The 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

Acoustic Matching by Embedding Impulse Responses

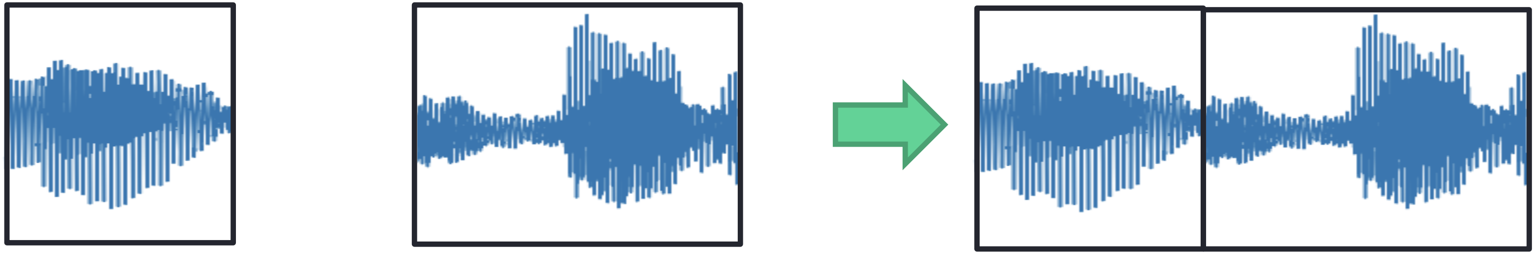
Jiaqi Su[1,2], Zeyu Jin[2], Adam Finkelstein[1]

[1] Princeton University
[2] Adobe Research

Motivation

Audio recordings with varying acoustic properties:

- **Variety of environment effects** in natural space
- **Variety of qualities** due to devices & recording setups



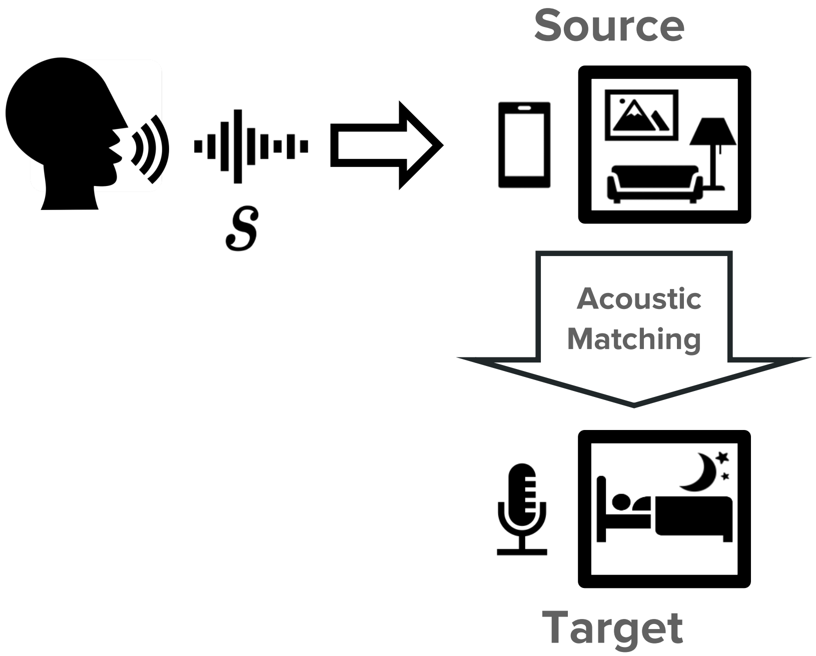
I record this part in kitchen;

And then record the rest in living room.

Goal: Acoustically match one part to the other, so that they sound seamless together.

Acoustic Matching

Transform **source** recordings to sound as if recorded in **target** environment



$$y_s = h_s * s + n_s$$

$$y_t = h_t * s + n_t$$

Previous Work: Acoustic Modeling

- **Acoustic Parameter Estimation**

  The ACE challenge [Eaton 2016]: Blind estimation of DRR & RT60 from recorded speech

  ➤ ***Simple approximation***

- **Impulse Response Estimation**
  - Assume knowing emitting source statistics [Florencio 2015] or having multiple channels [Crocco 2015]
  - Side-product of de-noising & de-reverb by NMD & NMF [Kagami 2018, Duan 2012]

  ➤ ***Under-constrained problem***

- **Impulse Response Generation**
  - The image method [Allen 1979]
  - Scene-Aware audio for 360° Videos [Li 2018]

