Multi-Sensor Voice Activity Detection based on Multiple Observation Hypothesis Testing

Theodoros Petsatodis\textsuperscript{1,2}, Fotios Talantzis\textsuperscript{2,3}, Christos Boukis\textsuperscript{4}, Zheng-Hua Tan\textsuperscript{5}, Ramjee Prasad\textsuperscript{1}

\textsuperscript{1} Center for TeleInFrastruktur (CTIF), Aalborg University Denmark
\textsuperscript{2} Athens Information Technology (AIT), Greece
\textsuperscript{3} Imperial College London, United Kingdom
\textsuperscript{4} Accenture Interactive, Greece
\textsuperscript{5} Department of Electronic Systems, Aalborg University Denmark

thp@es.aau.dk, fota@ait.gr, christos.boukis@accenture.com, zt@es.aau.dk, prasad@es.aau.dk

Abstract

Voice Activity Detection (VAD) in acoustic environments remains a challenging task due to potentially adverse noise and reverberation conditions. The problem becomes even more difficult when the microphones used to detect speech reside far from the speaker. An unsupervised VAD scheme is presented in this paper. The system is based on processing signals captured by multiple far-field sensors in order to integrate spatial information in addition to the frequency content available at a single-channel recording. To decide upon the presence or absence of speech the system employs a modified multiple observation hypothesis that tests at each sensor the probability of having an active speaker and then fuses the decisions. To minimize misdetections and enhance the performance of the hypothesis test a computationally efficient forgetting scheme is also employed. Simulations conducted in several artificial environments illustrate that significant improvements in performance can be expected from the proposed scheme when compared to systems of similar philosophy.

Index Terms: Speech processing, Voice Activity Detection, Likelihood Ratio, Hypothesis Testing

1. Introduction

Voice Activity Detection (VAD), is a core speech processing technology with application in several domains. It can be found integrated in several telecommunication systems used to reduce power consumption of transmitters and bandwidth utilization [1]. VAD is often combined with other speech-processing systems, such as Automatic Speech Recognition and Speaker Identification, to prevent their operation in the absence of speech aiming to error rate reduction [2].

Typically, VAD systems rely on the continuous observation of a specific metric to decide on the content of an audio signal. Such metrics can be the energy levels, zero-crossing rate, periodicity, linear prediction coding parameters, and mutual information [1, 3, 4, 5]. Recently introduced statistical VADs attempt to mathematically formulate the problem, by employing a Likelihood Ratio Test (LRT) as a decision criterion [6, 7, 8].

The performance of VAD systems depends strongly on various factors, including the discriminative ability of the classification criterion employed, the dynamics of the additive noise and the signal to noise ratio. Speech signals transmitted within reverberant enclosures and captured using far-field microphones are subject to superposition of reflected versions of the source signal. Additionally, the movement of the talking person is also affecting the characteristics of the captured audio signal.

Towards overcoming such adversities, related research focused on microphone arrays [9, 10]. These VAD systems have the advantage of utilizing spatial information while involving multiple and independent observations in contrary to single microphone based methods, which can only utilize time and/or frequency information. Nevertheless, most of microphone array based VAD require precise estimates of the direction-of-arrival (DOA) of speech signals in advance or assume that the speaker’s movement is limited [9, 10]. DOA estimation can seriously affect when audio signals are captured within reverberant enclosures or by directional noise sources.

An alternative approach proposed by Ramirez et al. [11] is Multiple Observation likelihood ratio test (MO-LRT) VAD. In MO-LRT the decision rule, that is based on the likelihood ratio of the Gaussian modeled conditioned speech absence and presence, is formulated over a sliding window consisting of a set of observation vectors around the frame for which the decision is being made. Nevertheless, this fact imposes a significant delay to the algorithm and increased computations that, for several applications, like real-time operating telecommunication systems can be a major disadvantage.

