OBJECTIVE AND SUBJECTIVE PERFORMANCE MEASURES FOR VOICE ACTIVITY DETECTORS

Beena Ahmed and W. Harvey Holmes
School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney, Australia
b.ahmed@unsw.edu.au, h.holmes@unsw.edu.au

ABSTRACT

The accurate performance of a Voice Activity Detector (VAD) is critical in several areas, including speech coders and speech recognition systems as well as in mobile telephony. Hence the need to comprehensively evaluate the performance of a VAD before integrating it into the application. We initially analyze the behaviour of VADs and their possible errors. Two separate measures are proposed which allow comparisons of the performances of different VADs to be made. The first measure is objective and uses the cross-correlation between the VAD output and the corresponding true speech/noise classification. The second measure estimates the perceptual effects of VAD errors on the overall speech quality felt by the listeners. Subjective MOS tests of VADs were carried out and it is shown that the proposed Perceptual Quality Measure (PQM) closely estimates the subjectively evaluated MOS scores.

1. INTRODUCTION

Voice activity detectors (VADs) have a wide range of applications, including uses in speech coders and speech recognition systems, digital mobile networks, packetized communication systems, and digital speech storage applications. GSM systems use VADs for discontinuous transmission in variable-rate speech coding to extend the unit's battery life and save channel capacity. They are also used in DSI systems where the channel bandwidth is allocated to a call based on the presence of speech activity.

A large number of VADs have been developed over the past three decades. Various parameters of speech signals are used to make the voice activity decisions, e.g. short term energy, spectral variations between speech and noise, autocorrelation coefficients and zero crossings. Given their wide-ranging applications, the accuracy and robustness of VADs are critical. Despite the large number of VADs available, general techniques for comparing and assessing their performance have not been published.

This paper investigates methods of quantifying the performance of VADs. First, the operation of a VAD was studied and the error types identified. Next a simple quantitative measure, based on the correlation between the VAD output and the true speech/noise classification of a signal, is discussed. However, the cross-correlation measure can be misleading, as it inaccurately represents the perceptual effect of these errors. The subjective performance of a VAD is affected by the frequency of its errors and their types.

Subjective experiments were carried out to assess the speech quality from imperfect VADs with different error types and combinations. Based on the results of these tests, the perceptual effects of VAD errors were modelled, and an objective perceptual quality measure (PQM) closely correlated with the subjective mean opinion score (MOS) results was developed.

2. VAD ERRORS

A voice activity detector is aimed at providing an accurate speech versus noise/silence representation of the input signal. A perfect VAD will have an output waveform identical to that obtained if the noise-free speech signal were classified frame by frame into speech and noise/silence.

A VAD can make two main categories of error:

1. Additive errors, when noise frames are classified as speech
2. Subtractive errors, when speech frames are classified as noise

Subtractive errors are less desirable than additive errors because, when frames of speech are removed, they essentially reduce the information content of the input speech signal. By contrast, additive errors do not remove any information from the speech signal, but add extra noise frames to the speech signal. This masks the true speech content, making it difficult or annoying for the listener to interpret the sentence and results in a longer data stream for transmission or a higher transmission bit rate.

Each of the two major error categories can be classified into three types of errors, as shown in Figure 1:

a) Additive errors:

1. Forward end error, when a noise frame preceding the start of a speech segment is declared speech.
2. Backward end error, when a noise frame following the end of a speech segment is declared speech.
3. Middle error, when a noise frame in the middle of a series of noise frames is declared speech.

b) Subtractive errors:

1. Forward end error, when the first frame of a speech segment is declared noise.
2. Backward end error, when the last frame of a speech segment is declared noise.
3. Middle error, when a frame in the middle of a speech segment is declared noise.

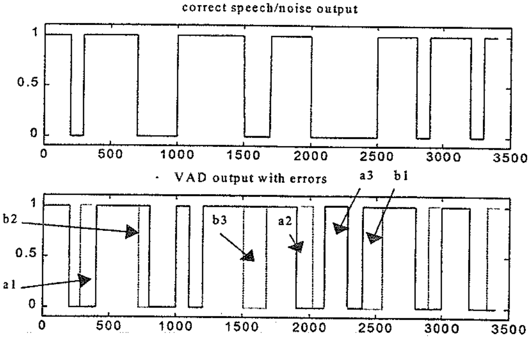

Figure 1. Possible errors in a VAD output compared to the correct speech/noise classification.

