Computer Speech & Language 65 (2021) 101137

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

# A neural network approach for speech activity detection for Apollo corpus

Vishala Pannala*, B. Yegnanarayana

*Speech Processing Laboratory, International Institute of Information Technology, Hyderabad, India*

ARTICLE INFO

*Article History:*
Received 14 July 2019
Revised 11 June 2020
Accepted 9 July 2020
Available online 30 July 2020

*Keywords:*
ANN model
Apollo corpus
Noisy speech
Single frequency filtering
Speech activity detection

ABSTRACT

This paper describes a new method for speech activity detection (SAD) based on the recently proposed single frequency filtering (SFF) analysis of speech signals and a neural network model. The SFF analysis gives instantaneous spectrum of the speech signal at each sampling instant. The frequency resolution of the spectrum is decided by the number of frequencies used in the SFF analysis, which in turn depends on the frequency spacing. Using a frequency spacing of 10 Hz and a sampling frequency of 8 kHz, a 401 dimensional spectrum, covering 0–4 kHz, is obtained at each sampling instant. This is used as a feature vector to train an artificial neural network (ANN) model to discriminate (noisy) speech and nonspeech (mostly noise). The output of the trained ANN model for a given test utterance gives speech/nonspeech decision at every sampling instant. Post processing of the decision is used for SAD. The system generated SAD is evaluated on the Apollo corpus for SAD task in terms of detection cost function (DCF). The DCF values of the proposed system on the development and evaluation datasets are 3.1% and 4.6%, respectively, whereas the DCF values of the reported baseline system are 8.6% and 11.7%, respectively.



## 1. Introduction

Detection of speech regions in audio signals is the first step in developing any speech processing system such as language identification, speech and speaker recognition. The speech activity detection (SAD) is a challenging task because the collected speech is generally degraded by environmental conditions such as noise, reverberation and other audio signals including background speech. The collected speech is also degraded by the communication and recording channels. One such degraded speech data is the Apollo speech corpus. The SAD task for the Apollo corpus described in (Fearless Steps Challenge, 2019; Kaushik et al., 2018) is addressed in this paper.

The SAD has been studied extensively, exploiting mostly the characteristics of speech, although in a few cases the characteristics of degradations, i.e., noise and channel, are also considered (Tanyer and Ozer, 2000; Morita, Shota and Unoki, Masashi and Lu, Xugang and Akagi, Masato, 2016; Sangwan and Zhu and Ahmad, 2005; Choi and Chang, 2014; Park et al., 2016). There are approaches based mainly on (a) signal processing (Aneeja and Yegnanarayana, 2015; Sadjadi and Hansen, 2013), (b) statistical characterization of speech and noise features (Sohn et al., 1999; Ghosh et al., 2011), and (c) neural network models (Hughes and Mierle, 2013; Zhang and Wu, 2013), for discriminating speech and noise. More recently, data driven deep neural networks (DNNs) are also explored for SAD (Kaushik et al., 2018; Zhang and Wu, 2013). Significant results have been obtained in all these cases. But the studies in most of the cases are limited either to standard databases provided for evaluation (Ziaei et al., 2015), or

*Corresponding author.
*E-mail addresses:* p.vishala@research.iiit.ac.in (V. Pannala), yegna@iiit.ac.in (B. Yegnanarayana).

https://doi.org/10.1016/j.csl.2020.101137
0885-2308/

on specific tasks as in DARPA RATS for different channel distortions (Thomas et al., 2015). The approaches for SAD task for naturalistic data such as Apollo corpus poses a different challenge due to different types of degradation (Kaushik et al., 2018; Ziaei et al., 2014).

Most SAD methods represent speech using standard features such as filterbank coefficients or mel frequency cepstral coefficients (MFCCs), computed for every frame of size 10− 30 msec with a frame shift of 5−10 msec (Vlaj et al., 2005). In a few cases features representing the excitation source, like the fundamental frequency ($F_0$) and formant contours are appended to the standard features (Sadjadi and Hansen, 2013). In Kinnunen et al. (2007), the MFCC features along with support vector machine classifier were used to discriminate speech and nonspeech regions. Although it was shown to give good performance, its performance was limited to the trained noise conditions. A new feature, called multi-resolution cochleagram (MRCG), was proposed for SAD under low signal-to-noise (SNR) conditions (Chen et al., 2014). The feature combines four cochleagrams at different spectral and temporal resolutions to capture the local and contextual information. The regions with dominant time-frequency (TF) unit in the multiple cochleagrams are considered as speech regions. The MRCG features with delta and double delta features were exploited with boosted DNN (bDNN) to perform SAD. The bDNN generates multiple predictions from a single DNN (Zhang and Wang, 2014). A stack of classifiers are explored in Zhang and Wang (2015), where each classifier explores the context in different resolutions targeting different predictions, which are integrated with other classifiers. In Wang et al. (2015), one DNN for feature mapping and one DNN for classification of speech and noise are integrated, to train jointly for speech nonspeech discrimination. In Kim and Hahn (2018), adaptation of the context information according to the noise type and noise level are studied. Some features are also derived using data-driven dimension reduction (Pitsikalis and Maragos, 2009). Fusion of multiple features and models are also adopted for developing SAD (Thomas et al., 2015; Zhang and Wu, 2013).

In this paper we propose a new representation of speech information using the single frequency filtering (SFF) analysis of speech signals (Aneeja and Yegnanarayana, 2015). This representation gives instantaneous spectral information, aggregated over the past data, as in filtering. The instantaneous spectral information captures both the excitation information in the form of harmonics and the system characteristics in the form of shape of the spectral envelope. Hence it is expected to preserve the speech characteristics well. The instantaneous spectra of noisy speech and of only noise (additive degradations) are discriminated by training an artificial neural network (ANN) model. Post-processing of the ANN output is done for the SAD task as per the guidelines for performance evaluation given in Fearless Steps Challenge Evaluation Plan v1.2 (2019). The proposed SAD method also illustrates a way of dealing with the inconsistencies in human annotations while generating the ground truth.

In general, the data driven DNNs are expected to extract the relevant features and then classifying in one trained model (Zhang and Wu, 2013; Hwang et al., 2015; Ferroni et al., 2015; Zazo et al., 2016). Hence standard features are used for representation of speech in these models. On the other hand, in the present study, the role of ANN is mainly for classification, and the objective is to study the effectiveness of the proposed instantaneous SFF spectrum for the SAD task on Apollo corpus data.

The paper is organized as follows. Section 2 describes the database, the proposed SAD challenge and the baseline system used for comparison. Section 3 gives details of the proposed method, which includes a brief description of the SFF analysis, and training the ANN model with small amount of data (10 minutes) from the development set (of 20 hours). Section 4 gives details of processing the output of the trained ANN model using the given ground truth on the development data, to arrive at suitable post processing for decision making for SAD. Section 5 discusses the performance of the proposed method of SAD on the complete development (dev) and evaluation (eval) data sets. The results are compared with those for the baseline method (Fearless Steps Challenge, 2019), and with a few other methods. Section 6 gives a summary of the studies reported in this paper, highlighting the specific advantage of using the SFF analysis and neural network classifier for such studies.

