# MICROPHONE ARRAY VOICE ACTIVITY DETECTION AND NOISE SUPPRESSION USING WIDEBAND GENERALIZED LIKELIHOOD RATIO

*Ilyas Potamitis*[†], *Eran Fishler*[±]

[†]Wire Communications Laboratory, Electrical and Computer Engineering Dept.,
University of Patras, 261 10 Rion, Patras, Greece, Tel:+30 2610 991722, Fax:+30 2610 991855
[±]Electrical Engineering Dept., Princeton University, USA, Tel: (609) 2586868, Fax: (609) 2583745
e-mail: potamitis@wcl.ee.upatras.gr, efishler@ee.princeton.edu

## Abstract

The subject of this work is the use of microphone arrays for speech activity detection and noise suppression in the case of a moving speaker. The approach is based on the generalized likelihood ratio test applied to the framework of far-field, wideband moving sources (W-GLRT). It is shown that under certain distributional assumptions the W-GLRT provides a unifying framework for evaluation of Direction of Arrival (DOA) measurements against spurious DOAs, probabilistic speech activity detection as well as noise suppression. As regards speech enhancement, we demonstrate the direct connection of W-GLRT with enhancement based on subspace methods. In addition, through the concept of *directive a-priori SNR* we demonstrate its indirect connection with Minimum Mean Square Error spectral (MMSE_SA) and log-spectral gain modification (MMSE_LSA). The efficiency of the approach is illustrated on a moving speaker where additive white Gaussian Noise (AWGN) is present in the acoustical field at very low SNRs.

## 1. Introduction

Microphone arrays have received considerable interest during the last decade as a potential replacement of headset and directive microphones for speech communication applications. Mostly because of the spatial selectivity of their receptive field they have the advantage over one-channel techniques to suppress interfering talkers, directive noise, and to certain extent reduce the reverberation. This makes them good candidates for a wide variety of applications including speech recognition of distant talkers and speech enhancement for coding applications (i.e. hands-free GSM).

The beamformer output is a filtered estimate of the time-aligned microphone outputs where coherence of signals is achieved by using the estimated DOA of the speakers. The DOAs can be estimated by high resolution spectral estimation techniques like MUSIC and Minimum Variance, or by finding the maximum power of a steered response over a range of angles or by first estimating the time delay of arrival between sets of microphones [1].

Beamforming on a *moving* speaker imposes some restrictions in the way microphone outputs are processed. In order to avoid unrealistic stationarity assumptions one has to allow the derivation of the spatial correlation matrix from few frames (snapshots). However, speech pauses, reverberation, background noise and the lack of long data segments to perform some kind of ensemble averaging of the microphone outputs leads to the estimation of unreliable angle fixes. Although the beamforming process as such provides some degree of enhancement against directive interference, it is beneficial to incorporate noise suppression through an adaptive noise cancellation stage or post-processing after the focusing procedure (see [1] and the references therein).

In this work we demonstrate a unifying treatment of DOA evaluation and speech enhancement by employing the wideband generalized likelihood ratio (W-GLRT).

The W-GLRT detector assumes that in each spectral bin both the source (if present) and the additive noise are zero-mean i.i.d complex Gaussian probability distributions with unknown variance. Assuming independence between spectral components the case of a single Gaussian point source embedded in additive Gaussian [2] is generalized to the case of wideband signals. In this work a broadband scanning scheme that estimates the DOA of a far-field moving source is employed. Subsequently, the DOA measurement is validated through the use of W-GLRT against the null hypothesis 'noise only case' and against the hypothesis that speech is present. In this way W-GLRT is used to either associate a DOA measurement to an active speaker and allow the beamforming procedure to carry on, or to reject the DOA measurement and switch off the microphones (in a software sense). It is also shown that the probability of speech absence/presence which serves as statistical voice activity detector (VAD) and as a gain modification rule (see [3-6] for the one-channel case) can be directly related to the W-GLRT.

Finally, it is demonstrated that under the distributional assumptions of W-GLRT the maximum likelihood estimation of the speech variance has a signal subspace interpretation allowing the direct reconstruction of the enhanced signal. Moreover, the introduction of the *directive a-priori SNR* concept allows the association of the powerful MMSE spectral [7] and log-spectral amplitude estimators [8] with the theory of microphone arrays. Therefore, the proposed technique combines the advantage of providing the spatially selective speech acquisition of a microphone array along with the non-tonal residual noise of the MMSE Spectral and log-Spectral Amplitude estimators.

