

Yuxuan Liu yul067@ucsd.edu  
 Haoran Sun hsun@ucsd.edu  
 Moyan Zhou moz006@ucsd.edu  
 Cai Chen cac005@ucsd.edu

## Abstract

The dynamic sound in the urban environment is still neglected aspects in urban planning and architectural design. Urban sound classification could have much contribution in that it helps to unify independent areas related to sound and environment. The problem of classifying sound is that the feature of audio data is more complex than visual objects' and how the feature is processed will have huge impact on the result. In this project, three models are trained on different features of audio clips, and the outputs are ten classes of audio with accuracy. Currently, the best accuracy is 91.1% produced by Recurrent Neural Network (RNN) model with combined features.

## Dataset

The dataset contains 8732 labeled environmental sound files in .wav format, and the file names are their IDs [7]. The CSV file in the come with the dataset includes each file ID and its corresponding folder and type of sound, and each sound file is around 1 to 4 seconds. There are 10 classes of sounds: Air conditioner, Dog bark, Engine idling, Children playing, Drilling, Gun shot, Jackhammer, Siren, and Street music.

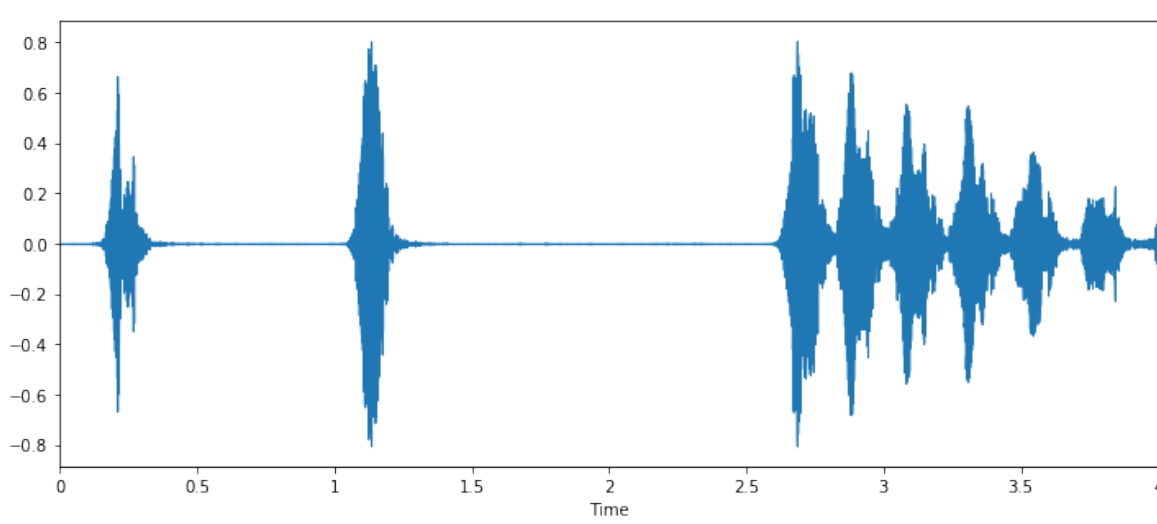


Figure 1: a dog barking sound clip in wave format

## Features

The original feature size varies due to different sample rate of .wav files, hence we use paddings or duplication to obtain feature vectors of desirable length empirically and extract some combinations of the following features.

- **Fast Fourier Transform (FFT)**: specifically the short time Fourier transform (STFT), applies windows onto the signal and transferring the signal from time domain to frequency domain in order to better characterize a sound; feature length is 2000.

