Voice Actitivity Detection with a Deep Neural Network

"THE HUMAN VOICE IS THE FIRST AND MOST NATURAL MUSICAL INSTRUMENT, ALSO THE MOST EMOTIONAL. ”

KLAUS SCHULZE

PISTER ALEXIS

NOVEMBER 2018

Contents

Introduction

Voice Activity Detection (VAD) is a technique of speech processing which aim to localize the presence of human voice in an audio sample. It is an important step of speech processing and enable the construction of various speech-based application such as speech recognition. Various techniques have been proposed to tackle this problem, such as unsupervised models that use thresholds upon certain feature values and supervised methods which train a classifier with different features. Support Vector Machines (SVM) and Gaussian Mixtures Models (MM) have both found success [Ryant et al., 2013], but deep learning methods seem to give higher scores, due to their high discrimination capability [Mendelev et al., 2015]. The process of VAD usually consists of 3 steps : features extraction, classification and decision smoothing [Ramirez et al., 2007]. The purpose of this document is to propose a voice activity detection system from the sampling of the audio data to the decision smoothing module. A multilayer perceptron (MLP) model is proposed as a classifier, which is a certain type of deep neural network (DNN).

1 Pre-processing

1.1 The data

957 audio files of different people speaking with some silences are used as dataset for the model. These files does not contain much noise and have a duration of around 10 seconds each. Times of speech are known and saved in a json format. It must be noted that this type of data facilitates the classification and does not necessarily reflect raw noisy data which is more often encountered in real life applications.

1.2 Sampling

As audio constitute an analog signal, there is the necessity of transposing it into digital data before any possibilities of analysis. Indeed, the neural network or any classifier can not receive the raw audio files directly as an input. The audio data must beforehand be sampled : at a timing interval, the amplitude of the wavelength describing the sound is saved as a digit. This enable to pass from a continuous unknown function into a discrete array. The frequency at which we sample the file is called the sampling frequency and is exprimed in Hertz. The present files have been sampled at a frequency of 16000 Hz. It means that we extract a value each $\frac{1}{16} = 0.0625$ millisecond. A 10 seconds audio file then corresponds of on array of 160000 values

1.3 Features extraction

Once the audio is sampled, it can not be fed to the network yet. Indeed, this type of data would be very weak-resistant to noise and the model may not differ human voice to other types of sounds. This is why the input signal must be split into multiples frames described by a computed feature

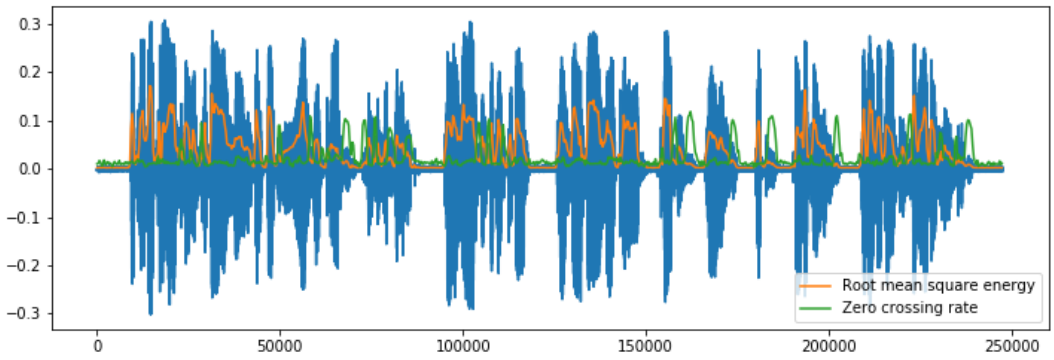


Figure 1: Root-mean-square energy and zero-crossing rate calculated along a 15 seconds audio signal of human voice with no noise

vector. Several types of features have been proposed in the literature : time domain, frequency domain, perceptual and windowing features. We used as a feature vector a combination of 13 cepstral coefficients (MFCC), the zero-crossing rate and the root-mean-square energy (rmse) calculated each 20 milliseconds. These have seemed to give good results in the literature [Qi and Hunt, 1993]. It can be noted that a frame of 20 milliseconds corresponds of 320 values with a sampling rate of 16000 Hz.