Implementation code and sample recordings can be found at http://slt.wcl.ee.upatras.gr/potamitis/GL.zip. This work was supported by the EC Project INSPIRE (IST-2001-32746).

## 2. Problem Formulation

### 2.1. Wideband GLRT

Due to the nonstationary nature of the speech the microphone outputs are segmented into fixed blocks, signal stationarity is assumed for each block and each FFT bin is considered a narrowband signal. The output from a linear microphone array

of $M$ sensors receiving a far-field wideband signal sampled at time instances $t$=1,..,T is:

$$\mathbf{x}(f,\mathrm{t})=\left[\mathrm{X}_1(f,\mathrm{t}),..,\mathrm{X}_\mathrm{M}(f,\mathrm{t})\right]^\mathrm{T} \in \mathbb{C}^{\mathrm{Mx1}}$$

where $\mathrm{X_m}(f,\mathrm{t})$ m=1,..,M is the DFT coefficient of the signal at bin $f$ received at the $m^{th}$ microphone. The array observation vector is assumed to satisfy:

$$\mathbf{x}(f,\mathrm{t})=\mathbf{a}(f,\theta_0)\mathrm{s}(f,\mathrm{t})+\mathbf{n}(f,\mathrm{t}) \quad (1)$$

where the $\mathbf{a}(f,\theta)$ is the steering vector of the source in the direction of $\theta_0$ for frequency $f$.

$$\mathbf{a}(f,\theta_0)=\left[1, e^{-\mathrm{j}\kappa d\cos(\theta_0)},.., e^{-\mathrm{j}\kappa d(\mathrm{M}-1)\cos(\theta_0)}\right]^\mathrm{T} \quad (2)$$

Let $\mathrm{N_m}(f,\mathrm{t})$ m=1,..,M be the DFT coefficients of the noise process at bin $f$ received at the $m^{th}$ microphone. $\mathbf{n}(f,\mathrm{t})$ is set as:

$$\mathbf{n}(f,\mathrm{t})=\left[\mathrm{N}_1(f,\mathrm{t}),..,\mathrm{N}_\mathrm{M}(f,\mathrm{t})\right]^\mathrm{T} \in \mathbb{C}^{\mathrm{Mx1}} \quad (3)$$

Let L be the window size and also the size of the FFT, then $f=\mathrm{f_s}k/\mathrm{L}$ $k$=0,..,L-1 where $\mathrm{f_s}$ is the sampling frequency, $\kappa=2\pi f/c$, $d$ the microphone spacing, and $c$ is the sound velocity ($c\approx343$ m/sec). Let $\mathbf{S}(\mathrm{t}) \in \mathbb{C}^{\mathrm{LxL}}$ and $\mathbf{N}(\mathrm{t})$, $\mathbf{A}(\theta_0) \in \mathbb{C}^{\mathrm{MxL}}$ be the matrices holding the DFT of the source, noise and steering vectors respectively in all frequencies, then

$$\mathbf{X}(\mathrm{t})=[\mathbf{x}(f_1,\mathrm{t}),..,\mathbf{x}(f_\mathrm{L},\mathrm{t})]^\mathrm{T}$$
$$\mathbf{S}(\mathrm{t})=\mathrm{diag}[\mathrm{s}(f_1,\mathrm{t}),..,\mathrm{s}(f_\mathrm{L},\mathrm{t})]$$
$$\mathbf{N}(\mathrm{t})=[\mathbf{n}(f_1,\mathrm{t}),..,\mathbf{n}(f_\mathrm{L},\mathrm{t})]^\mathrm{T}$$
$$\mathbf{A}(\theta_0)=[\mathbf{a}(f_1,\theta_0),..,\mathbf{a}(f_\mathrm{L},\theta_0)]$$

Under the null hypothesis $\mathrm{H_0}$ speech is absent ($\mathbf{X}(\mathrm{t})$ t=1,..,T is noise only), while under the $\mathrm{H_1}$ speech is present. That is

$$\mathrm{H_0}: \mathbf{X}(\mathrm{t})=\mathbf{N}(\mathrm{t})$$
$$\mathrm{H_1}: \mathbf{X}(\mathrm{t})=\mathbf{A}(\theta_0)\mathbf{S}(\mathrm{t})+\mathbf{N}(\mathrm{t}) \quad (4)$$

The GLRT is based on the following assumptions:

a) Each spectral component in each frame is considered independent from the others.
b) $\mathbf{x}(f,\mathrm{t})$, $\mathbf{n}(f,\mathrm{t})$, are zero mean, temporally white, uncorrelated complex Gaussian distributions with unknown variance.
c) Added noise is uncorrelated with the clean spectrum.