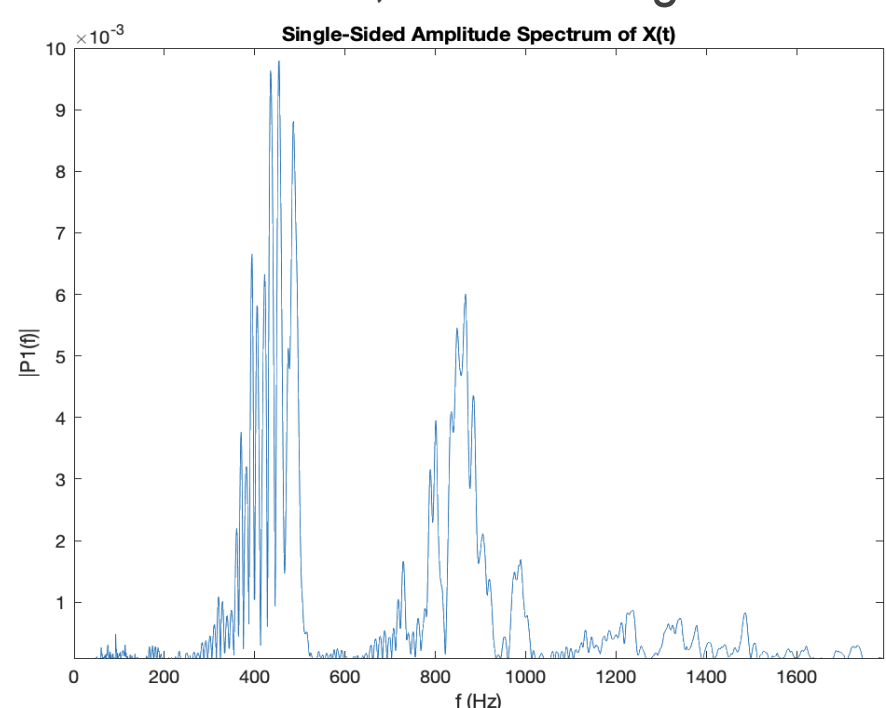


Figure 2: SFTF spectrogram of dog bark

- **Root Mean Squared Energy (RMSE)**: the energy of a signal corresponds to the total magnitude of the signal (energy could also help to recognize a sound); feature length is 100.

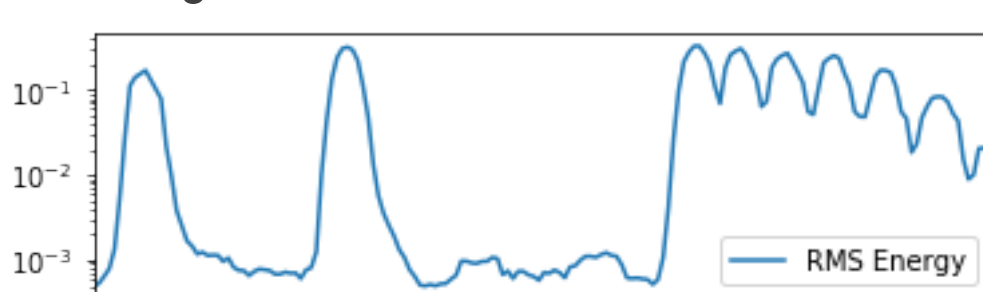


Figure 3: RMS Energy of dog bark

- **Mel Frequency Cepstral Coefficient (MFCC)**: resembles the human auditory system and is widely used in speech recognition [6]. DCT of the filtered power spectrum; feature length is 256.

