SIGNAL INDEPENDENT WIDEBAND ACTIVITY DETECTION FEATURES FOR MICROPHONE ARRAYS

Tuomo W. Pirinen[*] *and Ari Visa*

Institute of Signal Processing, Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland
{tuomo.pirinen, ari.visa}@tut.fi

ABSTRACT

Detection of activity is a key capability for microphone arrays. An array system should tell when a source of interest is present and evaluate the usability of the computed spatial estimates. This work proposes activity features that are computed from spatial data only, using time delays and direction of arrival. The features are validated with a loudspeaker experiment. Results show that the features are effective: tolerable detection errors are achievable with simple detection methods. In addition, direction of arrival estimation error is reduced down to one third when unreliable estimates are discarded.

1. INTRODUCTION

When microphone array direction of arrival (DOA) estimates are utilized, two questions arise. Firstly, is there a source of interest present in the estimated data? Secondly, are the spatial estimates from this source, and are they reliable? Answering these questions is vital in many applications of microphone arrays, including surveillance systems [1]; traffic monitoring [2], [3]; and speech applications [4], [5].

The first question can be partially answered with conventional methods that are based on signal models. These methods, such as voice activity detectors, utilize the time and frequency domain properties of signals to detect periods of activity. This kind of detection is accurate when assumptions of the underlying signal model are met. However, when received signals or background differ from the model or training conditions, detection accuracy suffers. This may corrupt further processing that relies on array information. For example, speech detection errors have an unwanted effect on array based speech enhancement [6]. If signals and background are varying, several models may be required to cover them satisfactorily. In addition, the models usually need to be trained with data that represent the operating conditions comprehensively. For some applications, such as surveillance systems, this can be difficult or even impossible.

Answering the second question is a much more complicated task. Time/frequency model based detectors provide little or no information on the spatial estimate, because they are not designed for this task. For example, a signal model based detector may label a frame of data inactive in low SNR, although an array processor is still able to perform spatial estimation. Or, in multipath conditions, spatial estimates can be largely erroneous even though signal based detection labels the data active.

An automatic system should be able to answer both of the presented questions by telling whether a source is present and is it *known* to be in the estimated direction. Achieving this goal requires detection to be robust against changing conditions and against deviations from the assumed signal model. Such a situation occurs, for example, when the observed signals are from a type of source that has not been encountered before.

One way to increase this kind of robustness is to use spatial features in combination with the time and frequency domain information [7], [8]. Temporal and spectral features are dependent on input signals, thus their utilization requires a signal model or distributional assumptions. This makes the methods prone to errors caused by the above described signal variability.

A more spatially oriented approach is to utilize error criteria from pairwise time delay estimation (TDE). Usually this is done in frequency domain with cross-spectral phase. Detection using the the residual error of phase regression based TDE was proposed in [4],[9], where a low value of error was interpreted as an indication of activity. In [10],[11], error criteria were computed from the pairwise generalized correlation function (GCC) [12]. Although these methods do not employ a signal model, they are affected by the signal and background properties through the cross-spectral phase.

This work addresses the detection robustness by proposing array level features that set restrictions only in the spatial domain by assuming an approximate local planar wave model. The proposed features utilize only the directional properties of received signals by looking at the estimated time delays and the propagation vector [13]. This makes the features usable also in applications involving wideband signals.

The remaining parts of the paper are organized as follows. Section 2 explains the required background theory and proposes the activity detection features. Data recording setup and experiment results are given in Section 3. Section 4 provides a commentary of the results. The paper is summarized in Section 5.

2. ACTIVITY DETECTION

2.1. Direction of arrival estimation

This subsection outlines briefly the local planar wave model for propagation vector estimation. This is done in order to develop the notation required for presenting the proposed features.

