METHODOLOGY FOR THE DESIGN OF A ROBUST VOICE ACTIVITY DETECTOR FOR SPEECH ENHANCEMENT

Virginie Gilg, Christophe Beaugeant, Martin Schönle, Bernt Andrassy†

Siemens AG, ICM Mobile Phones, Grillparzerstrasse 10-18, 81675 Munich, Germany
†Siemens AG, Corporate Technology IC 5, Otto-Hahn-Ring 6, 81730 Munich, Germany
{first name}.{last name}@siemens.com

ABSTRACT

We propose a general methodology to design a robust voice activity detector that suits the needs of the speech enhancement system it is dedicated to. More than imposing rules, we initiate ideas on how to perform the analysis of the requirements for the Voice Activity Detection (VAD) and how to choose a reference, and evaluate the performances of the explored solutions in order to choose the one that best fits. As an example, the methodology is then applied to evaluate five VADs based on features described in the literature in the scope of two typical speech enhancement applications.

1. INTRODUCTION

Voice activity detection is an outstanding problem for speech transmission, enhancement and recognition. The variety and the varying nature of speech and background noise makes it especially challenging. In the past years, many features emphasizing the differences between speech and noise have been proposed for their robustness. Different performance evaluation criteria can be used to assess the quality of the VAD based on these features. In speech recognition, the word error rate at different level of noise is compared [1]. In speech transmission, one usually conducts subjective tests leading to a mean opinion score or one derives a measure of the distortion due to the speech clipping by using psychoacoustic auditory models. In speech enhancement, the most widespread objective criterion is to count and classify the number of misdetected signal frames in error categories (e.g. FEC, MSC, OVER, NDS) for different SNR values [2]. The performance of the algorithm is then obtained through the comparison with a standard VAD.

In this paper, rather than describing a new voice activity detection method, we propose a methodology to choose among a set of VADs based on different features the one that best fits the needs of the application.

2. METHODOLOGY

2.1. Determination of the speech decision errors critical for the application

A VAD can be decomposed in two steps : the computation of metrics and the application of a classification rule. Independently from the VAD method, we have to operate a compromise between having voice detected as noise or noise detected as voice. Thus, as we are not able to design a perfect VAD, we should define what errors are ”fatal” for our application. After the analysis of the need of the application, we propose to use the error types that were defined in [3] to identify the critical errors. The table in Fig. 1 lists possible errors obtained by taking into account the context.

Error Type	1	2	3	4	5	6	7	8
Activity Inactivity								
VAD Decision								
Name	NDS	-	WC	FEC	OVER	-	-	MSC

Fig. 1. Types of errors considered for evaluating VAD algorithms taken from [3]. Standard Error types : NDS : Noise detected as speech, WC : Word Clipping, FEC : Front End Clipping, OVER: Prolongated detection of speech in noise, MSC : Midspeech Clipping.

Errors 3,4,7 and 8 can be regrouped under the errors ”Speech frames detected as Noise” (*SdN*) and errors 1,2,5 and 6 under the errors ”Noise frames detected as Speech” (*NdS*).

2.2. Choice of the reference

The hand-labelling of the frames is tedious and depends in some extent from the auditory perception of speech pauses. The use of an automatic labelling scheme implies that a certain threshold is set and thus a compromise is done again in the discrimination of the speech from the noise. Hence, it is not possible to find one absolute reference. The choice of one reference type already implicitly denotes the set of acceptable errors. Possible references, as shown in Fig. 2 could be:

a) A VAD decision derived from a speech recognition system on clean speech (a speech block ↔ a word)

b) An energy based VAD on clean speech (a speech block ↔ a high energy signal part)

c) A reference defined with soft values ([0...1]) according to the probability of the presence of speech.