## 2. Description of SAD task in Apollo corpus

### 2.1. Database description Fearless Steps Challenge (2019)

Apollo data is a naturalistic audio recordings lasting for 8 days 3 hours 18 minutes and 35 seconds, in challenging environments (Kaushik et al., 2018; Fearless Steps Challenge, 2019; Fearless Steps Challenge Evaluation Plan v1.2, 2019; Sangwan et al., 2013; Hansen et al., 2018b; 2018a). This data is a subset of the NASA's Apollo-11 space mission, reflecting Air-to-Ground (CAPCOM - Capsule Communicator) communications from the Astronauts. The data represents 30 individual analog communication channels with multiple speakers in different locations working in real-time, with each channel reflecting a communication loop, with multiple instances involving rapid switching of speakers containing 3 to 33 speakers over extended time periods. In some instances, the average duration of speech of a speaker is close to 0.5 sec, with 15 speakers speaking in turns in a span of 10 sec. In some instances more than 20 speakers were active at a time, with conversations lasting for 15 minutes at a stretch, and at some other instances the silence lasts for hours.

Although audio data of 19,000 hours is available, only 100 hours of data, sampled at 8 kHz, is proposed for the SAD task. Out of this 100 hours, 20 hours of data is provided with human verified ground truth labels for speech regions. This is called development data. About 20 hours of data is used for evaluation. The SNR levels in the data, range from 0- to 20 dB. The 100 hours of data contains data from 5 channels, namely, Flight Detector (FD), Mission Operations Control Room (MOCR), Guidance Navigation and Control (GNC), Network Controller (NTWK) and Electrical, Environmental and Consumables Manager (EECOM), reflecting the channel variability during the lift off, lunar landing and lunar walking. The degradations are mostly due to channel noises, system noise, signal attenuation, transmission noise, analog tape static noise, noise due to analog tape aging, babble noise, and environmental noises. The data was collected using headmounted as well as fixed far-field microphones. The data contains rapid

switching of speakers. Also, all the speech is spontaneous. The distribution of speech and nonspeech is quite uneven. The development data was divided into 39 signals and evaluation data into 40 signals, with each signal of 30 minutes duration.

### 2.2. Development of ground truth and scoring Fearless Steps Challenge Evaluation Plan v1.2 (2019)

The development of ground truth involves markings using the baseline system discussed in Section 2.3, followed by manual verification. The manual annotations have ambiguities, due to noisy speech. The ground truth is marked with a bias more towards the speech regions, as missing the speech regions is considered more serious.

In the scoring procedure (Fearless Steps Challenge Evaluation Plan v1.2, 2019), a collar of 0.5 sec on either side of the speech region is included, and the results in the collar region are not scored. This is to take into account the inconsistencies in human annotations. If a segment of nonspeech between collars is less than or equal to 0.1 sec, then the collars involved are expanded to include the segment with less than 0.1 sec of nonspeech into the collar. If the segment of nonspeech between the collars is greater than 0.1 sec, then it is left as it is as nonspeech. An illustration of scoring procedure as given in Fearless Steps Challenge Evaluation Plan v1.2 (2019) is shown in Fig. 1 (reproduced from Fearless Steps Challenge Evaluation Plan v1.2 (2019)).

### 2.3. Description of baseline system Ziaei et al. (2014)

The baseline system uses unsupervised Combo-SAD technique which was shown to be robust for noisy conditions Sadjadi and Hansen (2013). Here, a 1-dimensional feature vector per frame is obtained using a combination of perceptual spectral flux, harmonicity, clarity, prediction gain and periodicity measures. Using the normalized values of the feature vectors, a two-mixture GMM is trained where the mixture with larger mean corresponds to speech and the other corresponds to noise. Using speech data from various sources, an ideal speech mixture is evolved. The combination of the means of the two mixtures is used to determine a threshold, to discriminate speech and nonspeech of a test utterance.

The limitations of the baseline system are, (1) frame-based features, (2) usage of 1 minute segments to normalize the features, (3) inability to handle disproportion of speech and noise in a given segment, and (4) requirement of additional speech data to determine appropriate threshold for speech classification.

## 3. Proposed method for SAD

The proposed method for SAD consists of three stages: (a) Feature extraction using SFF analysis, (b) training ANN model for discrimination of (noisy) speech and nonspeech, and (c) post processing the output of the ANN model for decision making. The first two stages are described in this section.

### 3.1. Feature extraction using SFF analysis

The single frequency filtering (SFF) analysis provides the envelope of the signal as a function of time at each selected frequency (Aneeja and Yegnanarayana, 2015). The SFF is performed by first multiplying the signal $x[n]$ with a complex sinusoid

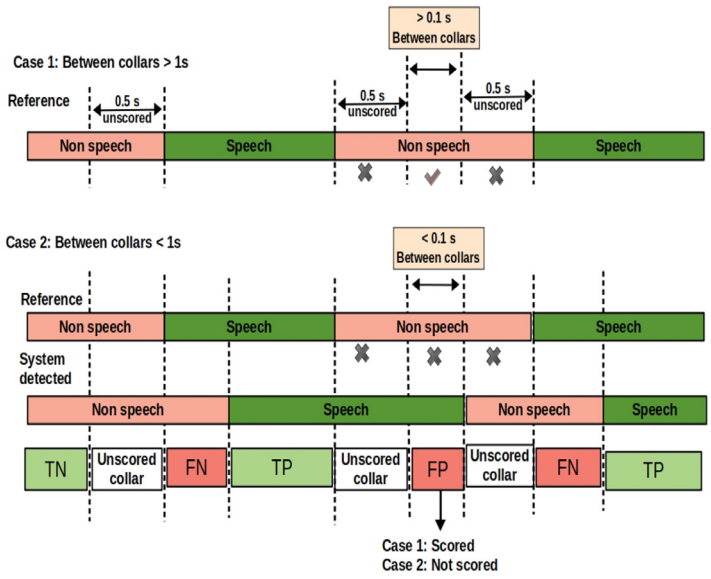


**Fig. 1.** Illustration of development of ground truth and scoring (reproduced from Fearless Steps Challenge Evaluation Plan v1.2 (2019)).

$e^{j\overline{\omega}_k n}$. The resulting frequency-shifted signal is given by

$$x_k[n] = x[n]e^{j\overline{\omega}_k n}, \tag{1}$$

where $\overline{\omega}_k = \pi - \omega_k = \pi - \frac{2\pi f_k}{f_s}$, $f_k$ is the selected frequency, and $f_s$ is the sampling frequency. The frequency-shifted signal $x_k[n]$ is passed through a single pole filter, whose pole is located on the negative real axis, close to the unit circle in the $z$-plane. This corresponds to a near ideal resonator at half the sampling frequency ($f_s/2$) of the given signal. The transfer function of the filter is given by

$$H[z] = \frac{1}{1 + rz^{-1}}, \tag{2}$$

where $r \approx 1$ for the root close to the unit circle. The output of the filter is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \tag{3}$$

The envelope $v_k[n]$ of the signal at the desired frequency $f_k$ is given by

$$v_k[n] = \sqrt{y_{kr}[n]^2 + y_{ki}[n]^2}, \tag{4}$$

where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary parts of $y_k[n]$.