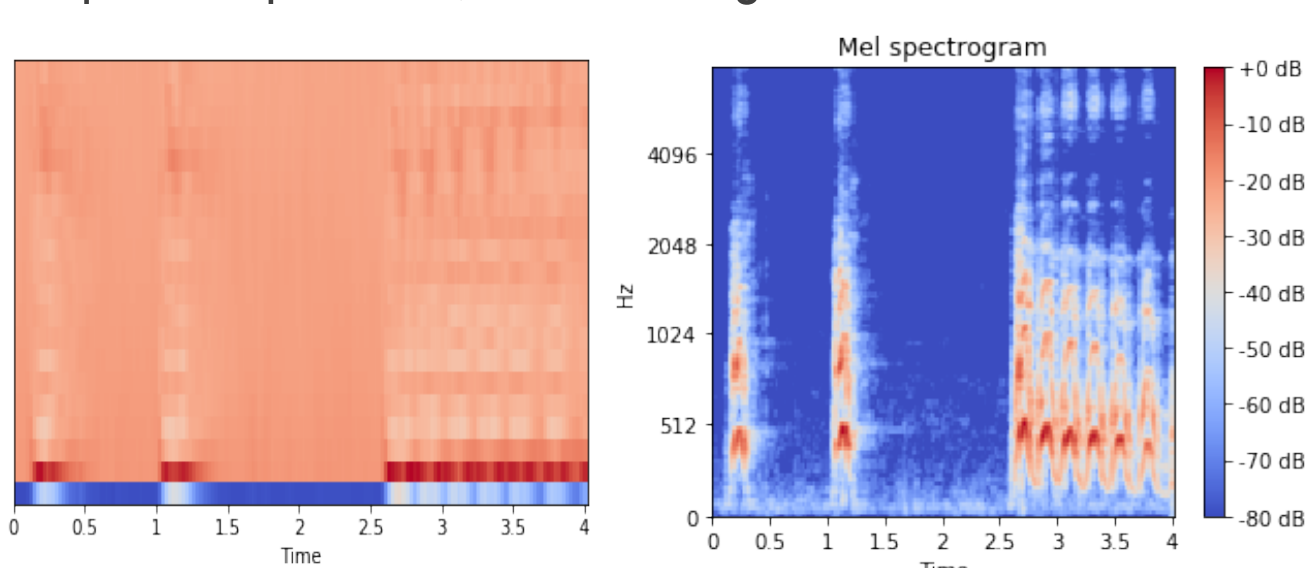


Figure 4: MFCC spectrogram and Mel-spectrogram of dog bark

- **Chromagram**: pitch-class profiles of a given audio [3].
- **Mel-spectrogram**: maps the spectrogram onto the mel-scale.
- **Cepstral contrast**: includes the spectral peak, the spectral valley, and their difference in each frequency sub-band features to enhance the performance [3].

## Models

### CNN Architecture

Our Convolutional Neural Network (CNN) model is composed of three kinds of layers: convolution, ReLU and pooling. The operations of three layers are repeated three times, with each layer learning to identify different features.

