Noise Robust Voice Activity Detection Based on an Iterative Approach

Gabriel Mittag, Friedemann Köster, Sebastian Möller
Quality and Usability Lab, Technische Universität Berlin, Deutschland,
Email: gabriel.mittag@tu-berlin.de, friedemann.koester@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

In this article, we present a new noise robust voice activity detection (VAD) for the application of speech quality estimation. Speech signals that are impaired by strong background noise contain frames with a high level of energy even if the speaker is silent (inactive). Because of this, traditional VADs often detect these silent segments as active. Consequently, quality indicators such as the background noise level are strongly underestimated. To overcome this problem, the proposed method exploits prior knowledge about the amount of speech pauses in the signal. The algorithm is based on the short term power of the signal and a simple iterative approach for finding the decision threshold. The iterative approach increases the threshold until a minimum number of frames is detected as silent. Thus, the method guarantees to find inactive segments even under extreme background noise conditions. To evaluate the method, the algorithm is applied to various databases and compared to other state-of-the-art VAD algorithms.

Introduction

Speech signals that are transmitted through a communication channel can be impaired by various types of degradations, such as packet loss, bandwidth limitations, and noise [1]. One major impairment is strong background noise, which can disturb the intelligibility of speech. Therefore, an accurate measurement of the background noise is vital for estimating the quality of transmitted speech.

In order to determine how these various impairments influence the service user, the *Quality of Experience* (QoE) [2] is of great interest. The QoE of transmitted speech is assessed in subjective tests with naïve participants that yield a *mean opinion score* (MOS) for each speech file [3]. In order to obtain diagnostic information about what caused the speech impairment, the overall perceived quality can be divided into so called *perceptual quality dimensions* [4]. One of these dimensions is the *noisiness* [5] that describes how noisy the speech signal was perceived (e.g. caused by background noise, circuit noise or coding noise). The *noisiness* is assessed by asking the test subjects to judge if the speech file was "not noisy" or "noisy" on a *absolute category rating* scale. The ratings are then averaged and result in the MOS_{noi}.

Unfortunately, these subjective tests are time and money consuming and therefore instrumental, signal-based models have been established. *Intrusive* full reference models use an unimpaired reference speech signal to compare it to the degraded signal under test, whereas *non-intrusive* single-ended models only use the degraded output signal to estimate the MOS. They have the advantage of being able to measure the speech quality of a communication system in operation.

One key point in the single-ended estimation of background noise is the *voice activity detection* (VAD). A VAD distinguishes between the non-speech (inactive) and speech (active) regions of a speech signal, where the inactive regions can include silence, environmental sounds or noise. A classification can be difficult particularly under strong background noise, and noisy segments without speech may be detected as active by the VAD. In literature one can find numerous VAD algorithms studies. Simple VADs are based on an energy measurement in combination with the zero-crossing rate. Others use features, such as mel-frequency ceptral coefficients (MFFCs), line spectral frequencies, cepstrum or linear prediction coefficients. Modern VADs are often based on a statistical model and use machine learning methods to draw a decision about the activity of speech.

In this paper, a VAD classifier that is based on the calculation of the short-term-power is presented, then this VAD is applied to an iterative approach that finds a fixed ratio of inactive speech segments. After this, the performance of the proposed VAD is compared to two other traditional VADs. The algorithm by Sohn is based on a statistical model [6] and the Segbroeck VAD [7] uses multiple signal feature streams in combination with a standard multilayer perceptron classifier. The evaluation is performed with two different methods. At first, the ability of the VAD to correctly classify a sample as active or inactive is analyzed in terms of classification metrics. To this end, a reference VAD, which uses the unimpaired speech signal is calculated and compared to the output of the analyzed VAD. Secondly, the VADs are used to calculate the noise level in the inactive speech segments and the correlation to the perceived noisiness is calculated and compared.