For ensuring stability of the filter, $r$ is chosen to be slightly less than 1. In this study $r = 0.998$ is chosen. The envelope is computed for every 10 Hz in the frequency range of 0 to $f_s/2$, i.e., 0 to 4000 Hz, resulting in 401 envelopes. At every instant there will be values at 401 frequencies, giving an instantaneous spectrum of 401 dimensions. Note that the frequency interval, and hence the number of frequencies in $0 - fs/2$, is a parameter that can be chosen depending on the desired dimension of the feature vector. In this study, the feature vector is the 401-dimension instantaneous SFF spectrum, i.e., $e_k[n]$, for $k = 0, 1, \ldots, 400$, for each $n$. The instantaneous spectral vector is normalized by dividing each value by the sum of the values across all the frequencies. This will reduce the dependency on the amplitude and energy of the signal.

An illustration of the SFF spectrogram, in comparison with the standard short-time Fourier transform STFT spectrogram is shown in Fig. 2. The STFT spectrogram, shown in Fig. 2(b), is obtained for each frame of 50 msec, using a frame shift of one sample. It can be seen clearly that the voiced speech regions exhibit harmonic structure, which is absent in the noise regions. It can also be seen that the speech information in terms of the harmonics and the formant information over the harmonics, is preserved in the SFF spectrogram (in Fig. 2(c)) as in the STFT spectrogram. The instantaneous SFF spectrum contains information about both the excitation (in the harmonic structure) and the vocal tract system (in the envelope) in the speech regions.

Note that due to infinite impulse response (IIR) characteristic of the filter, for low damping of the filter resonance, i.e., when $r = 0.998$, the silence regions are covered with pitch harmonics from the previous voiced segments, as can be seen in Fig. 2(c). The silence regions can be seen when a lower value, say $r = 0.992$, is used, as shown in the SFF spectrogram in Fig. 2(d). All spectrogram plots are suitably scaled for better display of spectral features in color.

The main difference between STFT and SFF is that STFT is a block processing approach, and SFF is a filtering approach. In the STFT-based analysis, the spectral values are obtained by projecting the signal values onto sinusoidal components. In the SFF-based analysis, the spectral value at each frequency is obtained by filtering. The response of the filter is a decaying IIR type, where the current sample value is given maximum weightage. The frequency resolution is controlled by a single parameter $r$, i.e., the location of the pole on the -ve real axis in the $z$-plane. This does not have the windowing effects of the block-based approach.

### 3.2. Description of the ANN model

An artificial neural network (ANN) model is proposed to capture the discrimination between the instantaneous SFF spectra of speech and noise. The number of hidden layers and the number of units at each layer are dependent on the problem being addressed and the dimensions of the input features Haykin (1994). Here the problem is a classification problem, and hence two output units are used. The input features are the 401 dimensional sum normalized instantaneous SFF spectra representing the signal. The network is to be trained to obtain a 2-class decision from the large 401 dimension feature vectors. At any hidden layer, if the number of units is less than the dimension of input vector, it results in compression of the input vectors. Conversely, if the number units is more than the input dimensions, an expansion of the feature vector happens. A gradual change of the number of units from one layer to another is used for the proposed ANN model. With 401 units in the input layer, the 1$^{st}$ hidden layer is chosen as an expansion layer with 601 dimensions, which is approximately equal to 1.5 times the dimension of the input layer. The number of units in the 2$^{nd}$ hidden layer is reduced to 1/6th of the dimension of the 1$^{st}$ hidden layer, i.e., 101 units, and to 1/3$^{rd}$ of 101, i.e., 31 units in the 3$^{rd}$ hidden layer. The output layer has two units corresponding to the two classes. The activation function of the units in the output layer is linear, and it is *tanhx* for the units in the 3 hidden layers. For each input feature vector, the desired output on each of the two units is either +1 and –1 or –1 and +1, depending on whether the input contains (noisy) speech or only nonspeech (noise), respectively. Although a single output unit is adequate for this 2-class problem, we use two units, just to verify whether they give complementary outputs for each class. We use linear units, to observe the ANN outputs at each instant, although we train the input using binary decision of +1 or –1. While the choice of the structure of the ANN model is somewhat flexible, the proposed structure is chosen with the number of units in the first hidden layer for expansion of the input dimension, and all other subsequent hidden layers to have progressively decreasing number of units (Haykin, 1994).

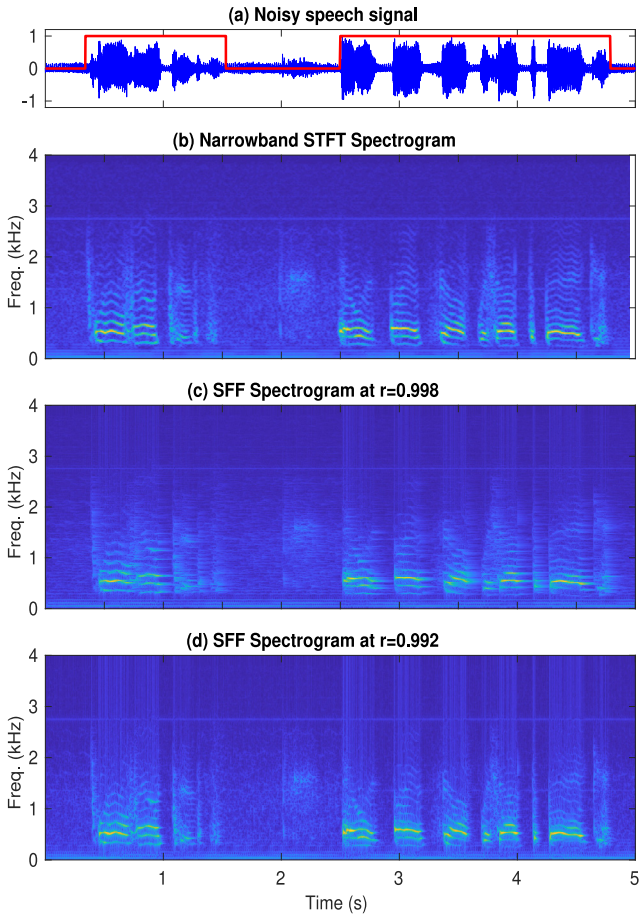


**Fig. 2.** Illustration of SFF spectrogram. (a) Noisy speech signal along with ground truth markings. (b) Narrowband STFT spectrogram. (c) SFF spectrogram for $r = 0.998$. (d) SFF spectrogram for $r = 0.992$. Note: All spectrogram plots are suitably scaled for better display of spectral features in color.