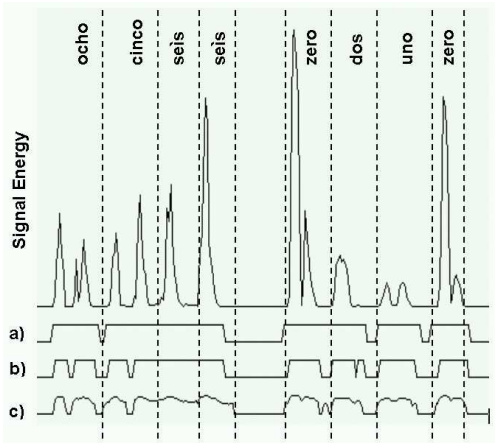

Fig. 2. Example of automatically-generated references.

2.3. Evaluation of the robustness and the suitability for the application

In the literature, the performance of a VAD is mostly measured in terms of error rate at different SNRs ([1, 3, 4]). The underlying definition of the robustness can be formulated as "a VAD is robust if it gives decisions close to a reference in quiet as well as in adverse environments".

We introduce a new definition claiming that a VAD is robust when it gives similar decisions for clean speech and noisy speech. The robustness can be estimated by taking the VAD's decision on clean speech as a reference and computing error statistics of the same VAD applied on noisy speech. The more robust the VAD, the scarcer the errors.

We also have to determine if the feature can be suitable for an application. Generally, the evaluation of the performance of a VAD is performed by counting the number of errors for each error type. This implies a strong confidence in the chosen reference. Possible alternatives could be:

a) counting the errors for each error type using two references : one emphasizing errors 1,2,5,6 and the other one emphasizing errors 3,4,7,8

b) using the probability of speech as a reference and weighting the errors statistics according to this "soft" value.

The choice of one VAD solution among others should be based on both the robustness and the suitability for the application.

3. APPLICATION TO SPEECH ENHANCEMENT

In speech enhancement systems, a reliable VAD is often a keystone component, for instance, for noise estimation and for adaptive echo cancellation. Let's apply the steps of the methodology to these two examples:

Step 1: Critical Speech Decision Errors

In noise estimation, we intend to update the noise energy as often as possible during a noise period. A VAD can trigger this adaptation. However, updating the noise estimation during speech periods can induce an important estimation error. The critical error is then to classify speech as noise and errors 3,4,7,8 (*SdN*) should be minimized.

Many echo cancellation algorithms need an estimation of the echo path. The adaptation of this estimation should happen exclusively during "echo-only" periods. Otherwise, a fast divergence can occur (i.e. during double-talk or in the absence of far-end signal). Besides other methods that control the adaptation step, one can use the VAD decisions of the far-end and near-end channel to determine when to freeze the adaptation. On the far-end, misclassifying noise as speech may lead to an adaptation during the absence of far-end speech. On the near-end, misclassifying speech as noise implies missing possible double-talk periods. Hence, errors 3,4,7,8 (*SdN*) should be minimized on the near-end and errors 1,2,5,6 (*NdS*) on the far-end.

Step 2: References choice

As a first experiment, we decided to use two references: a reference derived from a speech recognition system on clean speech and a reference generated with an energy-based VAD on clean speech. The first reference is emphasizing errors 3,4,7,8 and the second errors 1,2,5,6.

Step 3: Evaluation of the algorithm

We used the complete AURORA3 database to evaluate the algorithms. It is a subset of the SpeechDat-car database containing isolated and connected spanish and german digits in car driving conditions.

For the robustness measure, we derive a global percentage of false detections. The VAD is first applied on clean speech. The output is then used as a reference for the evaluation of the same VAD run on noisy speech (cf. Table 1)

The suitability is evaluated using the two references. For each reference, the statistics are composed of:

- the rate of speech frames detected as noise over the amount of speech frames
- the rate of noise frames detected as speech over the amount of noise frames
- a global percentage of false detection

Those statistics are given for clean speech, noisy speech and the whole database (cf. Table 2 and Table 4).

No fine optimization of the parameters was done on the algorithms.

4. DESCRIPTION OF THE EVALUATED ALGORITHMS AND RESULTS

We evaluated five VADs based on features taken from the literature :

Energy-based VAD in the time domain [5]:

The energy of the signal is compared with a threshold depending on the noise level. Speech is detected when the energy lies over the threshold. A hang-over of 2 frames is added to compensate for small energy gaps in the speech and to make sure the end of the utterance, often characterized by a decline of the energy, is not clipped.

