# Scaling Speech Enhancement in Unseen Environments with Noise Embeddings

Gil Keren[1], Jing Han[1], Björn Schuller[1,2]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

[2] GLAM – Group on Language, Audio & Music, Imperial College London, UK

# The challenge

Goal: speech denoising in unseen environments.

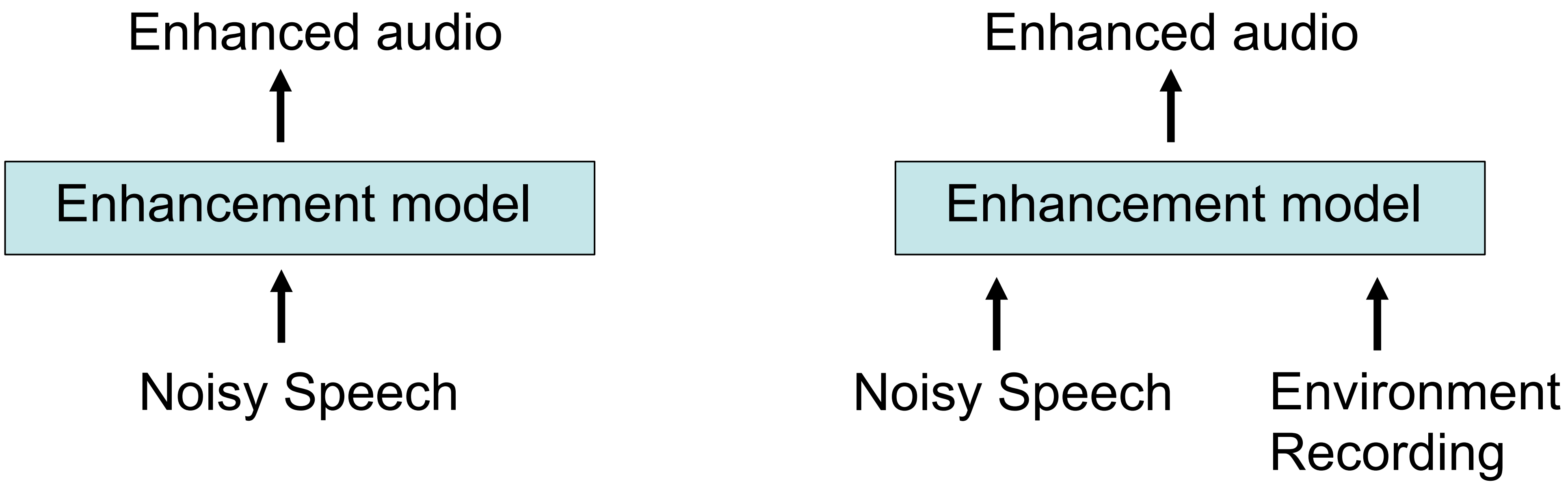
The real world contains a large variety of noises and environments. We cannot see all of them in training time.

# The challenge

- Every environment or sound may dictate different denoising “rules”.
- We need an adaptive model - a model that changes its enhancement behavior based on the environment.