An artificial neural network (ANN) model is proposed to capture the discrimination between the instantaneous SFF spectra of speech and noise. The number of hidden layers and the number of units at each layer are dependent on the problem being addressed and the dimensions of the input features Haykin (1994). Here the problem is a classification problem, and hence two output units are used. The input features are the 401 dimensional sum normalized instantaneous SFF spectra representing the signal. The network is to be trained to obtain a 2-class decision from the large 401 dimension feature vectors. At any hidden layer, if the number of units is less than the dimension of input vector, it results in compression of the input vectors. Conversely, if the number units is more than the input dimensions, an expansion of the feature vector happens. A gradual change of the number of units from one layer to another is used for the proposed ANN model. With 401 units in the input layer, the hidden layer is chosen as an expansion layer with 601 dimensions, which is approximately equal to 1.5 times the dimension of the input layer. The number of units in the hidden layer is reduced to 1/6th of the dimension of the hidden layer, i.e., 101 units, and to $1/3^{rd}$ of 101, i.e., 31 units in the hidden layer. The output layer has two units corresponding to the two classes. The activation function of the units in the output layer is linear, and it is *tanhx* for the units in the 3 hidden layers. For each input feature vector, the desired output on each of the two units is either $+1$ and $-1$ or $-1$ and $+1$, depending on whether the input contains (noisy) speech or only nonspeech (noise), respectively. Although a single output unit is adequate for this 2-class problem, we use two units, just to verify whether they give complementary outputs for each class. We use linear units, to observe the ANN outputs at each instant, although we train the input using binary decision of $+1$ or $-1$. While the choice of the structure of the ANN model is somewhat flexible, the proposed structure is chosen with the number of units in the first hidden layer for expansion of the input dimension, and all other subsequent hidden layers to have progressively decreasing number of units (Haykin, 1994).

### 3.3. Training the ANN model

For training the ANN model, a subset of (noisy) speech and nonspeech (noise) are taken from the development data, using the given ground truth to identify those regions. Approximately, the first 30 sec segment is chosen from each of the 39 signals of the development data. Although the proportion of speech and nonspeech in this subset data is about 28:72, equal amount of speech

and nonspeech data is taken for training the network. A total of approximately 5.46 min of speech regions and 4.58 min of noise regions are used for training the network. The network is trained for 150 epochs, where the training error is reduced by a factor of over 10. The error for validation data also reduced substantially after testing with the trained network, although it is slightly higher than the error for the training data. It is to be noted that the amount of data used for training the ANN model is approximately 10 min out of 20 h of the available development data.

### 3.4. Testing the ANN model

The output of the trained ANN model on a segment of the signal containing both speech and nonspeech regions is shown in Fig. 3. The output of the two units are shown in blue and red lines. The blue line dominates in the speech regions and the red line dominates in the noise regions. Using zero as the threshold on the output of unit 1 (blue line), the regions above the threshold are marked as +1, representing speech, and the other regions as –1, representing nonspeech regions. This output is further processed (as discussed in Section 4) for scoring.

It is interesting to note that the ANN output identifies even short segments of (noisy) speech regions, due to the instantaneous spectral representation of the speech utterance using the SFF analysis.

In Fig. 3(b), the beginning point of speech can be clearly seen in the blue line. Due to infinite impulse response (IIR) nature of the single pole filter, the end point of speech can not be identified precisely. This ambiguity is reduced by using the instantaneous SFF spectra derived using $r = 0.992$ for the SFF analysis. The output of the network tested with the instantaneous SFF spectra for $r = 0.992$, is as shown in Fig. 3(c). The effect of change in $r$ parameter can be seen at the locations marked with arrows in the figure. It can be observed that the end points are better reflected in Fig. 3(c), compared to those in Fig. 3(b).

## 4. Processing of ANN output for scoring

The output of the ANN is post processed by smoothing, and the collar is applied for scoring as described in the scoring procedure for the challenge (see Section 2.2).

### 4.1. Smoothing the decision

The speech regions identified in the output of the ANN for a test input are marked as +1, and the nonspeech regions as –1. This output data is post processed by smoothing the decisions using a window size of $N_w$ samples. This is implemented using an $N_w$-point mean smoothing at each sample. If the mean value at the current sample is above a threshold, then the decision at that sample is +1, otherwise the decision is –1. The threshold ($\alpha$) is varied between 0 and +1. Smoothing with long windows ( > 0.5 sec) is used to take into account the collar effect used in marking the ground truth.

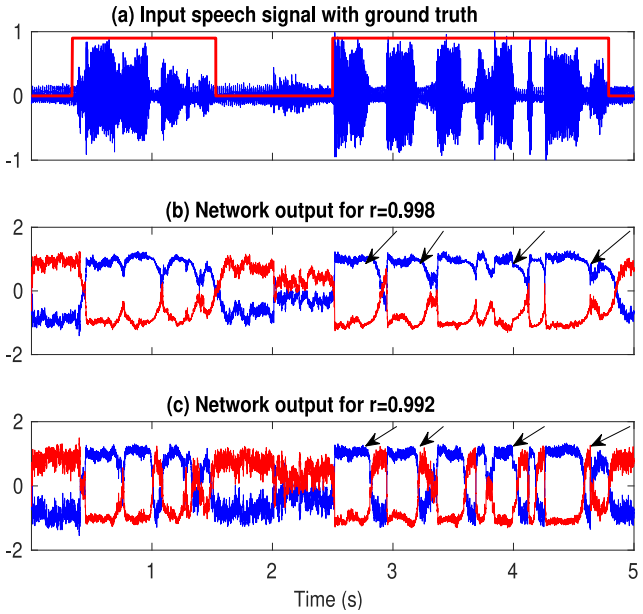


**Fig. 3.** Illustration of ANN output. (a) Test signal with ground truth. (b) Output of the ANN model tested using $r = 0.998$. The output of unit 1 is shown by blue line and of unit 2 by red line. (c) Output of the ANN model tested using $r = 0.992$. The output of unit 1 is shown by blue line and of unit 2 by red line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

### 4.2. Applying collar for scoring

The SAD challenge gives a scoring procedure (Fearless Steps Challenge, 2019), which includes a collar of max 0.5 sec on either side of the speech region. The collar regions are expanded as discussed in Section 2.2. The identified collar regions are not scored. For implementing this scoring procedure as given in (Fearless Steps Challenge, 2019), modified ground truth is generated, by making the nonspeech regions as –1, the collar regions as 0, and the speech regions as +1. In the modified ground truth sequence of +1, 0 and –1, the number of +1s give the total number ($N_1$) of samples labeled as speech, and the number of –1s give the total number ($N_2$) of samples labeled as nonspeech. The number of 0s correspond to the collar regions and are to be ignored.

### 4.3. Scoring

The smoothed ANN output decision consisting of +1 and –1 values is compared with the modified ground truth to arrive at the false negative (FN) and false positive (FP) regions. By subtracting the smoothed decision of the ANN output from the ground truth decision, we get samples with +2, 0 and –2 values. Note that the collar regions marked as 0 in the modified ground truth are ignored. All samples with +2 values correspond to false negatives (FN), as the system detected these samples as nonspeech, whereas the ground truth says it is speech. Likewise, all samples with –2 values correspond to false positives (FP), as the system detected these samples as speech, whereas the ground truth says it is nonspeech. Thus the count of the number of +2 values gives $M_1$, the number of samples of FNs, and the count of the number of –2 values gives $M_2$, the number of samples of FPs.

The performance is expressed in terms of detection cost function (DCF). The DCF in percentage is given by Fearless Steps Challenge (2019), Kaushik et al. (2018)

$$DCF\% = \left(0.75\frac{M_1}{N_1} + 0.25\frac{M_2}{N_2}\right) * 100. \tag{5}$$

It is to be noted that the false negatives are given higher weightage than the false positives. That is, treating the speech regions as nonspeech is considered more serious than missing the detection of nonspeech regions.