3. THE CROSS-CORRELATION MEASURE

A perfect VAD will have an output identical to the true speech/noise classification of a signal, giving two perfectly correlated waveforms. For any VAD output less than perfect, the correlation between the two will decrease. The normalized cross-correlation coefficient at zero lag between the VAD output and the true classification of a test sample is therefore a measure of the accuracy of the VAD. This is defined as

$$R(0)=R_{xy}(0)/\sqrt{R_{xx}(0)\cdot R_{yy}(0)}\ ,$$

where the cross-correlation is defined by

$$R_{xy}(k)=\frac{1}{N}\sum_{n=0}^{N-1-k}x(n+k)y(n),\quad k=0,1,\ldots,M\ ,$$

and similarly for R_{xx} and R_{yy}.

The stronger the similarity between the two waveforms, the higher will be the normalized cross-correlation. For a perfect VAD output, $R(0)$ will have the maximum possible value of one. For a completely incorrect VAD output, i.e. when it is the reverse of the true speech classification, $R(0)$ will have the minimum possible value of minus one. For practical VADs used in most systems, $R(0)$ ranges from 0.5 to 0.9. Two separate VADs can thus be compared using the cross correlation measure. The VAD with a higher $R(0)$ will have a superior representation of the true speech signal classification and a better performance. A sample VAD output, with a cross-correlation zero coefficient $R(0)$ of 0.744, is shown in Figure 2, including a plot of the normalized cross-correlation between the VAD output and the correct speech/noise classification.

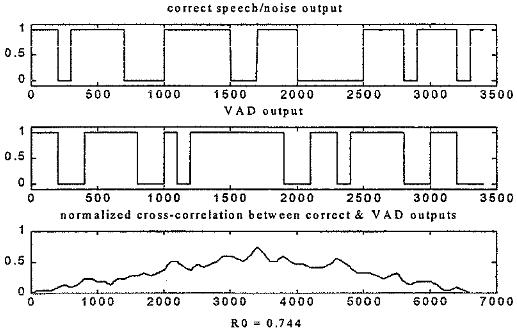

Figure 2. The cross-correlation between the VAD output and correct speech/noise classification.

4. FACTORS CONTROLLING VAD OUTPUT SPEECH QUALITY

4.1 Factors Controlling Errors in VADs

A voice activity detector affects the output speech quality by making errors in its speech/noise decision. The factors varying the resultant speech quality are those varying the errors themselves. These include:

1. Type and position of the error
   These include all six possible positions at which a VAD can make an error.
2. Quantity of errors
   The number of times that a VAD makes an error obviously has an effect on the final speech quality.
3. Length of the error
   It is possible that a VAD may make errors in two or more consecutive frames. Hence the need to study the effect that multiple frame errors can have as opposed to single frame errors.

4.2 Subjective Testing

Subjective MOS testing was carried out to study these factors. There were two sets of experiments, one for subtractive errors and the other for additive errors. The initial speech database consisted of 36 error conditions, plus one control (no errors). Sentences from 18 speakers (nine males and nine females) were used with unique samples for each of the cases. Testing was done in blocks of ten sentences per listener. Listeners between the ages 18 to 25 years were used.

For subtractive errors, listeners were asked to rank the quality of the sentences according to an opinion score value of 1 to 5, with 5 the best. From the results, MOS scores were obtained. As listeners cannot perceive additive errors in high SNR situations, extra noise at 10 dB SNR was added for tests with additive errors. To remove the bias in the listeners' judgements due to the presence of noise, they were asked to compare each test sentence with the same sentence repeated with the same amount of noise, but without errors. They ranked the difference according to a differential score of 1 to 5, with 5 the best, leading to DMOS scores. The MOS/DMOS scores obtained are given in Table 1.

An analysis of variance (ANOVA) was conducted on the test results to assess the individual and interaction effect of the factors identified on the scores obtained. This made it possible to isolate the sources of variation and develop a suitable model to replicate the subjective effect of these factors.