Let the joint pdf of the received vectors $\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)$ under hypothesis $\mathrm{H_i}$ over all frequencies, be:

$$p_i\left(\mathrm{X}(1),...,\mathrm{X}(\mathrm{T})\middle|\psi_i\right)=\prod_f p_{i,f}\left(\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)\middle|\psi_{i,f}\right)$$

where $\boldsymbol{\psi}_{\mathrm{i},f} \in \boldsymbol{\Psi}_{\mathrm{i},f}$ stands for the unknown parameters under hypothesis $i$ and $\boldsymbol{\Psi}_{\mathrm{i},f}$ denotes the parameter space of $\boldsymbol{\psi}_{\mathrm{i},f}$ all defined for each bin $f$. The W-GLRT based on T array snapshots is used for deciding whether to accept $\mathrm{H_0}$ or $\mathrm{H_1}$ and is set as:

$$T(\mathrm{X}(1),...,\mathrm{X}(\mathrm{T}))=\log\frac{\prod_f \max_{\psi_0\in\Psi_0} p_{0,f}\left(\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)\middle|\psi_{0,f}\right)}{\prod_f \max_{\psi_1\in\Psi_1} p_{1,f}\left(\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)\middle|\psi_{1,f}\right)}=\log\frac{\prod_f p_{0,f}\left(\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)\middle|\hat{\psi}_{0,f}\right)}{\prod_f p_{1,f}\left(\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)\middle|\hat{\psi}_{1,f}\right)} \quad (5)$$

where $\hat{\psi}_i$ stands for the ML estimates of the unknown parameters under hypothesis $i$. Under the null hypothesis, the unknown parameter vector $\boldsymbol{\psi}_{\mathrm{i},f}$ contains only the unknown noise level in spectral bin $f$, whereas under the alternative hypothesis, the unknown parameter vector contains the unknown source power and the unknown noise level. Finally $\boldsymbol{\psi}_\mathrm{i}$ is the vector of the unknown parameters over all frequencies under hypothesis $i$. That is,

$$\hat{\psi}_i=\arg\max_{\psi_{i,f}\in\Psi_{i,f}}\prod_f p_{i,f}\left(\mathbf{x}_1(f),..,\mathbf{x}_\mathrm{T}(f)\middle|\psi_{i,f}\right)$$

$$\boldsymbol{\psi}_0=\left[\sigma^2_{\mathrm{n},f}\right]^\mathrm{T},\ \boldsymbol{\psi}_1=\left[\sigma^2_{\mathrm{s},f},\sigma^2_{\mathrm{n},f}\right]^\mathrm{T},\ f=\frac{k}{L}f_s,\ k=0,..,\mathrm{L}\text{-}1$$

If $T(\mathrm{X}_1,..,\mathrm{X}_\mathrm{T})>\gamma$ (the detector threshold) then $\mathrm{H_0}$ is accepted, otherwise $\mathrm{H_1}$ is accepted instead. The threshold $\gamma$ is set to ensure the required probability of false alarm. In operational conditions, the detector's threshold is derived from the initial frames assumed to have only noise. One should note that due to possible movement of the speaker the number of snapshots on which GLRT is based is restricted. A small T leads to more spurious DOAs due to sparse data while a large T restricts the resolution of the likelihood to discern speech from noise. For 8kHz sampling T should be between 1 and 5. The source bearing $\theta_0$ can be estimated by any wideband DOA estimation method. Every spectral coefficient in each snapshot is a circular, zero mean, white, complex Gaussian vector with covariance matrix $\mathbf{R}_{\mathbf{x},f}=\sigma^2_{\mathrm{n},f}\mathbf{I}$ under the null hypothesis, and $\mathbf{R}_{\mathbf{x},f}=\sigma^2_{\mathrm{s}}\mathbf{a}(f,\theta_0)\mathbf{a}^\mathrm{H}(f,\theta_0)+\sigma^2_{\mathrm{n},f}\mathbf{I}$ under the speech presence hypothesis. The detection problem is now set as:

$$\mathrm{H_0}: \mathbf{X}(\mathrm{t}) \sim \prod_f N\left(\mathbf{0},\ \sigma^2_{\mathrm{n},f}\mathbf{I}\right)$$
$$\mathrm{H_1}: \mathbf{X}(\mathrm{t}) \sim \prod_f N\left(\mathbf{0},\ \sigma^2_{\mathrm{s},f}\mathbf{a}(f,\theta_0)\mathbf{a}^\mathrm{H}(f,\theta_0)+\sigma^2_{\mathrm{n},f}\mathbf{I}\right) \quad (6)$$

Under the Gaussian assumption and independence of spectral components, the W-GLRT as set in Eq. 5 takes the form of:

$$\begin{aligned}T(\mathrm{X}(1),..,\mathrm{X}(\mathrm{T}))&=-\sum_f \max_{\psi_0\in\Psi_0}\left\{\mathrm{N}\left(\log\left|\mathbf{R}_{\mathbf{x},f}(\psi_0)\right|+Tr\left\{\mathbf{R}_{\mathbf{x},f}(\psi_0)^{-1}\hat{\mathbf{R}}(f)\right\}\right)\right\}\\&\quad+\sum_f \max_{\psi_1\in\Psi_1}\left\{\mathrm{N}\left(\log\left|\mathbf{R}_{\mathbf{x},f}(\psi_1)\right|+Tr\left\{\mathbf{R}_{\mathbf{x},f}(\psi_1)^{-1}\hat{\mathbf{R}}(f)\right\}\right)\right\}\\&=-N\left\{\sum_f\left(\log\left|\mathbf{R}_{\mathbf{x},f}(\hat{\psi}_0)\right|+Tr\left\{\mathbf{R}_{\mathbf{x},f}(\hat{\psi}_0)^{-1}\hat{\mathbf{R}}(f)\right\}\right)\right\}\\&\quad+N\left\{\sum_f\left(\log\left|\mathbf{R}_{\mathbf{x},f}(\hat{\psi}_1)\right|+Tr\left\{\mathbf{R}_{\mathbf{x},f}(\hat{\psi}_1)^{-1}\hat{\mathbf{R}}(f)\right\}\right)\right\}\end{aligned} \quad (7)$$

where $\hat{\mathbf{R}}(f)=\frac{1}{T}\sum_{t=1}^{\mathrm{T}}\mathbf{x}(f,\mathrm{t})\mathbf{x}^\mathrm{H}(f,\mathrm{t})$ is the empirical correlation matrix at each bin $f$:

$$\hat{\psi}_\mathrm{i}=\arg\max_{\psi_i\in\Psi_i}\left\{\left(\log\left|\mathbf{R}_{\mathbf{x},f}(\hat{\psi}_\mathrm{i})\right|+Tr\left\{\mathbf{R}_{\mathbf{x},f}(\hat{\psi}_\mathrm{i})^{-1}\hat{\mathbf{R}}(f)\right\}\right)\right\}$$

Plugging the exact distributions of Eq. 6 in Eq. 7 results in:

$$\begin{aligned}T(\mathrm{X}(1),...,\mathrm{X}(\mathrm{T}))&=-NMTr\left\{\hat{\mathbf{R}}(f)\right\}-N\\&\quad+\sum_f\log\left(\sigma^2_{\mathrm{s},f}\mathbf{a}(f,\theta_0)\mathbf{a}^\mathrm{H}(f,\theta_0)+\sigma^2_{\mathrm{n},f}\right)\\&\quad+\sum_f Tr\left\{\left(\sigma^2_{\mathrm{s}}\mathbf{a}(f,\theta_0)\mathbf{a}^\mathrm{H}(f,\theta_0)+\sigma^2_{\mathrm{n},f}\mathbf{I}\right)^{-1}\hat{\mathbf{R}}(f)\right\}\end{aligned} \quad (8)$$

The variance of the spectral coefficients of speech $\sigma^2_{\mathrm{s},f}(\theta_0)$ is found by projecting the empirical correlation matrix on the subspace spanned by the steering vector in the direction of $\theta_0$ while noise estimation is based on its orthogonal counterpart. Since the noise exists in the steering vector subspace as well as in speech, one should subtract the variance of the amount of noise projected on this subspace from the spectral variance of speech (for the narrowband case see e.g. [9]).