## 5. Results and discussion

Performance of the proposed network is evaluated for both development (dev) and evaluation (eval) data provided for the SAD task in Apollo challenge Fearless Steps Challenge Evaluation Plan v1.2 (2019). Ideally, the DCF value should be zero, indicating perfect match of the obtained decision with the ground truth. It means that the false positives and false negatives should not exist. The baseline system for the challenge reported a DCF value of 8.6% for the development data and 11.7% for the evaluation data (Ziaei et al., 2014; Fearless Steps Challenge, 2019).

The DCF values for the development data are obtained for different values of $N_w$ and $\alpha$ for the network decisions. These values with and without applying the collar are given in Table 1(a) and (b), respectively. It can be seen that for any smoothing window size ($N_w$) and for any threshold ($\alpha$), the DCF value is improved by applying the collar. This helps in tuning the system to reduce the errors due to inconsistencies in manual markings of the ground truth, and also to align with the suggested scoring procedure.

As the smoothing window size ($N_w$) is increased from 801 to 4801, the DCF value decreases. This is due to gradual decrease in the false negatives, i.e., the number of speech regions misclassified as nonspeech are reduced. As the window size is further increased, beyond 4801, (i.e., duration greater than 0.5 sec), the DCF value increases. This is due to significant increase in the false positives, i.e., the nonspeech regions are also taken into the speech regions.

As the threshold ($\alpha$) value is increased, from 0 to 0.9, the DCF decreases, indicating reduction in the false negatives. But if the threshold ($\alpha$) value is increased beyond 0.9, then the false positives start contributing, thereby increasing the DCF value.

It can also be seen that the change in the DCF value is large due to changes in the window size ($N_w$) for a fixed threshold value ($\alpha < 0.9$). This is due to higher weightage of 0.75 given to the false negatives and more speech regions may be disappearing. But as the threshold ($\alpha$) value is changed in some range (0.9 – 0.95), the change in the DCF value is relatively small, due to less weightage of 0.25 given to the false positives. A threshold ($\alpha$) value of +1 for any smoothing window means that the whole signal is considered as speech region. It can be seen in the last column of the Table 1, the DCF is 25%. A threshold ($\alpha$) value of –1 gives a DCF score of 75%, indicating that the whole signal is considered as nonspeech region.

The best DCF value of 3.28% (Table 1(b)) is obtained for a threshold of $\alpha = 0.92$ and for a smoothing window size of $N_w = 8001$ samples (i.e., 1 sec duration), when evaluated for the complete development data. The optimum window size of 1 sec duration may have something to do with 0.5 sec collar on either sides of a speech region in the scoring procedure. Interestingly, nearly similar values for the parameters are obtained by testing even with a small amount (about 30 min) of development data.

It is to be noted that there is hangover in the blue line after every large amplitude voiced segment. This is shown by arrows for a few cases in Fig. 3(b). This is due to the infinite impulse response (IIR) nature of the filter in the SFF analysis. To reduce this effect, a slightly lower value of $r = 0.992$ is used for the SFF analysis on the test data. The DCF values of the ANN output using $r = 0.992$ for different $N_w$ and $\alpha$ values, with and without collar, are given in Table 1(c) and (d), respectively. The results show a slight reduction in the best DCF value to 3.1% for $r = 0.992$ (Table 1(d)) from 3.28% for $r = 0.998$ (Table 1(b)). But in the case of $r = 0.992$, the optimum window size $N_w = 8001$ and the threshold $\alpha = 0.9$. Fig. 4 shows the DCF values for different thresholds for each window size and for the case of SFF analysis using $r = 0.992$. This corresponds to the values given in Table 1(d). The figure

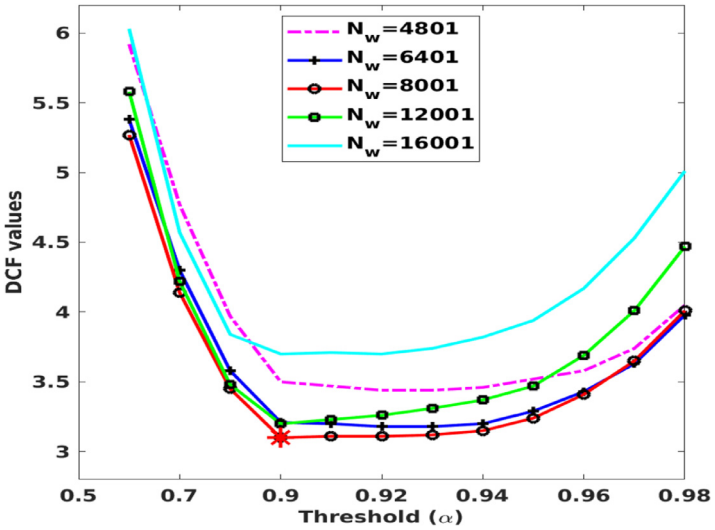


**Fig. 4.** Illustration of plots of DCF values for $r = 0.992$ for different smoothing window ($N_w$) sizes and thresholds ($\alpha$). (See Table 1(d)).

shows the variations of DCF as the threshold is changed. It appears that use of any window size $N_w$ in the range of 4001 to 12001 seem to give low DCF values for the threshold range of 0.9 to 0.94.

After fixing the optimum parameter values, the decisions for both the development data and evaluation data were submitted to the Fearless Steps Challenge. The results are given in Table 2 along with the results for the baseline system given in Fearless Steps Challenge (2019). As can be seen from Table 2, the proposed system gives significantly better performance than the baseline system for both the development data and evaluation data.

The results of submissions for the challenge are summarized in Hansen et al. (2019). The proposed system, as discussed in this paper, ranked 3 out of 27 submissions made for the challenge. The DCF results of the best system are 2.89% for Dev data and 3.32% for the Eval data. These are given in Table 2 as 2D CRNN (Vafeiadis et al., 2019). The best system used STFT spectrograms as 2-dimensional input features, and speech and nonspeech discrimination is made using convolutional recurrent neural networks. In another method reported for this challenge (Sharma et al., 2019), the speech activity detection was performed using a multi-level combination of modulation spectrum and Hilbert envelope of linear prediction (LP) residual. The DCF results for this method (mlasad) are 5.75% for Dev data and 7.35% for Eval data, as given in Table 2.

## 6. Summary and conclusions

This paper presents a novel approach for the SAD task for Apollo corpus. The method uses the instantaneous spectra derived using the SFF analysis. An ANN model is used to capture the discriminating features of speech and nonspeech present in the SFF spectra. The output of the ANN model is used to make the SAD decision. The decisions were submitted to the Fearless Steps Challenge to obtain the scores for both the development data and the evaluation data. The results show better performance for the proposed method compared to the results for the baseline system.

The novel features in the proposed system are the representation of speech in terms of instantaneous SFF spectrum, ANN model for capturing the discrimination between (noisy) speech and nonspeech (noise), and the decision at each sampling instant instead of at the frame level. Note that only the sum normalized spectra are used. The system development and evaluation is for the SAD task specified for the Apollo corpus. For the proposed system, only a small amount (10 min) of the development data was used out of 20 h of data. But the performance is evaluated for the complete 20 h of the development data.

While the difference in interpretation between the proposed ANN classifier and DNN is subtle, the focus here is mainly to study the effectiveness of the proposed instantaneous SFF spectra for the SAD task in the Apollo corpus.

