# Augmentation

Data augmentation is the process of synthesising new labelled data by varying the base sample by introducing slight permutation in the base data. The objective of this task is to make the model invariant to the additional permutation and to increase generalization. The new synthesised data will conserve the base label.

Due to limited quality labelled audio data corpus, the classification models trained on these data corpus tend to overfitting and poor generalization. In this paper we use classical audio augmentation techniques such as white noise addition, stretching, rolling and shifting the pitch. Also, we have applied sample mixing based augmentation over the new audio labelled corpus. Finally, we have experimented with Neural Synthesis based synthesis and interpolation of new labelled data.

# Adding white noise:

In this method a silent sample is created for the length of the input sample. Then white noise is added to this new sample. Finally the base noise and the white noise sample is weighted averaged. The weights assigned are 0.05 for the white noise and 0.95 for the original audio. This way the initial sound is preserved and help in generalizing the model.

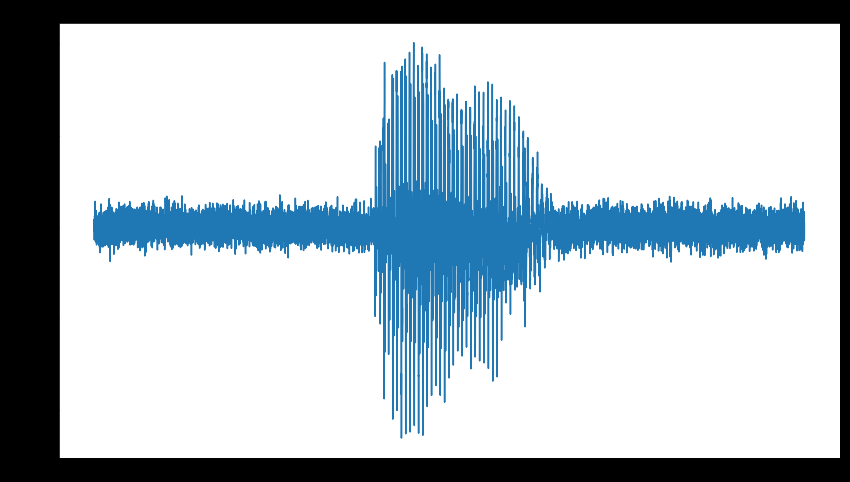
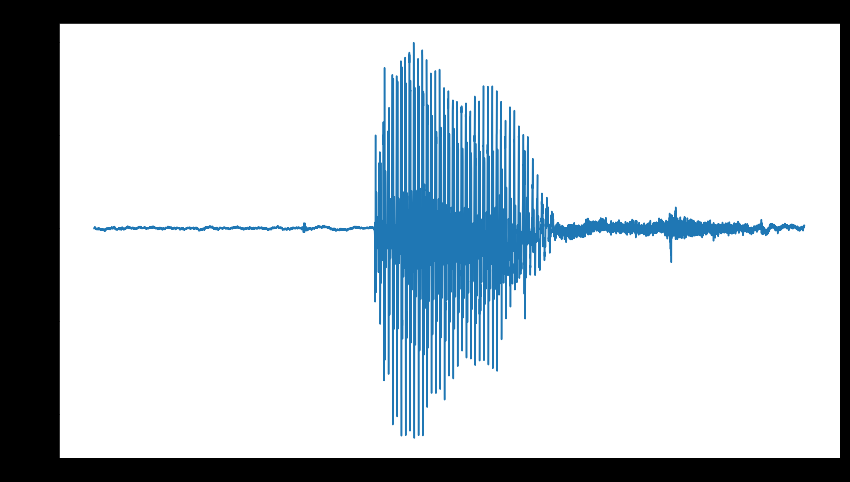


