ViVoVAD: a Voice Activity Detection Tool based on Recurrent Neural Networks

Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

Voice Input Voice Output Laboratory (ViVoLab)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: {*pablogj, ivinalsb, ortega, amiguel, lleida*}*@unizar.es*

Abstract

Voice Activity Detection (VAD) aims to distinguish correctly those audio segments containing human speech. In this paper we present our latest approach to the VAD task that relies on the modelling capabilities of Bidirectional Long Short Term Memory (BLSTM) layers to classify every frame in an audio signal as speech or non-speech.

Introduction

VAD is broadly applied in different speech processing applications such as Automatic Speech Recognition (ASR), speaker identification or speech enchancement. A large number of different techniques have been proposed for this task, from unsupervised approaches based on energy [1], or based on long-term spectral divergence (LTSD) [2], to supervised approaches using Gaussian Mixture Models (GMM) [3]. More recently, deep learning approaches have attracted great research interesest [4], going from Convolutional Neural Networks (CNNs) to Recurrent Neural Networks (RNNs).

RNNs are able to capture temporal dependencies introducing feedback loops between the input and the output of a neural network. The Long Short Term Memory (LSTM) network [5] is a special kind of RNN that has become quite popular due to its capability to capture simultaneously long and short term dependencies. Specifically, Bidirectional LSTMs (BLSTMs) combine two different LSTM networks working on the same sequence: the first one carrying out a causal analysis and the second one carrying out an anticausal analysis. Some previous research work has proven the performance of these kind of models in the VAD task [6].

System Description

Our proposed VAD system uses Bidirectional LSTMs as the main component of the system. The task is treated as a binary classifier, with our approach consisting of two different blocks: a first feature extraction step, and the RNN classifier. Both of them are described below:

Feature extraction

The input features for the neural network consist of log Mel filter bank energies. Furthermore, the log energy of each frame can be also considered. Features are extracted every 10 ms using a 25 ms window. Feature mean and variance normalization is applied at file level.

Recurrent Neural Network

The neural architecture proposed can be seen in the Figure 1. As shown, it is mainly made of one or more stacked BLSTM layers. The final BLSTM layer output is then indepently classified by a linear layer sharing their weights for all time steps. In order to reduce the delay of the dependencies, training and evaluation is performed with limited length sequences of 300 frames (3 seconds). However, a VAD label is emited for every processed frame, which is equivalent to one label every 10 ms in our case.

Adaptative Moment Estimation (Adam) optimizer is chosen due to its fast convergence properties. Furthermore, an exponential decay learning rate is implemented to ensure smooth convergence. Data will be shuffled in each training iteration aiming to improve model generalization capabilities. All the neural architectures have been evaluated using the PyTorch [7] toolkit.

Experimental Results

To evaluate the performance of our proposed VAD system, two different datasets are considered: The RNN is trained with the Albayzin 2010 evaluation TV3 dataset [8]. This data consists of around 84 hours of broadcast news from the Catalan public television. Final metrics and performance is reported in the Albayzín 2018 RTVE database, broadcast data with emissions from the Spanish public broadcast company.

Results obtained on the RTVE 2018 dataset can be found in Table 1. For this set of experiments, we are using 32 Mel log energies and the log energy of the frame. As it can be seen, increasing the number of BLSTM layers and neurons does not show a significative improvement in performance.

Table 2 shows a comparison between our proposed BLSTM VAD and two non-supervised techniques: an energy based VAD extracted using the Kaldi ASR state-of-the art toolkit [9], and a LTSD based system, tradionally used in speaker recognition applications. Our proposal outperforms both systems, achieving a relative improvement of 40.1% compared to the energy based system.

Conclusions

In this paper we have presented our latest approach to the VAD task using RNNs as the main component of our system. The system has been tested on a broadcast environment with competitive results, significantly outperforming other techniques previously proposed in the literature.

REFERENCES

[1]. WOO, Kyoung-Ho, et al. Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters*, 2000, vol. 36, no 2, p. 180-181.

[2]. RAMIREZ, Javier, et al. Voice activity detection with noise reduction and long-term spectral divergence estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004. p. ii-1093.

[3]. NG, Tim, et al. Developing a speech activity detection system for the DARPA RATS program. In *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[4]. ZHANG, Xiao-Lei; WANG, DeLiang. Boosting contextual information for deep neural network based voice activity detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2016, vol. 24, no 2, p. 252-264.

[5]. HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. *Neural computation*, 1997, vol. 9, no 8, p. 1735-1780.

[6]. EYBEN, Florian, et al. Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013. p. 483-487.

[7]. PASZKE, Adam, et al. Automatic differentiation in PyTorch. In *Advances in Neural Information Processing Systems*, 2017. p.1-4

[8]. BUTKO, Taras; NADEU, Climent. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011, vol. 2011, no 1, p. 1.

[9]. POVEY, Daniel, et al. *The Kaldi speech recognition toolkit*. IEEE Signal Processing Society, 2011.

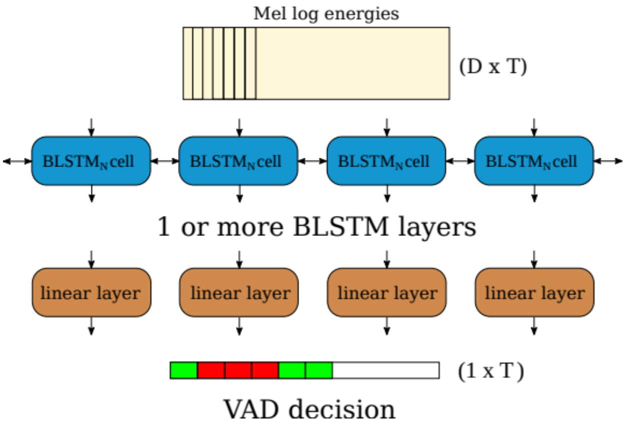

Figure 1. Neural architecture description of the proposed Voice Activity Detection system

Table 1. BLSTM VAD results on the RTVE 2018 eval dataset for different number of BLSTM layers and neurons

	Error (%)		
	False alarm	Miss	Total
1BLSTM 128 neurons	2.50	1.40	3.90
1BLSTM 256 neurons	2.40	1.50	3.90
2BLSTM 128 neurons	2.60	1.40	4.00
2BLSTM 256 neurons	2.30	1.90	4.20

Table 2. VAD results on the RTVE 2018 eval dataset using different non-suppersived VAD system compared to our BLSTM approach

	Error (%)		
	False alarm	Miss	Total
Energy VAD	4.2	2.4	6.60
LTSD VAD	2.30	16.80	19.10
BLSTM VAD	**2.40**	**1.50**	**3.90**