Performance improvement may occur by tuning the ANN model. For example, training the model with 250 epochs reduced the DCF value on the development data to 3.06 from 3.10. It is also possible to build multiple ANN models for the 2-class problem, and combine the evidence to improve the performance. Other SFF spectral features like using low $r$ value and SFF phase also can be explored for further improvement.

One of the main issues with the SAD task of the Apollo corpus is that the ground truth is not reliable. One can easily see this by observing the waveform and the marked ground truth on it. The output of the ANN model seems to match better with the speech regions observed in the waveform. Thus there is scope for preparing a better ground truth for the Apollo data than what is given.

**Table 1**
DCF values for different window sizes ($N_w$) and threshold values ($\alpha$).

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **r=0.998** | **(a) without collar** | $N_w/\alpha$ | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **0.91** | **0.92** | **0.93** | **0.94** | **0.95** | **0.96** | **0.97** | **0.98** | **0.99** | **1** |
| | | **801** | 15.60 | 14.80 | 14.03 | 13.32 | 12.63 | 11.96 | 11.34 | 10.76 | 10.18 | 9.61 | 9.56 | 9.50 | 9.43 | 9.38 | 9.33 | 9.27 | 9.20 | 9.15 | 9.00 | 25.00 |
| | | **1601** | 14.61 | 13.39 | 12.31 | 11.30 | 10.40 | 9.58 | 8.84 | 8.17 | 7.58 | 7.05 | 7.00 | 6.94 | 6.89 | 6.87 | 6.83 | 6.80 | 6.78 | 6.82 | 6.89 | 25.00 |
| | | **3201** | 13.71 | 11.95 | 10.41 | 9.08 | 8.00 | 7.12 | 6.44 | 5.88 | 5.46 | 5.16 | 5.15 | 5.13 | 5.12 | 5.11 | 5.10 | 5.11 | 5.17 | 5.28 | 5.59 | 25.00 |
| | | **4801** | 13.38 | 11.20 | 9.45 | 8.09 | 7.03 | 6.17 | 5.58 | 5.20 | 4.93 | **4.80** | **4.80** | 4.82 | 4.83 | 4.86 | 4.89 | 4.92 | 4.97 | 5.13 | 5.49 | 25.00 |
| | | **6401** | 13.69 | 11.27 | 9.25 | 7.73 | 6.69 | 5.94 | 5.41 | 5.10 | 4.94 | 4.90 | 4.91 | 4.92 | 4.94 | 4.98 | 5.02 | 5.08 | 5.15 | 5.26 | 5.68 | 25.00 |
| | | **8001** | 14.25 | 11.54 | 9.49 | 7.95 | 6.74 | 5.98 | 5.52 | 5.20 | 5.08 | 5.13 | 5.13 | 5.13 | 5.15 | 5.18 | 5.23 | 5.30 | 5.39 | 5.48 | 5.89 | 25.00 |
| | | **12001** | 16.13 | 12.98 | 10.60 | 8.82 | 7.49 | 6.61 | 5.96 | 5.64 | 5.53 | 5.58 | 5.61 | 5.64 | 5.69 | 5.72 | 5.76 | 5.82 | 5.93 | 6.06 | 6.27 | 25.00 |
| | | **16001** | 18.31 | 14.93 | 12.06 | 9.83 | 8.22 | 7.15 | 6.46 | 6.04 | 5.92 | 5.99 | 6.04 | 6.06 | 6.10 | 6.17 | 6.25 | 6.31 | 6.39 | 6.54 | 6.69 | 25.00 |
| | | **32001** | 23.51 | 19.34 | 15.93 | 13.17 | 10.92 | 9.03 | 7.87 | 7.23 | 7.02 | 7.21 | 7.27 | 7.31 | 7.34 | 7.44 | 7.59 | 7.74 | 7.90 | 8.03 | 8.20 | 25.00 |
| | **(b) with collar** | $N_w/\alpha$ | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **0.91** | **0.92** | **0.93** | **0.94** | **0.95** | **0.96** | **0.97** | **0.98** | **0.99** | **1** |
| | | **801** | 15.36 | 14.54 | 13.76 | 13.02 | 12.31 | 11.63 | 10.99 | 10.38 | 9.77 | 9.17 | 9.11 | 9.05 | 8.98 | 8.92 | 8.87 | 8.81 | 8.73 | 8.68 | 8.52 | 25.00 |
| | | **1601** | 14.38 | 13.13 | 12.02 | 10.97 | 10.04 | 9.18 | 8.39 | 7.67 | 7.03 | 6.43 | 6.38 | 6.32 | 6.26 | 6.23 | 6.19 | 6.15 | 6.13 | 6.16 | 6.23 | 25.00 |
| | | **3201** | 13.46 | 11.65 | 10.07 | 8.67 | 7.52 | 6.54 | 5.77 | 5.12 | 4.59 | 4.19 | 4.16 | 4.13 | 4.11 | 4.09 | 4.07 | 4.07 | 4.13 | 4.23 | 4.55 | 25.00 |
| | | **4801** | 13.10 | 10.86 | 9.01 | 7.54 | 6.37 | 5.39 | 4.67 | 4.15 | 3.74 | 3.49 | 3.48 | 3.48 | 3.49 | 3.50 | 3.52 | 3.55 | 3.59 | 3.75 | 4.13 | 25.00 |
| | | **6401** | 13.36 | 10.85 | 8.71 | 7.05 | 5.85 | 4.94 | 4.25 | 3.77 | 3.46 | 3.30 | 3.29 | 3.30 | 3.31 | 3.34 | 3.38 | 3.43 | 3.50 | 3.61 | 4.06 | 25.00 |
| | | **8001** | 13.86 | 11.04 | 8.83 | 7.13 | 5.74 | 4.78 | 4.13 | 3.63 | 3.36 | 3.29 | 3.29 | **3.28** | 3.29 | 3.32 | 3.37 | 3.44 | 3.53 | 3.64 | 4.09 | 25.00 |