Consider a pair of microphones m that receive a signal propagating as a planar wave. The angle of arrival of the wavefront θ determines the time delay τ_m between microphone outputs:

$$\tau_m = \frac{d_m}{c}\cos(\theta), \tag{1}$$

where d_m is the distance between the microphones and c is the wave propagation speed. Now let $\boldsymbol{x}_m$ be the vector connecting the micro-

[*] This work was funded in part by the Nokia Foundation and the Finnish Air Force.

phones of the pair. In addition, define the propagation vector of the planar wave by $\boldsymbol{k} := c^{-1}\boldsymbol{n}$, where $\boldsymbol{n}$ is the unit normal of the wavefront, pointing to the propagation direction. Using these definitions, we have $d_m = \|\boldsymbol{x}_m\|$ and $c^{-1} = \|\boldsymbol{k}\|$. Now (1) can be presented in vector form using dot product [13]

$$\tau_m = \boldsymbol{x}_m \cdot \boldsymbol{k}. \tag{2}$$

Equations for all microphone pairs in the array can be combined as a group of linear equations

$$\boldsymbol{\tau} = \boldsymbol{X}\boldsymbol{k}, \tag{3}$$

where

$$\boldsymbol{\tau} = \left[\ \tau_1, \tau_2, \ldots, \tau_M\ \right]^T, \boldsymbol{X} = \left[\ \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M\ \right]^T, \tag{4}$$

and M is the number of microphone pairs. Direction of arrival (DOA) estimation requires (3) to be solved. This can be done, for example, with the least squares solution [14]

$$\hat{\boldsymbol{k}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\tau}}. \tag{5}$$

The estimated pairwise time delays $\hat{\boldsymbol{\tau}}$ can be obtained with *any* method, e.g., with the one based on regression using cross-spectral phase mentioned in [4],[9] or GCC-based methods described in [12].

It should be emphasized, that the above method has been explained only to render the background needed for the proposed features. Any other method can be used for DOA or location estimation. The above method is just used to provide the activity features, which are proposed in the following section.

2.2. Features for activity detection

Three error statistics, and their variances, are proposed as features for activity detection: a lower bound of the DOA estimation error, a test on linear dependence of time delays, and the residual distance of time delays from the DOA estimate. In comparison to previous approaches, the proposed features differ by operating only on the values of the estimated time delays and the propagation vector. Use of a specific TDE method, or knowledge on its statistical properties are not required. The features are discussed below in more detail, and their computation is summarized as a block diagram in Fig. 1.

If the microphone array is three dimensional (i.e., the rank of matrix $\boldsymbol{X}$ is 3), the wave propagation speed is not needed to solve (3) for $\boldsymbol{k}$. Consequently, prior information about the propagation speed can be compared to the norm of the propagation vector. In [15], it was shown that this kind of comparison gives a lower bound for the DOA estimation error. This lower bound is proposed as the first feature

$$\mathrm{F}_1 := \left| c\|\hat{\boldsymbol{k}}\| - 1 \right|. \tag{6}$$

If the array contains at least four microphones, time delays can be tested for errors with *confidence factors* [16]. These factors utilize the linear dependence of time delays by summing delays on closed paths in the array, e.g., time delays in a triangle formed by three microphones. The actual confidence is obtained by combining all closed paths (e.g., all triangles) that include a given time delay.

For activity detection, the proposed feature is the mean of confidences of all time delays available from the array

$$\mathrm{F}_2 := \frac{1}{M}\sum_{m=1}^{M}\Pi_m, \tag{7}$$

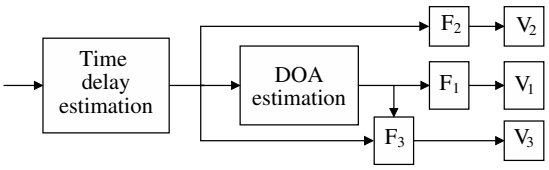

Fig. 1. *Block diagram of feature computation.*

where M is the number of microphone pairs and Π_m is the normalized confidence factor (NCF) for time delay τ_m. The triangular NCF are computed with [16]

$$\Pi_{(i,j)} = \frac{1}{N-2}\sum_{\substack{n=1\\ n\neq i,j}}^{N}\frac{c\left|\,\hat{\tau}_{(i,j)} + \hat{\tau}_{(j,n)} + \hat{\tau}_{(n,i)}\,\right|}{2\left(\|\boldsymbol{x}_{(i,j)}\| + \|\boldsymbol{x}_{(j,n)}\| + \|\boldsymbol{x}_{(n,i)}\|\right)}, \tag{8}$$

where N is the number of microphones. Indices i and j refer to the microphones of pair m in (7).