In this paper we propose the modification of the MO-LRT towards the development of a multiple sensor VAD. The proposed scheme takes advantage of the additional information provided by microphone arrays. It operates without the need of DOA estimation or additional delay compared to previous multi-microphone VAD technologies.

The paper is organized as follows. In Section II the modification of LRT and the proposed VAD are described in detail. Section III discusses the results of the experiments performed. Section IV concludes the work.

2. System description

Assuming that speech is generated by a single speaker (source), the reverberated speech signals captured by the distant microphone array, bearing \(M\) microphones, at time \(t\) are given by

\[x_m(t) = h_m(t) * s(t) + n_m(t)\]  \hspace{1cm} (1)

where \(x_m\) denotes the signal captured by the \(m^{th}\) microphone \(s(t)\) the source speech signal at time \(t\), \(h_m(t)\) the corresponding acoustic impulse response, \(n_m(t)\) the additive noise, and \(*\) denotes convolution.
2.1. Single Microphone Binary Hypothesis Testing

Voice activity detection can be expressed as the likelihood ratio of two hypotheses stating speech presence and absence. Assuming additive noise the two hypotheses \( H_0 \) and \( H_1 \) that indicate speech presence and speech absence are accordingly:

\[
H_0 : \text{speech absence : } X(t) = N(t) 
\]

\[
H_1 : \text{speech presence : } X(t) = S(t) + N(t) 
\]

where \( X(t) = [X_0(t), X_1(t), \ldots, X_{M-1}(t)]^T \), \( S(t) = [S_0(t), S_1(t), \ldots, S_{M-1}(t)]^T \), and \( N(t) = [N_0(t), N_1(t), \ldots, N_{M-1}(t)]^T \) are the noisy captured speech, reverberated speech, and noise frequency components.

Real and imaginary parts of noise and speech frequency spectrum are assumed to be zero mean Gaussian distributed. The probability densities for the noise and speech components with \( k \) denoting the frequency bin are given by

\[
f_{n}^G(N_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{n,k}^2}} e^{-\frac{|N_k(t)|^2}{2\sigma_{n,k}^2}} 
\]

\[
f_{s}^G(S_k(t)) = \frac{1}{\sqrt{2\pi\sigma_{s,k}^2}} e^{-\frac{|S_k(t)|^2}{2\sigma_{s,k}^2}} 
\]

where \( \sigma_{n,k}^2, \sigma_{s,k}^2 \) are the slowly varying variances of the Gaussian distributed noise and speech respectively estimated by employing Eq.(14) for the \( k^{th} \) frequency component. The probability density functions conditioned on \( H_0 \) and \( H_1 \) are given by

\[
p(X|H_0) = \prod_{k=0}^{K-1} \frac{1}{\pi\lambda_{n,k}} \exp\left\{ -\frac{|X_k|_\lambda_{n,k}^2}{} \right\} \tag{6} 
\]

\[
p(X|H_1) = \prod_{k=0}^{K-1} \frac{1}{\pi\lambda_{s,k}} \exp\left\{ -\frac{|X_k|_\lambda_{s,k}^2}{} \right\} \tag{7} 
\]

where \( \lambda_{n,k} \) and \( \lambda_{s,k} \) denote the variances of \( N_k \) and \( S_k \) respectively.

2.2. Single Microphone LRT (SM-LRT)

In the case of single microphone VAD scheme the likelihood ratio for the \( k^{th} \) frequency bin is defined as

\[
\Lambda_k \equiv \frac{p(X|H_0)}{p(X|H_1)} \frac{1}{1 + \xi_k} \exp\left\{ \frac{\gamma_k}{1 + \xi_k} \right\} \tag{8} 
\]

where \( \xi_k \equiv \lambda_n/k/\lambda_{s,k} \) and \( \gamma_k \equiv |X_k|^2/\lambda_{s,k} \) the a priori and a posteriori signal and noise ratios\([12]\).