Fig: (a) Initial fall Data (b) After adding White noise

# Shifting the pitch:

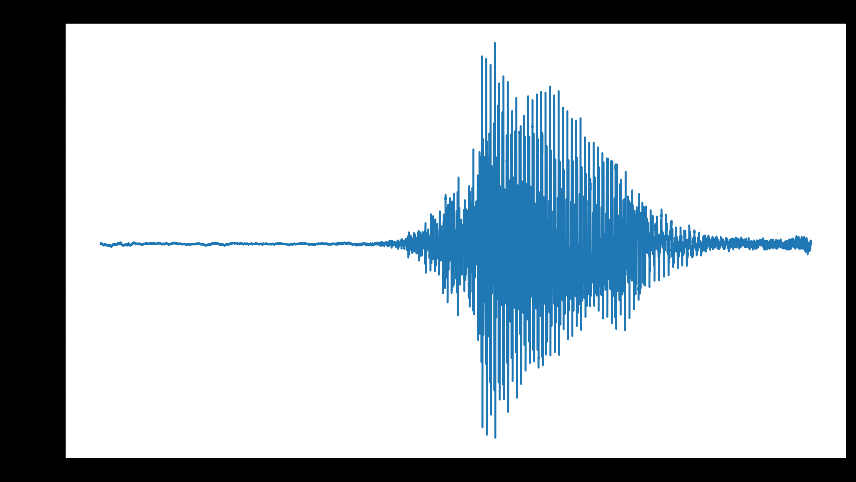
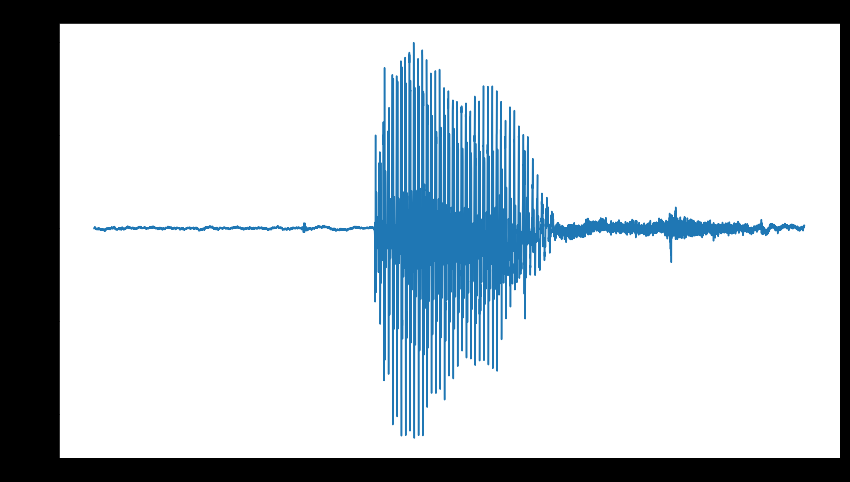


Fig: (a) Initial fall Data (b) After shifting

Pitch shifting is a audio technique in which the original pitch of a sound is raised or lowered. Pitch shifters are units which raises or lowers the pitch by a pre-designated interval (transposition).

In this process both pitch and speed are varied. This is done by speeding up or slowing down the recording speed.

Stretching:

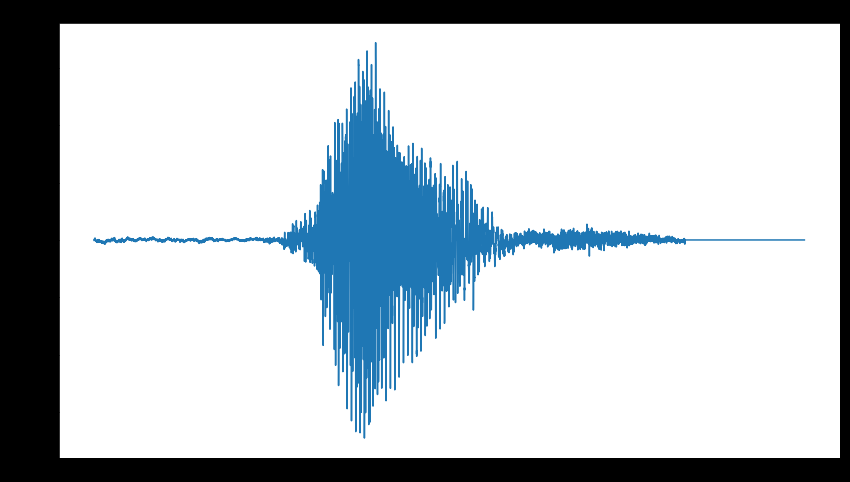
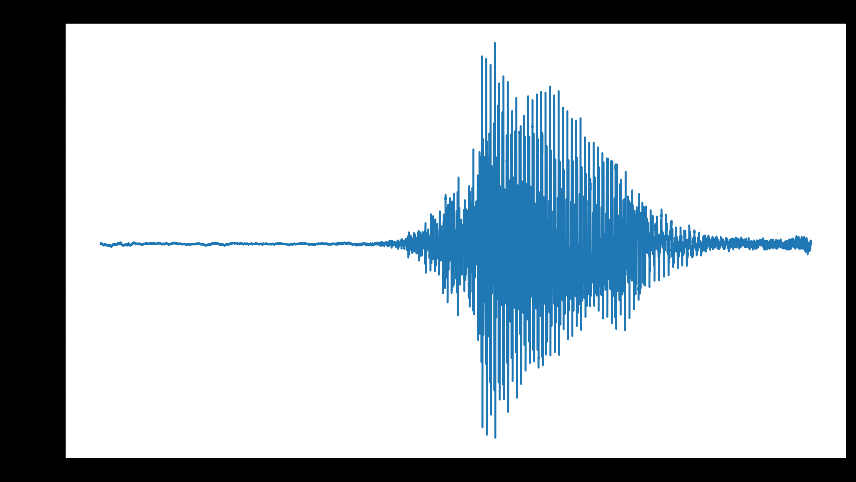
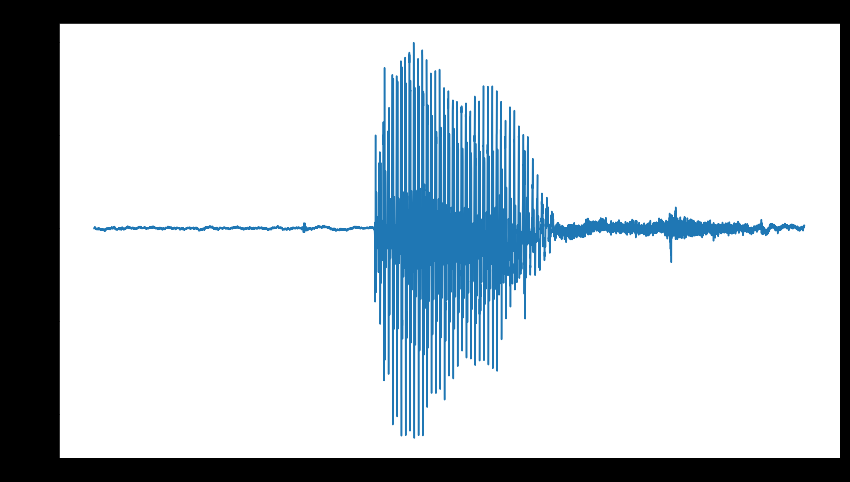
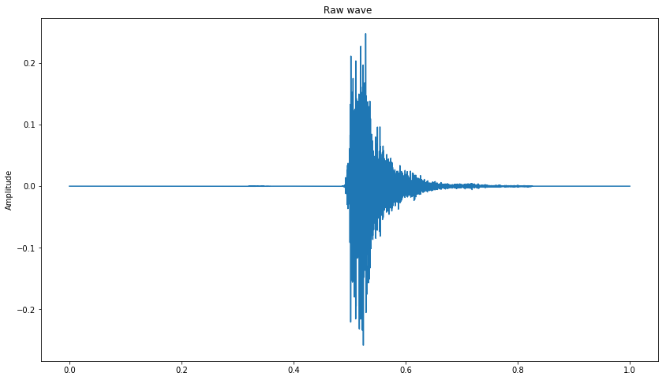
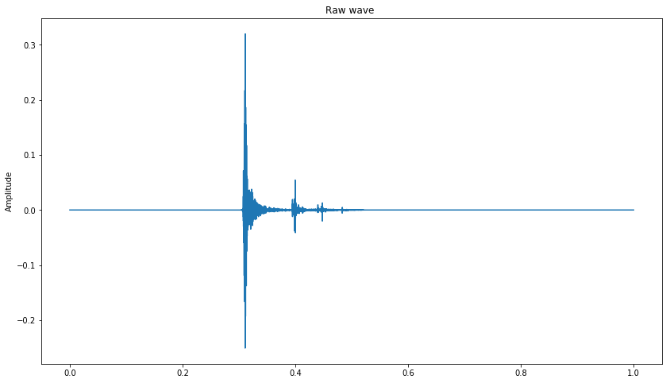


Fig: (a) Initial fall Data (b) After Stretching low frequency (C) Stretching high frequency

Time stretching is changing the duration or speed of a audio without affecting the pitch.

Here we do high freq time shifting at the rate of 1.2 and low frequency time shifting at the rate of 0.8.

# Audio Mixing:



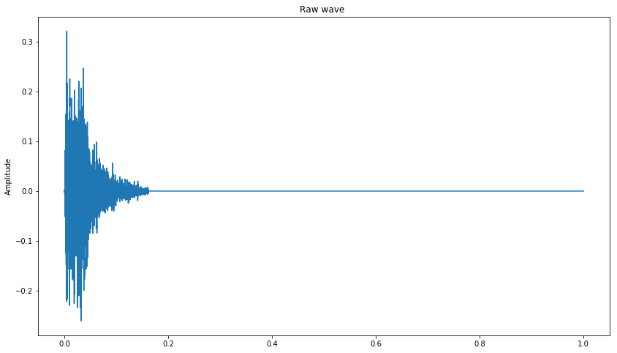


Fig: (a) Initial fall Data (b) Second Fall Data (c) Mixed Fall data

In this process initially the silent part of the audio preceding and in successive of the target audio (fall sound) is removed. For this the DBfs (DB full scale) = -0.5 is fixed as a threshold.

Then the overlapping length of a audio pair is obtained. Finally, both of this audio pair is averaged. This was we create a new audio sample for each pairs, and this new audio sample will retain .