  ➤ ***Performance Gap between synthetic IR and real IR***

Previous Work: Equalization Matching

- Source-differentiated equalization matching [Germain 2016]
  Address mismatched coloration and background noise

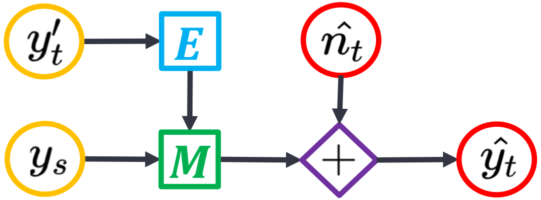
- Approximate specific equalization targets [Ramirez 2018]

- Mapping between different microphones [Mathur 2019]

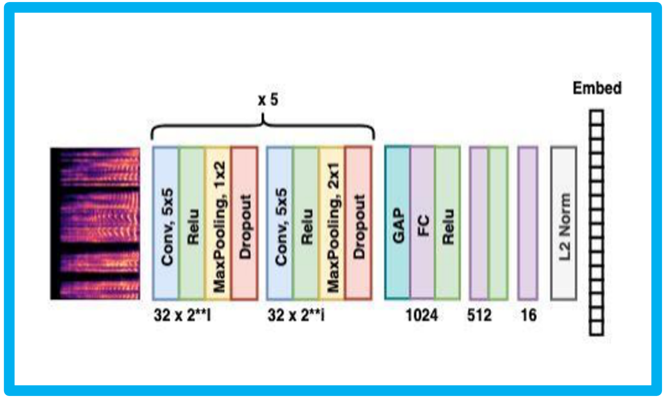
➤ ***Matching all of reverberation, equalization and noise remains an open problem***

Method: Overview

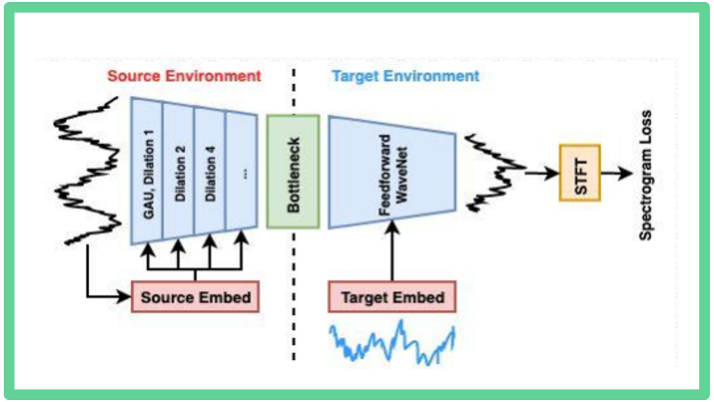
A **generic one-shot** acoustic matching method



