A Comparison of Boosted Deep Neural Networks for Voice Activity Detection

Harshit Krishnakumar and Donald S. Williamson

Department of Computer Science, Indiana University, USA
harkrish@iu.edu, williads@indiana.edu

***Abstract*—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. In this paper, 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. We evaluate our approach using various speech signals that are captured in different amounts of noise. Our results show that a multi-resolution ensemble approach using LSTM recurrent neural networks performs best. This is demonstrated for seen and unseen testing scenarios.**

***Index Terms*—Voice activity detection, multi-resolution stacking, deep neural networks, ensemble learning**

I. Introduction

Voice activity detection (VAD) has been a topic of interest for several decades, where telephone companies originally wanted to detect the presence of speech in audio signals. This technique has been used in a wide range of applications including telemarketing, conference calling and digital voice assistants. Given the recent growth of Voice over Internet Protocol (VoIP) applications and the burst of connected devices that are increasingly getting voice calling functionality, accurately detecting speech has piqued in interest amongst researchers.

One of the first VAD implementations was used for digital mobile telephone service, where long and short-term filters were used to identify periods of noise and block the transmission of those parts [1]. The authors found that in a typical conversation, a speaker talks for only about 40% of the time, so the noise parts were removed from transmission. This approach was trained and tested in mobile-phone related noise conditions, such as moving cars. Human subjects validated the outputs using different scoring metrics. The approach performed better than the existing technology for high noise conditions and it became the standard for Pan-European digital-mobile telephone systems. Srinivasan and Gersho [2] developed VAD models for cellular networks in stationary vehicular noise and time-varying babble noise. The authors use the VAD model from [1], but they include additional features, such as, energy-level comparisons in individual frequency sub-bands, measurement of spectral flatness of the output signal after noise suppression and an adaptive hangover period. The varying-noise VAD model uses energy levels and the percentage of energy in the low-frequency bands. The authors found that the stationary-noise model performed better in low signal-to-noise ratio (SNR) conditions than the existing methods. Haigh and Mason used cepstral features for VAD [3]. The authors proposed using a form of cepstral analysis, Perceptual Linear Prediction, that looks for variations between speech and noise cepstra. The results of this model are shown to be better than comparison models, since this approach achieves similar performance and generalizes better without knowing specific speech and noise levels.

Deep neural networks (DNN) have been used recently for VAD. In [4], the authors explore deep belief networks (DBNs) for VAD, where the authors argue that conventional machine learning models cannot capture the nonlinear properties of speech. A DBN with nonlinear hidden layers account for non-linearity, overfitting and local maxima problems. This model was tested under different seen and unseen noise profiles like factory, vehicular, street and white noise, and it outperformed alternative machine-learning based models. Zhang and Wu propose a denoising DNN model for VAD [5], where a DNN is initially pre-trained using an unsupervised algorithm. The model is then fine-tuned using supervised back propagation. The authors found better performance than traditional DBN based models, and observed increasing performance as the number of hidden layers increased. Recurrent neural networks have also been used for voice activity detection [6], [7]. Boosted deep neural networks (bDNN) with multi-resolution cochleagram (MRCG) features [8] and multi-resolution stacking (MRS) [9] generated further performance gains [10]. The authors propose a three-tiered model for boosting contextual information in voice activity detection. The top level employs an ensemble learning framework with MRS. The middle level uses boosted DNNs, which modify the input feature and output training target of a feed-forward neural network, to account for temporal correlations. The bottom level uses signal-processing techniques for feature extraction. The authors trained and tested on different types of seen and unseen noises. They observed great generalization performance, especially in unseen test scenarios.

Boosted DNNs for VAD is a promising direction, but it is

978-1-7281-2723-1/19/$31.00 ©2019 IEEE

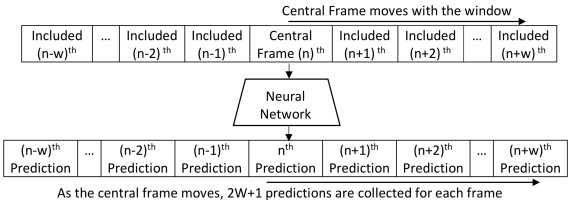

Fig. 1. An example one-dimensional input passed into a boosted DNN

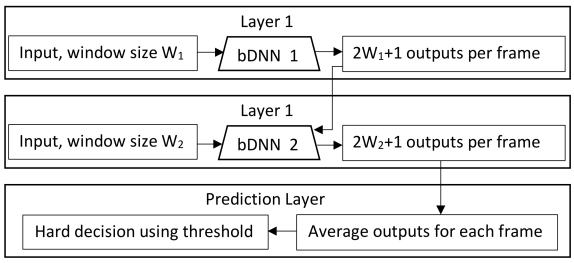

Fig. 2. A two-layered Multi-Resolution Stacking model with a boosted DNN as the base predictor. The inputs are boosted with different window lengths, and outputs are passed across layers

unclear on whether a feed-forward deep neural network is the best architecture. In this paper we use a similar three-tiered model, but we experiment with different types of neural network base classifiers. More specifically, we explore different boosted and multi-resolution configurations of convolutional neural networks (CNNs), long short-term memory (LSTM) recurrent neural networks (RNN), and dilated CNNs. We additionally test using several noise environments. We use the area under the ROC curve (AUC) to evaluate each model. Our results show that more sophisticated models that better capture temporal correlations offer additional performance gains over traditional feed-forward architectures, even when boosting and multi-resolution data are used.