$$\sigma^2_{\mathrm{n},f}(\theta_0)=\frac{1}{M-1}Tr\left\{\mathbf{I}-\left(\mathbf{a}(f,\theta_0)\mathbf{a}^\mathrm{H}(f,\theta_0)/\mathbf{a}^\mathrm{H}(f,\theta_0)\mathbf{a}(f,\theta_0)\right)\hat{\mathbf{R}}(f)\right\} \quad (9)$$

$$\sigma^2_{\mathrm{s},f}(\theta_0)=\left\{Tr\left\{\mathbf{a}^\mathrm{H}(f,\theta_0)\hat{\mathbf{R}}(f)\mathbf{a}(f,\theta_0)\right\}-\sigma^2_{\mathrm{n},f}(\theta_0)\right\}/\mathbf{a}^\mathrm{H}(f,\theta_0)\mathbf{a}(f,\theta_0) \quad (10)$$

### 2.2. Speech Enhancement

#### *2.2.1. Signal Subspace Method*

The enhanced speech signal in each spectral bin can be directly reconstructed from Eq. 10. After the subspace subtraction has been applied to all frequencies, when steered in direction $\theta_0$ the beamformer returns a filtered output for each $\theta_0$. The beamformer's output is given by:

$$\hat{\mathbf{x}}(f,\mathrm{t})=\mathbf{w}^\mathrm{H}(f,\theta_0)\mathbf{x}(f,\mathrm{t}) \quad (11)$$

The proposed technique is independent of the beamforming method applied. However, in this work we applied minimum variance beamforming [1] and, therefore:

$$\mathbf{w}(f,\theta_0)=\frac{\hat{\mathbf{R}}^{-1}(f)\mathbf{a}(f,\theta_0)}{\mathbf{a}^{H}(f,\theta_0)\hat{\mathbf{R}}^{-1}(f)\mathbf{a}(f,\theta_0)}$$

Based on the assumption that the human ear is relatively insensitive to phase distortion, we focus on the short-time amplitude of the speech signal, leaving the noisy phase of the beamformer's output unprocessed. The noisy phase of the beamformer's output is added to the enhanced output. That is,

$$s(f,t)=\sigma_{s,f}(\theta_0)\exp\left(j\angle\hat{\mathbf{x}}(f,t)\right) \quad (12)$$

and subsequently the underlying frame is reconstructed by using inverse FFT and the weighted overlap-and-add method.

*2.2.2. Calculation of Speech Absence/Presence Probability*

We show that conditioned on the current output of the beamformer $\hat{\mathbf{X}}(t)=\left[\hat{\mathbf{x}}(f_1,t),..,\hat{\mathbf{x}}(f_L,t)\right]^T$ we can derive the speech absence probability which is based on the likelihood of Eq. 5, calculated for each snapshot (i.e. T=1). By direct application of the Bayes rule and assuming independence between frequency bins we derive:

$$p\left(H_0|\hat{\mathbf{X}}(t)\right)=\frac{p(H_0)\prod_f p_{0,f}\left(\hat{\mathbf{x}}(f,t)|H_0\right)}{p(H_0)\prod_f p_{0,f}\left(\hat{\mathbf{x}}(f,t)|H_0\right)+p(H_1)\prod_f p_{1,f}\left(\hat{\mathbf{x}}(f,t)|H_1\right)}$$

and finally:

$$p\left(H_0|\hat{\mathbf{X}}(t)\right)=\frac{1}{1+q\left(\frac{\prod_f p_{0,f}\left(\hat{\mathbf{x}}(f,t)|H_0\right)}{\prod_f p_{1,f}\left(\hat{\mathbf{x}}(f,t)|H_1\right)}\right)^{-1}} \quad (13)$$

where $q=\frac{p(H_1)}{p(H_0)}=\frac{1}{0.0625}$, is set the same for every *f*.

