USING INFORMATION THEORY TO DETECT VOICE ACTIVITY

Fotios Talantzis[*]

Athens Information Technology
Autonomic & Grid Computing Group
0.8km Markopoulo Av., 19002, Athens, Greece

Anthony G. Constantinides

Imperial College London
Electrical & Electronic Engineering
Exhibition Rd., SW72AZ London, UK.

ABSTRACT

Voice Activity Detection systems attempt to discriminate between voice and other ambient sounds. Most systems use a single microphone approach and rely on training prior to employment. The performance of these systems relies heavily on reverberation and noise levels. In this paper we present an unsupervised Voice Activity Detection system that uses pairs of microphones to discern between a coherent acoustic source and spatially diffuse noise of low coherence. Measurement of coherency is performed using an information theoretic metric that integrates means to filter out more effectively the effect of reverberation and noise. Using extensive experiments, the performance of the system is investigated. Based on the conditions imposed by the experimental environments it is shown that the proposed system remains more robust than its counterparts in all cases.

Index Terms— Speech analysis, Speech processing, Array signal processing,Information theory.

1. INTRODUCTION

Voice Activity Detection (VAD) is important in a class of applications ranging from telecommunications to speech enhancement. A VAD system uses at least one microphone to make recordings and decide on the presence or silence of a speech source. Performance is generally a function of noise and reverberation levels as well as the distance between the source and the microphone.

VAD can be performed using algorithms that require training prior to employment or simpler systems that operate without supervision. In supervised systems training is typically performed using Gaussian mixture models [1]. Unavoidably these systems become dependent on the spectral characteristics of the user and the environmental conditions. Unsupervised methods are conceptually simpler. They often assume that there is access to a few seconds where there is no speech activity in order to initialize the system parameters. The short-term energy of the signal along with a simple thresholding is one of the signal features used in early VAD systems [2]. Recent approaches integrate statistical model-based features [3] where a likelihood ratio is developed and a statistical hypothesis test is conducted.

Concepts that originate from information theory have only been recently considered for VAD. In [4], authors used the entropy measure to distinguish between speech and silence as a robust extension to the 3GPP standard. Nevertheless, the system assumes close-talking microphones and during tests ignores the effect of reverberation. For increased robustness VAD measures can be applied on the average value of multiple recordings from microphones residing at different locations. Extending the multi-microphone approach, authors in [5] used phase correlation between a pair of microphones as the discriminating feature between speech and noise/silence.

The work in this paper is a direct extension of [4] and [5]. The system uses a multi-microphone approach and a new coherence measure in order to discern between a coherent acoustic source and spatially diffuse noise of low coherence. The coherence measure is a function of the mutual information (MI) between pairs of microphones and integrates means of reducing the effect of reverberation and noise.

In Section II we start by presenting the system model and the systems used at a later stage for comparison purposes. In Section III the MI-based VAD system is presented. Section IV presents the the performance measures used to evaluate the systems and discusses extensive results from experiments. Finally, conclusions are drawn in Section V.

2. SYSTEM MODEL

Consider the employment of M microphones in a reverberant and noisy environment. VAD uses at least one of these microphones to make recordings and decide on the presence of an acoustic source. VAD is a process that is required to often operate repeatedly and in real-time using short recordings. Thus, data is collected over t frames of L samples which for the t^{th} frame are converted into the frequency domain using an L-point Short Time Fourier Transform (STFT). This is performed over a set of discrete frequencies ω_l with $l = 0, 1, ...L-1$. Let the microphones be arranged in P pairs. The frame generated by the m^{th} microphone ($m = 1, 2$) of

[*] This work has been partly sponsored by the European Union, under the FP7 project HERMES. F. Talantzis is also with Dept. of Electrical & Electronic Engineering, Imperial College London.