VAD using the global SNR computed with the Kurtosis [6]:

The global SNR is implemented as in [6] except that we use an FFT instead of filter banks. The frame classification is done by applying an adaptive threshold on this feature. A hangover scheme of two frames is finally added to obtain the final voice activity decision.

VAD using the distance to the average cepstrum in noise [4]:

The euclidian distance between the actual cepstral coefficients and the average cepstrum in noise is much greater in speech period. Speech is detected when this distance is higher than an adaptive threshold, depending on the distance in non-speech period and a floor value of the feature.

However, the log scaling after the FFT dilates the components with lower energy and compresses the components with higher energy. Therefore the difference between noise and speech is less emphasized. More generally, any scaling assumes a certain level of noise. This kind of assumptions may lead to extra errors if the noise level is different from the supposed level.

VAD using the distance to the average spectrum in noise [4]:

The principle of this VAD is exactly the same as the VAD using the distance to the average cepstrum in noise. The spectrum was used instead of the cepstrum to overcome the drawback of the log scaling described above. In this case, the properties associated with the cepstrum are not valid anymore but this feature allows a more convenient discrimination of the noise.

Voicing Detection using the formant's shape and stability [1]:

As described in [1], the shape of the formants can indicate the presence of voiced speech. Furthermore, assuming that the speech is stationary in voiced-speech periods, the position of the formants should stay quite stable from one frame to another. After taking the FFT of the LPC coefficients of the signal, we use 3 heuristic criteria to detect voicing: the stability of the formants, the height of the peaks compared to the minima and the sharpness of the peaks.

4.1. Evaluation of the robustness

VAD	Global Error Rate
Energy	**12.80** %
Global SNR with Kurtosis	**11.35** %
Distance to Cepstrum	16.75 %
Distance to Spectrum	**11.50** %
Formants'shape and stability	20.00 %

Table 1. Global Detection Error Rate on noisy speech with a reference obtained with the same algorithm applied on clean speech

Table 1 shows that the most robust algorithms are the VADs using the global SNR computed with the Kurtosis, the average spectrum in noise and the energy in the time domain. These algorithms should be preferred to the others less robust.

4.2. Evaluation of the suitability for an application

Ref. Speech Recognition	**SdN**	NdS	Total
Energy			
Clean Speech	9.75 %	2.40 %	12.15 %
Noisy Speech	14.60 %	4.05 %	18.65 %
Total	12.17 %	3.23 %	15.40 %
Global SNR with Kurtosis			
Clean Speech	8.45 %	2.75 %	11.20 %
Noisy Speech	**11.40** %	4.75 %	16.15 %
Total	**9.93** %	3.75 %	**13.68** %
Distance to Cepstrum			
Clean Speech	3.35 %	7.75 %	11.10 %
Noisy Speech	**12.80** %	4.75 %	17.55 %
Total	**8.08** %	6.25 %	**14.33** %
Distance to Spectrum			
Clean Speech	11.70 %	1.80 %	13.50 %
Noisy Speech	15.80 %	2.95 %	18.75 %
Total	13.75 %	2.38 %	16.13 %
Formants'shape and stability			
Clean Speech	36.95 %	1.90 %	38.85 %
Noisy Speech	41.45 %	2.35 %	43.80 %
Total	39.20 %	2.13 %	41.33 %

Table 2. Statistics obtained with the reference derived from Speech Recognition: Percentage of Speech frames detected as Noise (*SdN*), Noise frames detected as Speech (*NdS*) and overall wrong detected frames

As explained in section 3, the VAD dedicated to the noise estimation and the echo cancellation on the near-end path should minimize the number of speech frames detected as noise (*SdN*). Hence, one should observe the statistics obtained with the reference emphasizing this type of errors

(i.e. the Speech Recognition reference) and especially the percentage of SdN in this table. Good results were obtained by the VAD using the distance to the cepstrum and the one using the global SNR. Even though the total percentage of SdN errors is somehow better for the first, we prefer the VAD using the global SNR, as it yields better results in noisy speech and a lower overall error rate.

