Pitch Insensitive Speech Recognition Liquid State Machine

Prateek, Sahil & Pranava

EE746 Course Project
Guide: Prof. Udayan Ganguly, Vivek Saraswat

November 26, 2022

Summary

1. Background
2. Motivation
3. Existing Literature
4. Our Contribution
   - Idea
   - Experiments
   - Results
5. Further Work

Task

- This project aims to build a brain-inspired speech recognition system called a liquid state machine (LSM).
- LSMs differ from artificial neural networks in that the usual ‘learned’ fully-connected layer is preceded by a ‘liquid’ layer consisting of random recurrent connections with fixed weights which allow it to extract time-series features and perform time-series classification effectively.
- Our aim is to build an LSM that is pitch-insensitive i.e. it recognizes class of the speech irrespective of its pitch.

Liquid State Machine

- LSMs consist of a large recurrent network of randomly connected spiking neurons called the reservoir.
- The LSM consists of LIF Neurons with second-order synaptic dynamics. The neuronal dynamics for a LIF Neuron is given by:

$$\frac{\partial V_i}{\partial t} = -\frac{V_i}{\tau_{\text{Neu}}} + \sum_i \sum_j w_{ij} v_j$$

$$V_i > V_{th} \rightarrow V_i = 0 \;\forall\; t_s < t < t_s + T_{rp}$$

where,

$$v = \frac{1}{\tau_1 - \tau_2}\left(e^{-\frac{t-t_s}{\tau_1}} - e^{-\frac{t-t_s}{\tau_2}}\right)H(t - t_s)$$

- LSM is randomly connected where the probability of connection between two neurons N1 and N2 is governed by the following function of their separation [1]:

$$P(N_1, N_2) = K \cdot e^{-\frac{D^2(N_1,N_2)}{\lambda^2}}$$

Preprocessing

Lyon Passive Ear Model

Speech preprocessing stage for the LSM consists of a human ear-like model called Lyon's Auditory Cochlear model. It consists of a cascade of second order band pass filters to produce a response for each channel, where it is rectified and low-pass filtered to get a smooth signal at the output.

Ben Spiking Algorithm [3]

BSA is an algorithm for encoding an analog input signal $f(t)$ as a spike train $x(t) = \sum_{k=0}^{\infty} \delta(t-k)$ such that $f(t) \approx x(t) * h(t)$ for some FIR reconstruction filter $h(t)$. This is a variation of rate coding known as stimulus estimation.

Pitch Shift-Invariance

- The speech recognition system has to learn to classify digits spoken by a variety of speakers with different vocal characteristics and pitches. We want our network to be insensitive to these variations while classifying the content of the speech.
- We can make the simplifying assumption that different speakers saying the same digit only cause a vertical shift in the spectrogram (i.e. along the frequency axis). We ignore the other pitch variations between different speakers. Thus, we now come up with a way of processing the spectrograms from the Lyon Passive ear model to produce a shift-invariant representation. We will make use of the DFT magnitude for this purpose.

Using DFT

- $\text{DFT}\{x[n]\} := X[k] = \sum_{n=0}^{N-1} x[n]e^{-\frac{j2\pi kn}{N}}$
- $\text{DFT}\{x[(n-n_0)_N]\} = X[k]e^{\frac{j2\pi kn_0}{N}}$. Thus, DFT magnitude is invariant to circular shift by n_0.
- By adding zeros to a finite sequence $x[n]$ we can increase the resolution of the DFT since the size of DFT matches its input.
- A real signal has symmetric DFT magnitude about zero.

Motivation

A liquid state machine has several recurrent connections and since these are randomly generated we can extract features pertaining to correlations across various timescales. Thus, the output of the LSM layer can be then put through a fully connected layer to classify the input.
This feature makes an LSM very suitable for our task, that is classifying spoken digits in the TI-46 dataset.

Existing Literature

Hardware-Friendly Synaptic Orders and Timescales in Liquid State Machines for Speech Classification [2]

1. authors - Vivek Saraswat, Ajinkya Gorad, Anand Naik, Aakash Patil and Udayan Ganguly
2. This paper classifies a subset of the TI-46 dataset consisting of 5 female speakers. Our model is exactly identical to this except for preprocessing and changes in network parameters
3. The subset of female speakers has a much smaller pitch variation compared to the full TI-46 dataset (which also includes male speakers)
4. This paper achieves an accuracy of 99% using the second order synaptic model (which our simulation also uses)

Another paper [4] tries to use LSMs for classification on the TI-46 dataset as well

Idea

Our key contribution is to come up with a shift-invariant encoding to scheme for the input spectrogram (i.e. the output of the Lyon Passive Filter stage). We first zero pad the spectrogram to 4 times the input size, perform DFT and then extract the low to mid frequency DFT magnitude output. We neglect DC components as they have an overwhelming presence compared to higher frequency components. We choose our transformation such that the input and output size are same. This transformation is similar to the idea of cepstrum.