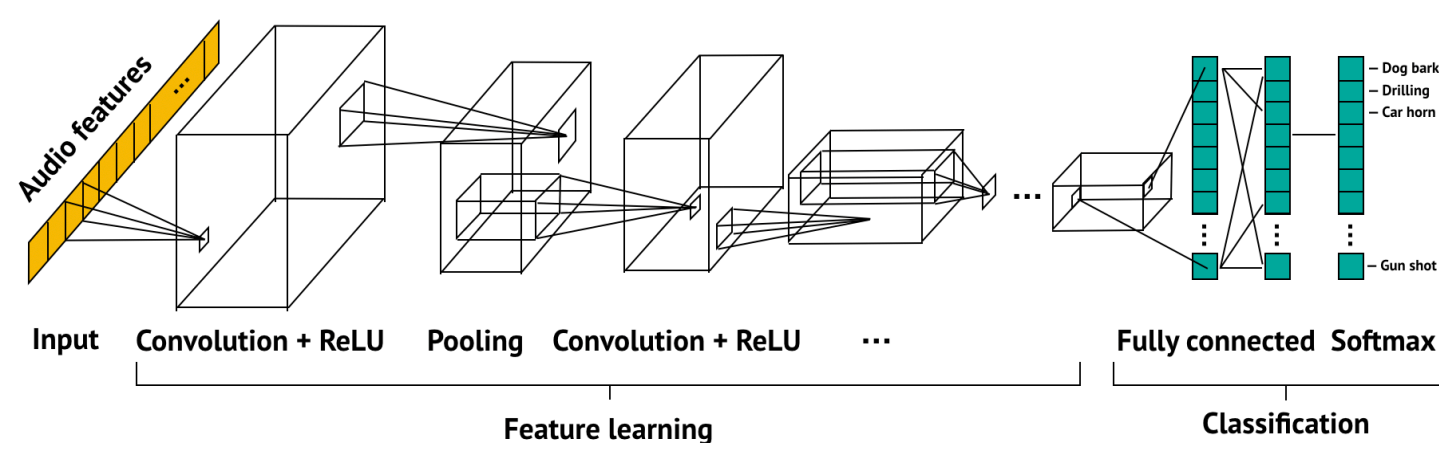


Figure 6: example of CNN network with some convolutional layers

### Kernel SVM Model

Support vector machine finds a hyperplane in a N-dimensional space which can maximize the margin among different data clusters. By using the grid search to choose a type of kernels and its parameter C and gamma in kernel function to optimize the result. We found radial basis function kernel( with C = 32 and gamma = 128) trained on FFT with 2000 features gives the best result. Let w be the weight vector of SVM and  $\phi(x)$  be some kernel function,  $w = \sum_{j=1}^n \alpha_j y^{(j)} \phi(x^{(j)})$ . The kernel SVM (dual form) is:

$$\max_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \alpha_i - \sum_{i,j=1}^n \alpha_i \alpha_j y^{(i)} y^{(j)} (\phi(x^{(i)}) \cdot \phi(x^{(j)}))$$

$$\text{s.t. } \sum_{i=1}^n \alpha_i y^{(i)} = 0, 0 \leq \alpha_i \leq C$$

## Results

Using combined features, MFCC, mel-spectrogram (Mel), cepstral contrast (Contrast), chromagram (Chroma), on RNN produces the best validation accuracy.

Table 1: Classification Accuracy of Different Features on Different Models

Feature	FFT	FFT	RMSE	MFCC	MEL	Contrast	Chroma	MFCC+ Mel	MFCC+ Contrast	MFCC+ Chroma	Combined
Model	CNN	SVM	CNN	RNN	RNN	RNN	RNN	RNN	RNN	RNN	RNN
Validation Accuracy	85.5%	63.3%	43.9%	88.8%	74.3%	40.2%	29.6%	88.7%	89.0%	89.2%	91.1%

## Discussion

- For MFCC, mel-spectrogram, spectral contrast, and chromagram features, they are not discriminative on CNN and not linearly separable on SVM.
- CNN network does not preserve ordering of input data and not consider the time-domain correlation of the input features.
- Using the combined feature on RNN gives the best accuracy since the combined feature covers more characteristics of an audio clip.
- For instance, although the chromagram itself gives a low classification accuracy, some types of sounds with distinguishable pitches can be better classified by using more general features along with the chromagram.
- RMSE captures the energy/magnitude of a sound clip; however, magnitude alone cannot be used to distinguished different kinds of sound, and noises in the clip can significantly lower the quality of this measurement.

