Improved Performance Measures for Voice Activity Detection

Simon Graf$^1,2$, Tobias Herbig$^1$, Markus Buck$^1$, Gerhard Schmidt$^2$

$^1$Acoustic Speech Enhancement Research, Nuance Communications Deutschland GmbH, 89077 Ulm, Germany
$^2$Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel, 24143 Kiel, Germany
Email: simon.graf@nuance.com
Web: $^1$www.nuance.com; $^2$www.dss.tf.uni-kiel.de

Abstract

Voice activity detection is an essential part of many speech processing algorithms. The requirements of the speech application determine the design of voice activity detection. Some applications need low-latency results whereas the accuracy of speech detection is more important for other applications. The performance is generally evaluated by Receiver Operating Characteristic (ROC) curves, which perform a static analysis averaged over speech and non-speech segments, respectively. We adopt the ROC curves but evaluate them for specific speech classes, e.g., voiced or unvoiced speech, to describe the overall accuracy of speech detection. In addition, we present a new measure for the dynamic behavior that considers the delay and latency of speech on- and offset detection. Finally, we present a unified measure for both aspects. This measure may be used to find appropriate voice activity detection features for a given application. An automotive noise scenario is employed to demonstrate the measures as it contains stationary and non-stationary noise.

1 Introduction

Many speech processing algorithms require Voice Activity Detection (VAD) for separating noisy speech utterances from segments with only noise. Various VAD strategies have been developed subject to different optimization criteria: For speech signal enhancement, accurate estimates for speech and background noise are required to scale the aggressiveness of noise reduction. The latter affects the audio quality of the processed audio signal. Low delay of VAD is targeted, especially in case of speech onsets. For speech recognition, a larger latency can be accepted. The confidence of speech segmentation is more important for the recognition accuracy. Non-intended input, such as babble noise, has to be rejected so that voice recognition is only performed on speech segments [1].

The performance of VAD algorithms is frequently evaluated by Receiver Operating Characteristic (ROC) curves under the objective to avoid false triggers, e.g., caused by non-stationary noise, and to achieve high detection rates on average in case of speech activity. The dynamic behavior, e.g., the detection of speech onsets, is typically not considered by ROC curves. Since many features of VAD algorithms are generally quite complex and span different time-frequency contexts, we expect differences not only in terms of the ROC but also the delay and latency of onset detection.

In this paper, we present a more fine-grained representation of the standard ROC curve for specific phoneme groups. We introduce a measure for dynamic properties of VAD algorithms, e.g., onset detection. Then we discuss how to combine both representations before we present our experimental results. Finally, we close with a summary and a conclusion.

2 Voice activity detection and evaluation measures

Detection of speech in a VAD algorithm is usually achieved by considering different speech characteristics. Based on features extracted from the noisy signal a classification scheme determines the presence of speech. In this section, we describe the general detection procedure and summarize measures for evaluation of the results.

Several features have been employed: Simple features might be based on short-term energy, zero-crossing rate, or signal-to-noise ratio. The harmonic structure of voiced phonemes is a reliable indicator for presence of speech and is therefore incorporated in many features [2]. Modulation-based and other long-term features that capture the temporal structure of speech significantly increase the detection performance [3]. However, the long window lengths prevent immediate detection of speech onsets.

In a VAD algorithm, one or more features are selected and a classification scheme

$$
\text{VAD} : [\mathbb{F} \times \mathbb{H}] \mapsto \{0, 1\}
$$

is applied that maps a feature \( ftr[u,n] \in \mathbb{F} \) to a binary decision. The classification is controlled by a parameter \( \eta \in \mathbb{H} \). The detected presence of speech at frame index \( n \) and utterance \( u \) is indicated by \( “1” \) whereas \( “0” \) denotes absence of speech.