Four mixed-band databases (DAT1-DAT4) are available for the evaluation in terms of correlation with the perceived noisiness. They consist of different sentences (double sentences, duration: 8-10s) and between 4-12 speakers. In total the databases contain 840 speech files with 210 different conditions (e.g. ambient background noise, temporal clipping, different codecs, packet loss, frequency distortions). All databases were used to validate a now standardized speech quality model [8]. The evaluation in terms of classification metrics is performed with the NOIZEUS database [9], which contains 30 IEEE sen-

tences with eight different real-world noise conditions.

Method

In order to measure the perceived background noise of a speech signal it is crucial to use a reliable VAD that guarantees to exclude active speech segments. If active speech signals are used in the measurement of the background noise the noise level will be artificially increased. On the other hand, the VAD should be able to find inactive speech segments even in extreme cases in which the energy of the noise may be greater than the energy of the actual speech. In these extreme cases, most traditional VADs fail to find any inactive speech segments and therefore it is not possible to adequately measure the background noise level.

To avoid these two problems, the proposed VAD makes use of prior knowledge about the amount of speech pauses in the signal. Speech pauses during a telephone conversation are generally between $0.1-3\,\mathrm{s}$ long and contribute to around 30% of the total speech duration if a text is read out over a period of 4 minutes [10, 11]. Based on this knowledge a VAD is proposed that iteratively increases the amount of as inactive detected speech segments until a certain ratio of inactive speech samples is achieved. The *inactive speech ratio* β can be calculated as follows:

$$\beta = \frac{N_{\mathrm{Inactive}}}{N_{\mathrm{Total}}}, \tag{1}$$

where N_{Inactive} is the number of as inactive detected speech samples and N_{Total} the total amount of samples in the speech signal.

Inactive Speech Ratio VAD (ISR-VAD)

The proposed *inactive speech ratio based voice activity detection* (ISR-VAD) is based on a simple power threshold: If the short-term-power of the signal exceeds a certain threshold, the signal is classified as active speech, otherwise it is classified as inactive speech. Then, the threshold is increased until the desired amount of samples is detected as inactive. In order to obtain the VAD, the energy of the speech signal is calculated and then smoothed with help of a moving average window of size $N_{\mathrm{MM}} = 50\,\mathrm{ms} \cdot f_s + 1$ as follows:

$$P_x(k) = \frac{1}{N_{\mathrm{MM}}} \sum_{i=k-\frac{N_{\mathrm{MM}}-1}{2}}^{k+\frac{N_{\mathrm{MM}}-1}{2}} x^2(i), \tag{2}$$

where $P_x(k)$ denotes the short-term-power at sample k. The moving average window size is shrinked near the endpoints to include only the available samples. This results in a short-term-power value for each speech signal sample. The output of the VAD is given as:

$$voi(k) = \begin{cases} 1 & \text{if } P_x(k) > P_{\mathrm{th}} \\ 0 & \text{else} \end{cases}, \tag{3}$$

with P_{th} being the power threshold.

The goal of the ISR-VAD is to obtain a certain amount of inactive speech samples. To this end, first the VAD is calculated for a very low power threshold P_{th}, then β (1) is calculated. If β is below the inactive speech ratio threshold $\beta < \beta_{\mathrm{th}}$ the short-term-power threshold P_{th} is increased and consequently β will increase as well. This is repeated until the desired inactive speech ratio β_{th} is reached.

Reference VAD

In order to evaluate the proposed ISR-VAD in terms of the ability to detect inactive speech segments, a reference VAD was calculated. To this end, a fixed power threshold P_{th} is used for all speech files. Due to the fact that this VAD is applied to clean, unimpaired speech signals only, which therefore have very low energy in the inactive speech sample segments, it is assumed that this simple approach is sufficient to detect the active voice samples.