A plot of these features along a 15 seconds audio signal of human voice with very low noise is shown in figure 1. It shows that there is certainly a high correlation between these features and the presence of human voice.

2 Decision module

The deep neural network architecture have been inspired by [Qi and Hunt, 1993] and [Ryant et al., 2013]. It is made up of 2 hidden layers each composed of 20 ReLU units and one output layer using a softmax activation function. The output layer is composed of 2 neurons, one for each label : speech or non-speech frame. The network is fed with the features vectors of size 15. All layers are fully linearly connected to their inputs and possess a bias. All weights are initialized with an the follow uniform law : $w_i \sim \mathcal{U}(-\frac{1}{\sqrt{size_i}}, \frac{1}{\sqrt{size_i}})$. w_i and $size_i$ correspond of the weights and the size of the input of the i^{th} layer. The model was trained with 80 % of the available files, corresponding of 472480 features vectors. It used back-propagation with mini-batch gradient descent with a mini-batch size of 10, a learning rate of 0.005 and a momentum of 0.9. It was tested on the 20% remaining files.

3 Decision smoothing

Before comparing the output of the network with the target labels (speech or non speech), a 3-point median-filter is applied to the output vector. Each value is transformed as the median of

itself and its 2 side values. It correspond to the decision smoothing of the VAD and is meant to correct errors of the decision module.

4 Results

4.1 Classification rates

Three error metrics have been computed to evaluate the model :

- Error rate (ER) : ratio of misclassification
- Miss rate (MR) : ratio of speech frames classified as non-speech frames
- False alarm rate (FAR) : ratio of non-speech frames classified as speech frames

ER	MR	FAR
0.10	0.04	0.06

Table 1: Results of the classification

The errors metrics of the test set is given in Table 1. Only 10 % on the frames were misclassified which seem to be a good result. Results of the classification on a example file is shown in Figure 2.

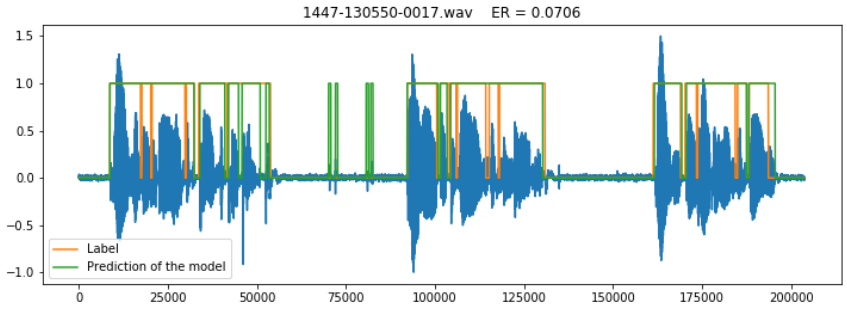


Figure 2: VAD result of an example audio file. 1 correspond to 'speech' and 0 to 'non-speech'

4.2 Impact of the decision smoothing

It is hard to get the best neural network possible for a particular task given the high number of hyper-parameters possibilities. However the impact of the decision-smoothing module on the classification can be easier to evaluate. The impact of the three-points median filtering is visualized on a single audio file in Figure 3. In this example it is particularly visible than 2 isolated non-speech frames had been misclassified by the model around the 75000th value. The smoothing process corrected these misclassifications given the context, enhancing the error rate from 0.084 to 0.078.