The remaining portion of this paper is organized as follows. We describe boosted deep neural networks and multi-resolution stacking in section II. Our proposed approach is described in section III. Our experiments, results, and conclusions are given in sections IV and V.

II. Boosted Deep Neural Networks and Multi-Resolution Stacking

A boosted deep neural network is an ensemble learning model that makes multiple predictions for each temporal input frame, where each frame of the input and output includes information from neighbouring frames. In the first step, each data frame is expanded to $2W + 1$ frames on the time axis by adding W frames to the left and right of the current frame. W is a user-defined half-window size. The expanded input data is fed to a neural network where the output dimension is also $2W + 1$, to account for the temporal context in training targets across multiple frames. The label of each frame is then used to train the neural network. In the testing phase, the $2W + 1$ predictions of each frame are averaged and a threshold is applied to perform classification. Fig.1 shows a bDNN example for a simple one-dimensional input.

Multi-resolution stacking builds off of bDNNs by incorporating varying number of neighbouring frames with layered bDNNs for better contextual prediction. Getting multi-resolution information adds value that improves prediction performance. Each classification task may be modeled best with a specific resolution (e.g. window size), but it is not practical to empirically determine the appropriate window size for each problem. Multi-resolution stacking eliminates this problem, as boosted DNNs with different window lengths are stacked in a layered manner [9]. Additionally, the outputs of each layer are combined with the original input before passing on to the next layer, see Fig.2.

III. Proposed Approach

We propose to use different neural network architectures for the boosted DNN. These architectures include a convolutional neural network (CNN), dilated CNN, and long short-term memory (LSTM) recurrent neural network (RNN). We compare against the original feed-forward DNN approach from [10]. All the proposed neural networks use Adaptive Moment estimation (Adam) [11] as the optimization strategy and a mean square error loss function. Adam optimization is computationally effective and better for handling large datasets in complicated models, and hence well suited for this problem.

A. Convolutional Neural Network

We first consider a CNN model, since they capture information across small time-frequency segments and they have performed well for classification tasks [12], [13]. The CNN consists of one input convolutional layer, one hidden convolutional layer, a hidden dense layer and the final output layer. Each convolutional layer has a varying number of filters (32, 64 and 128) with a kernel of size 3×3. The second convolutional layer is followed by a max pooling function with a pool size of 2×2. The ReLU activation function [14] is applied after each convolutional layer. The hidden dense layer has 512 units with ReLU activation functions, whereas the output layer uses a sigmoid function. Each layer, except the input and output, has a 20% dropout rate to prevent over fitting. Different experiments were conducted to determine these parameter values and network configuration, where the best performing approach is used.

B. Dilated Convolutional Neural Network

Dilated CNNs were introduced first by Fisher and Vladlen [15] as a way to incorporate global contextual information in convolutional layers without exponentially increasing model complexity. It builds on a regular convolutional layer and introduces a new parameter called dilation rate. Dilation rate indicates the degree of spatial separation between the samples considered for convolution. For example, a convolutional layer with a dilation rate of 2 will select every alternate sample from the input. By introducing spaces between selected samples,

a larger frame of the input (time-frequency representation of sound) is considered, which adds global context to the model while also keeping the model complexity low. Typically a regular convolutional layer is followed by one or more dilated layers to incorporate additional contextual information.

We use a dilated CNN model that consists of one convolutional input layer, two dilated convolutional layers, a hidden dense layer and the output layer. Each of the convolutional layers contains 32 filters with a 3×3 kernel size, followed by ReLU activations. Max pooling is performed after each dilated layer, where a 2×2 pool size is used. A 20% dropout rate is used in all layers, except the input and output layers. The hidden dense layer has 512 units and ReLU activation functions. The output layer uses a sigmoid activation function. Different experiments were conducted, where the dilation rate for both layers was held variable in each round and the model accuracy was measured.