The decision criteria is based on evaluating the geometric mean of the likelihood ratios for the individual frequencies and is given by

\[
\log \Lambda_k = \frac{1}{K} \sum_{k=0}^{K-1} \log \Lambda_k \geq H_1 
\]

and elaborating on Eq.(8)

\[
\log \Lambda = \frac{1}{K} \sum_{k=0}^{K-1} \{\gamma_k - \log (\gamma_k - 1)\} \geq H_1 \tag{9} 
\]

where \( \eta \) denotes the threshold of decision.

2.3. Multiple Microphone LRT (MM-LRT)

Using multiple observations to enhance the likelihood of a VAD system has shown good properties in previous studies \([11]\). In the MO-LRT system the decision rule is formulated over a sliding window consisting of \( 2D + 1 \) observation vectors around the frame for which the decision is being made. The likelihood ratio for MO-LRT is given by \([11]\)

\[
\log \Lambda_{MO} = \frac{1}{K(2D+1)} \sum_{d=1}^{2D+1} \log \left\{ \gamma_{k,d} - \log (\gamma_{k,d} - 1) \right\} \geq H_1 
\]

Nevertheless, this approach imposes a \( D \)-frame delay to the algorithm and \( 2D + 1 \) times increased computations that, for several applications, like real-time operating telecommunication systems can be a major disadvantage.

Nowadays, microphone arrays have become a commodity in commercial (mobile phones, VOIP terminals) and research environments (smart rooms). The modification of MO-LRT we propose, takes advantage of the additional information, required to enhance the decision of an LRT based VAD, that can be retrieved by the available microphones rather than using past information through sliding windows that increase the overall delay. This modification on MO-LRT to a Multiple Microphone LRT (MM-LRT) based VAD relies on the following ratio test

\[
\log \Lambda_{MM} = \frac{1}{KM} \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \{\gamma_{m,k} - \log (\gamma_{m,k} - 1)\} \geq H_1 \tag{12} 
\]

where \( M \) denotes the number of available microphones.

2.4. Combining MO-LRT and MM-LRT

For the cases that the VAD system doesn’t need to operate real-time we propose the combination of MM-LRT Eq.(12) and MO-LRT Eq.(11) to a Multiple Microphone Multiple Observation LRT (MM-MO-LRT). This combination can potentially enhance even further the performance of such systems, given that the conditions of operation allow for the increased delay.

The combined likelihood ratio is given by

\[
\log \Lambda_{MM-MO} = \frac{1}{KM(2D+1)} \sum_{m=0}^{M-1} \sum_{d=1}^{2D+1} \sum_{k=0}^{K-1} \{\gamma_{m,d,k} - \log (\gamma_{m,d,k} - 1)\} \geq H_1 \tag{13} 
\]

where \( M \) and \( D \) indicate the number of the employed microphones and the introduced delay respectively.

2.5. SNR Estimation

The values of speech and noise power spectrum have to be continuously tracked for accurate VAD performance. The methodology of \([6]\) is followed, namely Predicted Estimation (PD). According to PD method, the a priori SNR is estimated on the power spectrum of noise \( \lambda_{n,k}(t) = \sigma_{n,k}(t)^2 \) and speech \( \lambda_{s,k}(t) = \sigma_{s,k}(t)^2 \) which are given by

\[
\hat{\lambda}_{n,k}(t+1) = \xi_n \hat{\lambda}_{n,k}(t) + (1 - \xi_n) E[|N_k(t)|^2 |X_k(t)|] 
\]

\[
\hat{\lambda}_{s,k}(t+1) = \xi_s \hat{\lambda}_{s,k}(t) + (1 - \xi_s) E[|S_k(t)|^2 |X_k(t)|] 
\]

\[
(14) 
\]
where \( \hat{\lambda}_{n,k}(t), \hat{\lambda}_{s,k}(t) \) are estimates of \( \lambda_{n,k}(t), \lambda_{s,k}(t) \) and \( \zeta_n, \zeta_s \) are smoothing parameters both set to 0.99. Following the considerations in [6], Eq.(14) can be further analyzed as:

\[
\hat{\lambda}_{n,k}(t + 1) = \zeta_n \hat{\lambda}_{n,k}(t) + (1 - \zeta_n) \left[ p(H_0|X_k)|X_k(t)|^2 + \left( \frac{\hat{\lambda}_{s,k}(t)}{1 + \xi_{n,k}^{PD}(t)} \right)^2 p(H_1|X_k) \right]
\]

\[
\hat{\lambda}_{s,k}(t + 1) = \zeta_s \hat{\lambda}_{s,k}(t) + (1 - \zeta_s) \left[ p(H_0|X_k)|X_k(t)|^2 + \left( \frac{\hat{\lambda}_{n,k}(t)}{1 + \xi_{s,k}^{PD}(t)} \right)^2 p(H_1|X_k) \right]
\]

respectively, where the a priori SNR \( \xi_{n,k}^{PD} \) at time instant \( t \) is estimated as

\[
\xi_{n,k}^{PD}(t) \equiv \frac{\hat{\lambda}_{n,k}(t)}{\lambda_{n,k}(t)}
\]

and the speech absence probability is

\[
p(H_0|X_k) = \frac{1}{1 + \frac{P(H_1|X_k)\Lambda_{k}}{P(H_0|X_k)}}
\]

The speech presence probability is therefore given by

\[
p(H_1|X_k) = 1 - p(H_0|X_k)
\]

2.6. Decision Smoothing

In order to enhance the performance of the hypothesis tests the following forgetting scheme is employed

\[
\Phi(t) \equiv (1 - \lambda_\Lambda)\Phi(t-1) + \lambda_\Lambda \log \Lambda(t)
\]

where \( \lambda_\Lambda \) a smoothening factor and \( \Phi(t) \) the smoothed likelihood.

3. Experimental Setup

To evaluate the performance of the proposed MM-LRT VAD, the following metrics were employed:

- **Speech Detection Error Rate** (\( P_e \)) : the ratio of the incorrect decisions at speech segments over the total time of speech segments (voice clipping).
- **Non-speech Detection Error Rate** (\( P_f \)) : the ratio of the incorrect decisions at non-speech segments over the total time of non-speech segments (false alarm).
- **Average Detection Error Rate** (\( P_e \)) : the average error rate estimated as the mean of \( P_e \) and \( P_f \).

The \( P_e \) and \( P_f \) were evaluated using the speech recordings performed in the anechoic chamber of Aalborg University Denmark using a close talking microphone (Section II). For the anechoic data collection, 13 participants (7 males and 6 females) were recorded at 16kHz, speaking at mother-languages (Arabic, Bulgarian, Chinese, Greek, Italian, Portuguese, Urdu, and Turkish) for approximately 15 min each, reading sentences and words presented to them with random pause intervals. Eight different languages appear in the data set. The participants were also recorded speaking English for 15 additional minutes under the same pattern. Speech intervals occupy half of the recording time. The recordings have been annotated manually.

Speech data were contaminated artificially with white and vehicular noises from NOISEX-92 database [13]. The microphone array data were artificially generated using the Image Method [14] for a reverberation time of \( T_{60} = 0.15 \text{sec} \) and room dimensions [4.4, 5.8, 2.6]m. The speaker was 2.5m away from the linear array. The input data were sampled at 8 kHz and segmented into overlapping frames of 40 msec duration (10 msec step size).

The performance was evaluated under several scenarios and has been compared to SM-LRT, MO-LRT, the proposed MM-MO-LRT combination and to the standard ITU-T G.729 Annex B VAD. For a fair evaluation of the systems the same \( \lambda_\Lambda \) = 0.04 and frame/step sizes have been used. By examining the detection performance under a variety of noisy conditions, a set of thresholds \( \eta \) for each scheme and noise scenario has been heuristically defined.