The third proposed feature is the distance between the estimated time delays and the time delays projected by the solution (5). There are several ways to compute the distance, but for consistency we use the Euclidian distance

$$\mathrm{F}_3 := \|\hat{\boldsymbol{\tau}} - \boldsymbol{X}\hat{\boldsymbol{k}}\|, \tag{9}$$

that is minimized by (5). The distance of time delays from the solution has been proposed as an error metric earlier [4]. However, because the distance is closely related to solving (3) and it describes the "fit" of the estimated propagation vector to the TDE data, it is proposed here as an activity feature.

Rapid changes of these three features are an indication of DOA estimation errors. Therefore, the short-term variances of k past values of features are used as activity features:

$$\mathrm{V}_i(n) := \mathrm{var}\{\mathrm{F}_i(n), \mathrm{F}_i(n-1), , \ldots, \mathrm{F}_i(n-k+1)\}, \tag{10}$$

where i is the feature index and n is the frame index.

3. EXPERIMENTS

The proposed features were tested using data recorded in a hall-like room with dimensions 5.8 m × 8.3 m × 3.5 m. Data was acquired with a four microphone tetrahedron array, with edge length of 0.36 m. The array was located 1.25 m above the center of the floor. Recording sampling rate was 48 kHz.

Eight loudspeakers, located in approximate elliptic fashion around the array, were used as sound sources. Distances of loudspeakers from the array ranged from 2.2 m to 3.5 m, and loudspeakers were 0.39–0.84 m above the array base. Room reverberation times (to 60 dB attenuation) were, depending on the location of source and microphone, between 0.46–0.56 s.

Source signals consisted of three noise samples, three music samples, and six speech samples. Noise samples were maximum length sequences (MLS), that are pseudorandom sequences whose autocorrelation function approximates the unit impulse [17]. The speech samples were generated by combining utterances from the TIMIT database [18]. Three of them contained a female speaker, and three a male speaker. Duration of all used samples was 15 seconds. Each of the samples was played through all of the loudspeakers individually with three second pauses in between. This resulted into a test signal including 96 distinct samples and 1725 seconds of data.

Feature		Noise	Music	Speech	Overall
F_1	fn	9.9	18.6	24.3	19.0
	fp	22.1	19.2	21.7	21.2
F_2	fn	0.8	8.5	18.7	11.2
	fp	6.0	5.1	7.0	6.4
F_3	fn	0.3	3.2	7.2	4.3
	fp	40.7	41.8	44.7	43.4
V_1	fn	1.8	13.3	31.3	18.7
	fp	10.4	11.9	21.6	17.5
V_2	fn	5.1	11.3	30.8	18.8
	fp	11.0	9.8	26.1	20.0
V_3	fn	1.9	12.7	26.2	16.2
	fp	8.8	7.9	11.7	10.4

Table 1. *Performance of proposed features in activity detection by signal type. Performance is measured with relative amount (%) of false negative (*fn*) and false positive (*fp*) detections.*

Detector	Noise	Music	Speech	Overall
No detection	3.7	14.7	23.0	17.6
F_1	1.0	7.6	9.5	7.4
F_2	1.0	5.0	7.0	5.3
F_3	2.2	11.9	21.2	15.6
V_1	1.0	4.5	14.1	9.2
V_2	1.7	6.9	14.4	9.9
V_3	1.0	5.2	18.0	11.9

Table 2. *Angular RMS error in degrees for two dimensional direction of arrival estimation using correctly detected active frames.*

In addition to distortions caused by room reverberation, signal quality was artificially degraded by adding zero-mean, uncorrelated, Gaussian white noise to the microphone data during the estimation process. This lowered the SNR average to 14 dB. SNR variation was, depending on the frame signal content, between 0–25 dB.