C. Long Short Term Memory Network

LSTMs are a type of recurring neural network that has better memory retention capabilities than a feed forward neural network. Our LSTM network has one LSTM input layer, one hidden LSTM layer, two hidden dense layers and an output layer. Both LSTM layers have 32 units each (dropout of 20%) and use hyperbolic tangent (tanh) activations. The two hidden dense layers each have 512 units with ReLU activations. The sigmoid activation function is used in the output dense layer. We experimented with different configurations of LSTM networks and found this setup to give maximum accuracy on development data.

D. Input Features and Multi-Resolution Stacking

We use the multi-resolution cochleagram (MRCG) features as inputs to our approach, which was also done in [16]. A cochleagram provides a time-frequency representation of an audio signal, and it mimics the processing that the human ear performs. A cochleagram is greatly influenced by the selection of frame size, as this influences time and frequency resolution. Using a smaller frame size increases time resolution, but compromises frequency resolution. Similarly, using a bigger frame size increases frequency resolution, but reduces temporal resolution. The MRCG feature combines cochleagrams with two different resolutions (big and small), to ensure good temporal and frequency resolution. These two cochleagrams are each smoothed using an averaging function.

We use a two-layered MRS model, where the half-window lengths (e.g. W_1 and W_2) are equal in each layer (see Fig. 2). In the first layer, the input features are first augmented using the window length. These are provided to a bDNN (see Fig. 1), where the bDNN is either based on a CNN, LSTM, or dilated CNN model. We use an ensemble of eleven bDNNs in the first layer, where each one uses a different window length to account for varying degrees of temporal information. The eleven window lengths of W_1 are $\{1, 7, 11, \text{ and } 19 : 4 : 51\}$, where MATLAB notation (e.g. $x : y : z$) is used to take all values between 19 and 51, with increments of 4. As in [10], a different dilation rate is also used for each window length (11 dilation rates in total) to reduce computational complexity. The dilation rates are chosen so that 7 samples are contained in each window. Each of the eleven bDNNs is trained independently. The eleven bDNN outputs are then each concatenated to another copy of the input features that are augmented with the same half window length that was previously used, resulting in eleven new input features that are provided to the second MRS layer. The second MRS layer, hence, consists of eleven additional bDNNs (one for each window length), where each gets its corresponding input from the first MRS layer. The second layer's bDNN model then makes a prediction based on the modified input. The final voice activity prediction is made by averaging the predictions from the ensemble bDNNs in the second MRS layer. Note that the selected parameter values are based on recommendations from [10].

IV. Experimental Setup and Results

A. Dataset Generation

We use two different datasets to test our approach. First, we use the AURORA2 [17] speech corpora, where 700 utterances from male and female speakers are randomly chosen. These speech signals are then mixed with babble noise at a -5 dB signal-to-noise ratio (SNR). MRCG features are computed from the noisy speech data using the same parameters as in [16]. The data is then randomly split into 300 training and 400 testing utterances. The VAD labels are generated by applying the statistically-based VAD algorithm proposed by Sohn *et al.* [18] to the clean speech corpora. The mixed audio signals are merged into one long file each for training and testing. Testing is performed on partially unseen data, since we randomly split the data into training and testing sets.

The second dataset is generated from the IEEE male and female speech corpus [19]. 1440 speech utterances are split randomly into smaller batches of 144 files, and then used for training and testing the VAD model. The utterances are mixed with babble, restaurant, factory, traffic and train noises. Additionally, to test the model's performance on unseen noises, the testing clean speech data is also mixed with unseen helicopter and radio static noises. The signals are mixed at -5, 0 and 5 dB SNRs. Labels are generated for both seen and unseen datasets by passing the clean speech signals through the VAD algorithm proposed by Sohn *et al.* [18] and implemented by Kim [20].

B. Results and Discussions

Our approach is evaluated with the area under the ROC Curve (AUC), as it is often used to assess VAD performance [7], [10]. The different neural network configurations defined above are tested, where we compare against the original deep neural network (DNN) based approach as defined in [10]. The DNN in this approach contains two-hidden layers with 512 ReLU units in each layer.

Table I shows the results for the baseline DNN approach and the proposed approaches that use different neural network

TABLE I
COMPARING VAD PERFORMANCE WITH DIFFERENT NEURAL NETWORK CONFIGURATIONS ON THE -5 DB, AURORA2 DATASET

Approach	AUC
DNN [10]	81.7%
CNN	81.9%
Dilated CNN	82.9%
LSTM	**83.5%**

architectures. These systems are trained and tested with the AURORA2 noisy speech dataset, where the SNR is -5 dB. Note that each proposed configuration improves performance over the baseline DNN, as measured by AUC. More specifically, the CNN offers a slight improvement (e.g. 0.2%) over the DNN approach. This likely occurs because the CNN better captures temporal and frequency correlations during the convolutional feature mapping stage. The LSTM boosted-DNN approach results in the largest performance gains, where the AUC score is 83.5%. This is considerably better than all other approaches. Voice activity detection is highly correlated across time, and the LSTM-based ensemble approach adequately captures these temporal correlations. The dilated CNN approach also performs better than all other approaches, except the LSTM approach. This is interesting because this indicates that dilation, which expands temporal range using the same amount of context data, performs better than a CNN. This indicates that longer range information also helps with voice activity detection.