In the following, we assume that only one scalar parameter \( \eta \in \mathbb{R} \) is available. The examples we present in the last section are based on a simple thresholding scheme

$$
\text{VAD}(ftr[u,n]) = \begin{cases} 
1, & \text{if } ftr[u,n] \geq \eta, \\
0, & \text{else}.
\end{cases}
$$

It is advantageous to consider the requirements of a certain application in the evaluation. We therefore discuss known measures and their applicability to reflect the dynamic behavior:

- The ROC curve plots the probability of correct detection \( P_d \) vs. the probability of false alarm \( P_f \) for a varying algorithm parameter [1]. As it is based on averaging over speech and non-speech segments respectively, only the static behavior is analyzed.
- To consider transient effects, a measure was introduced that selects four intervals at the begin and the end of speech [4, 5] instead of averaging over the complete utterance.
- The endpoint accuracy [6] was evaluated in terms of histograms of the differences between manually labeled and detected speech onsets / offsets.

In this paper a measure is proposed that takes into account both the static and dynamic behavior of VAD algorithms. By choosing the weighting coefficients according to the requirements of the application at hand, the measure allows for a simple comparison of different algorithms or tuning of algorithm parameters.
3 Extended performance measures

The measures discussed so far do not explicitly estimate the delay and latency of detection. As we expect delay and latency to provide valuable information about the dynamic properties of VAD features, we incorporate them into a new measure in addition to the well-known ROC curve.

Subsequently, we rely on a ground truth as reference for evaluation. The time-aligned phonetic transcription provided with the TIMIT database [7] is well suited for this task as it contains hand labeled, high quality data. We make use of the transcription to select phonemes and to find speech on- and offsets.

3.1 Phoneme-specific ROC curve

We evaluate the feature performance for different phoneme categories. Based on the transcription included in the TIMIT database we select intervals of the signal corresponding to specific speech classes. The mapping of TIMIT symbols $l[u,n]$ onto speech classes is given in Table 1.

<table>
<thead>
<tr>
<th>Speech class</th>
<th>TIMIT symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech pause</td>
<td>hfl, pau</td>
</tr>
<tr>
<td>word pause</td>
<td>epi, bcl, dcl, gcl, pcl, tck, kcl, tcl</td>
</tr>
<tr>
<td>speech</td>
<td>unvoiced, voiced, all other</td>
</tr>
</tbody>
</table>

Table 1: TIMIT symbols mapped onto speech classes

For each feature $ftr[u,n]$, we estimate the probability of correct detection

$$P_d = \hat{P} \left( \text{VAD}(ftr[u,n]) = 1 \mid l[u,n] \in \text{class}^{\text{speech}} \right)$$

(3)

and the probability of false alarm

$$P_{fa} = \hat{P} \left( \text{VAD}(ftr[u,n]) = 1 \mid l[u,n] \in \text{class}^{\text{pause}} \right)$$

(4)

where we define class$^{\text{speech}}$ as a subset of speech symbols, e.g., voiced or unvoiced, and class$^{\text{pause}}$ as speech pause. The ROC curve is based on empirical probabilities

$$\hat{P}(\text{VAD}(ftr[u,n]) = 1) = \frac{\# \{ (u,n) \mid \text{VAD}(ftr[u,n]) = 1 \}}{\# \{ (u,n) \}}$$

(5)

that are calculated over utterances and the time intervals corresponding to the specified speech classes. The cardinality of the sets is denoted by #.

Short intervals before the begin of speech (80 ms) and after the end of speech (640 ms) are excluded from the ROC evaluation in order to consider only the static behavior of algorithms and neglect transient effects. An example for the final selection is shown in Figure 1.

For generating the ROC curve, the parameter $\eta$ is varied and $P_d(\eta)$ vs. $P_{fa}(\eta)$ is plotted as shown in Figure 2.

The area under ROC curve (AUC) is commonly used to condense the ROC curve to a scalar value. The worst case AUC = 0.5 corresponds to a diagonal line, whereas the optimal value AUC=1 describes a ROC curve through $P_d = 1, P_{fa} = 0$.

3.2 Measure of dynamic behavior

As already discussed, the ROC curve only shows the static behavior of the algorithm. Therefore, we introduce a measure in addition to the ROC curve that considers the time dependency of the algorithm. We first determine the dynamic characteristic of detection and then extract time estimates that quantify on- and offset reaction times.