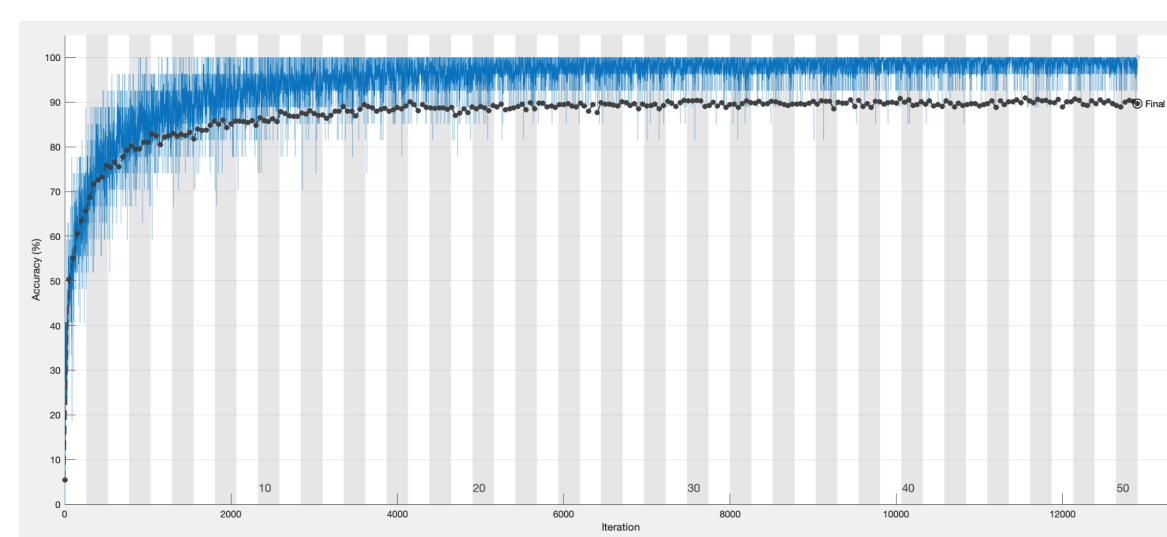


Figure 9: Convergence example

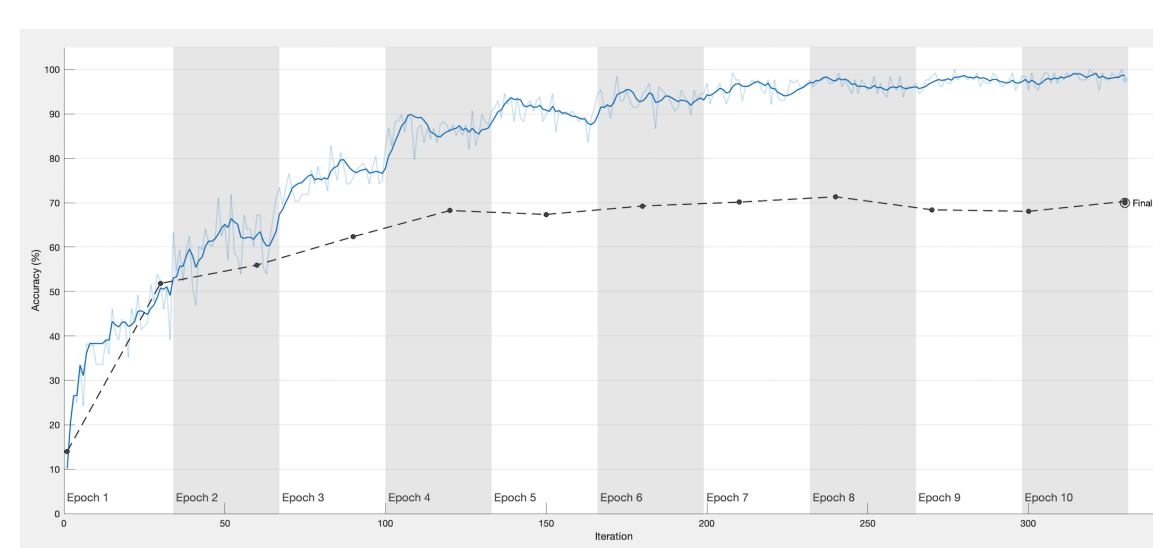


Figure 10: Overfit example

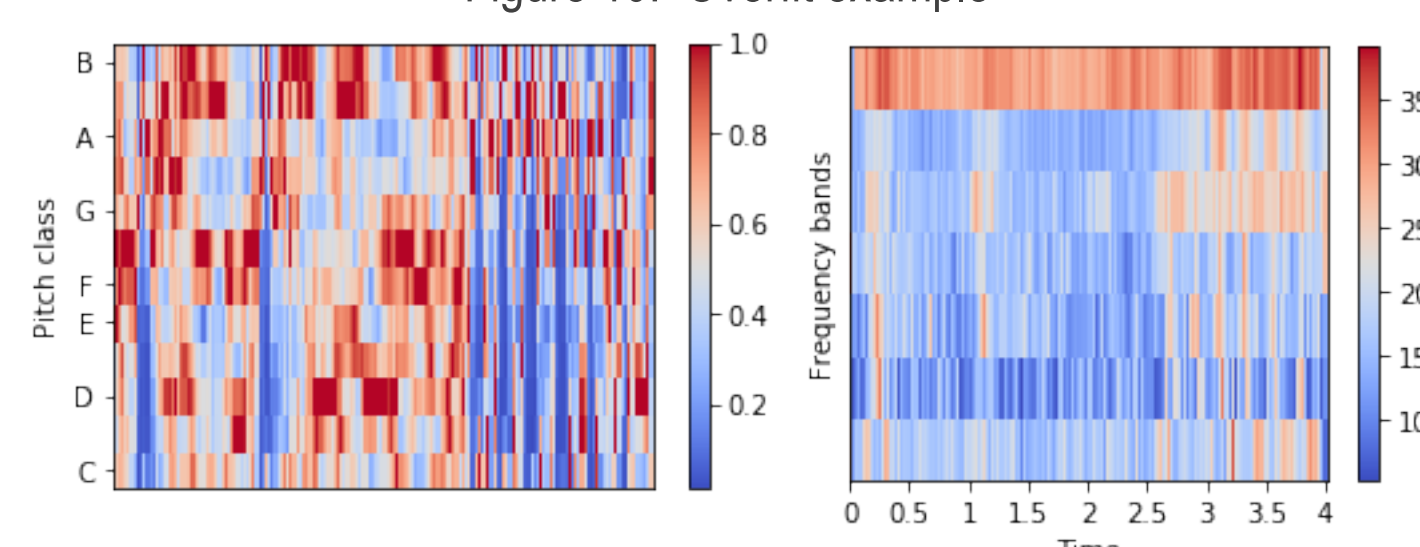


Figure 5: Chromagram and cepstral contrast of dog bark

### RNN Architecture

An Long Short-Term Memory (LSTM) is a type of RNN that can be better used for time series data classification. We use five LSTM layer to receive sequence or time series data, then the network also ends with a fully connected layer, a softmax layer and a classification output layer to predict the labels. The diagram in Figure 8 illustrates the flow of a time series X with C features of length S through an LSTM layer.  $h_t$  and  $c_t$  denote the output and the cell state at time step t, respectively.