The CNN model has more than two layers, however, so performance gains over the DNN may be attributed to this. Thus, we trained a DNN with varying number of layers. Table II shows how DNN VAD performance varies with the number of layers. The results show that increasing the number of layers to 4 gives the best results, which is now slightly better than the CNN VAD performance, but still worse than the dilated CNN and LSTM approaches.

We additionally show results for the different architectures by varying key parameters. In particular, we vary the number of filters in the CNN and the dilation rate in the dilated CNN. As Table II shows, increasing the number of filters in the CNN does not improve VAD performance. Likewise, only slight inconsistent changes occur with a change in the dilation rate.

We further train and test the best performing LSTM, DNN and Dilated CNN models under more noise and SNR conditions. We use the IEEE male and female speech utterances that are mixed with seen and unseen noise profiles. The clean speech corpora are mixed with babble, forest, restaurant, traffic and train noise profiles, and this data is divided into seen training and testing sets. To test the model's performance on unseen noise profiles, the testing clean speech data is mixed with helicopter and radio static noise, while keeping the training signals constant. This data is generated at -5, 0 and 5 dB SNRs. We measure the AUC under different noise profiles and average the results across the respective SNR levels. The results of the VAD models on seen and unseen noise profiles are presented in Table III. As a general trend, we note that higher SNR levels have better VAD performance, which is

TABLE II
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%				

TABLE III
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%

expected, since the detection will be more accurate with higher levels of speech in the mixture. Upon comparing the results of different neural network types, we observe that the LSTM has the best overall performance. More specifically, the LSTM performs best for the seen noise case at 0 and 5 dB SNRs, and it performs best at all SNRs for the unseen testing cases. However, the performance of Dilated CNN is close to that of the LSTM model, especially in lower SNR levels. The DNN performs worse among the three comparison approaches, in all categories, where the performance gap compared to the LSTM is greater under unseen noise conditions.

V. CONCLUSION

We have proposed the use of different ensemble neural network configurations for voice activity detection. We obtained a significant performance increase when compared to a comparison approach [10]. The use of an LSTM ensemble provided the best results among all other types of networks, showing that retaining contextual information is best suited for this problem. Our LSTM model also demonstrated excellent generalization performance on both seen and unseen results at different signal to noise levels. This opens up the path for further work to explore alternative recurrent-ensemble approaches and for attention-based models to obtain further performance gains, since they have been recently shown to handle contextual information better.

ACKNOWLEDGMENT

The authors would also like to thank X. Zhang, D. Wang and E. Kim for answering questions and providing software.

REFERENCES

[1] D. Freeman, G. Cosier, C. Southcott, and I. Boyd, "The voice activity detector for the pan-european digital cellular mobile telephone service," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1989, pp. 369–372.

[2] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Proceedings, IEEE Workshop on Speech Coding for Telecommunications*, 1993, pp. 85–86.

[3] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *IEEE Region 10 International Conference on Computers, Communications and Automation*, vol. 3, 1993, pp. 321–324.

[4] I. Hwang and J. Hyuk Chang, "Voice activity detection based on statistical model employing deep neural network," in *Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2014, pp. 582–585.

[5] X. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 853–857.

[6] P. Sertsi, S. Boonkla, V. Chunwijitra, N. Kurpukdee, and C. Wutiwiwatchai, "Robust voice activity detection based on LSTM recurrent neural networks and modulation spectrum," in *Processing of the Association Annual Summit and Conference (APSIPA ASC)*, 2017.

[7] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 483–487.

[8] X. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014, pp. 1534–1538.

[9] ——, "Multi-resolution stacking for speech separation based on boosted dnn," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 1745–1749.

[10] ——, "Boosting contextual information for deep neural network based voice activity detection," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, 2016, pp. 252–264.

[11] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015, pp. 1–15.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Conference on Computer Vision and Pattern Recognition*, 2009.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, 2012, pp. 1097–1105.

[14] G. Xavier, B. Antoine, and B. Yoshua, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15, 2011, pp. 315–323.

[15] Y. Fisher and K. Vladlen, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations*, 2016, pp. 1–13.

[16] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, 2014, pp. 1993–2002.

[17] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the ISCA workshop ASR2000*, vol. 4, 2000, pp. 29–32.

[18] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," in *IEEE Signal Processing Letters*, vol. 6, 1999, pp. 1–3.

[19] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[20] E. Kim, "A statistical model-based voice activity detector," https://github.com/eesungkim/Voice_Activity_Detector, 2018.