As expected, for subtractive errors the ANOVA results indicated all three factors studied had significant effects on the quality as perceived by the listeners. Of the three types of errors, middle errors scored the best, followed by backward and then by forward errors, with close means. Increasing quantity and length of errors decreased the perceptual quality. There was also a significant interaction effect between the length of speech removed and quantity of errors present. However for additive errors, the ANOVA results indicated that only the total length of the errors had a significant effect on the perceptual quality. Neither type nor quantity caused any consistent variation. It was thus possible to group additive errors into a single category, independent of the position of the errors.

Table 1. The subjective MOS recorded for subtractive and additive errors

Subtractive error means						Additive error means				
		quantity					quantity			
length	type	20%	50%	80%	mean	type	20%	50%	80%	mean
	forward	4	3.63	3.31	3.65	forward	4.31	4.25	4.19	4.25
single	backward	4.19	3.69	3.47	3.78	backward	4.06	4.38	3.81	4.08
	middle	4.63	4.25	3.5	4.13	middle	4.25	4.25	3.63	4.04
	mean	4.27	3.85	3.43	3.85	mean	4.21	4.29	3.88	4.13
	forward	3.19	2.5	1.63	2.44	forward	3.63	3.88	3.38	3.63
triple	backward	3.25	2.44	1.81	2.5	backward	4.13	4.31	3.63	4.02
	middle	3.63	2.75	1.88	2.75	middle	4.00	3.25	3.63	3.63
	mean	3.35	2.56	1.77	2.56	mean	3.92	3.81	3.54	3.76

5. PERCEPTUAL QUALITY MEASURE (PQM) FOR SUBTRACTIVE ERRORS

5.1 Proposed Model

To estimate the perceived quality of speech of the output of a VAD, the model shown in Figure 3 is used. The PQM of a VAD is calculated using the quantity of the various types of errors it makes. The number of errors made by the VAD is obtained by running it on test samples whose true speech/noise characteristics are available and quantifying the errors made in them.

The PQM is obtained in two steps. First the subscores, S_f, S_b and S_m are calculated for the three types of errors. Errors of different frame lengths are combined into a single sub-score. These sub-scores are then combined into a final score for subtractive errors, the PQM. Details of the model are given in section 5.3.

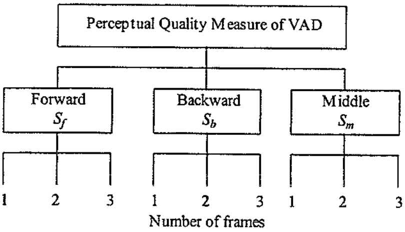

Figure 3. Model for calculating the PQM of a VAD.

5.2 Subjective Testing

To develop the PQM model, a second round of subjective testing was carried out. In it the interaction effects of different error conditions (the various lengths and types) was studied when combined in a speech sample, Here a total of 52 error combinations of the different error types identified, plus one control, were tested for. Sentences from 26 speakers (13 male, 13 female) were used, with unique samples for each of the cases. The structure of the tests was similar to that used earlier, with listeners between 18-25 years ranking the samples according to a MOS of 1 to 5. There were four sets of experiments in this round, examining the following areas:

- Combinations of various lengths of forward errors (i.e. the number of frames in the error)
- Combinations of various lengths of backward errors
- Combinations of various lengths of middle errors
- Combinations of the three types of errors

5.3 Results

Data from the first three sets of experiments indicated that simple linear regression equations could be used to obtain a composite score for the different lengths of errors for each type present. As the number of frames of the errors increased, so did the coefficients proportionally. This was in line with our earlier result in Section 4.3, that there was a significant interaction between the quantity of errors and the number of frames in the errors. It was thus possible to greatly simplify the equations. The forward, backward and middle composite subscores (MOS estimates) are obtained using the regression formulas:

$$S_f = 4.163 - 1.153\left(f_1 + 2f_2 + 3f_3 + \cdots\right)$$

$$S_b = 4.073 - 0.979\left(b_1 + 2b_2 + 3b_3 + \cdots\right)$$

$$S_m = 4.545 - 1.323\left(m_1 + 2m_2 + 3m_3 + \cdots\right)$$

The variables f_k, b_k and m_k in these equations are the quantities of errors for the three respective types. The subscripts denote the number of frames in each occurrence of the error. Quantity is defined as the number of error frames in the signal as a percentage of the total speech segments (two or more speech frames) in the sentence. Note that the maximum possible values of these subscores are less than 5 because these formulas are derived from regression fits of the speech corrupted by VAD errors.

