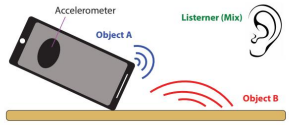
Feeling To Listen: Separating Simultaneous Impact Sounds Using Multi-Modal Learning

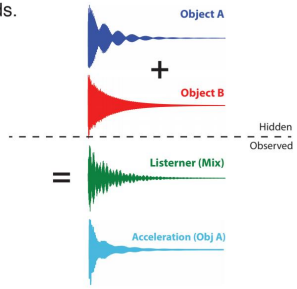
Jui-Hsien Wang, Stanford University

jw969@stanford.edu

Introduction

- Blind source separation for impact sounds.
- Hard problem:
  - Almost simultaneous
  - Mixed/overlapped frequencies
  - Sounds can have equal power



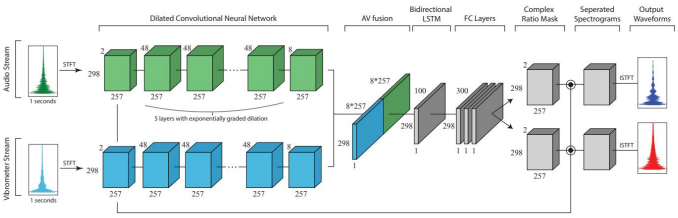


Data Collection

- Difficult to collect real-world data -> synthetic dataset with simulation.
  - Modal vibrational model + accurate acoustic radiation solver.
  - 90/5/5 split.
  - 120,000 sound clips (4 objects), each 1 second long at 48 kHz.

Model

- Network architecture (impl. Keras) inspired by Ephrat et al. (SIGGRAPH 2018)



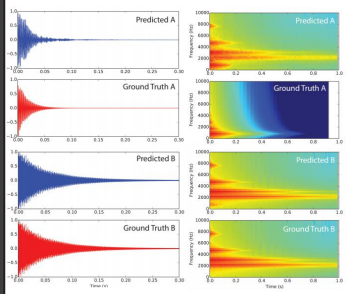
Model (cont.)

- Loss: Mean Squared Error (MSE) on predicted spectrogram
- Dilated CNN, Bi-LSTM, FC; Batch normalizations, Adam optimizer
- Learning rate: 0.001 (10 epochs), 0.0002 (10-20), 0.0001 (20-50).
- 50 epochs training takes ~8.5 days on 4 Tesla K80s (4.2m parameters).

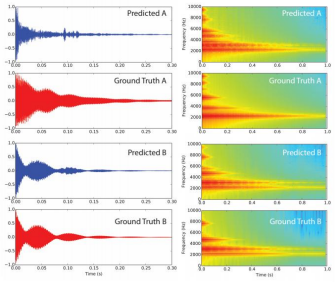
Results

- Test set average loss: 0.009 (Pascals).
- Adding the acceleration data helps break the symmetry.
- Frequency crossover, possibly due to STFT frequency resolution (~ 93 Hz).
- Explore larger architecture and other learning rate schedules (small diff).

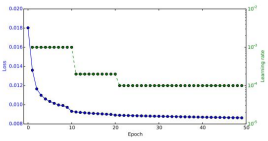
Good example (loss = 0.0013):



Worst example (loss = 0.088):



Convergence of training loss:



Future Work

- Better normalization of acceleration data.
- Investigate other loss functions.
- Gen. to unknown object (larger dataset).
- More efficient network architecture.