Noise Level

The noise level is the main indicator for perceived background noise within a speech signal and is based on the *Power Spectral Density* (PSD) $S_{xx}(\mu)$. It is measured in dB and describes the intensity of the noise in inactive speech segments. In this paper, the noise level NL is used to evaluate the ability of the proposed ISR-VAD to detect the inactive speech samples that correlate with the perceived background noise.

In order to calculate the noise level, the speech signal is divided into Hann windows with length $N_{\mathrm{NL}} = 32\,\mathrm{ms} \cdot f_s$ and an overlap of $N_{\mathrm{NL}}/2$. Only if all samples of a window are classified as inactive by the ISR-VAD the PSD $S_{xx}(\mu)$ is calculated. The average over all inactive PSDs then gives $\overline{S}_{xx}(\mu)$ and is used for the calculation of the noise level as follows:

$$NL = 10\,\log\left(\frac{1}{N_S}\sum_{\mu=1}^{N_S}\overline{S}_{xx}(\mu)H_{\mathrm{ph}}(\mu)\right), \tag{4}$$

where μ describes the frequency component, $N_S = N_{\mathrm{NL}}/2 + 1$ the number of frequency components and $H_{\mathrm{ph}}(\mu)$ the 'A' weighting curve that follows the ANSI S1.42 standard [12].

Evaluation

Because the ISR-VAD is used for measuring background noise we are interested in the amount of samples that are falsely detected as inactive although they contain speech. The falsely detected inactive segments artificially increase the measured noise level. Furthermore we want to know the amount of inactive samples that were not detected as such by the ISR-VAD. When inactive samples are not detected, the noise level measurement may be inaccurate because there are not enough measurement points. To evaluate this the VAD output voi_{ISR} is compared to the reference VAD output voi_{Ref}.

The *false omission rate* (FOR) is used to evaluate the amount of falsely detected inactive samples and describes

how many samples are falsely detected as inactive in relation to the total amount of inactive samples in the signal. It is defined as:

$$\mathrm{FOR} = \frac{\sum \text{False negative}}{\sum \text{Reference VAD negative}}. \quad (5)$$

The *Specificity* describes how many of the inactive speech samples were found in relation to the total amount of inactive speech samples. It is defined as:

$$\text{Specificity} = \frac{\sum \text{True negative}}{\sum \text{Reference VAD negative}}. \quad (6)$$

The application of the ISR-VAD is to detect inactive speech segments, which are then used for measuring perceived background noise. Therefore, the noise level NL is calculated in the samples that were detected as inactive by the VAD. The *Pearson correlation coefficient* ρ between the noise level NL and the auditory MOS_{noi} of the quality dimension *Noisiness* then gives an indication on how accurately the ISR-VAD found inactive speech segments that contain background noise (or silence if no noise is present).

Results and Discussion

The performance of the proposed ISR-VAD was first evaluated in terms of classification metrics and then secondly in terms of correlation ρ between the noise level NL and the auditory MOS_{noi}.

Classification Metrics

To evaluate the method in terms of false omission rate (Eq.(5)) and specificity (Eq.(6)) the ISR-VAD was applied to the NOIZEUS database, which is solely impaired by background noise. As the other databases contain a variety of different degradations, the reference and impaired speech files are not aligned and therefore the VAD outputs voi_{ISR} and voi_{Ref} can't be compared. Hence, it was only possible to calculate classification performance metrics for the NOIZEUS database. Note that this analysis depends on the reference VAD, which is based on a simple power threshold and not on perceptive active/inactive voice auditory tests. The FOR and specificity was calculated for each speech file and then averaged over all files.

Figure 1 presents the resulting FOR and specificity over changing β, compared to two other traditional VADs. A specificity of 1 means that all inactive samples have been found by the VAD and thus a high specificity is desired. A FOR of 0.5 means that half of the detected inactive samples are in fact active speech samples and have been falsely detected as inactive; thus, a low FOR is desired. This means there is a tradeoff between specificity and FOR, in regards to noise level measurement with a higher specificity more measuring samples are available but the FOR should be as small as possible to avoid false measurement samples, which are in fact active.

