Automatic Switching Between Noise Classification and Speech Enhancement for Hearing Aid Devices

Fatemeh Saki, Student Member IEEE, and Nasser Kehtarnavaz, Fellow IEEE

Abstract—This paper presents a voice activity detector (VAD) for automatic switching between a noise classifier and a speech enhancer as part of the signal processing pipeline of hearing aid devices. The developed VAD consists of a computationally efficient feature extractor and a random forest classifier. Previously used signal features as well as two newly introduced signal features are extracted and fed into the classifier to perform automatic switching. This switching approach is compared to two popular VADs. The results obtained indicate that the introduced approach outperforms these existing approaches in terms of both detection rate and processing time.

I. INTRODUCTION

It is well known that the performance of the signal processing pipelines deployed in hearing aid devices degrade significantly in noisy environments. This in turn negatively impacts the hearing experience of users of these devices in noisy environments. Initial attempts to address this issue have involved adjusting the enhancement setting manually by users on these devices. More recently, attempts have been made to automatically adjust the enhancement setting depending on the noise types encountered by users. For example, in [1] a real-time speech processing pipeline for cochlear implants was developed that allowed an automatic on-the-fly classification of different background noise types for the purpose of tuning the speech enhancement parameters to the classified noise type.

Such noise adaptive solutions require a voice activity detector (VAD) to be used at the frontend of the pipeline in order to identify the input sound status as unvoiced or voiced, that is as pure noise or speech plus noise. As illustrated in Fig.1, the noise classifier is activated when the VAD identifies the presence of pure noise and the speech enhancement is activated when the VAD identifies the presence of speech, noting that speech may appear as clean speech with no noise or more realistically as speech plus noise. It is important to note that the VAD should be able to identify pure noise irrespective of its type. In this work, our objective has been to develop a computationally efficient VAD that can cope with different types of noise as part of a noise adaptive speech enhancement pipeline.

VAD constitutes an essential component in many speech processing applications such as speech enhancement and speech recognition for the purpose of distinguishing speech from noise. The literature includes many studies on VADs. A typical VAD consists of two modules: a feature extractor and a classifier. The first module extracts signal features that allow discrimination between voice and unvoiced signals. The second module is a decision making one to separate features of voiced signals from unvoiced signals. VADs can be categorized into two categories based on their decision making approach: statistical-based and machine learning-based. The statistical-based ones establish probability density functions for noise and speech classes and then a data driven decision rule is applied to classify speech signal segments or frames from noise signal segments or frames [2-4]. More recently, machine learning-based VADs have been developed generating improved performance [5-12]. Table I provides an overview of existing VAD approaches.

The existing approaches have mostly concentrated on one type of noise with often stationary statistical characteristics. In this work, we consider different types of noise (stationary, semi-stationary, and non-stationary). Furthermore, we have found many of the features used previously are computationally demanding relative to the features we have considered for separating voiced from unvoiced signals. The thrust of this work is thus on a computationally efficient VAD so that it can be deployed as an automatic switch as part of a real-time noise adaptive speech processing pipeline in hearing devices. From a practical standpoint, this switch is designed in such a way that the switching between voiced and unvoiced situations only takes place in the presence of sustained type of sound environments, i.e. switching is not done frequently in response to noise or voiced frames rather in response to a large number of frames in a majority voting manner.

The rest of the paper is organized as follows. In section II, the modules of the developed automatic switch or VAD are discussed. This is followed by the experimental results in section III. Finally, the conclusion is stated in section IV.

II. DEVELOPED AUTOMATIC SWITCH OR VAD

We have considered the following computationally efficient features to separate pure noise signals from speech signals that may also contain noise: band-periodicity, band-entropy, spectral flux, subband short-time energy deviation (STED) and subband power spectral deviation (SPSD). These features are chosen here as they are found to exhibit good discriminatory power between pure noise signals and speech signals while at the same time they are computationally efficient to compute. A feature vector consisting of the above feature components for a sound frame at time $t$ is computed over the period $[t-S, t]$.

Spectral flux has been widely used to separate speech from noise signals. As defined in [13], spectral flux denotes...
where \( H_{b,m} \) represents the entropy of the \( m^{th} \) frame during the period \([t-S, t]\) in band \( b \).