DOA estimation and feature computation was performed using 8192 sample windows with 50 % overlap. Pairwise time delays were estimated using the GCC-PHAT estimator [12]. This method was chosen because it is simple and it has been used in prior studies, e.g., [11]. However, the proposed features do not require the TDE method to be GCC-based and therefore, recalling the comment made at the end of Section 2.1, *any* other TDE method could have been used.

Time delays were estimated with the precision allowed by the sampling rate, without interpolation, in order to verify the usability of the features when accurate TDE is not available. Direction of arrival was estimated framewise with (5) and the proposed features were computed for each frame. Length of history for variance based features was $k = 5$ frames, which corresponds to 512 ms.

Activity detection was performed on each frame using the proposed features. Detection threshold was variable, and it was set using exponentially weighted moving average over the feature history. This simple method was chosen in order to test and illustrate the capabilities of the features.

Measured detection performance criteria were *false negative* (fn), the relative amount of active frames detected as inactive; and *false positive* (fn), the relative amount of inactive frames detected as active. Reference activity of speech samples was obtained using the ITU-T P.56 speech level metric [19]; activity threshold was set to the level of the noise added in the TDE process. Performance of the proposed features in detection is given in Table 1.

To evaluate the significance of activity detection for DOA estimation, angular RMS error for two-dimensional DOA estimation was computed from correctly detected active frames. The error values are given in Table 2. The false positive frames are not included in the DOA error assesment, because there is not a reference direction for periods of inactivity. This is not an unfair scoring, because the contribution of false positive detections was already included in the detection accuracy analysis in Table 1.

As seen in Table 1, all features, except F_3, can provide overall detection error rates smaller than 22 %. Within this dataset, speech signals are the most difficult to detect correctly. Even the best feature, F_2, has fn rate of 18.7 % for speech. The variance based features do not improve the performance on speech, except for V_3. This is not surprising, because the performance of F_3 was bad, but it suggests that the single threshold detection is not suitable for use with F_3 and V_3.

The importance of array based detection is clearly visible in the DOA estimation errors in Table 2. Feature F_2, that had the best detection performance, can lower the estimation error below one third of the original for speech and overall.

4. DISCUSSION

The purpose of this work was to introduce a framework for signal independent spatial domain activity detection. The results suggest that the proposed signal independent features carry information about spatial activity. They could be used to support activity detection in microphone arrays, especially when other signal or source specific detection methods are unreliable. However, to achieve an acceptable level of detection performance, further work is required to develop the features and especially detection. Clearly, a more sophisticated detection method is needed, for instance, one that uses the features and their dependencies jointly through a feature-space approach. Also, combining the features with other detection techniques is a matter worth studying.

The detection performance is at most moderate in comparison to other approaches, specifically [8], where detection error rates below 10% were reported for speech at 15 dB SNR. However, keeping in mind that in this work the detection method was simple and not aggressively tuned, the results are promising.

The reduction achieved in RMS errors for DOA estimation is a clear illustration of the significance and the need for spatial activity detection. With this kind of detection, an array processing system can not only give the localization estimate, but also indicate that the estimate is accurate, and from a real source.

The proposed features did not assume a certain type of signals (except for the spatial behavior). This makes them usable in general, also in scenarios involving wideband signals. The features are not strictly bound to acoustic properties of signals, which enables them to be used also beyond the audio domain.

Computational load and memory requirements of propagation vector and feature estimation are minimal in comparison to time delay estimation. Furthermore, the feature computation can be performed without delay, as several passes through the data are not required. For these reasons, the proposed features can easily be used in a real-time array processing system without additional or more complex hardware.

This usability is further supported by the fact that the features can be used with *any* TDE method, not just the GCC-based approach taken in these experiments. In fact, the GCC-based methods are

known to perform poorly in reverberation [20]. Therefore, depending on the application, some other methods (e.g., [4]) may be more suitable. At any rate, the performance and accuracy of TDE affects the activity detection capability, thus selection of a proper method is important. However, the proposed features will not break down completely with erroneous TDE. Instead, they will indicate that the computed time delay or DOA estimates are unreliable. The length of the processing window is also an application dependent performance factor. The window length used in the experiments (171 ms) is quite long, e.g., for speech applications that require fast updates.