Combining Eq. 5 and Eq. 13 yields:

$$p\left(H_0|\hat{\mathbf{X}}(t)\right)=\frac{1}{1+q\exp\left(-T\left(\hat{\mathbf{X}}(t)\right)\right)} \quad (14)$$

The assumption of spectral independence is also used in [6] to form a global soft decision about speech presence. However, in our case the log-likelihood is calculated from signal subspace techniques for the multichannel case. Eq. 14 allows the calculation of speech presence probability (SPP); $p\left(H_1|\hat{\mathbf{X}}(t)\right)=1-p\left(H_0|\hat{\mathbf{X}}(t)\right)$ the aim of which is threefold:

a) It serves as a voice activity detector (VAD).
b) It allows the estimated noise power to be updated regardless whether speech is present or not (see [5-6]).
c) It can be used as an additional gain function for speech enhancement (see paragraph 2.2.3 and [3- 7]).

*2.2.3. MMSE Spectral Amplitude and log-Spectral Amplitude estimation*

The W-GLRT is based on the assumption that the DFT coefficients of speech and noise are independently distributed as complex Gaussian pdfs with a time-varying unknown variance as summarized in Eq. 15, Eq. 16 and Eq. 17.

$$p\left(\hat{\mathbf{x}}(f,t)|H_0\right)=\frac{1}{\pi\sigma^2_{n,f}(\theta_0)}\exp\left(\frac{-|\hat{\mathbf{x}}(f,t)|^2}{\sigma^2_{n,f}(\theta_0)}\right) \quad (15)$$

$$\sigma^2_{d,f}(\theta_0)=\sigma^2_{s,f}(\theta_0)\mathbf{a}^H(f,\theta_0)\mathbf{a}(f,\theta_0) \quad (16)$$

$$p\left(\hat{\mathbf{x}}(f,t)|H_1\right)=\frac{1}{\pi\left(\sigma^2_{d,f}(\theta_0)+\sigma^2_{n,f}(\theta_0)\right)}\exp\left(\frac{-|\hat{\mathbf{x}}(f,t)|^2}{\sigma^2_{d,f}(\theta_0)+\sigma^2_{n,f}(\theta_0)}\right) \quad (17)$$

The suppression gain functions proposed by Ephraim and Malah for the problem of one channel speech enhancement are based on the same assumptions. The latter gain functions are constructed to minimize the mean-square error estimates of spectrum [7] and of log-spectrum [8], (the latter known to be better associated with speech perception). The reader is referred to [7-8] for a detailed derivation. Here we briefly present both enhancement rules, primarily with a view to associate them with microphone arrays through the concept of *directive a-priori SNR* which is DOA dependent. As firstly stated in [3] and later in [4-7] the noise suppression rule can be modified based on the knowledge of speech presence.

$$\begin{aligned}s(f,t)&=E\left\{s(f,t)|\hat{\mathbf{x}}(f,t)\right\}\\&=E\left\{s(f,t)|\hat{\mathbf{x}}(f,t),H_0\right\}p\left(H_0|\hat{\mathbf{x}}(f,t)\right)\\&\quad+E\left\{s(f,t)|\hat{\mathbf{x}}(f,t),H_1\right\}p\left(H_1|\hat{\mathbf{x}}(f,t)\right)\\&=E\left\{s(f,t)|\hat{\mathbf{x}}(f,t),H_1\right\}p\left(H_1|\hat{\mathbf{x}}(f,t)\right)\end{aligned} \quad (18)$$

The rules are summarized in their corresponding gain functions $G^{SA}$ (Eq. 19) and $G^{LSA}$ (Eq. 20) applied to $|\hat{\mathbf{x}}(f,t)|\cdot$

$$E\left\{s(f,t)|\hat{\mathbf{x}}(f,t),H_1\right\}=G_f\left(\xi_f(\theta_0,t),\gamma_f(\theta_0,t)\right)\hat{\mathbf{x}}(f,t)$$

$$G_f^{SA}\left(\xi_f(\theta_0,t),\gamma_f(\theta_0,t)\right)=\frac{\sqrt{\pi}}{2}\sqrt{\frac{\xi_f(\theta_0,t)}{\gamma_f(\theta_0,t)\left(\xi_f(\theta_0,t)+1\right)}}M\left(\frac{\gamma_f(\theta_0,t)\xi_f(\theta_0,t)}{1+\xi_f(\theta_0,t)}\right) \quad (19)$$

$$M(z)=\exp\left(-\frac{z}{2}\right)\left\{(1+z)I_0\left(\frac{z}{2}\right)+zI_1\left(\frac{z}{2}\right)\right\}$$