To compute the band-periodicity features, the cross-correlation between every two adjacent frames in each band is computed and then the peak value of the cross-correlation denoted by \( \rho_{b,m} \) is used to define the band-periodicity features in band \( b \) as follows [13]:

\[
BP_b = \frac{1}{M} \sum_{m=1}^{M} \rho_{b,m}, \quad b = 1, \ldots, B
\]

Energy-based features have shown promising results for identifying the presence of speech. Hence, the energy level is also used here as a feature. Assuming that in a sustained noise environment, the level of background noise remains more or less constant, in the presence of speech, the energy level goes higher. In other words, on average, the deviation in the energy level between the highest (when a person talks) and the lowest energy level (gaps between speech frames) for speech and background noise is higher than that for pure noise as captured by a microphone. It is understood that there are exceptions to this assumption but in general this assumption holds in many practical situations. The deviation in the energy level of the input sound signal STED is computed in different frequency bands as follows:

\[
STED_b = \frac{\mu_b - \gamma_b}{\mu_b}, \quad b = 1, \ldots, B
\]

where \( \mu_b \) and \( \gamma_b \) are the average and the minimum energy of \( M \) frames during the period \([t-S, t]\) in band \( b \). Here, the average value is considered instead of the maximum value in order to capture noise which is sustained and to avoid capturing transient noise. The difference between the average and the minimum value for sustained noise is expected to be lower, while in the presence of noisy speech or clean speech is expected to be higher. Fig. 2 presents the distributions of this feature for machinery noise and speech plus machinery noise at different SNRs. It can be seen that the distributions for the speech plus noise shift to the right of the pure noise distribution.

Another feature which is introduced and used here is the

**Table 1. Overview of existing VAD approaches**

<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Features</th>
<th>Decision rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al. [5]</td>
<td>2016</td>
<td>likelihood ratios</td>
<td>deep belief neural network</td>
</tr>
<tr>
<td>Zou et al. [7]</td>
<td>2014</td>
<td>PCA- mel-frequency cepstral coefficients (MFCC)</td>
<td>support vector machine (SVM)</td>
</tr>
<tr>
<td>Zhang et al. [8]</td>
<td>2014</td>
<td>multi-resolution cochleagram (MRCG), pitch, discrete Fourier transform (DFT), MFCC, linear predictive coding (LPC), relative-spectral perceptual linear predictive analysis (RASTA-PLP), and amplitude modulation spectrograms (AMS)</td>
<td>boosted deep neural network</td>
</tr>
<tr>
<td>Zhang et al. [9]</td>
<td>2013</td>
<td>likelihood ratio</td>
<td>deep belief networks</td>
</tr>
<tr>
<td>Wu et al. [10]</td>
<td>2011</td>
<td>multiple-observation maximum probability (MO-MP) features/ multiple-observation SNR(MO-SNR) features</td>
<td>multiple kernel SVM</td>
</tr>
<tr>
<td>Kinnunen et al. [12]</td>
<td>2007</td>
<td>MFCC, delta-MFCC and double delta-MFCC</td>
<td>SVM</td>
</tr>
</tbody>
</table>
difference between two adjacent bands in the average power spectral density of sound signals over a long duration. The shape of the average power spectral density over a long duration provides an indication of which frequency regions of the spectrum are more affected by noise distortion and which ones are least affected. Fig. 3 shows an example of the average power spectral density of clean speech, noisy speech and noise. As shown in this figure, the averaged difference in the power spectrum between the first and the second frequency bands for the clean speech is more noticeable compared to the pure noise and noisy speech. This feature is computed as follows:

$$SPSD_b = \frac{1}{(K/B)} \left\{ \sum_{k=1}^{u_b} \sigma_{b,1}(k) - \sum_{k=1}^{u_b} \sigma_{b,0}(k) \right\}, \quad b = 1, \ldots, B$$  \hspace{1cm} (5)$$

$$\sigma = 10 \log_{10} \sum_{m=1}^{M} \overline{P}$$  \hspace{1cm} (6)$$

where $\sigma$ is the sum of power spectral density of frames, denoted by $\overline{P}$, during the period $[t-S, t]$, and $u_b$ is the number of frames in the band $b$.

As stated earlier, for classification, an RF classifier is used since it is computationally efficient and provides good classification performance. An RF classifier is an ensemble of $T$ number of decision trees. Each tree is trained independently using a randomly selected subset of the training set. Then, the tree is built in such a way that the entropy is decreased as tree levels are added until the tree reaches its leaves. In the recall mode, an input sound is assigned to the most voted class by all the trained trees in the RF. More details on RF classification can be found in [16].

