | 0.004 | 0.548 | 0.54 | 4.13 | 0.734 |
| | | FiNS | 1.157 | 0.041 | 0.005 | 0.646 | 0.43 | 4.45 | 0.721 |
| FRL | ACE | Wave-U-Net | **1.119** | 0.006 | 0.004 | 0.495 | -0.58 | 5.55 | 0.625 |
| | | FiNS (D) | 1.137 | 0.034 | 0.006 | 0.479 | 0.50 | 5.14 | 0.661 |
| | | FiNS | 1.183 | 0.057 | 0.008 | 0.540 | 0.50 | 6.29 | 0.625 |

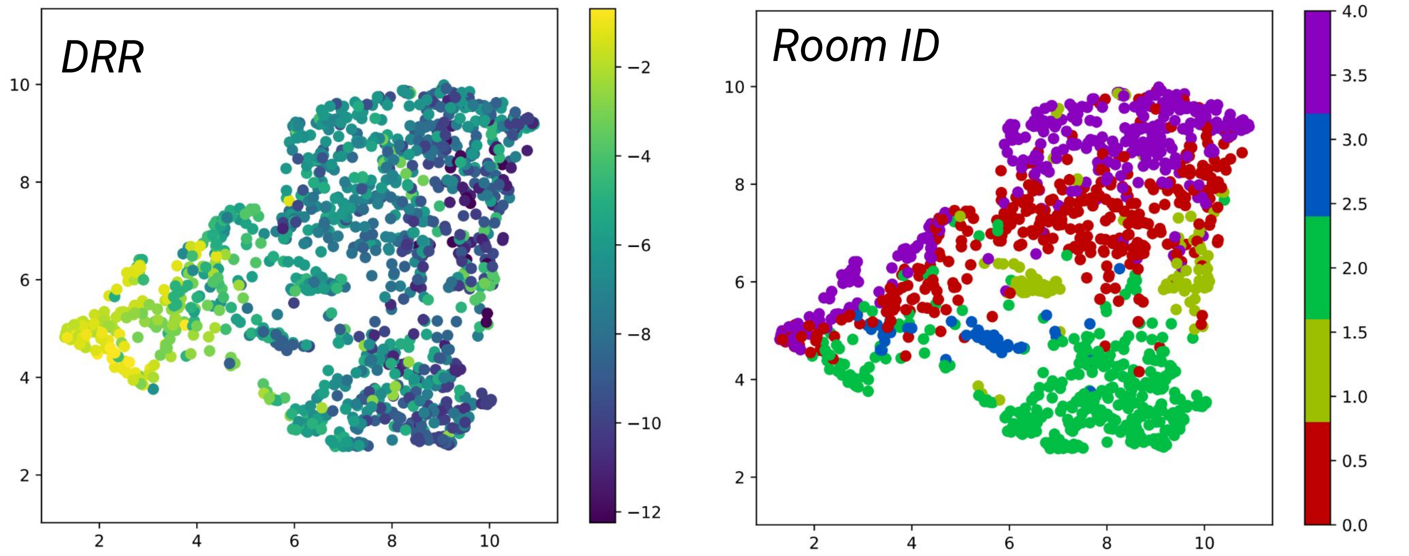
- All models are capable of estimating RIRs with accurate $T_{60}$ and DRR
- Generalizes to unseen speech from VCTK and ACE datasets
- **Listening indicates FiNS (D) and Wave-U-Net produce ringing artifacts**

*“Listeners rated RIRs produced by FiNS the most similar to the reference, yet they could still differentiate among them.”*




MUSHRA design with 15 listeners

# Encoder implicitly captures room characteristics

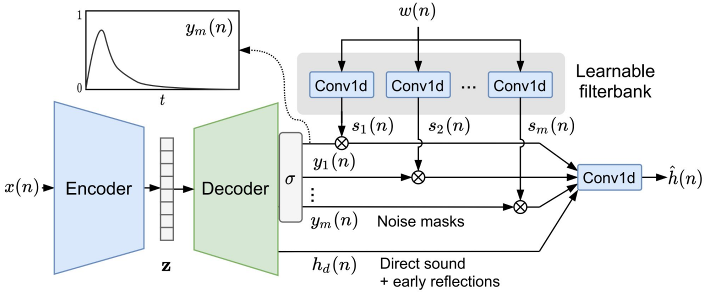



2D UMAP projections of 128 dim encoder embeddings

| Utterance | Clean | Reference | Anchor | Wave-U-Net | FiNS (D) | FiNS |
|---|---|---|---|---|---|---|
| F1 VCTK Speech | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| | RIR | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| F2 VCTK Speech | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| | RIR | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| M1 VCTK Speech | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| | RIR | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| M2 VCTK Speech | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |
| | RIR | 0:00 | 0:00 | 0:00 | 0:00 | 0:00 |

https://facebookresearch.github.io/FiNS


Christian J. Steinmetz
Centre for Digital Music, Queen Mary University of London
c.j.steinmetz@qmul.ac.uk


# **FiNS**: **Fi**ltered **N**oise **S**haping network

- Analyze reverberant speech to estimate time domain RIR
- Model RIR as sum of decaying filtered noise signals
- Operate at 48 kHz for use in high fidelity audio processing
- Outperforms other approaches in listening test