It can be seen that for inactive speech ratios greater than $\beta > 0.33$ the specificity outperforms the specificity of

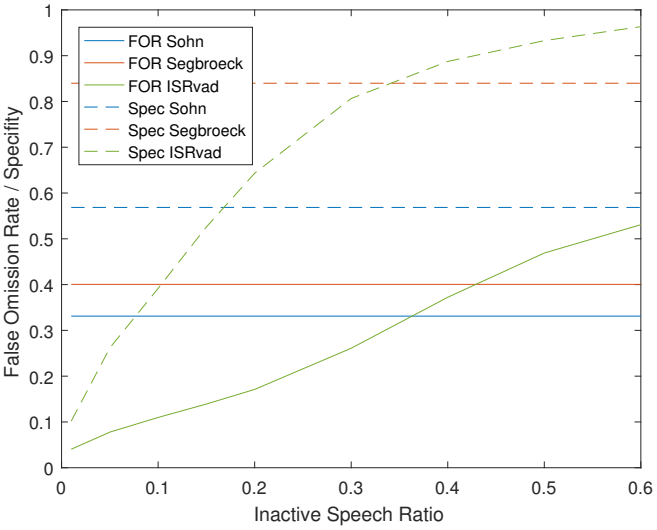

Figure 1: Average false omission rate and specificity over inactive speech ratio for the ISR-VAD applied to the NOIZEUS database.

the Segbroeck VAD while the false omission rate is still lower than one obtained with the VAD by Sohn. Generally, the false omission rate of the proposed ISR-VAD seems to perform better than the two other traditional VADs, which follows our expectations since the amount of as inactive detected samples is limited by the method and thus less active samples with high energy are detected as inactive. As a consequence the method seems to efficiently avoid measuring voice as background noise. The specificity of the proposed ISR-VAD is fairly small. However, if we assume static background noise that is constant over time, a small amount of inactive segments should be sufficient to measure the background noise accurately.

Correlation between noise level and perceived noisiness

The ISR-VAD was applied to the four databases for which auditory scores of the perceived noisiness in terms of MOS_{noi} are available. The classification output voi_{ISR} of the ISR-VAD was then used to calculate the noise level NL in inactive speech segments. The noise level was calculated for each speech file and then averaged per condition. The correlation ρ of the condition averaged noise level NL and the auditory MOS_{noi} scores was then calculated for each database and for different inactive speech ratios β. It should be noted that the databases contain various types of noisiness conditions, such as signal correlated noise or interruptions that may be perceived as noisiness. Thus ρ is not only influenced by how accurately the background noise is measured, but also by low MOS_{noi} scores that are caused by other noisiness conditions.

The results are presented in Figure 2. As the MOS_{noi} increases with perceived quality and the noise level NL increases with higher noise intensity a negative correlation ρ indicates a good accuracy. It can be seen that for databases DAT1 - DAT4 and inactive speech ratios

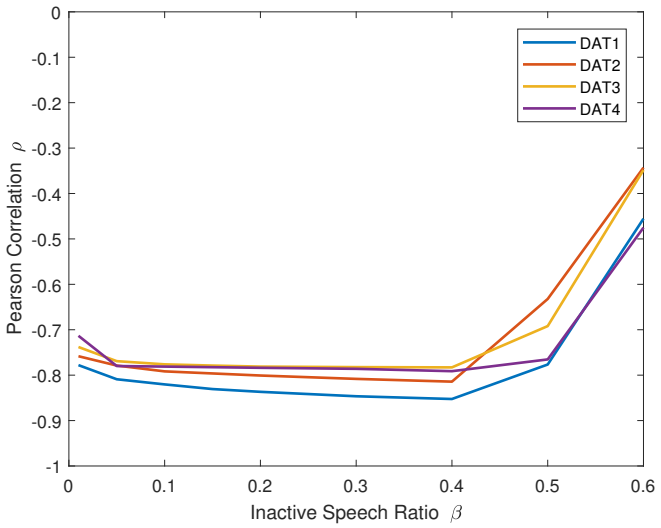

Figure 2: Pearson correlation between NL and MOS_{noi} over inactive speech ratio calculated with the ISR-VAD applied to several databases.