the p^{th} pair is then given as:

$$\mathbf{X}_{mp}^{[t]} = [X_{mp}(\omega_0), X_{mp}(\omega_1), \ldots, X_{mp}(\omega_{L-1})], \tag{1}$$

where $p = 1, 2, ...P$ and

$$X_{mp}(\omega_l) = A_{mp}(\omega_l)S(\omega_l) + N_{mp}(\omega_l) \tag{2}$$

in which $S(\omega_l)$ is the STFT of the source signal, $A_{mp}(\omega_l)$ is the room transfer function between the source and the m^{th} microphone of the p^{th} pair, and $N_{mp}(\omega_l)$ is additive white Gaussian noise which is assumed to be uncorrelated with the source signal. Since the analysis will be independent of the time frame we have omitted t from Eq. (2). Thus, we can also drop t to express frames simply as $\mathbf{X}_{mp}$. The task then for any audio-based VAD algorithm is to use only $\mathbf{X}_{mp}$ and decide on whether the acoustic source is active or not, during that period of time. The energy of a frame is the typical criterion used to indicate the presence of speech [2]. Nevertheless, in [4] authors introduced the use of spectral entropy as a measure and demonstrated that robustness increases when compared to energy methods. The use of more than one microphones in the VAD process adds spatial information to the domain of our frequency analysis. Thus, using a microphone array allows the introduction of a method to discern between a coherent acoustic source and spatially diffuse noise of low coherence. Such VAD systems assume the sound source to be in the far-field of the microphones. In [5] authors used phase correlation between a pair of microphone recordings as the discriminating feature. This concept can then be extended to more than one pair of microphones.

3. VAD USING MUTUAL INFORMATION

The VAD system proposed in this work assumes the recording setup of Sec. 2. Extending the architectures of [4],[5] the system will use more than one microphones to detect the presence of speech and MI as the coherence function.

The MI of two variables is an information theoretical measure (function of the entropy measure used in [4]) that represents the difference between the measured joint entropy of the two variables (in our case these are the microphone signals) and their joint entropy if they were independent. Since the microphones of each pair p reside in different spatial locations, their corresponding recordings will be delayed with respect to each other by a relative time delay τ_p. If we assume that the microphone recordings exhibit normal distribution then for any set of frames in any pair p, the MI between the two microphones is:

$$I_N = -\frac{1}{2} \ln \frac{\mathbf{det}[\mathbf{C}(\tau)]}{\mathbf{det}[\mathbf{C}_{11}]\mathbf{det}[\mathbf{C}_{22}]} \tag{3}$$

where τ is a time delay at which we calculate I_N. The joint covariance matrix $\mathbf{C}(\tau)$ is a concatenation of frames $\mathbf{X}_{1p}$ and $\mathbf{X}_{2p}$ shifted by different amounts in samples:

$$\mathbf{C}(\tau) \approx \Re\left\{ \begin{bmatrix} \mathbf{X}_{1p} \\ \mathcal{D}(\mathbf{X}_{1p}, 1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_{1p}, N) \\ \mathcal{D}(\mathbf{X}_{2p}, \tau f_s) \\ \mathcal{D}(\mathbf{X}_{2p}, \tau f_s + 1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_{2p}, \tau f_s + N) \end{bmatrix} \begin{bmatrix} \mathbf{X}_{1p} \\ \mathcal{D}(\mathbf{X}_{1p}, 1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_{1p}, N) \\ \mathcal{D}(\mathbf{X}_{2p}, \tau f_s) \\ \mathcal{D}(\mathbf{X}_{2p}, \tau f_s + 1) \\ \vdots \\ \mathcal{D}(\mathbf{X}_{2p}, \tau f_s + N) \end{bmatrix}^H \right\} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12}(\tau) \\ \mathbf{C}_{21}(\tau) & \mathbf{C}_{22} \end{bmatrix} \tag{4}$$

where the $\Re\{.\}$ operation returns only the real part of its argument and f_s denotes the sampling frequency. Function $\mathcal{D}(\mathbf{A}, n)$ shifts the frequency components contained in frame $\mathbf{A}$ by n samples.