III. EXPERIMENTAL RESULTS AND DISCUSSION

This section reports the experimental results of switching when using the developed VAD. HINT sentences [17] are widely used by audiologists to measure a person’s ability to hear speech in noisy background environments. These sentences were used to serve as clean speech sounds. Noisy speech sounds were generated by adding three different noise types: machinery (stationary type), driving car (semi-stationary type) and babble (non-stationary type) to the HINT sentences at different SNR levels. The noise files are made available for public use and can be downloaded from the link http://www.utdallas.edu/~kehtar/VAD-dataset.

Frames of 10 ms durations with 50% overlap were considered for feature extraction. Feature vectors were extracted for a majority voting period of $S=200$ ms. Four subbands, or $B=4$, were used for the subband features and 10 trees, or $T=10$, were used in the RF classifier. One half of the dataset was randomly chosen for training and the remaining half with no overlap was used for testing. Among the extracted subband features, the four band-periodicity features, the first two band-entropy features and the first band STED feature were found to be the most effective features by carrying out an MRMR feature selection analysis as discussed in [18]. Thus, together with the spectrum flux and SPSD features, 9 features were used in total.

The developed VAD was evaluated in terms of speech hit rate or true positive rate (TPR) and false alarm rate (FAR). True positive rate here means the ratio of number of correctly detected speech frames to the number of true speech frames and false alarm rate is defined based on non-speech detection hit rate (NHR) (FAR = 100-NHR), where non-speech hit rate denotes the ratio of number of correctly detected noise frames to number of true noise frames.

Table II provides a comparison of the performance of our developed VAD with the standard G.729 annex B [19] and Sohn’s VAD [3] in terms of TPR and NHR. These results reflect the outcome after applying majority voting over 200 ms. The same majority voting was also considered for the other two VADs. As can be seen from this table, our approach outperformed these approaches by more than 25% in terms of TPR. The FAR for our VAD was about 5% while for G.729 and Sohn, it was much higher; 40.8% and 57.8%, respectively. Fig. 4 shows an example of sound segments passed through the developed VAD. This example consists of the concatenation of seven files consisting of pure noise and speech plus noise. As shown in this figure, it is important to note that the switching, denoted by vertical arrows, only occurred when the noise was sustained for more than 200 ms and the silent gaps between speech sounds did not get detected as noise. The gray area indicates the 200 ms latency associated with switching after the sound environment was changed. In terms of computation time, it took 7.5 ms to compute our features for 10 ms frames using a PC with 2.67 GHz clock compared to 13.4 ms and 11.1 ms for the features used in the G.729 and Sohn VADs.
Figure 4. Illustration of switching between sustained noise and speech presence, small vertical arrows on top indicate the occurrence of switching, the grey area shows the latency associated with switching after transitioning to a new sound environment (for a 200 ms majority voting decision buffer).

IV. CONCLUSION

This paper has presented an automatic switching mechanism between speech enhancement and noise classification for deployment in hearing devices. A total of nine computationally efficient features have been extracted and fed into a random forest classifier to identify the presence of speech in different types of noise. The results obtained have indicated that the developed automatic switch or voice activity detector outperforms two other popular voice activity detectors in terms of both detection rate and processing time.

ACKNOWLEDGMENTS

This work was supported in part by the National Institute of the Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under the award number 5R56DC014020-02. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

<table>
<thead>
<tr>
<th>Environments</th>
<th>SNR (dB)</th>
<th>Developed VAD</th>
<th>G.729</th>
<th>Sohn’s VAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real noisy speech</td>
<td>-</td>
<td>98.8</td>
<td>79.1</td>
<td>77.6</td>
</tr>
<tr>
<td>Simulated noisy speech</td>
<td>0</td>
<td>94.1</td>
<td>55.0</td>
<td>69.0</td>
</tr>
<tr>
<td>(machinery, driving car and babble</td>
<td>5</td>
<td>95.9</td>
<td>74.1</td>
<td>70.0</td>
</tr>
<tr>
<td>noises)</td>
<td>10</td>
<td>96.9</td>
<td>68.8</td>
<td>70.6</td>
</tr>
<tr>
<td>Clean speech</td>
<td>15</td>
<td>98.4</td>
<td>70.6</td>
<td>73.5</td>
</tr>
<tr>
<td>Pure noise</td>
<td>-</td>
<td>99.5</td>
<td>56.2</td>
<td>67.8</td>
</tr>
<tr>
<td>(machinery, driving car, babble</td>
<td></td>
<td>95.0</td>
<td>59.2</td>
<td>42.2</td>
</tr>
<tr>
<td>noises)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. REFERENCES