$[0.01 < \beta < 0.40]$ excellent correlations ρ are achieved. It can be assumed that for $\beta > 0.4$ speech is falsely measured as background noise.

From the results of Figure 2 we recommend to use an inactive speech ratio of $\beta = 0.1$, which yields the best and most robust overall results. Table 1 shows that for $\beta = 0.1$ the two traditional VADs are outperformed for all database, except for DAT4. In the case of a speech signal segment that does not contain any voice, this method will lead to measuring only the 10% samples with the lowest energy. However, 10% seem to be enough to measure an average noise level intensity of static background noise. On the other hand, even in the case of a talker that makes few pauses while speaking, the speech signal should contain at least 10% pauses if the signal is long enough. To investigate this further, more databases with longer speech signals and varying speech pauses are needed.

Table 1: Pearson correlation between NL and MOS_{noi}

Database	Pearson correlation ρ		
	ISR-VAD ($\beta = 0.1$)	Sohn VAD	Segbroeck VAD
DAT1	-0.86	-0.69	-0.30
DAT2	-0.76	-0.70	-0.43
DAT3	-0.74	-0.59	-0.20
DAT4	-0.75	-0.79	-0.29

Conclusion

In this paper we presented a new VAD for the application of measuring background noise of speech signals. The method is based on the assumption that each speech signal has a minimum amount of pauses in which it is possible to efficiently measure the noise level intensity of static background noise. The motivation of this approach is to overcome shortcomings of traditional VADs that often either measure voice as background noise or fail to find any inactive speech segments. The proposed VAD was analyzed in terms of the classification metrics FOR and specificity and furthermore it was evaluated on the ability to classify inactive speech samples that are relevant for measuring the perceived noisiness. It was shown that the proposed ISR-VAD outperforms the two traditional VADs for the application of measuring perceived background noise. The inactive speech ratio of $\beta = 0.1$ resulted in the most robust noise level measurements.

References

[1] Alexander Raake, *Speech Quality of VoIP Assessment and Prediction*, John Wiley & Sons, Chichister, West Sussex, 2006.

[2] Sebastian Möller and Alexander Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, Berlin, 2014.

[3] ITU-T Recommendation. P.800, "Methods for subjective determination of transmission quality," 1996.

[4] Ingwer Borg and Patrick Groenen, *Modern Multidimensional Scaling - Theory and Applications*, Springer Series in Statistics, New York, NY, 2005.

[5] Marcel Wältermann, Alexander Raake, and Sebastian Möller, "Quality Dimensions of Narrowband and Wideband Speech Transmission," Acta Acustica united with Acustica, 2010, pp. 1090–1103.

[6] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[7] Maarten Van Segbroeck, Andreas Tsiartas, and Shrikanth S. Narayanan, Eds., *A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice: Proceedings of InterSpeech*, 2013.

[8] ITU-T Recommendation. P.863, "Perceptual Objective Listening Quality Assessment," 2011.

[9] Yi Hu and Philipos C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, 2007.

[10] Hans Lazarus, Charlotte Sust, Rita Steckel, Marko Kulka, and Patrick Kurtz, *Akustische Grundlagen sprachlicher Kommunikation*, Springer, Berlin, Heidelberg, 2007.

[11] C. Bordone-Sacerdote and G. G. Sacerdote, "Distribution of Pauses as a Characteristic of Individual Voices," *Acta Acustica united with Acustica*, vol. 34, no. 4, pp. 245–247, 1976.

[12] ANSI S1.42, *Design Response of Weighting Networks for Acoustical Measurements*, American National Standard, 2001.