# An Additional Input

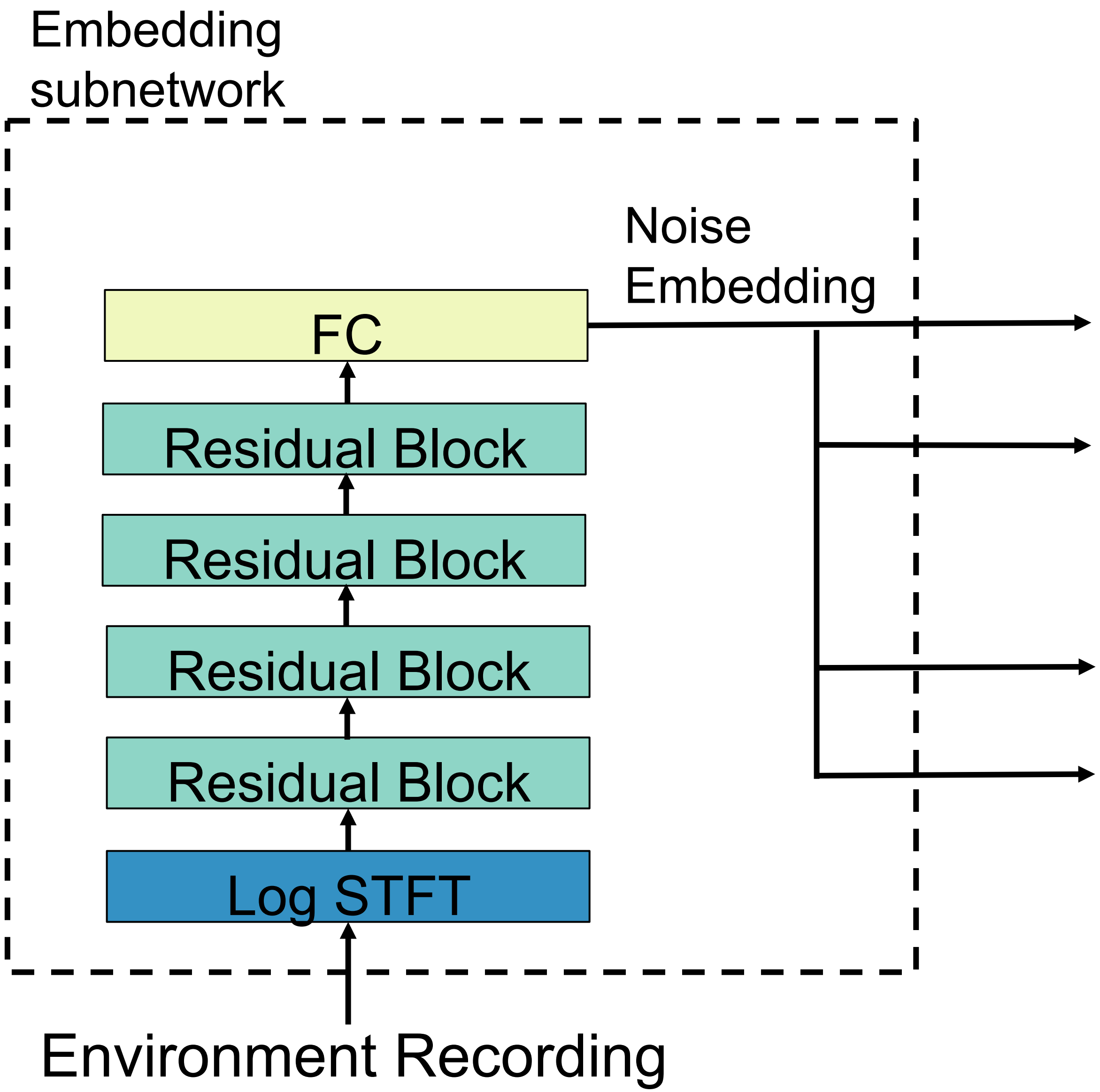
Additionally conditioning the model on a sample recording of the environment alone (no speech).