$$G_f^{LSA}\left(\xi_f(\theta_0,t),\gamma_f(\theta_0,t)\right)=\frac{\xi_f(\theta_0,t)}{\xi_f(\theta_0,t)+1}\exp\left(\frac{1}{2}\int_{u_f(\theta_0,t)}^{\infty}\frac{e^{-t}}{t}dt\right) \quad (20)$$

where $\xi_f(\theta_0,t)$ is the *a-priori SNR ratio* which is dependent on the DOA and is calculated by plugging Eqs. 21 in Eq. 19 and Eq. 20. One should note that *a-priori SNR ratio* is derived from signal subspace projections and not by using the standard method of [4-8]. $\gamma_f(\theta_0,t)$ is the *a-posteriori SNR ratio* and $I_0$ and $I_1$ are the modified Bessel functions of zero and first order.

$$\xi_f(\theta_0,t)=\frac{\sigma^2_{d,f}(\theta_0)}{\sigma^2_{n,f}(\theta_0)},\ \gamma_f(\theta_0,t)=\frac{|\hat{\mathbf{x}}(f,t)|^2}{\sigma^2_{n,f}(\theta_0)},\ u_f(\theta_0,t)=\frac{\gamma_f(\theta_0,t)\xi_f(\theta_0,t)}{\xi_f(\theta_0,t)+1} \quad (21)$$

Finally a time-smoothed version of *directive a-priori* SNR is employed, conducive to the elimination of musical noise [10]:

$$\hat{\xi}_f(\theta_0,t)=a\frac{s^2(f,t)}{\sigma^2_{n,f}(\theta_0)}+(1-a)\max\left(0,\gamma_f-1\right),\quad \alpha=0.98 \quad (22)$$

## 3. Evaluation

### 3.1. Simulation Experiments

The simulation is based on the method of images [11] and takes place in a typical 6.8m×4.55m×3m room. We made use of an array of *M*=8 omni-directional microphones with *d*=0.1m spacing between microphones with its center located at (3.4, 4.3, 1.6)m. During his movement the speaker is continuously active uttering randomly selected digits from the NOISEX-92 database. The speaker starts from (6.4, 0,25, 1.6)m at t=0 and performs a circular movement with angular velocity 0.1257 rad/sec. He walks for 6.25 seconds and stops at 45°. At t=8.25 sec he moves on and at 14.5 sec stops

moving at 90° with regard to the endfire of the array. At t=17.5 sec he continues the circular motion until t=30 sec. The DOAs were calculated by using wideband MUSIC from one hamming-windowed frame (512 samples) with 50% overlap at 8kHz sampling, using 512 points FFT. We conducted experiments that demonstrated the high robustness of our technique against noises that diverged from Gaussianity but due to space limitations we show the results on the additive white Gaussian (AWGN) noise case.

In *Fig.1 First row* one can see the estimated DOAs of the moving speaker by using wideband MUSIC on a per frame basis. DOAs marked with circles are the DOAs which were evaluated by using the W-GLRT and they were rejected. The evaluation process of DOA measurements is clearly demonstrated in *Fig.1 Second Row* where their corresponding log-likelihoods are compared against the threshold derived from the initial frames of the speech signal. One can see that the vast majoriry of rejected DOAs belong to the silence period of the speaker. In *Fig. 1 Third Row* the VAD decision as estimated from the noisy speech is applied to the clean signal for demonstration purposes. The VAD decision is based on the W-GLRT. However, Speech Presence Probability can serve as a voice detector as well.

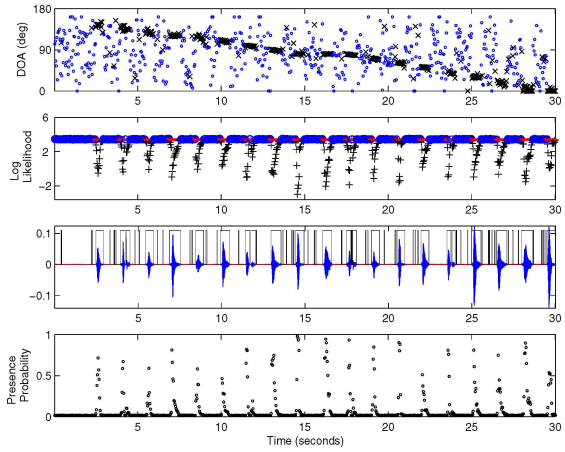


**Fig. 1**. *Fisrt row:* DOA measurements (speech corrupted by AWGN at 10 dB). *Second row:* Log-Likelihood scores (circles are rejected measurements). *Third row:* Clean speech signal and VAD. *Fourth row:* Speech Presence Probability.