| | | **12001** | 15.57 | 12.27 | 9.70 | 7.71 | 6.15 | 5.03 | 4.16 | 3.67 | 3.46 | 3.52 | 3.55 | 3.59 | 3.65 | 3.69 | 3.74 | 3.82 | 3.95 | 4.11 | 4.36 | 25.00 |
| | | **16001** | 17.61 | 14.06 | 10.99 | 8.53 | 6.68 | 5.37 | 4.51 | 3.99 | 3.84 | 3.98 | 4.03 | 4.06 | 4.12 | 4.21 | 4.31 | 4.39 | 4.50 | 4.68 | 4.87 | 25.00 |
| | | **32001** | 22.59 | 18.26 | 14.66 | 11.70 | 9.29 | 7.27 | 6.02 | 5.34 | 5.15 | 5.40 | 5.48 | 5.53 | 5.59 | 5.71 | 5.89 | 6.07 | 6.26 | 6.43 | 6.63 | 25.00 |
| **r=0.992** | **(c) without collar** | $N_w/\alpha$ | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **0.91** | **0.92** | **0.93** | **0.94** | **0.95** | **0.96** | **0.97** | **0.98** | **0.99** | **1** |
| | | **801** | 26.15 | 24.21 | 22.36 | 20.63 | 19.02 | 17.47 | 16.06 | 14.71 | 13.36 | 11.98 | 11.84 | 11.69 | 11.54 | 11.38 | 11.22 | 11.06 | 10.87 | 10.67 | 10.38 | 25.00 |
| | | **1601** | 23.44 | 20.64 | 18.23 | 16.18 | 14.39 | 12.78 | 11.39 | 10.19 | 9.10 | 8.07 | 7.97 | 7.86 | 7.76 | 7.67 | 7.60 | 7.52 | 7.45 | 7.49 | 7.56 | 25.00 |
| | | **3201** | 22.72 | 19.19 | 16.05 | 13.38 | 11.13 | 9.30 | 7.85 | 6.74 | 5.90 | 5.34 | 5.31 | 5.28 | 5.26 | 5.23 | 5.22 | 5.26 | 5.35 | 5.56 | 6.07 | 25.00 |
| | | **4801** | 23.63 | 19.47 | 15.71 | 12.54 | 9.97 | 8.02 | 6.61 | 5.63 | 5.00 | 4.72 | 4.71 | **4.70** | 4.72 | 4.76 | 4.83 | 4.90 | 5.06 | 5.36 | 5.98 | 25.00 |
| | | **6401** | 24.72 | 20.03 | 16.00 | 12.54 | 9.79 | 7.68 | 6.30 | 5.43 | 4.91 | 4.73 | 4.74 | 4.74 | 4.76 | 4.79 | 4.88 | 5.02 | 5.21 | 5.54 | 6.23 | 25.00 |
| | | **8001** | 25.87 | 20.91 | 16.58 | 12.84 | 10.09 | 7.85 | 6.39 | 5.51 | 5.05 | 4.89 | 4.91 | 4.93 | 4.94 | 4.98 | 5.07 | 5.22 | 5.44 | 5.76 | 6.41 | 25.00 |
| | | **12001** | 28.84 | 23.49 | 18.56 | 14.26 | 11.04 | 8.70 | 7.09 | 6.02 | 5.50 | 5.31 | 5.34 | 5.37 | 5.41 | 5.45 | 5.53 | 5.71 | 5.96 | 6.35 | 6.78 | 25.00 |
| | | **16001** | 31.24 | 25.63 | 20.53 | 16.13 | 12.45 | 9.52 | 7.73 | 6.50 | 5.91 | 5.78 | 5.77 | 5.76 | 5.78 | 5.85 | 5.94 | 6.12 | 6.41 | 6.82 | 7.19 | 25.00 |
| | | **32001** | 37.43 | 30.90 | 25.04 | 20.18 | 15.91 | 12.44 | 9.75 | 7.93 | 7.16 | 7.00 | 7.03 | 7.05 | 7.05 | 7.11 | 7.21 | 7.39 | 7.71 | 8.16 | 8.61 | 25.00 |
| | **(d) with collar** | $N_w/\alpha$ | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **0.91** | **0.92** | **0.93** | **0.94** | **0.95** | **0.96** | **0.97** | **0.98** | **0.99** | **1** |
| | | **801** | 26.04 | 24.08 | 22.22 | 20.47 | 18.84 | 17.26 | 15.82 | 14.43 | 13.03 | 11.59 | 11.43 | 11.27 | 11.11 | 10.94 | 10.78 | 10.61 | 10.41 | 10.20 | 9.90 | 25.00 |
| | | **1601** | 23.34 | 20.52 | 18.09 | 16.01 | 14.18 | 12.54 | 11.11 | 9.84 | 8.66 | 7.52 | 7.40 | 7.29 | 7.17 | 7.07 | 6.98 | 6.88 | 6.81 | 6.84 | 6.91 | 25.00 |
| | | **3201** | 22.62 | 19.06 | 15.87 | 13.16 | 10.84 | 8.93 | 7.39 | 6.17 | 5.19 | 4.47 | 4.41 | 4.37 | 4.33 | 4.28 | 4.25 | 4.27 | 4.35 | 4.56 | 5.08 | 25.00 |
| | | **4801** | 23.52 | 19.31 | 15.49 | 12.24 | 9.56 | 7.48 | 5.92 | 4.77 | 3.97 | 3.50 | 3.47 | 3.44 | 3.44 | 3.46 | 3.52 | 3.58 | 3.74 | 4.05 | 4.71 | 25.00 |
| | | **6401** | 24.58 | 19.83 | 15.73 | 12.16 | 9.26 | 6.97 | 5.38 | 4.30 | 3.58 | 3.21 | 3.20 | 3.18 | 3.18 | 3.20 | 3.29 | 3.43 | 3.63 | 3.98 | 4.73 | 25.00 |
| | | **8001** | 25.69 | 20.67 | 16.22 | 12.35 | 9.42 | 6.96 | 5.27 | 4.14 | 3.45 | **3.10** | 3.11 | 3.11 | 3.12 | 3.15 | 3.24 | 3.41 | 3.65 | 4.01 | 4.74 | 25.00 |
| | | **12001** | 28.57 | 23.11 | 18.03 | 13.54 | 10.08 | 7.48 | 5.58 | 4.22 | 3.48 | 3.20 | 3.23 | 3.26 | 3.31 | 3.37 | 3.47 | 3.69 | 4.01 | 4.47 | 4.99 | 25.00 |
| | | **16001** | 30.88 | 25.13 | 19.85 | 15.24 | 11.32 | 8.10 | 6.03 | 4.57 | 3.84 | 3.70 | 3.71 | 3.70 | 3.74 | 3.82 | 3.94 | 4.17 | 4.53 | 5.01 | 5.47 | 25.00 |
| | | **32001** | 36.89 | 30.19 | 24.15 | 19.08 | 14.58 | 10.91 | 8.04 | 6.09 | 5.27 | 5.15 | 5.18 | 5.21 | 5.22 | 5.31 | 5.43 | 5.66 | 6.05 | 6.60 | 7.13 | 25.00 |