We can notice that the VAD using the distance to the average spectrum and the VAD based on the formants have very low NdS errors. They may correctly detect speech frames that were ignored by the other algorithms and introduce very few false detection. By logically combining the VAD using the global SNR with the VAD using the distance to the spectrum (OR operation), we obtain an improvement of the percentage of errors SdN and the global error rate. The results are presented in Table 3.

Ref. Speech Recognition	**SdN**	NdS	Total
Kurtosis+Spectrum			
Clean Speech	7,65 %	3,00 %	10,65 %
Noisy Speech	**10,45** %	5,25 %	15,70 %
Total	**9,05** %	4,13 %	**13,18** %

Table 3. Combination of the VAD using the global SNR computed with the Kurtosis and the VAD based on the distance to the average spectrum in noise - Reference derived from Speech Recognition

As the VAD used for the echo cancellation on the far-end path should minimize the number of noise frames detected as speech (*NdS*), we focus on the percentage of NdS in the statistics with the energy-based VAD as reference (Table 4). The VAD using the formants' shape and stability gives then the best results but the overall error rate is low. However, the algorithms based on the energy in the time domain and the distance to the spectrum give quite close results.

5. CONCLUSION

After analysis of the results, we can now deduce that the combination of the VAD using the global SNR computed with the Kurtosis and the VAD using the distance to the spectrum best meets the constraints of robustness and suitability for the noise estimation and the near-end speech activity detection.

Concerning the far-end speech activity detection for the adaptation freeze, two algorithms can be recommended: the VAD based on the energy in the time domain and the VAD using the distance to the average spectrum in noise.

Ref. Energy	SdN	**NdS**	Total
Energy			
Clean Speech	3.50 %	2.15 %	5.65 %
Noisy Speech	9.00 %	4.75 %	13.75 %
Total	6.25 %	**3.45** %	**9.70** %
Global SNR with Kurtosis			
Clean Speech	4.50 %	4.85 %	9.35 %
Noisy Speech	7.30 %	6.85 %	14.15 %
Total	5.90 %	5.85 %	11.75 %
Distance to Cepstrum			
Clean Speech	0.15 %	10.55 %	10.70 %
Noisy Speech	8.10 %	6.15 %	14.25 %
Total	4.13 %	8.35 %	12.48 %
Distance to Spectrum			
Clean Speech	5.40 %	1.60 %	7.00 %
Noisy Speech	10.05 %	3.30 %	13.35 %
Total	7.73 %	**2.45** %	**10.18** %
Formants' shape and stability			
Clean Speech	30.45 %	2.00 %	32.45 %
Noisy Speech	35.10 %	2.60 %	37.70 %
Total	32.78 %	**2.30** %	35.08 %

Table 4. Statistics obtained with the reference derived from the energy-based VAD

6. REFERENCES

[1] J.D.Hoyt and H.Wechsler, "Detection of human speech in structured noise," *Proc. ICASSP*, vol. 2, no. 2, pp. 237–240, 1994. 1, 2, 3

[2] F. Beritelli, S. Casale, and G. Ruggeri, "A psychoacoustic auditory model to evaluate the performance of a voice activity detector," *Elsevier - Signal Processing*, vol. 80, pp. 1393–1397, 2000. 1

[3] J. Rosca, R. Balan, N.P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," *Proc. of EUSIPCO*, vol. 1, pp. 251–254, Sept. 2002. 1, 2

[4] S.E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition," *Proc. of ICASSP*, vol. 4, pp. 3808–3811, 2002. 2, 3

[5] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," *Proc. of the IEEE Speech Coding Workshop*, pp. 85–86, Oct. 1993. 3

[6] R. Goubran, E. Nemer, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Processing Letters*, vol. 6, no. 7, pp. 171–174, 1999. 3