The loudspeaker experiment discards some imperfections that are present in real life conditions, including source movement, changes in orientation, and radiation patterns. This makes the estimation considerably easier than in real life scenarios. This must be kept in mind when evaluating the results. A comprehensive performance evaluation would require a more challenging experimental setup, with real sources such as vehicles, human subjects etc. The assumption on planarity is required for feature computation. Although in these experiments sources were relatively close to the array, with distances 6-10 times the size of the array, this did not cause any problems. This suggests that the features could also be used in combination with localization methods assuming spherical waves. Studying the achievable benefits in such combinations is a topic left for further research. However, in situations where source distances are small compared to the array size, wavefront curvature is expected to cause degradations in detection performance.

5. SUMMARY

This work proposed new activity detection features for microphone arrays. The features were computed using spatial information only, from estimated propagation vector and time delays. This made the features free from assuming a certain type of signals, and usable in wideband applications. In addition, the features are suitable for delay-constrained real-time operation. The performance of the proposed features was tested using recorded data. Given the simplicity of the used detection method, the achieved level of performance is acceptable. Although there is room for improvement, results suggest that the proposed features can be used to detect activity and to evaluate estimate reliability.

6. REFERENCES

[1] J. C. Chen, Y. Kung, and R.E. Hudson, "Source localization and beamforming," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30–39, 2002.

[2] S. Chen, P. Sun, and B. Bridge, "Automatic traffic monitoring by intelligent sound detection," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems (ITSC)*, 1997, pp. 171–176.

[3] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle detection and tracking using acoustic and video sensors," in *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, vol. III, pp. 793–796.

[4] M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech & Language*, vol. 11, no. 2, pp. 91–126, 1997.

[5] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approcah," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[6] A. Spriet, M. Moonen, and J. Wouters, "The impact of speech detection errors on the noise reduction performance of multi-channel Wiener filtering and Generalized Sidelobe Cancellation," *Signal Processing (Article in Press)*, 2005.

[7] Y. Hioka and N. Hamada, "Voice activity detection with array signal processing in the wavelet domain," in *Proceedings of the XI European Signal Processing Conference (EUSIPCO)*, 2002, pp. 255–258.

[8] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Processing Letters*, vol. 11, no. 12, pp. 956–959, 2004.

[9] Y. Chan, R. Hattin, and J. Plant, "The least squares estimation of time delay and its use in signal detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 217–222, 1978.

[10] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, vol. IV, pp. 985–988.

[11] K. Varma, T. Ikuma, and A. A. Beex, "Robust TDE-based DOA-estimation for compact audio arrays," in *Proceedings of the Second IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2002, pp. 214–218.

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

[13] S. Haykin and J. Justice, *Array signal processing*, Academic Press, 1985.

[14] J. Yli-Hietanen, K. Kalliojärvi, and J. Astola, "Low-complexity angle of arrival estimation of wideband signals using small arrays," in *Proceedings of the 8th IEEE Signal Processing Workshop on Statistical Signal and Array Signal Processing*, 1996, pp. 109–112.

[15] T. Pirinen, P. Pertilä, and A. Visa, "Toward intelligent sensors - reliability for time delay based direction of arrival estimates," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, vol. V, pp. 197–200.

[16] T. Pirinen, "Normalized confidence factors for robust direction of arrival estimation," in *Proceedings of the 2005 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2005.

[17] F. MacWilliams and N. Sloane, "Pseudorandom sequences and arrays," *Proceedings of the IEEE*, vol. 64, no. 12, pp. 1715–1729, 1976.

[18] *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1-1.1*, 1990.

[19] International Telecommunications Union Telecommunication Standardization Sector (ITU-T), *Recommendation P.56: Objective measurement of active speech level*, 1994.

[20] B. Champagne, S. Bédard, and A. Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.