If N is chosen to be greater than zero the elements of $\mathbf{C}(\tau)$ are themselves matrices. In fact for any value of τ, the size of $\mathbf{C}(\tau)$ is always $2(N+1)\times 2(N+1)$. We call N the *order* of the coherence function. N is really the parameter that controls the robustness of the VAD against reverberation. In the above equations and in order to estimate the information between the microphone signals, we actually use the marginal MI that considers jointly N neighboring samples (thus the inclusion of delayed versions of the microphone signals). This way the function of Eq. (3) takes into account the spreading of information due to reverberation and returns more accurate estimates. To avoid possible large variations of I_N over time a median filter over a set of estimated values can be used.

Ideally, for a given pair p we would like to calculate I_N at $\tau = \tau_p$. τ_p is dependent on the source location. In the forthcoming simulations a VAD decision is made either by providing this set of delays for all P pairs or estimate it using a TDE function. We also show how the use of more than one microphone pairs can improve performance further. In order to decide the presence or absence of speech during frame t the value of I_N is then compared to a threshold γ_t. In the experiments to follow we assume that threshold γ_t can be initialized by having access to some frames T_o where the source is not active. The initial value of γ_t is then initialized to be $\gamma_o = \frac{1}{T_o}\sum_{t=1}^{T_o} I_N^{[t]}$ where $I_N^{[t]}$ denotes the value of Eq. (3) for time frame t. As in [4] we further keep estimating γ_t for each of the later frames in real-time. A fixed threshold would lead to decreased robustness in varying acoustic environments. Thus, at frame t, γ_t adapts to a possible environment change as $\gamma_t = \frac{1}{2}(\mathbf{m}_0(\gamma_t, k) + I_N^{[t]})$ where $\mathbf{m}_0(\gamma_t, k)$ denotes the median value of the last k values of γ_t during which no speech was detected. Again, by using the median we avoid the effect of sporadicly large variations in the value of $I_N^{[t]}$ during frames that were detected as silent. In order to decide the presence or absence of speech at a given time frame the value of $I_N^{[t]}$ is compared to γ_t. If $I_N^{[t]} > \gamma_t$ then the system assumes the presence of speech and a binary decision variable α_t is set to 1. If $I_N^{[t]} < \gamma_t$ then $\alpha_t = 0$.

A hangover scheme is important in a real-time VAD system since it allows us to filter out false detections among frames with the correct ones. Suppose for example that for a relatively long period of time silence is being detected. The system then detects a single frame where speech is present and immediately after a series of new frames without speech. For the examined frame sizes this single speech detection should be filtered out since speech utterances can not be that short. Thus, a hangover scheme requires an initial state and two time constants δ_0 and δ_1 that determine after which amount of time we should switch to the no-speech or speech states respectively. In general, $\delta_1 < \delta_0$ to allow for quick transition from no-speech to speech state and to restrict the opposite. This is in line with the fact that speech utterances are generally highly correlated with time [4].

4. PERFORMANCE ANALYSIS

There are four VAD metrics that are used to compare the systems in this paper. For their calculation we test the VAD estimate at each time frame against the ground truth (annotated manually), for the total duration of the test signals. The metrics are: 1. **Mismatch Rate** (MR) is the ratio of the incorrect decisions over the total time of the tested segment, 2. **Speech Detection Error Rate** (SDER) is the ratio of incorrect decisions at speech segments over the total time of speech segments, 3. **Non Speech Detection Error Rate** (NDER) is the ratio of incorrect decisions at non speech segments over the total time of non speech segments, 4. **Average Detection Error Rate** (ADER) is the average of SDER and NDER. All of the above metrics are presented as percentages. The lower their values, the better the performance of the system.

In this section, we compare the performance of the multi-microphone MI-based VAD system as presented in Section 3 with the two systems of [5] and [4]. We will refer to these algorithms as $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ respectively.

One of the most typical applications of VAD systems is PC-based video-conferencing. To demonstrate the improvement in robustness when $\mathcal{A}_1$ is used we recorded four segments of ten minutes duration each. This was done with a typical web-camera [6] (Creative Live Cam Voice) that includes two microphones, placed 7 cm apart. Recordings involved four different speakers (two male and two female) speaking naturally in front of the microphones at a distance of approximately 1 m. The text read was chosen to have different pauses between phrases while the recording also included noise contribution from air-conditioning and the PC used, estimated to result into an SNR of approximately 15 dB. During recordings speakers were allowed small movements in front of the camera. The reverberation time of the room was measured to be approximately $T_{60} = 0.35$ sec.