Spectrograms

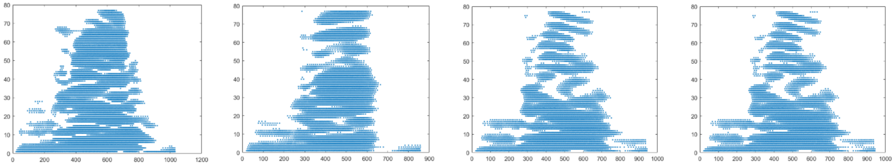

Figure: LSM input spike train for spoken digit 2 with shift-invariant preprocessing for 4 different speakers (2 male, 2 female)

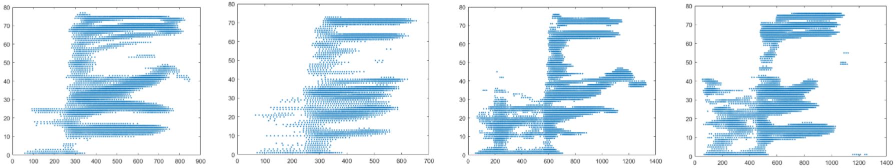

Figure: LSM input spike train for spoken digit 2 without shift-invariant preprocessing for 4 different speakers (2 male, 2 female)

Spectrograms

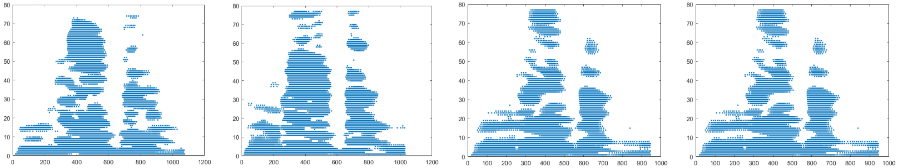

Figure: LSM input spike train for spoken digit 8 with shift-invariant preprocessing for 4 different speakers (2 male, 2 female)

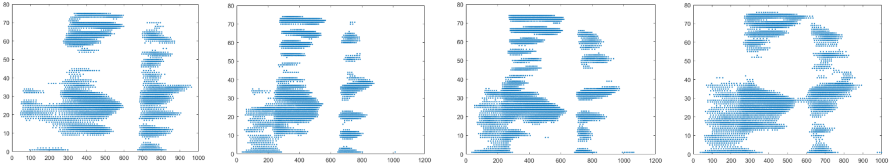

Figure: LSM input spike train for spoken digit 8 without shift-invariant preprocessing for 4 different speakers (2 male, 2 female)

Experiments Conducted

After performing preprocessing we tried to perform hyperparameter search over $\alpha_{G_{in}}$ and $\alpha_{G_{res}}$. The optimal values of hyperparameters were chosen to give minimum classification error (or maximum classification accuracy). We also tried to analyse the effect of LSM capacity on performance. We took a grid of 5x5x5 neurons and compared it with a grid of 7x7x7 neurons. We compared performance with baseline (i.e. without the shift invariant preprocessing) as done in Stage 1.

Results

	accuracy
train	94.39%
test	91.95%

Table: Accuracy for 5x5x5 LSM without preprocessing (baseline)

	accuracy
train	82.32%
test	80.50%

Table: Accuracy for 5x5x5 LSM with preprocessing

	accuracy
train	88.21%
test	83.58%

Table: Accuracy for 7x7x7 LSM with preprocessing

Results

Thus, we experimentally see that the shift-invariant preprocessing actually degrades network performance. Thus the DFT magnitude is probably not a good choice for a shift-invariant transform. However, we might need to do a more extensive hyperparameter search to establish this claim. Our approach is very similar to the popularly used cepstrum analysis for extracting features from spoken audio.

Further, it is also important to consider the implementation difficulties of this processing stage in a neural circuit. DFT magnitude computation involves multiplication and addition of complex numbers (which have to be handled in terms of separate real values). We have currently done this stage at the output of the Lyon Passive Ear model which is in the analog valued domain (rather than spiking domain) and thus, doesn't allow efficient implentation using neural circuits.

Further Work

We can mathematically show using properties of BSA and DFT that this transformation can also be applied in spiking domain. We can possibly fix synaptic weights as a means of implementing multiplication with the DFT matrix. We do not have a concrete implementation in mind but we can discuss this in the QnA.

Bibliography

- A. Gorad, V. Saraswat **and** U. Ganguly. 2019. Predicting performance using approximate state space model for liquid state machines. **in***2019 International Joint Conference on Neural Networks (IJCNN)* 1–8. DOI: 10.1109/IJCNN.2019.8852038.

- Vivek Saraswat, Ajinkya Gorad, Anand Naik, Aakash Patil **and** Udayan Ganguly. 2021. Hardware-friendly synaptic orders and timescales in liquid state machines for speech classification. **in***2021 International Joint Conference on Neural Networks (IJCNN)* 1–8. DOI: 10.1109/IJCNN52387.2021.9534021.

- B. Schrauwen **and** J. Van Campenhout. 2003. Bsa, a fast and accurate spike train encoding scheme. **in***Proceedings of the International Joint Conference on Neural Networks, 2003.* **volume** 4, 2825–2830 vol.4. DOI: 10.1109/IJCNN.2003.1224019.

- D. Verstraeten, B. Schrauwen, D. Stroobandt **and** J. Van Campenhout. 2005. Isolated word recognition with the liquid state machine: a case study. *Information Processing Letters*, 95, 6, 521–528. Applications of Spiking Neural Networks. DOI: https://doi.org/10.1016/j.ipl.2005.05.019.