# Neural Audio Synthesis

Traditional audio synthesis will generate audio from oscillators and wavetables. Nsynth factorizes sound as timber and dynamics. This is an autoencoder approach for audio generation.

Here we fine tune a pre trained wave net encoder using our fall data. The last layer of the encoder and the decoder are retrained. Other layers are frozen.

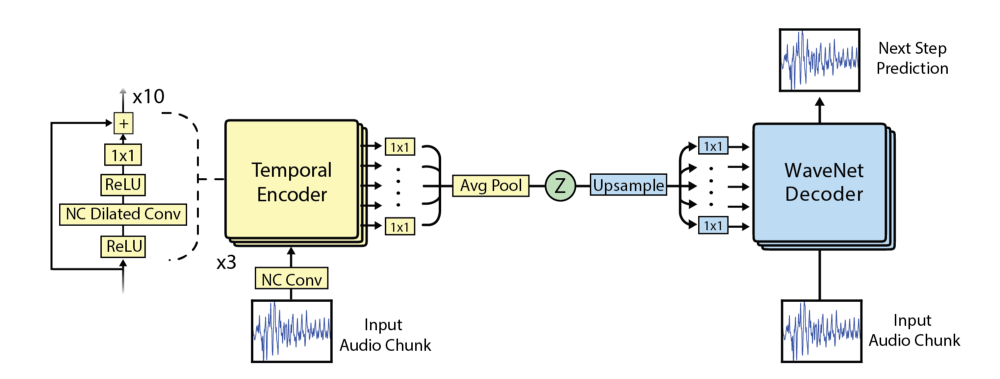


Fig: Wavenet Autoencoder

Here a Wave Net based autoencoder model is used as a conditional auto-regressive decoder on temporal codes derived from audio waveforms. This is a data driven approach for augmentation.

# Encoded representation (16 Channels)

After encoding the data is represented as 16 channels (Reduced embedding

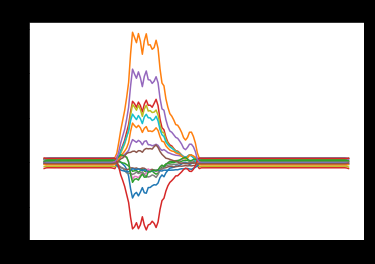
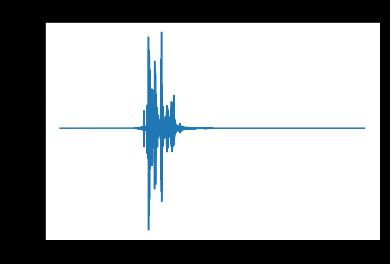


Fig: (a) Initial fall Data (b) Encoded 16 Channel Representation

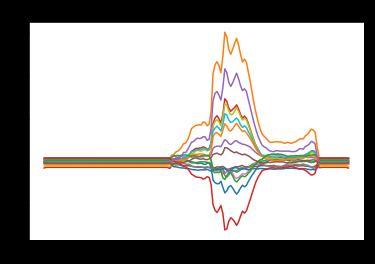
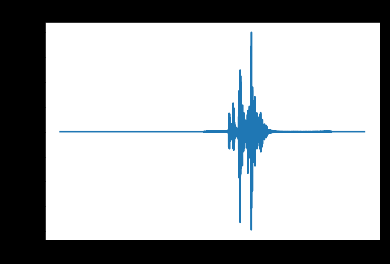


Fig: (a) Initial fall Data (b) Encoded 16 Channel Representation

## Reconstruction samples:

The audio sample Is passed through the pretrained wave net variational autoencoder. This process retains the basic structure of the audio sample but does not recreate the sample exactly. The encodings are not varied. Thus, this could be used as augmented data to increase the no of samples. We can see that the encoded 16 channel output follows the overall pattern of the input audio. When decoded this produces a slightly permuted audio.

Also samples are created by reconditioning the “Z” latent representation of the input audio. This is done by changing the pitch.

## Interpolation of encoded values (Midpoints):

In this method similar to the audio mixing method we will be combining all combinations of our target class. A pair of fall sound is passed through the encodder and this creates the latent embeddings. Then the average of these embedings are found. This is different from adding the two sounds or like listining to it in the same time. Here we are averaging the represention of the timber,tonality and change over time thus resulting in a new audio sample. This gives us superior results compared to simple averaging.

Ref:

* SAMPLE MIXED-BASED DATA AUGMENTATION FOR DOMESTIC AUDIO TAGGING <https://arxiv.org/pdf/1808.03883.pdf>
* Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification

<https://arxiv.org/pdf/1608.04363.pdf>

* Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders

<https://arxiv.org/pdf/1704.01279.pdf>

* Wavenet

<https://arxiv.org/pdf/1609.03499.pdf>

* Sample Mixed-Based Data Augmentation for Domestic Audio Tagging

<https://arxiv.org/abs/1808.03883>