Parameter γ_o was estimated by allowing 2 sec of silence before the first utterance of speech. Additionally, the relative delay between the two microphones was first estimated using [7] and then used in the $\mathcal{A}_1$ VAD system. Correspondingly, the algorithm of [8] was used prior to $\mathcal{A}_2$. During all tests, f_s was chosen to be 22.05 KHz and $L = 4096$. The hangover scheme applied is retained identical for all algorithms with $\delta_0 = 0.74\ sec$ and $\delta_1 = 0.37\ sec$. In the case of $\mathcal{A}_3$ only one of the microphones was used.

Tables 1(a)-1(c) show the resulting values for all performance measures when the three examined algorithms are used upon the video-conferencing audio data. The tables show that the proposed system remains more robust in detecting speech for the specific speakers and environmental conditions.

To demonstrate the robustness of the MI-based VAD system we also conducted experiments for the scenario of a meeting with far-field recordings. These were performed in the laboratory of Athens Information Technology (AIT). This is s typical reverberant meeting room equipped with microphones and cameras. An overview of the rooms can be seen in Fig. 1. A, B, C, D denote the arrays used in our experiments. Figure 1 also contains the relative geometry of the microphone arrays. Recordings were conducted in presence of ambient noise from both air-conditioning and personal computers (resulting in an SNR estimated to be approximately 10 dB).

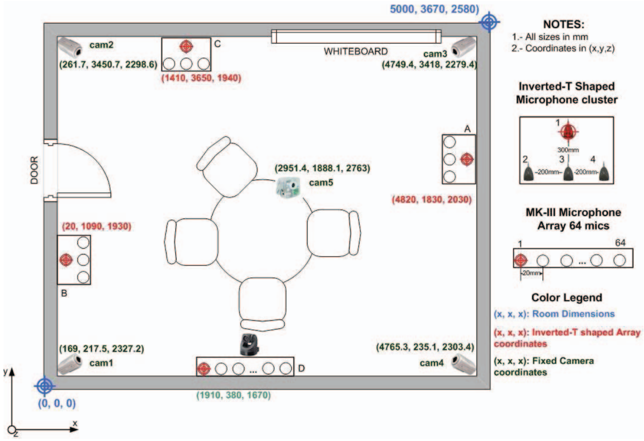

Fig. 1. Overview of AIT's laboratory. Microphone-array topology and geometry is also shown.

A total of four recordings were performed in the laboratory. Each seminar consists of a presentation to a group of 3 attenders. There exists significant interaction between the presenter and the audience, with numerous questions and often a brief discussion amongst participants. This type of seminar recordings provide data with rich acoustic activity. A significant number of acoustic events is generated to allow more meaningful evaluation of the VAD systems. From each of these seminars two five-minute segments were used. The choice of segments and their duration was performed by [9]. The data are annotated by humans to provide the speech activity of each segment. These annotations are considered to be the ground truth for the measurement of the system performance. For $\mathcal{A}_1$ and $\mathcal{A}_2$ three pairs of microphones from the 64-channel linear array (denoted as D) and one pair from each of the other arrays (denoted as A, B, C) were used. Analytically, the pairs used are: $A_2 - A_3$, $B_2 - B_3$, $C_2 - C_3$, $D_{24} - D_{39}$, $D_{17} - D_{32}$ and $D_{33} - D_{48}$. Subscripts denote the microphone number of the corresponding array. The choice of pairs (and the corresponding microphone distances) ensures that far-field assumptions are not violated.

As before TDE algorithms [7] and [8] were used prior to $\mathcal{A}_1$ and $\mathcal{A}_2$ respectively. A majority voting scheme amongst the decision for every pair was used to determine the presence or absence of speech during time frame t. Tables 2(a)-2(c) show the resulting values for all performance measures. Results indicate that $\mathcal{A}_1$ performs better during all segments. Still the SDER and NDER values indicate that all systems fail more often during the silent segments of the recordings.