### *3.1.1. Speech Enhancement Evaluation*

We performed subjective mean opinion score (MOS) tests in order to assess the quality of perception of the enhanced signals. Comparative evaluation of the proposed algorithm was conducted and the quality and impairment of the enhancement results were scored by 10 speakers. The results were averaged over 10 utterances from 5 male and 5 female speakers taken from the TIMIT database. All speakers followed the scenario of par. 3.1. Table 1 shows the MOS results provided by the enhancement techniques when the array VAD is applied (noise only frames are nulled and noisy speech are enhanced). As regards the quality of perception, in general, the noise reduction of the proposed array-VAD combined with the enhancement rule of Eq. 20 effected smaller percentage of non-tonal noise than the subspace technique (although it inflicted an echo-like artifact) and at low SNRs it consistently outperformed the latter. However, at higher SNRs the subspace subtraction had a small advantage.

| SNR (dB) | Subspace Subtract. | MMSE SA | MMSE LSA | Unprocessed Signal |
|---|---|---|---|---|
| -10 | 2.25 | 1.65 | **2.32** | 1.15 |
| -5 | 2.72 | 2.90 | **2.95** | 1.81 |
| 0 | 3.05 | 3.22 | **3.55** | 2.25 |
| 5 | 3.75 | 3.93 | **3.95** | 2.82 |
| 10 | 4.21 | 4.11 | **4.29** | 3.25 |
| 15 | **4.62** | 4.13 | 4.34 | 3.83 |
| 20 | **4.82** | 4.32 | 4.45 | 4.45 |

**Table 1:** MOS results of the proposed array-VAD combined with noise suppression for Gaussian noise corruption (circularly moving speakers). *Subspace Subtraction* refers to Eq. 10, *MMSE_SA* refers to Eq, 19, MMSE_LSA to Eq. 20.

## 4. Conclusions

We demonstrated that W-GLRT provides a means to combine the spatial selectivity of microphone arrays along with signal subspace and MMSE Spectral and log-Spectral Amplitude enhancement techniques. W-GLRT serves as a unifying statistical framework for VAD using microphone arrays and proved robust at very low SNRs.

## 5. References

[1] Johnson D., Dudgeon D., "Array Signal Processing: Concepts and Techniques," *Prentice Hall,* 1993.

[2] Friedmann J., Fishler E., Messer H., "General Asymptotic Analysis of the Generalized Likelihood Test for a Gaussian Point Source Under Statistical or Spatial Mismodeling," *IEEE Transactions on Signal Processing,* vol. 50, pp. 2617-2631, 2002.

[3] McAulay R., Malpass M., "Speech Enhancement Using a Soft Decision Noise Suppression Filter," *IEEE Trans. Speech & Audio Proc*., vol. 28, no. 2, pp. 137-145, 1980.

[4] Malah D., Cox R., Accardi A., "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," *Proc. IEEE ICASSP*, vol. I, pp. 789-792, 1999.

[5] Soon I., Koh S., Yeo C., "Improved Noise Suppression Filter Using Self-Adaptive Estimator of Probability of Speech Absence," *Signal Proc.,* v. 75, pp. 151-159, 1999.

[6] Kim N., Chang J., "Spectral Enhancement Based on Global Soft Decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, 2000.

[7] Ephraim Y., Malah D., "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109-1121, 1984.

[8] Ephraim Y., Malah D., "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, pp. 443-445, 1985.

[9] Stoica P., Nehorai A., "On the Concentrated Stochastic Likelihood Function in Array Signal Processing," *Circuit Systems & Signal Proc*., vol 14, no. 5, pp. 669-674, 1995.

[10] Cappé O., "Elimination of the Musical Noise Phenomenon Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Speech and Audio Proc.*, vol. 32, pp. 345-349, 1994.

[11] Allen J., Berkley D., "Image Method for Efficiently Simulating small-room Acoustics," *Journal of the Acoust. Society of America*, vol. 65, no. 4, pp. 943-950, 1979.