4. Performance Discussion

Figure 1 depicts the difference in likelihood ratio when employing 2 and 7 microphones in Eq.(12) at 10dB of vehicular noise. In the latter case the likelihood ratio is significantly enhanced. The LRT value of short silence intervals within words at intervals of speech has been increased. This results in a system the likelihood ratio of which is more uniform within speech segments assisting the overall behavior of the system towards clipping error reduction.

This type of performance enhancement can be evaluated as a function of speech detection rate versus the normalized value of threshold that is employed every time. To do this the range of values for the likelihood ratio are normalized to 1.
Table 1: Performance Results under Various Types of Noise

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR</th>
<th>SM-LRT</th>
<th>MF-LRT (m=7)</th>
<th>MO-LRT (d=3)</th>
<th>MO-MO-LRT (m=7, d=1)</th>
<th>G.729</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c</td>
<td>f</td>
<td>e</td>
<td>f</td>
<td>e</td>
<td>f</td>
</tr>
<tr>
<td>AWGN</td>
<td>20dB</td>
<td>7.85</td>
<td>6.46</td>
<td>9.24</td>
<td>7.54</td>
<td>7.50</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>13.89</td>
<td>19.97</td>
<td>7.81</td>
<td>9.85</td>
<td>10.17</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>16.50</td>
<td>19.80</td>
<td>13.20</td>
<td>9.72</td>
<td>9.83</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>19.64</td>
<td>18.58</td>
<td>20.71</td>
<td>9.50</td>
<td>8.47</td>
</tr>
<tr>
<td>Vehicle</td>
<td>20dB</td>
<td>6.56</td>
<td>6.26</td>
<td>6.85</td>
<td>5.52</td>
<td>5.37</td>
</tr>
<tr>
<td></td>
<td>15dB</td>
<td>8.55</td>
<td>9.45</td>
<td>7.65</td>
<td>6.80</td>
<td>6.49</td>
</tr>
<tr>
<td></td>
<td>10dB</td>
<td>14.74</td>
<td>11.72</td>
<td>17.77</td>
<td>11.78</td>
<td>10.72</td>
</tr>
<tr>
<td></td>
<td>5dB</td>
<td>18.53</td>
<td>19.57</td>
<td>17.49</td>
<td>16.44</td>
<td>13.71</td>
</tr>
</tbody>
</table>

Figure 2 depicts the performance gain when increasing the number of microphones in Eq.(12) for 10dB of vehicular noise. As shown, the system’s performance is significantly enhanced by just introducing a second microphone. Additional microphones have a positive effect to the response of the system.

Figure 3 illustrates the performance of the previously discussed VAD systems under 5dB of vehicular noise. MM-LRT and MO-LRT are compared under the same computational complexity in terms of iterations performed to evaluate Eq.(11) and Eq.(12) respectively. For a fair evaluation the two systems where set to operate under the same number of iterations thus, \( M = 2D + 1 \). A number of \( M = 7 \) microphones has been selected for this case that results in increasing the delay by \( D = 3 \) frames. The results show that MM-LRT performs better or equal to MO-LRT with the same complexity and significantly lower delay. The scenario of MM-MO-LRT has been also evaluated for the least frame delay increment \( D = 1 \) showing that it outperforms the systems with the cost of additional delay.

The performance of the proposed systems has been also evaluated under white noise as illustrated in Table 1, showing similar properties to vehicular noise operation.

5. Conclusions

A statistical VAD, which relies on a multiple microphone likelihood ratio test has been proposed in this paper. The system is based on processing signals captured by far-field microphone arrays. This way the proposed scheme is taking advantage of the spatial information provided by multiple sensors without assuming knowledge of direction-of-arrival estimates. In scenarios that are not real-time critical the system can be further extended to include additional observations employing a sliding window around the currently processed frame. Through simulations we have demonstrated that the proposed system remains more robust than a set of related counterparts.

6. References