We evaluate the dynamic behavior of VAD algorithms by estimating the detection performance w.r.t. on- and offset position. Hereby, we determine the probability of detection

$$P_{on}(\Delta t_{on}) = \hat{P}(\text{VAD}(ftr[u, \Delta t_{on} + \eta \max \{|u,n|\}) = 1)$$

(6)

for a certain frame with time-lag $\Delta t_{on}$ relative to an onset. The onset position, given by the TIMIT transcription, of utterance $u$ is denoted by $\eta \max \{|u,n|\}$. By calculating the empirical probability over utterances we estimate $P_{on}(\Delta t_{on})$. The offset behavior $P_{off}(\Delta t_{off})$ is evaluated analogously.
lay and latency by normalizing the dynamic characteristics
\[
P_{\text{on/off}}(\Delta t_{\text{on/off}}) = \frac{P_{\text{on/off}}(\Delta t_{\text{on/off}}) - P_{\text{min}}}{P_{\text{max}} - P_{\text{min}}},
\]
(7)
with \(P_{\text{min}} = \min(P_{\text{on}}(\Delta t_{\text{on}}))\) and \(P_{\text{max}} = \max(P_{\text{on}}(\Delta t_{\text{on}}))\).

A linear interpolation between the integer valued frame indexes is applied to determine the time estimates \(t_{\text{on}}^{10\%}\), \(t_{\text{on}}^{90\%}\), and \(t_{\text{off}}^{10\%}\) by
\[
P_{\text{on}}(t_{\text{on}}^{10\%} \cdot f_{\text{fr}}) = 0.1, \quad P_{\text{on}}(t_{\text{on}}^{90\%} \cdot f_{\text{fr}}) = 0.9, \quad \text{and}
\]
\[
P_{\text{off}}(t_{\text{off}}^{10\%} \cdot f_{\text{fr}}) = 0.1.
\]
(8)
with frame-rate \(f_{\text{fr}}\).

The time estimates can be interpreted in terms of latency, delay and hang-over: A systematic latency or look-ahead (for non-causal systems) is reflected by \(t_{\text{on}}^{10\%}\). The delayed reaction on onsets is expressed by
\[
\Delta t_{\text{on}} = t_{\text{on}}^{90\%} - t_{\text{on}}^{10\%}.
\]
(10)
The hang-over time after a speech offset
\[
\Delta t_{\text{off}} = \max(t_{\text{off}}^{10\%} - t_{\text{on}}^{10\%}, 0)
\]
(11)
is limited to non-negative values, e.g., for features with decaying detection between on- and offset. The latency is already covered by \(t_{\text{on}}^{10\%}\), therefore it is excluded from \(\Delta t_{\text{off}}\).

### 3.3 Fusion of static and dynamic measures

The measures presented so far allow us to analyze static and dynamic behavior independently. Now, we combine both aspects to a single measure
\[
m_{\text{comb}} = \frac{2 \cdot |\text{AUC} - 0.5|}{1 + \gamma_{\text{on}} \cdot \Delta t_{\text{on}} + \gamma_{\text{off}} \cdot \Delta t_{\text{off}}}
\]
(12)
under the objective of optimizing the VAD algorithm for a certain application. By choosing the weighting coefficients \(\gamma_{\text{on}}\) and \(\gamma_{\text{off}}\), the influence of on- and offset delay can be controlled separately. For weighting coefficients \(\gamma_{\text{on}} = \gamma_{\text{off}} = 0 s^{-1}\) only the AUC is used.

### 4 Experimental setup and results

We present results obtained with our performance measures in an automotive environment: We use 100 clean TIMIT utterances, convolve them with measured car impulse responses and add noise. We use the UTD-CARNOISE database [8] which comprises stationary noise (e.g. highway noise) and non-stationary noise (e.g. blinder). A soft high-pass filter (3-dB cut-off frequency \(\approx 500\) Hz) is applied in addition to simulate the characteristic of a cardiod microphone and to reduce the influence of low-frequency noise on the SNR. The impulse response, which reflects the sound propagation from the speaker’s mouth to the microphone, is extracted from audio recordings with close-talk and sun-visor microphones.

The speech signal is rescaled before signal mixing to achieve a certain SNR. The speech power is estimated according to ITU-T P.56 [9], whereas the noise power is based on root-mean-square estimation.