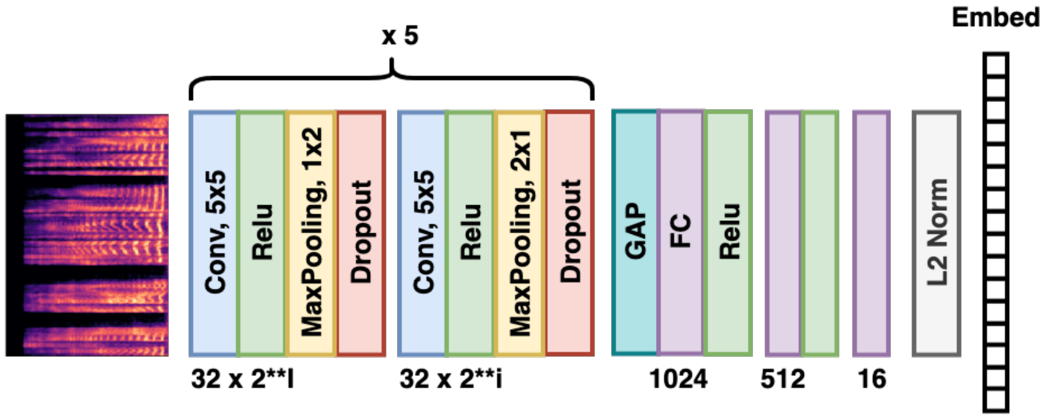
Embedding Network E



Acoustic Matching Network M



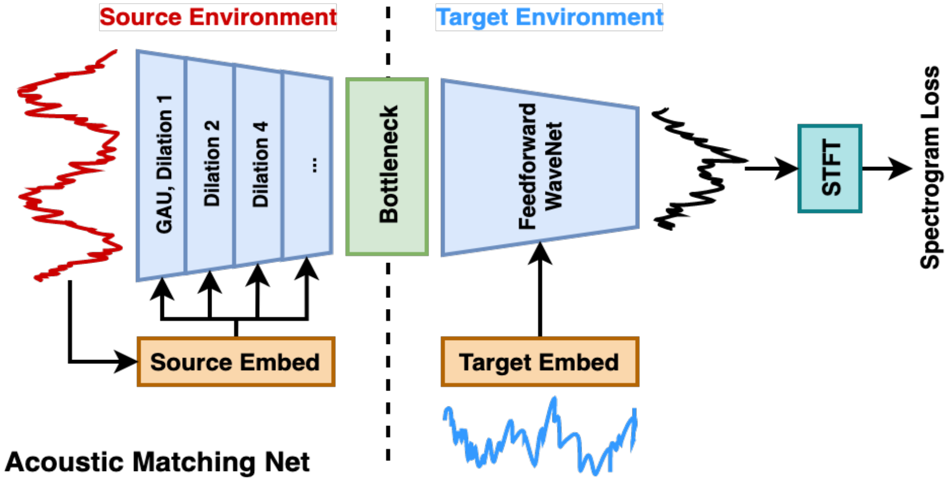
Method: IR Embedding



Embedding Net

- 16-dim embedding to encode IR information of recordings
- Pre-trained with triplet loss
- Nearest Neighbor IR Search

Method: End-to-end Acoustic Matching



- A stack of two feedforward WaveNets with bottleneck in between
- Globally conditioned on IR embedding of an example target recording
- IR embedding co-trained with acoustic matching task
- Perceptually-motivated spectrogram loss in place of sample loss

Data

Synthetic Data (training & evaluation)

- Clean speech: Device and Produced Speech (DAPS) clean set [Mysore 2015]
- IR: MIT Impulse Response Survey Dataset [Traer 2016]
- Noise: the Reverb Challenge [Kinoshita 2013] and the ACE Challenge [Eaton 2016]
- Data augmentation on speech, IR and noise:
  - Speaker voices (speed, pitch & volume)
  - DRR and RT60 of reverb
  - EQ distortion
  - Noise coloration and SNR

Real Data (evaluation)

- Device and Produced Speech (DAPS) Dataset
  - Recordings of high-quality speech re-recorded under different rooms environments

Evaluations

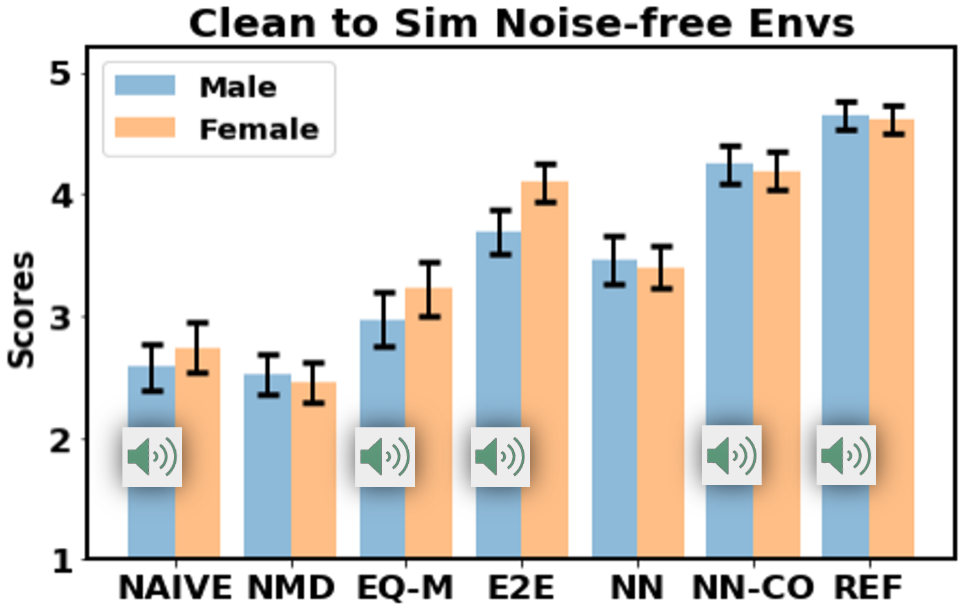
- Amazon Mechanical Turk
- Conducted 3000 HITs with 12 rating questions each.
    - Audio clips created by stitching two consecutive utterances from two different environments and acoustically matching one to the other.
    - Subjects rate how seamless the audio clips sound on a scale from 1=*very different* to 5=*seamless*

