A Comparison of Boosted Deep Neural Networks for Voice Activity Detection

Harshit Krishnakumar and Donald S. Williamson
Department of Computer Science, Indiana University, Bloomington, IN, USA

Problem Statement

Voice activity detection (VAD) is an integral part of speech processing for real world problems, and a lot of work has been done to improve VAD performance. Of late, deep neural networks have been used to detect the presence of speech and this has offered tremendous gains. Unfortunately, these efforts have been either restricted to feed-forward neural networks that do not adequately capture frequency and temporal correlations, or the recurrent architectures have not been adequately tested in noisy environments.

Proposed Idea

We investigate different neural network configurations for voice activity detection. More specifically, we explore solutions that incorporate multi-resolution stacking and ensemble learning using convolutional, long short- term memory (LSTM), and dilated convolutional neural network architectures.

Boosted Deep Neural Network (bDNN)

1. bDNN is an ensemble based model that makes multiple predictions for each frame and averages the results
2. Each frame in the input is expanded into 2W+1 frames surrounding the central frame and passed into DNN

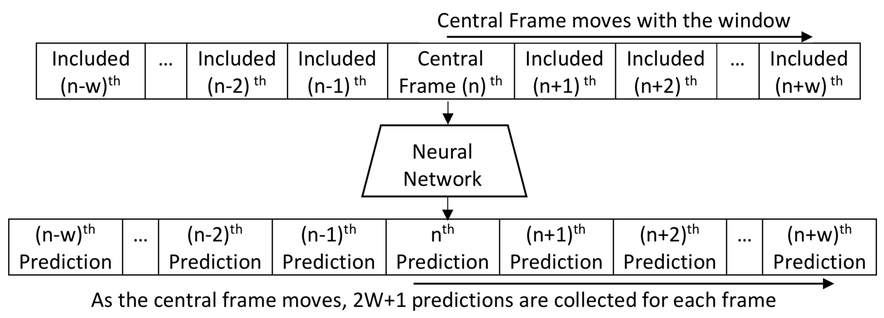

Figure 1: An example one-dimensional input passed into a boosted DNN

Multi-Resolution Stacking (MRS)

- MRS incorporates varying number of neighboring frames with layered bDNNs for better contextual prediction
- We use two MRS layers. The original input is combined with predictions when passing into the second layer

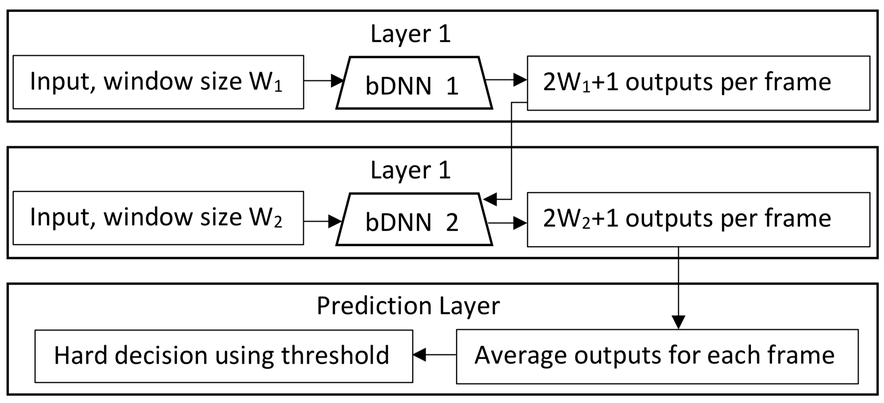

Figure 2: A Multi-Resolution Stacking model with a boosted DNN as the base predictor. The inputs are boosted with different window lengths, and outputs are passed across layers

Proposed MRS Network Architectures

- We propose to use different neural network architectures for the boosted DNN (CNN, Dilated CNN and LSTM)
  - CNN with 1 input and 2 hidden convolutional layers and 1 output layer (128 filters)
  - Dilated CNN model with similar architecture as the CNN
  - LSTM model with 1 input and 1 hidden LSTM layers, 32 units each, 2 hidden dense layers and 1 output layer
- CNN models capture information across small time-frequency segments
- LSTM captures temporal correlations

Experimental Data

- Two Datasets are used: AURORA2 [17] and IEEE [19] speech corpora
- AURORA2 uses 700 random male and female speaker utterances
  - Speech and babble noise at a -5 dB signal-to-noise ratio (SNR)
  - Data is randomly split into 300 training and 400 testing utterances
  - VAD labels: statistically-based algorithm (Sohn *et al.* [18]) is applied to clean speech.
- IEEE corpus contains 1440 speech utterances
  - *Training Noises*: babble, restaurant, factory, traffic and train noises
  - *Unseen Testing Noises*: helicopter and radio static noises
  - *SNRs*: -5, 0 and 5 dB

Experimental Results

1. **Task 1**: Comparing different boosted DNN configurations
   - AURORA2 noisy speech dataset

Table 1: Area under the curve (Receiver operating characteristic) results.

Approach	AUC
DNN	81.7%
CNN	81.9%
Dilated CNN	82.9%
LSTM	**83.5%**

2. **Task 2**: Results when varying key parameters.
   - The number of layers in the DNN
   - The number of filters in the CNN
   - The dilation rate in the dilated CNN

Table 2: Results when varying key parameters

DNN		CNN		Dilated CNN	
# Layers	AUC	# Filters	AUC	Dilation	AUC
1	81.4%	32	**81.9%**	2	**82.9%**
2	81.7%	64	81.6%	4	82.7%
3	81.3%	128	81.6%	6	**82.9%**
4	**82%**				

3. **Task 3**: Unseen Noises and Multiple Signal-to-Noise Ratios (SNRs)
   - IEEE dataset for training and testing

Table 3: VAD Results using IEEE Speech Corpora, unseen noises and multiple SNRs

Model Type	SNR (dB)	Seen Noises	Unseen Noises	
			Helicopter	Radio
LSTM	-5	79.5%	**73.6%**	**74.8%**
	0	**87.1%**	**83.4%**	**84.8%**
	5	**91.4%**	**89.9%**	**90.7%**
Dilated CNN	-5	**80%**	**73.6%**	73.1%
	0	86.9%	82.8%	81.1%
	5	91%	89.5%	88.4%
DNN	-5	78.3%	67.8%	69.7%
	0	86.4%	77.2%	78.2%
	5	90.9%	86.2%	85.5%

Conclusion

- We have proposed the use of different ensemble neural network configurations for voice activity detection.
- An LSTM ensemble approach provided the best results, showing that retaining contextual information is key.
- Our LSTM model also demonstrated excellent generalization performance on both seen and unseen results.
- This opens up the path for further work to explore alternative recurrent-ensemble approaches and attention-based models.

williads@indiana.edu