# AUDIO SIGNAL ANALYSIS, INDEXING AND TRANSFORMATION

**Samuel Guilluy - MVA**

**ABSTRACT**

In this brief review of [1], we will describe a method for environmental audio tagging using unsupervised feature learning based on deep models. This model can be decomposed into two parts : a feature learning model using a denoising auto-encoder and an acoustic model using a Deep Neural Network (DNN).

## 1 Introduction

The objective of environmental audio tagging is to detect the presence or absence of certain acoustic events known as labels. DNNs are acknowledge as powerful models that perform remarkably well across a range of prediction tasks, ranging from language translation to image classification. Thus it is natural to try to use it for audio tagging.

However, audio tagging faces many challenges for DNNs : only the chunk-level instead of frame-level labels are available in the audio tagging task; multiple acoustic events could occur simultaneously with interfering background noise; the tags are weakly labeled and not accurate through the multiple voting scheme; there are lots of related audio files without labels on the web.

Thus, it is still not clear yet what would be the appropriate input features, the objective functions and the model structures for deep learning based audio tagging.

The article [1] proposed to learn feature with an asymmetric or asymmetric deep denoising auto-encoder (syDAE or asyDAE) based unsupervised method to generate a new feature from the basic features. Then, it will be possible to use a DNN acoustic model composed of a large set of contextual frames of the chunk which are fed into the DNN to perform a multi label classification for the expected tags

Our presentation will follow the structure of the architecture presented in figure 1 by first presenting the feature learning and then the acoustic modelling.

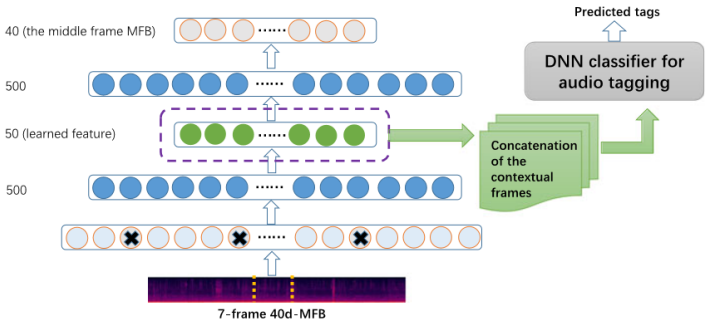


Figure 1: Presentation of the architecture of the solution (figure 3 from [1] )

## 2 Feature learning

In order to extract features from sound, we propose to first apply melfilbanks techniques to our signal and then reduce its dimension by using a auto-encoder Neural Networks.

### 2.1 MFCC and Melfilterbanks

MFCC and Mel filter banks are commons techniques in order to extract features from sound[1].

**The mel-frequency cepstrum (MFC)** is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

**Melfilterbank (MFB)** applies the Mel-frequency scaling, which is perceptual scale that helps to simulate the way human ear works. It corresponds to better resolution at low frequencies and less at high. Using the triangular filter-bank helps to capture the energy at each critical band and gives a rough approximation of the spectrum shape, as well as smooths the harmonic structure.

### 2.2 Deep Auto-Encoder

We propose to use a symmetric or asymmetric deep denoising auto-encoder to generate new data-driven features from the logarithmic Mel-filter banks features. It is composed of two parts : an encoding module : the encoder and a decoding module : the decoder.

The encode takes as input the MFB features as input and produce a smaller representation of the data as output. These data are then transfer to the decoder which try to reproduce the original MFB features. Finally, it is the output of the encoder which will serve us as input for the acoustic modelling. In addition, we add a Dropout technique which randomly selects nodes to set their values to zero in order to force the hidden layer to be more robust and prevent it from simply learning the identity function.

The new features learned by the DAE enable to mitigate the impact of background noise and compact representation of the contextual frames is needed for the reason that only chunk-level labels are available.

## 3 Acoustic modelling

Once the new features are produced by the encoder, we can use a more classical Deep Learning Architecture in order to process our classification task.

### 3.1 Deep Learning Architecture

They choose a classical architecture for multi labels classification tasks with noised input : a bootleneck shape for the hidden layer size to reduce the dimension through deep, binary cross-entropy loss, SGD + mini batches optimize, dropout.

To enable this noise adaptation, the DNN is fed with the main audio features appended with an estimate of the background noise. In this way, the DNN can utilize extra on-line information of background noise to better predict the expected tags. The specificity of this network is in the use of a background noise aware training in order to reduce the impact of the noise on the prediction.

## 4 Conclusion

The focus of the article [1] is on the denoising and the dimension reduction of their data. It leads them to good results on audio tagging with weak labels. The specificity of their network is not on the architecture of their neural network but rather on the good pre processing of their data which combines the Mel filterbank with a Deep auto-encoder and a background noise aware training in order to increase to robustness of their data and decrease the impact of the noise. However this article was published in 2016 which can be seen as "old" in regards to the rapidity of the competitivity in DNN research. Indeed, the same team achieve to have better results last year : [2]. They use an attention mechanism [3] in order to learn from the sequence order of the frames.

[1] I use MFCC and MFB in another project of classification of segmented voice for the course of speechs and text algorithms : `https://colab.research.google.com/drive/1yI-VacoAmKdxxVZeLWfiLOj_gGLL9aNs` `https://www.overleaf.com/read/yyxtvwtwcxvv`

## References


[1] Yong Xu, Qiang Huang, Wenwu Wang, Peter Foster, Siddharth Sigtia, Philip J. B. Jackson, and Mark D. Plumbley. Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1230–1241, June 2017.

[2] Qiuqiang Kong, Changsong Yu, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D. Plumbley. Weakly Labelled AudioSet Tagging with Attention Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1791–1802, November 2019. arXiv: 1903.00765.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017.