For testing, we use 16 kHz sampling rate and Hann windows of 256 samples with 50% overlap.

In the next sections one long-term and one short-term feature are briefly summarized before their performance is evaluated in terms of the new measures.

---

1. In contrast to the usual definition, e.g., [2], the flatness in LSFM is calculated along the time dimension.
The spectral flatness measure [2] is then calculated by

\[ SFM(n) = -\log_8 \frac{GM_{\omega}(n)}{AM_{\omega}(n)} \]

with

\[ GM_{\omega}(n) = \left( \frac{K_0}{K_0 - K_f} \right) \frac{1}{\sum_{k=K_i}^{K_f} S_{prep}(n, \omega_k)} \]

and

\[ AM_{\omega}(n) = \frac{1}{K_0 - K_f + 1} \sum_{r=K_0}^{K_f} S_{prep}(n, \omega_k), \]

where the lower bound is \( K_i = 10 \) (±562 Hz) and the upper bound is \( K_f = 80 \) (±4938 Hz).

Again we expect the feature to assume high values on harmonic peaks in the speech spectrum that significantly exceed the envelope. Low values can be expected for speech pauses.

### 4.3 Results

For our evaluation, we apply the thresholding scheme given by (2) to generate detection results for LSFM and SFM features. An SNR of 0 dB is used in our experiments to obtain a realistic scenario for automotive applications.

First, in an analysis of the static behavior, we evaluate the influence of phoneme groups on the AUC. We conclude from the results shown in Table 2 that the performance of LSFM appears to be independent from the speech class, whereas SFM performs better on voiced speech.

<table>
<thead>
<tr>
<th>Feature</th>
<th>voiced</th>
<th>unvoiced</th>
<th>speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFM</td>
<td>0.76</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>LSFM (M=10, R=5)</td>
<td>0.84</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>LSFM (M=10, R=40)</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 2: AUC for class\( \text{speech} = \text{voiced} / \text{unvoiced} / \text{speech} \)

In addition to the static analysis we now discuss the dynamic measures. The results are depicted in Figure 4. We fix \( P_{\text{th}} = 0.1 \) for all features. If we alter the window length of LSFM, we can observe that the hang-over time \( \Delta_{\text{off}} \) increases linearly with \( R \). The onset delay \( \Delta_{\text{on}} \) is much shorter than \( \Delta_{\text{off}} \) and is affected by \( R \), too. No hang-over is noticeable for SFM.

\[ \Delta_{\text{on}} = \frac{\gamma_{\text{on}}}{\gamma_{\text{off}}} \]

\[ \Delta_{\text{off}} = \frac{\gamma_{\text{off}}}{\gamma_{\text{on}}} \]

\[ \gamma_{\text{on}} = \gamma_{\text{off}} = 0 \text{s}^{-1} \]

Figure 4: Delay estimates for LSFM with varying window length \( R \) (blue/red), and SFM (green). The delay increases when more temporal context is considered.

The combined measure based on static and dynamic behavior is illustrated in Figure 5. For the static behavior we evaluate the AUC for \( \text{speech} \). LSFM with large window length (\( R=40 \)) shows the best static performance (\( \gamma_{\text{on}} = \gamma_{\text{off}} = 0 \text{s}^{-1} \)). However, as it incorporates delays, the measure decreases when we consider the dynamic behavior by increasing \( \gamma_{\text{on}} \) and \( \gamma_{\text{off}} \). For applications that require low delays, we choose higher \( \gamma_{\text{on}} \) and \( \gamma_{\text{off}} \). In this case, short-term features are preferred.

![Figure 5: Combined measure for varying weighting coefficients \( \gamma_{\text{on}} \) and \( \gamma_{\text{off}} \).](image-url)

### 5 Conclusions

In this paper, we presented a new measure to evaluate the performance of voice activity detection (VAD) features and algorithms. The measure includes both static and dynamic behavior, which are determined from AUC, and delay and latency estimates, respectively. By weighting both influences according to a particular application we can compare different features, or tune algorithm parameters. Our experiments for an automotive scenario confirmed that both characteristics of VAD features are reflected. It became evident that future developments of VAD algorithms can benefit from these performance measures.

### References