Evaluations

- Two sets of environment pairs:
  - Clean to synthetic noise-free environment ⟹ **Eval the quality of IR embedding**
  - Between real environments ⟹ **Eval generic acoustic matching**

- Method conditions:
  - **NAIVE:** No acoustic matching
  - **REF**: Ground truth
  - Baselines
    - **NMD**: IR estimated via NMD [Baby 2016]
    - **EQ-M**: Source-differentiated EQ matching [Germain 2016]
  - Ours
    - **E2E**: Our end-to-end acoustic matching network
    - **NN**: Our NN IR retrieved from the pre-trained embedding
    - **NN-CO**: Our NN IR retrieved from the co-trained embedding

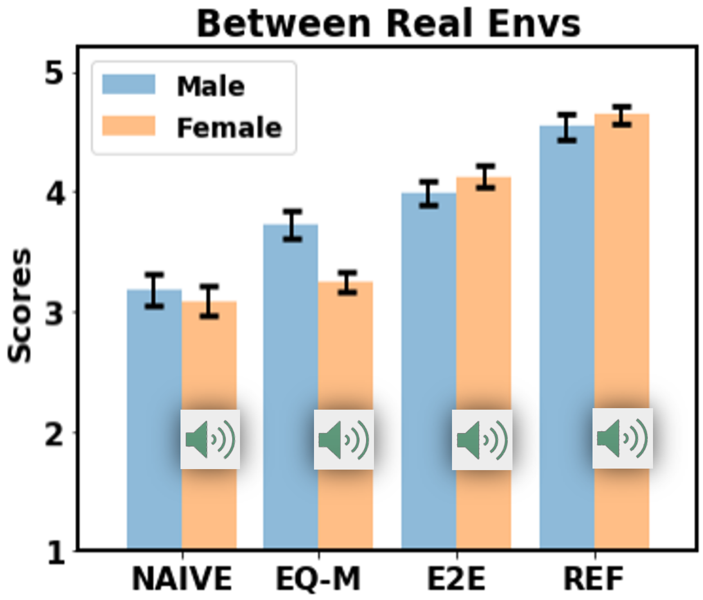
Evaluations: IR embedding



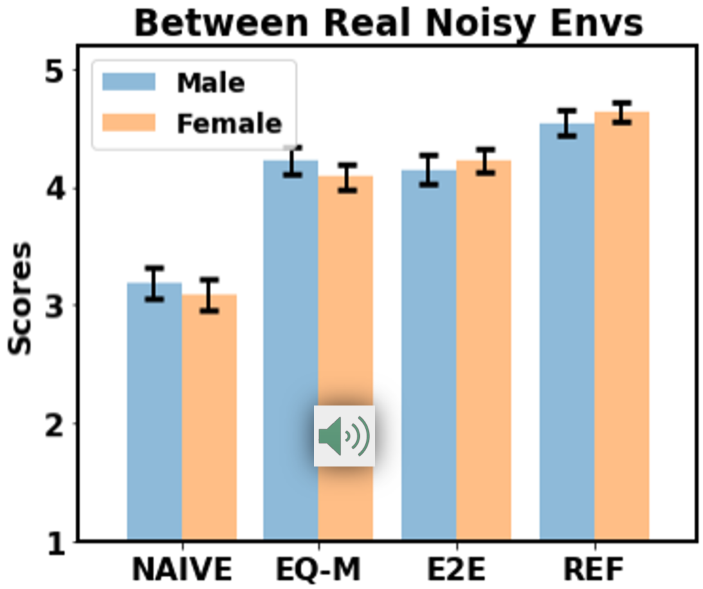
Takeaways:

- **NN-based** and **E2E** beat all the baselines
- **NN-CO** almost perfectly reproduces the desired acoustic effects

Evaluations: Acoustic Matching



⊃



Takeaways:

- **E2E** outperforms other approaches, being robust to noise
- **EQ-M** catches up when converting from noisy env to noisy env, due to noise masking

More audio examples

https://daps.cs.princeton.edu/projects/matching/

Conclusions

- An embedding space for acoustic impulse responses independent of speaker and speech content
- A generic one-shot waveform-to-waveform acoustic matching network based on the embedding.
- A simple and high-quality clean-to-environment matching solution based on nearest neighbor search in the embedding space.
- A human listening study on both real and synthetic data.

Thanks for watching!