Multi-View Networks for Multi-Channel Audio Classification

Jonah Casebeer, Zhepei Wang, Paris Smaragdis

UNIVERSITY OF ILLINOIS @ URBANA-CHAMPAIGN

Proposal

We propose an architecture that allows us to train a neural net on a fixed number of channels and deploy it on an arbitrary number of channels at test time.

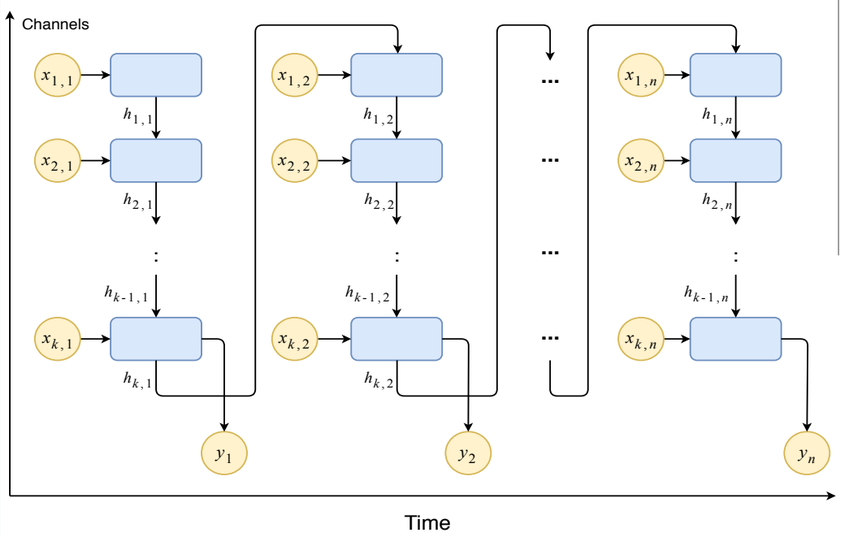
Setup

Here we focus on sound classification across a varying number of acoustic sensors with recordings of different quality

Standard neural net methodologies do not facilitate learning for deployment using arbitrary number of input channels.

Multi-View Networks (MVN)

Our model observes all available channels' data at the same time frame before making a prediction, using an RNN-like forward pass. The last channel's hidden state feeds into the first channel of the next time step. This model generalizes to unseen numbers of channels at test time just like RNNs for sequences of arbitrary length.



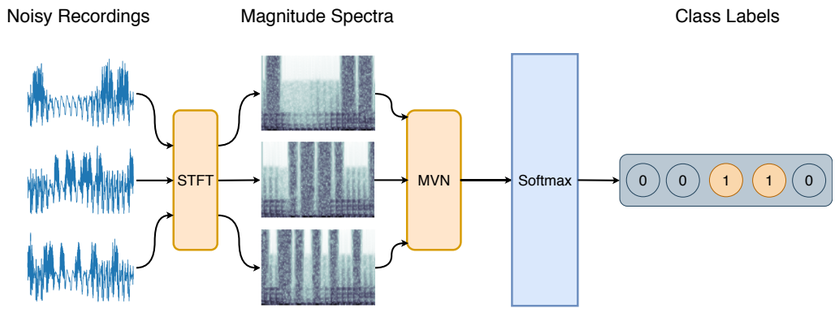
Recurrence Relation

MVNs learn a set of weights $\mathbf{W}_h, \mathbf{W}_x, \mathbf{U}_h$ through this recurrence

$$\mathbf{h}_{k,t} = \begin{cases} \sigma(\mathbf{W}_h \mathbf{x}_{k,t} + \mathbf{U}_h \mathbf{h}_{k,t-1}), k = 1 \\ \sigma(\mathbf{W}_h \mathbf{x}_{k,t} + \mathbf{U}_h \mathbf{h}_{k-1,t}), 1 < k \leq K \end{cases} \qquad \mathbf{y}_t = \sigma(\mathbf{W}_x \mathbf{h}_{K,t})$$

where K is the number of channels, $\mathbf{h}_{k,t}$ is the hidden state for channel k at time t, and $\mathbf{y}_t$ is the predicted label at time t. This recurrence allows the model to aggregate information from all channels at each time step.

Proposed Pipeline



Take each channel's magnitude Short-Time Fourier Transform (STFT). Unroll across each STFT frame and predict background noise or event of interest.

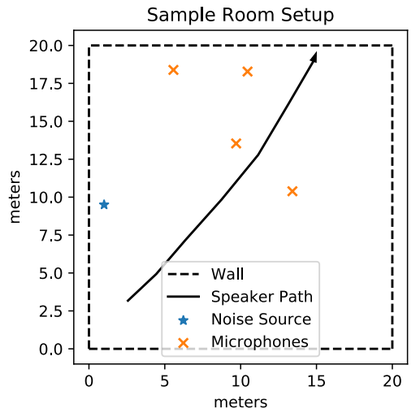
Training Setup

Data

- 4-channel, 2-sec recordings of intermittent speech and noise
- SNR linearly spaced between -5 and 5dB
- Recognition task was voice activity detection

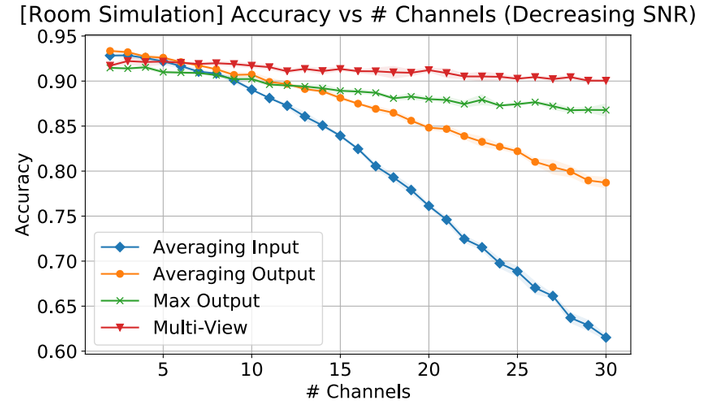
Room Simulation

- 20m x 20m reverberant room
- Moving point speech source
- Diffuse noise source
- Stationary microphones

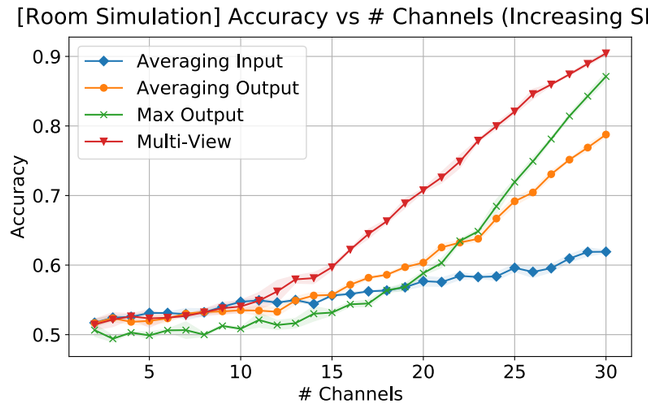


Experiments

We evaluate the performance of the MVN against three baseline models based on GRUs (averaging input data, averaging output labels, and taking output with highest confidence)



- Each additional channel has a lower SNR than the prior channels with SNR range between 0 and -29dB
- MVN less affected by channels with poor signal quality



- Each new channel has a higher SNR than the prior channels with SNR increasing from -29 to 0dB.
- MVN more effective at collecting information from few clean channels among many noisy channels

Conclusion

- Good performance when SNR varies a lot across channels
- Processing is largely invariant to order of input channels
- Can be deployed in settings with arbitrary number of sensors