The correlation coefficients between the actual MOS scores and the predicted values were computed (Table 2), with all three greater than 0.93 and standard errors of estimate less than 0.4. The scatter plots given in Figures 4a-c further support the good performance of these equations.

Table 2. The correlation coefficients and standard errors of the three MOS estimates and actual MOS

Error type	Correlation coefficient of estimate with true MOS	Standard error of estimate
Forward	0.975	0.2099
Backward	0.972	0.1890
Middle	0.939	0.3907

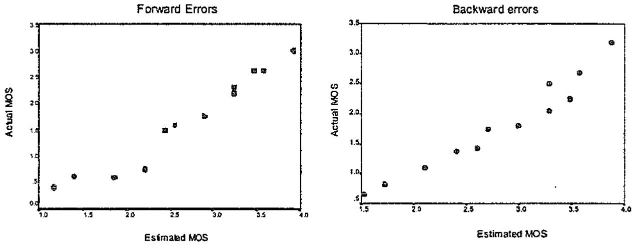

Figure 4a-b. Scatter plots of estimated MOS vs actual MOS for forward and backward errors

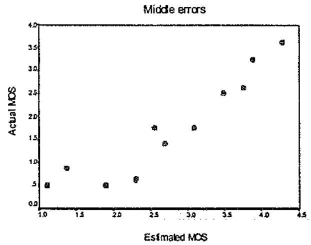

Figure 4c. Scatter plots of estimated MOS vs actual MOS for middle errors

The final set of subjective experiments developed in Section 5.2 studied the interaction between mixed error types. Using the data collected and based on how partial MOS scores must be combined when errors of different types are present, an equation was developed to combine the individual scores for the three types of subtractive errors. The result is the proposed perceptual quality measure (PQM) for subtractive errors in a VAD, given by:

$$\frac{4}{PQM-1}=\frac{4}{S_f-1}+\frac{4}{S_b-1}+\frac{4}{S_m-1}-2\ ,$$

where the subscores S_f, S_b and S_m are obtained as developed previously.

Estimates for the various error combinations using the proposed formula and estimates from other alternatives, including regression equations were compared with the data collected. The formula above was found to give the best results. A scatter plot comparing the true MOS with estimates obtained using this formula is given in Figure 5. The correlation between the estimated score (PQM) and the actual MOS was 0.973, and the standard error of the estimates was 0.21. These values indicate that the proposed measure can accurately predict the output speech quality as perceived by the listener for combinations of various types of subtractive errors.

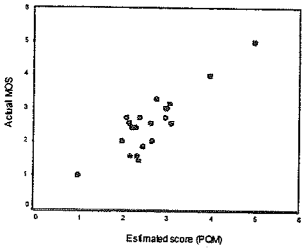

Figure 5. Scatter plot of PQM vs actual MOS for various combinations of subtractive error types

6. CONCLUSIONS

We have examined the performance of a voice activity detector and identified the possible errors in its operation. The cross-correlation between a VAD output and the true speech/noise characteristic provided a simple quantitative measure that gave an initial indicator of the accuracy of a VAD.

Subjective testing showed that for subtractive errors, the type, quantity and length of the errors controlled the performance of a VAD. The postulated perceptual quality measure (PQM) allowed the performance of VADs to be estimated objectively for the first time. The PQM had excellent performance when compared with the MOS data.

For the less important additive errors, only the length, not their positions, is important, and they are noticeable only in low SNR environments. Further work is required to quantify the subjective effect of additive errors and to develop an overall PQM for VADs that completely models all categories of errors.

7. REFERENCES

1. Basburg F., Nandkumar S. and Swaminathan K. (1999), "Robust voice activity detection for DTX operation of speech coders", *IEEE Speech Coding Workshop*, Finland, pp. 58-60

2. Freeman, Cosier G., Southcott .CB. and Boyd I. (1989), "The voice activity detector for pan-European digital cellular mobile telephone service", Proc. *ICASSP*, pp. 369-372.

3. Howell D., (1997) *Statistical Methods for Psychology.* (Wadsworth Publishing, Belmont).

4. Mead R., (1988) *The Design of Experiments.* (Cambridge University Press, Cambridge).

5. Thorpe L. and Yang W., (1999) "Performance of current perceptual objective speech quality measures", *IEEE Speech Coding Workshop*, Finland, pp. 144-146.