(a) $\mathcal{A}_1$

$_{Metric}\backslash^{Experiment}$	Male 1	Male 2	Female 1	Female 2
MR	16.85	13.38	17.01	17.66
SDER	15.48	12.12	16.48	17.05
NDER	18.75	15.5	18.21	18.36
ADER	17.11	14.85	17.89	17.99

(b) $\mathcal{A}_2$

$_{Metric}\backslash^{Experiment}$	Male 1	Male 2	Female 1	Female 2
MR	23.75	18.85	23.90	24.82
SDER	21.67	17.02	23.25	23.91
NDER	26.27	21.73	25.58	25.83
ADER	23.99	20.79	25.12	25.35

(c) $\mathcal{A}_3$

$_{Metric}\backslash^{Speaker}$	Male 1	Male 2	Female 1	Female 2
MR	30.57	24.09	30.98	32.15
SDER	28.05	22.14	29.95	30.85
NDER	34.10	28.07	32.84	33.40
ADER	31.10	26.97	32.36	32.40

Table 1. Experimental results for all VAD systems. Results are shown for different speakers. $L = 4096$ samples.

5. CONCLUSION

In the present work a novel multi-microphone VAD system was presented. To discern between an acoustic source and ambient noise the system uses a information theoretic measure that integrates means of filtering out reverberation and noise. Through the use of more than one microphones the system also utilizes the spatial diversity of sensors placed in different physical positions. The system requires no training prior to employment but its performance is subject to TDE estimators and access to a number of silent frames in order to initialize the system. Extensive real experiments were performed. $\mathcal{A}_1$ remained more robust than any of its counterparts for all experiments. It is worth noting that $\mathcal{A}_1$ and $\mathcal{A}_2$ are subject to TDE inaccuracies. Thus, exact knowledge of the TDEs would improve the robustness of these VAD systems further.

(a) $\mathcal{A}_1$

$_{Metric}\backslash^{Sem.-Seg.}$	AIT 1-1	AIT 1-2	AIT 2-1	AIT 2-2
MR	28.50	33.49	11.34	19.29
SDER	21.58	23.24	8.07	15.87
NDER	40.20	39.64	29.65	31.26
ADER	30.89	31.44	18.86	23.56

(b) $\mathcal{A}_2$

$_{Metric}\backslash^{Sem.-Seg.}$	AIT 1-1	AIT 1-2	AIT 2-1	AIT 2-2
MR	42.10	49.39	16.61	25.71
SDER	28.53	30.06	9.81	21.21
NDER	51.55	68.20	39.26	45.27
ADER	37.95	41.80	24.47	29.39

(c) $\mathcal{A}_3$

$_{Metric}\backslash^{Sem.-Seg.}$	AIT 1-1	AIT 1-2	AIT 2-1	AIT 2-2
MR	43.09	54.14	37.16	36.04
SDER	33.42	36.22	33.62	24.09
NDER	77.82	79.96	50.74	50.37
ADER	53.66	56.27	42.18	43.53

Table 2. Performance measures for the two segments of each of the four recordings for all VAD systems.

6. REFERENCES

[1] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 1, pp. 84–91, 2005.

[2] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Proc. of Speech Signals*, Macmillan Publishing, London, 1993.

[3] J. Ramirez, J.C. Segura, C. Benitez, and L. Garcia A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689 – 692, 2006.

[4] R.V. Prasad, R. Muralishankar, S. Vijay, H.N. Shankar, P. Pawelczak, and I. Niemegeers, "Voice activity detection for voip-an information theoretic approach," *IEEE Global Telecommunications Conference*, pp. 1–6, 2006.

[5] M. Brandstein and D.B. Ward (Eds.), *Microphone Arrays Signal Proc. Techniques and Applications*, Springer, London, 2001.

[6] http://gr.europe.creative.com/, ," .

[7] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Proc. Letters*, vol. 12, no. 8, pp. 561 – 564, 2005.

[8] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustics Speech and Signal Proc.*, vol. 24, no. 4, pp. 320–327, 1976.

[9] National Institute of Standards and Technology, "Classification of events, activities and relationships (clear)," http://www.clear-evaluation.org/, 2007.