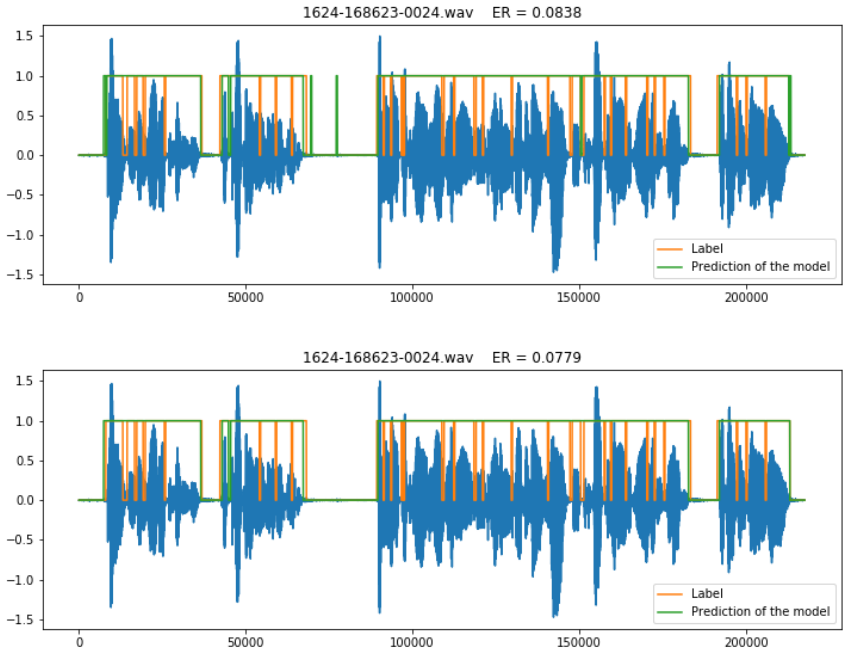
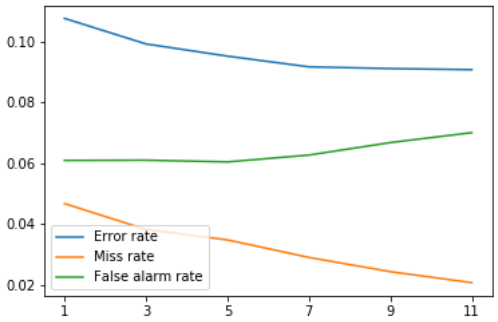


Figure 3: VAD result of an audio file before (top) and after (bottom) the application of a 3-points median filter

The application of a 3 points median-filter seem to reduce the error rate. The impact of the window of the filter has been evaluated by testing different values on the complete test set. The results are shown in Figure 4. Increasing the median filter window from 1 to 11 seem to reduce the global error rate from 0.106 to 0.090 (reduction of 1.6 %). Increasing more the window would probably not reduce the ER anymore at the view of the curves.



Window	ER	MR	FAR
1	0.106	0.043	0.063
11	0.090	0.022	0.068

Figure 4: Impact of the median filter window on the error metrics

5 Discussion

The voice activity detection pipeline proposed here is overall effective with a maximum error rate of 0.090. However, the data used was pretty much noise-free and the results of some audio files tend to the hypothesis that the model proposed is not very noise-resistant (figure 2). Some methods have been proposed to increase the robustness of deep learning models such as maxout activation functions [Mendelev et al., 2015]. Moreover, this model takes as input the features computed frame by frame individually. The global context of the frame evaluated is not taken into account and it can be hard to evaluate if a frame corresponds of speech on such a small scale. Some models aim to use this type of information into their classification module [Ryant et al., 2013] [Liao, 2013]. In the future, these types of modifications could be added to the model. Furthermore, other types of decision smoothing could be used such as minimum speech time and hangover scheme, and a exploration of the best neural network hyper-parameters (weights initialization, number of layers, activation functions, momentum ...) could improve the model classification.

References

[Liao, 2013] Liao, H. (2013). Speaker adaptation of context dependent deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7947–7951. IEEE.

[Mendelev et al., 2015] Mendelev, V. S., Prisyach, T. N., and Prudnikov, A. A. (2015). Robust voice activity detection with deep maxout neural networks. *Modern Applied Science*, 9(8):153.

[Qi and Hunt, 1993] Qi, Y. and Hunt, B. R. (1993). Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. *IEEE Transactions on Speech and Audio Processing*, 1(2):250–255.

[Ramirez et al., 2007] Ramirez, J., Górriz, J. M., and Segura, J. C. (2007). Voice activity detection. fundamentals and speech recognition system robustness. In *Robust speech recognition and understanding*. InTech.

[Ryant et al., 2013] Ryant, N., Liberman, M., and Yuan, J. (2013). Speech activity detection on youtube using deep neural networks. In *INTERSPEECH*, pages 728–731.