**Table 2**
Performance comparison of proposed method with other methods.

| DCF | Dev data(%) | Eval data(%) |
|---|---|---|
| Proposed | 3.1 | 4.6 |
| Baseline Fearless Steps Challenge (2019) | 8.6 | 11.7 |
| 2D CRNN Vafeiadis et al. (2019) | 2.89 | 3.32 |
| mlasad Sharma et al. (2019) | 5.75 | 7.35 |

## Declaration of Competing Interest

We wish to confirm that there are no known conflicts of interest associated with this submission.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided current, correct email addresses of all the authors. For reference, they are provided here again.

## Acknowledgments

The first author would like to thank MeitY, Govt. of India, for granting PhD Fellowship under Visvesvaraya PhD Scheme. The second author would like to thank the Indian National Science Academy for their support.

## References

Aneeja, G., Yegnanarayana, B., 2015. Single frequency filtering approach for discriminating speech and nonspeech. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (4), 705–717.
Chen, J., Wang, Y., Wang, D., 2014. A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Trans. Audio Speech Lang. Process. 22, 1993–2002.
Choi, J., Chang, J., 2014. Dual-microphone voice activity detection technique based on two-step power level difference ratio. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (6), 1069–1081.
Fearless Steps Challenge. 2019http://fearlesssteps.exploreapollo.org/.
Fearless Steps Challenge Evaluation Plan v1.2. 2019https://exploreapollo-audiodata.s3.amazonaws.com/fearless_steps_challenge_2019/v1.0/Fearless_Step_Evaluation_Plan_v1.2.pdf.
Ferroni, G., Bonfigli, R., Principi, E., Squartini, S., Piazza, F., 2015. A deep neural network approach for voice activity detection in multi-room domestic scenarios. International Joint Conference on Neural Networks (IJCNN), pp. 1–8.
Ghosh, P.K., Tsiartas, A., Narayanan, S.S., 2011. Robust voice activity detection using long-term signal variability. IEEE Trans. Audio Speech Lang.Process. 19 (3), 600–613.
Hansen, J., Sangwan, A., Joglekar, A., E. Bulut, A., Kaushik, L., yu, C., 2018. Fearless steps: apollo-11 corpus advancements for speech technologies from Earth to the Moon. INTERSPEECH, pp. 2758–2762.
Hansen, J., Sangwan, A., Kaushik, L., yu, C., 2018. Fearless steps: Advancing speech and language processing for naturalistic audio streams from Earth to the Moon with Apollo. The Journal of the Acoustical Society of America 143.1868−1868
Hansen, J.H.L., Joglekar, A., Shekhar, M., Kothapally, V., Yu, C., Kaushik, L., Sangwan, A., 2019. The 2019 inaugural fearless steps challenge: a giant leap for naturalistic audio. INTERSPEECH, pp. 1851–1855.
Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. Prentice Hall PTR.
Hughes, T., Mierle, K., 2013. Recurrent neural networks for voice activity detection. ICASSP, pp. 7378–7382.
Hwang, I., Sim, J., Kim, S., Song, K., Chang, J.-H., 2015. A statistical model-based voice activity detection using multiple DNNs and noise awareness. INTERSPEECH, pp. 2277–2281.
Kaushik, L., Sangwan, A., Hansen, J.H.L., 2018. Speech activity detection in naturalistic audio environments: fearless steps Apollo Corpus. IEEE Signal Process Lett 25 (9), 1290–1294.
Kim, J., Hahn, M., 2018. Voice activity detection using an adaptive context attention model. IEEE Signal Process. Lett. 25 (8), 1181–1185.
Kinnunen, T., Chernenko, E., Tuononen, M., Fränti, P., Li, H., 2007. Voice activity detection using MFCC features and support vector machine. Int. Conf. on Speech and Computer (SPECOM07), vol. 2, pp. 556–561.
Morita, S., Unoki, M., Lu, X., Akagi, M., 2016. Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments. J. Signal Process. Syst. 82 (2), 163–173.
Park, J., Jin, Y.G., Hwang, S., Shin, J.W., 2016. Dual microphone voice activity detection exploiting interchannel time and level differences. IEEE Signal Process. Lett. 23 (10), 1335–1339.
Pitsikalis, V., Maragos, P., 2009. Analysis and classification of speech signals by generalized fractal dimension features. Speech Commun. 51 (12), 1206–1223.
Sadjadi, S., Hansen, J., 2013. Unsupervised speech activity detection using voicing measures and perceptual spectral flux. IEEE Signal Process. Lett. 20 (3), 197–200.
Sangwan, A., Kaushik, L., Yu, C., Hansen, J.H.L., Oard, D.W., 2013. 'Houston, we have a solution': using NASA apollo program to advance speech and language processing technology. INTERSPEECH. ISCA, pp. 1135–1139.
Sangwan, A., Zhu, W.P., Ahmad, M.O., 2005. Improved voice activity detection via contextual information and noise suppression. IEEE International Symposium on Circuits and Systems, vol. 2, pp. 868–871.
Sharma, B., Das, R.K., Li, H., 2019. Multi-level adaptive speech activity detector for speech in naturalistic environments. INTERSPEECH, pp. 2015–2019.
Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1), 1–3.
Tanyer, S.G., Ozer, H., 2000. Voice activity detection in nonstationary noise. IEEE Trans. Speech Audio Process. 8 (4), 478–482.
Thomas, S., Saon, G., Segbroeck, M.V., Narayanan, S.S., 2015. Improvements to the IBM speech activity detection system for the DARPA RATS program. ICASSP, pp. 4500–4504.
Vafeiadis, A., Fanioudakis, E., Potamitis, I., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., Hamzaoui, R., 2019. Two-dimensional convolutional recurrent neural networks for speech activity detection. INTERSPEECH, pp. 2045–2049.

Vlaj, D., Kotnik, B., Horvat, B., Kačič, Z., 2005. A computationally efficient mel-filter bank VAD algorithm for distributed speech recognition systems. EURASIP J. Adv. Signal Process 487–497.
Wang, Q., Du, J., Bao, X., Wang, Z.-R., Dai, L.-R., Lee, C.-H., 2015. A universal VAD based on jointly trained deep neural networks. INTERSPEECH. ISCA, pp. 2282–2286.
Zazo, R., Sainath, T.N., Simko, G., Parada, C., 2016. Feature learning with raw-waveform CLDNNs for voice activity detection. INTERSPEECH, pp. 3668–3672.
Zhang, X., Wu, J., 2013. Deep belief networks based voice activity detection. IEEE Trans. Audio Speech Lang. Process. 21 (4), 697–710.
Zhang, X.-L., Wang, D., 2014. Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection. INTERSPEECH, pp. 1534–1538.
Zhang, X.-L., Wang, D., 2015. Boosting contextual information for deep neural network based voice activity detection. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (2), 252–264.
Zhang, X.-L., Wu, J., 2013. Denoising Deep Neural Networks Based Voice Activity Detection. ICASSP, pp. 853–857.
Ziaei, A., Kaushik, L., Sangwan, A., Hansen, J.H.L., Oard, D.W., 2014. Speech activity detection for NASA apollo space missions: challenges and solutions. INTERSPEECH, pp. 1544–1548.
Ziaei, A., Sangwan, A., Kaushik, L., Hansen, J.H.L., 2015. Prof-life-log: analysis and classification of activities in daily audio streams. ICASSP, pp. 4719–4723.

**Vishala Pannala** is currently pursuing Ph.D. at International Institute of Information Technology (IIIT) Hyderabad. She holds a B.Tech degree in the department of Electrical and Electronics Engineering, from Jawaharlal Nehru Technological University (JNTU), Hyderabad. Her research interests include speech signal processing, artificial neural networks, speech/nonspeech separation, detecting spoofing attacks in speech, spotting of predefined words. The research focus is on dealing degraded practical environmental conditions.

**Dr. Bayya Yegnanarayana** is currently INSA Senior Scientist at IIIT Hyderabad. He was Professor Emeritus at BITS-Pilani Hyderabad Campus during 2016. He was an Institute Professor from 2012 to 2016 and Professor & Microsoft Chair from 2006 to 2012 at the IIIT Hyderabad. He was a professor at IIT Madras (1980 to 2006), a visiting associate professor at Carnegie-Mellon University, Pittsburgh, USA (1977 to 1980), and a member of the faculty at the IISc, Bangalore, (1966 to 1978). He received BSc from Andhra University in 1961, and BE, ME and PhD from IISc Bangalore in 1964, 1966, and 1974, respectively. He was the General Chair for Interspeech2018 held in Hyderabad, India, during September 2018. His research interests are in signal processing, speech, image processing and neural networks.