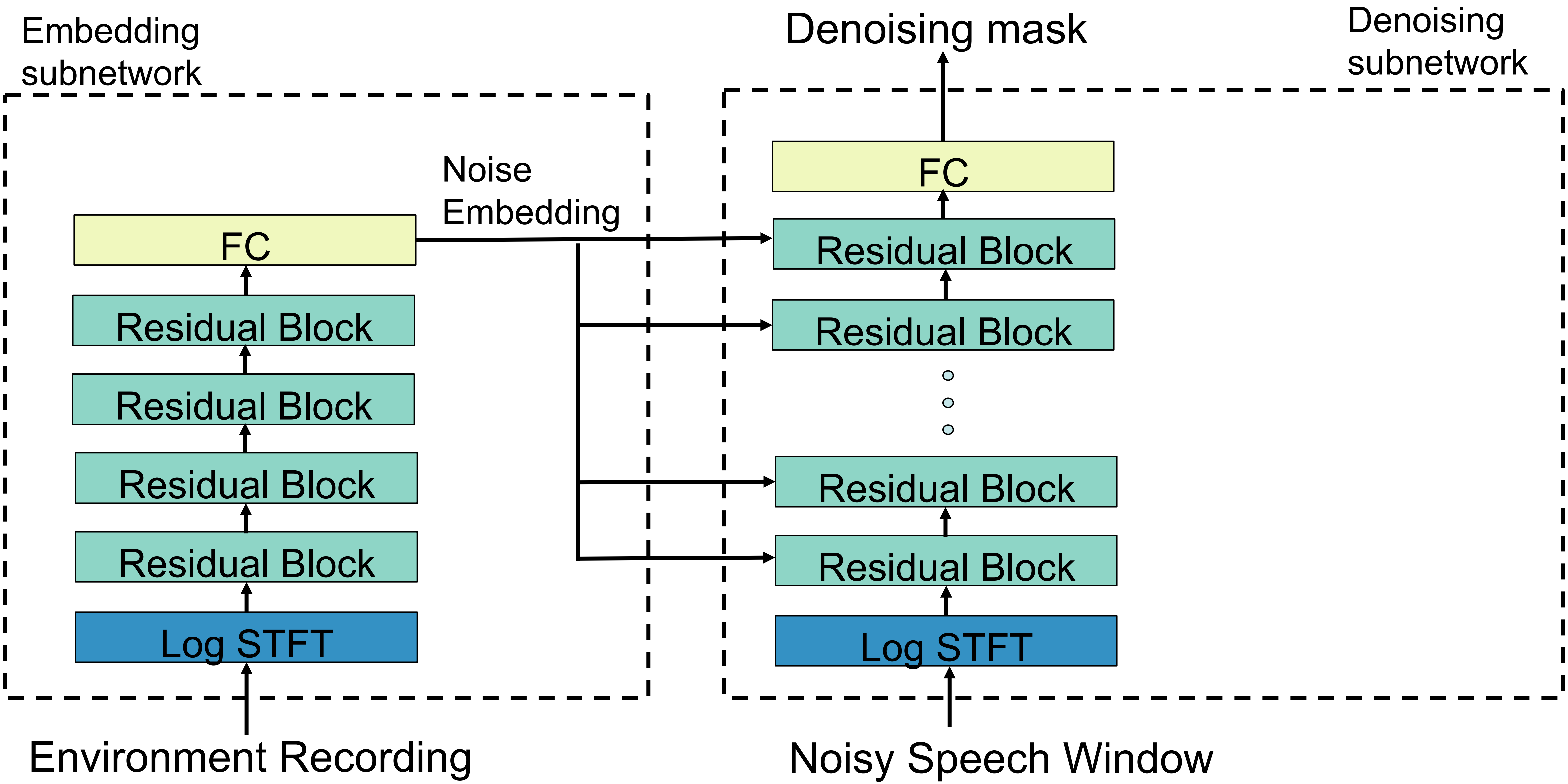
# An Additional Input

- A realistic setting: we can record a few seconds of the environment alone, before speech starts.
- Isolating the environment can help learning what frequency components to denoise.