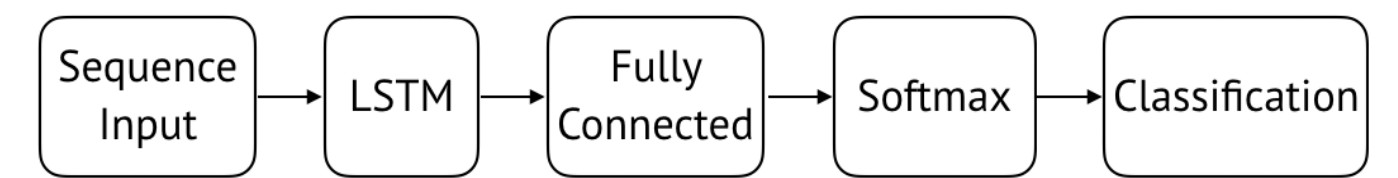


Figure 7: architecture of a simple LSTM network for classification

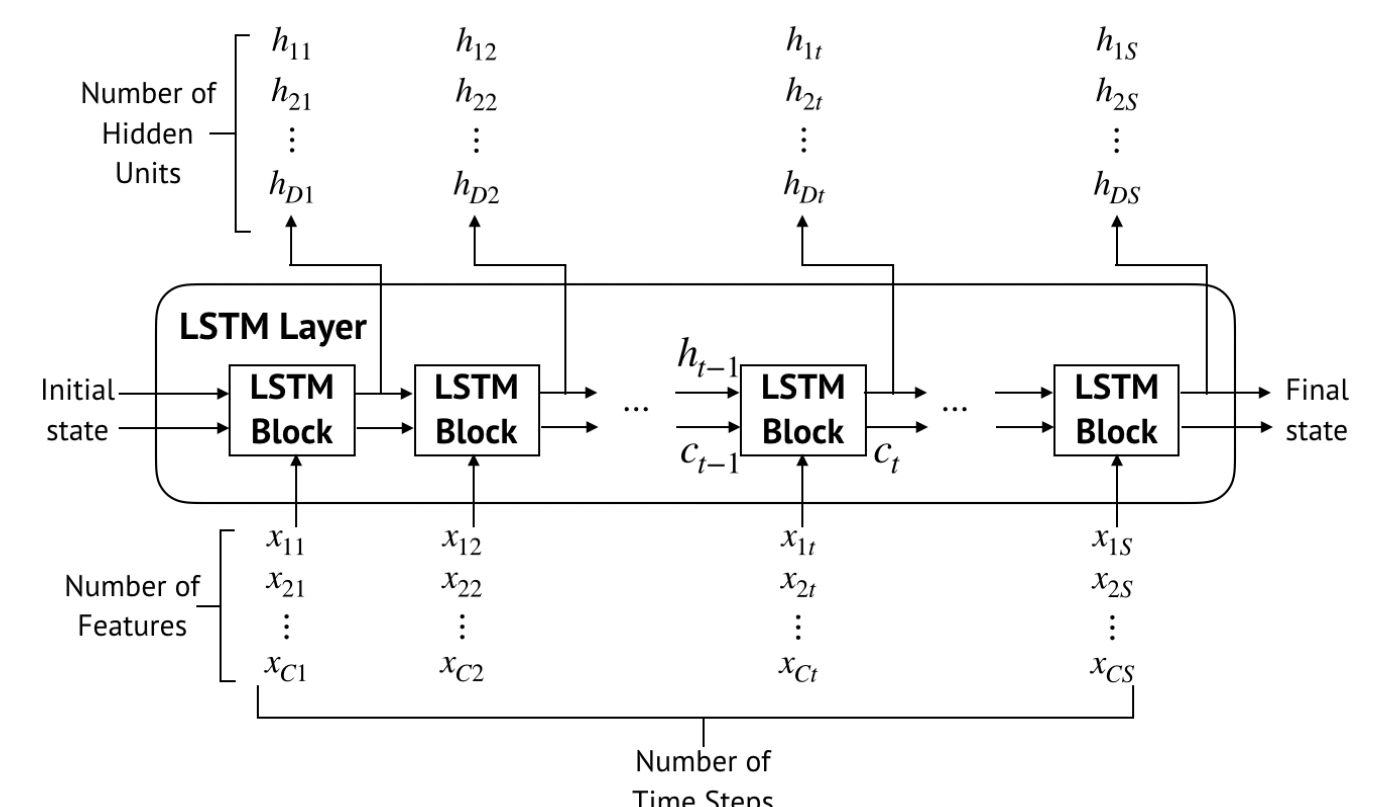


Figure 8: flow chart of LSTM layer

## Future Work

- We aim to compare other neural networks such as DNN and add more filters, such as the Gabor filter, to input signals before extracting.
- We plan to use the same features and models on different dataset, such as the ESC-50 dataset.
- Since the training process shows that using CNN may lead to overfit, we could use Principal Component Analysis(PCA) to reduce the input data's dimension.

## References

- [1] Boddapati, V., Petef, A., Rasmusson, J. and Lundberg, L. (2019). "Classifying environmental sounds using image recognition networks," in *Progr. Comput. Sci. Appl. Logic*, 112 (2017), pp. 2048-2056
- [2] D. M. Agrawal, H. B. Sailor, M. H. Soni, and H. A. Patil, "Novel TEO-based gammatone features for environmental sound classification," in *European Signal Processing Conf. (EUSIPCO)*, Kos island, Greece, 2017.
- [3] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai. Music type classification by spectral contrast feature. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Lausanne (Switzerland), August 2002.
- [4] H. Zhou, Y. Song and H. Shu, "Using deep convolutional neural network to classify urban sounds," *TENCON 2017 - 2017 IEEE Region 10 Conference, Penang*, 2017, pp. 3089-3092.
- [5] Piczak, K.J. "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia*, 26-30 October 2015; ACM: New York, NY, USA, 2015; pp. 1015-1018.
- [6] S. Chachada and C.-C. Jay Kuo, "Environmental Sound Recognition: A Survey," in *Proceedings of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013.
- [7] J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", *22nd ACM International Conference on Multimedia, Orlando USA*, Nov. 2014.