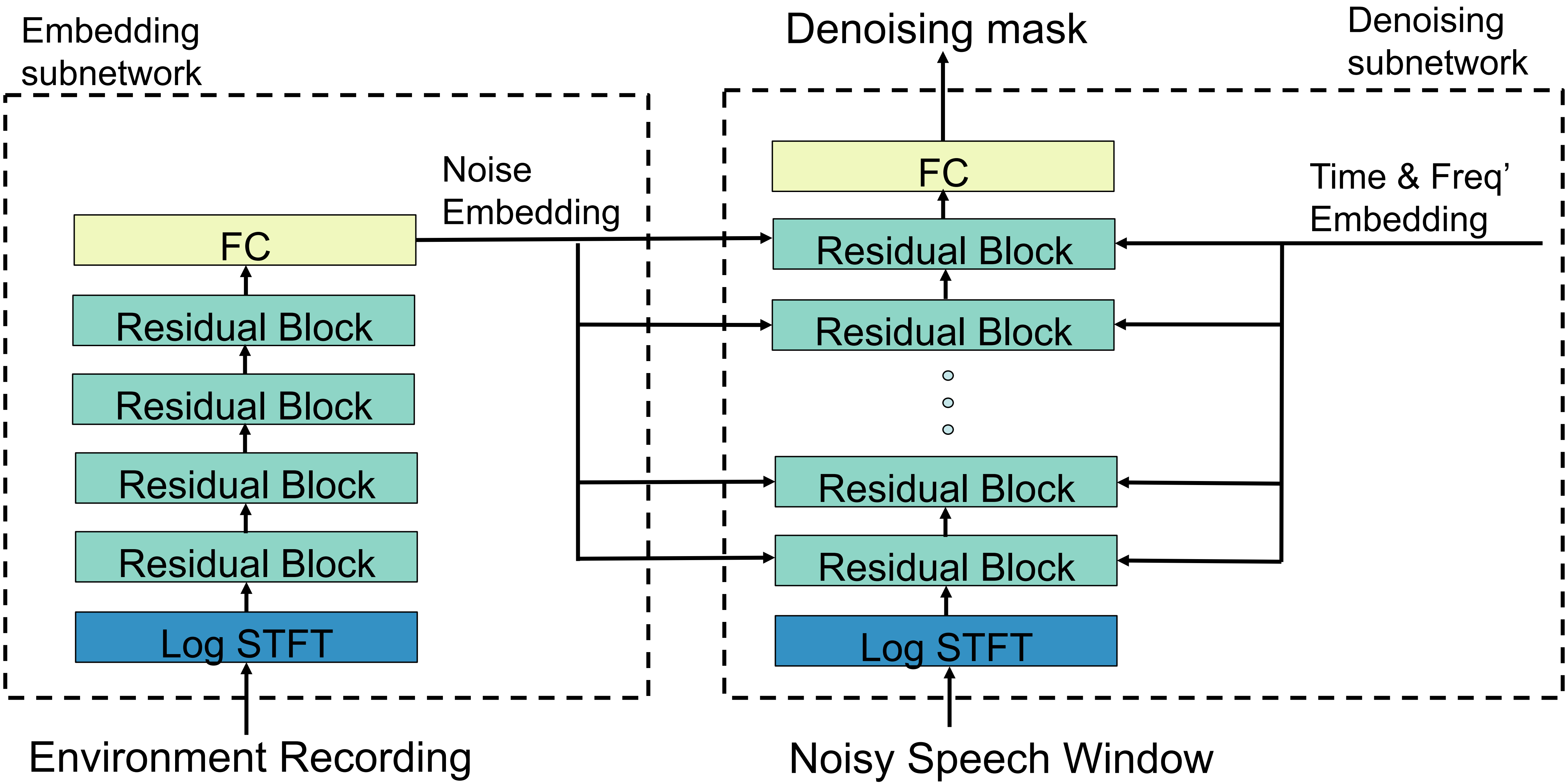
# Architecture



# Architecture



# Architecture



# Generalising to unseen environments

- To enhance well in unseen environments, we need to generalise to unseen points in the space of environments.

- We therefore consider a distribution over environments $p(e)$, where environments are the data points.

- Given a large training sample from $p(e)$, we may observe generalisation in the space of environments.

A recipe for generalizing to unseen categories (one-shot learning) [1]:

- Consider a distribution over categories $p(c)$.
- Design a model that is conditioned on raw representations of categories c, not their id.
- Train the model with a dataset containing a large training sample from $p(c)$.

[1] G. Keren, M. Schmitt, T. Kehrenberg, and B. Schuller, “Weakly supervised one-shot detection with attention Siamese networks,” arXiv preprint arXiv:1801.03329, 2018.

# Datasets

Universität
Augsburg
University

- Audio Set: 16,784 different training noise environments (656 for validation and 740 for test).

- Librispeech: 360 hours of clean speech (5.4 hours for validation/test).

- Random mixing at training time with 0dB-25dB.
- The model is unlikely to see many example twice.

# Evaluation Metrics

- Speech Recognition WER: Using a pretrained ‘Deep Speech’ system [1].

[1] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates et al., “Deep speech: Scaling up end-to-end speech recognition,” arXiv preprint arXiv:1412.5567, 2014.

# Results

Test set results for unseen environments, speakers and utterances.

| Method | WER [%] |
|---|---|
| Clean Speech | 4.21 |
| Noisy Speech | 34.04 |
| Log-MMSE [1] | 35.38 |
| Noise Aware [2] | 25.30 |
| No Embedding – 200 noises | 21.51 |
| No Embedding – 1000 noises | 20.54 |
| No Embedding – 16K noises | 16.78 |
| With Embedding | **15.46** |

[1] Ephraim & Malah, IEEE Trans. Acoustics, Speech, and Signal Processing, 1985 [2] Seltzer et al., ICASSP 2013

# Evaluation Metrics

- Perceptual Evaluation of Speech Quality (PESQ): industry standard for objective voice quality testing.

- Segmental Signal-to-Noise Ratio (SegSNR).

- Log-Spectral Distortion (LSD).

# Results

Test set results for unseen environments, speakers and utterances.

| Method | PESQ | SegSNR | LSD |
|---|---|---|---|
| Clean Speech | – | – | – |
| Noisy Speech | 2.59 | 7.02 | 0.94 |
| Log-MMSE [1] | 2.66 | 7.12 | 0.91 |
| Noise Aware [2] | 2.96 | 11.01 | 0.54 |
| No Embedding – 200 noises | 3.12 | 10.03 | 0.53 |
| No Embedding – 1000 noises | 3.13 | 10.00 | 0.52 |
| No Embedding – 16K noises | 3.25 | 11.71 | 0.48 |
| With Embedding | **3.30** | **12.99** | **0.45** |

[1] Ephraim & Malah, IEEE Trans. Acoustics, Speech, and Signal Processing, 1985 [2] Seltzer et al., ICASSP 2013

# Results

- A deep residual network performs better than an MLP.
- Scaling the number of training noise environments has a critical role.
- Explicitly embedding the noise further improves enhancement ability.
- Consistent across all SNRs.

# Qualitative Evaluation

Universität
Augsburg
University

| | | |
|---|---|---|
| 25 dB: | Original: | Enhanced: |
| 20 dB: | Original: | Enhanced: |
| 15 dB: | Original: | Enhanced: |
| 10 dB: | Original: | Enhanced: |
| 5 dB: | Original: | Enhanced: |
| 0 dB: | Original: | Enhanced: |

# Future Work and Applications

- Psychoacoustics motivated loss function: allow the model to focus on the important things.

- Embedding speakers for source separation.

- Embedding environments for audio localisation in beamforming.

- Exploring the embedding space.

# Summary

Audio enhancement in unseen environments by:

- Condition the model on learned environment embeddings:

  → Learned adaptation to unseen environments.

- Collecting a large training sample from the environments distribution

  → Generalisation to unseen environments.

- Contact: **gil.keren@informatik.uni-augsburg.de**