Audio Engineering Society

Convention Paper 8328

Presented at the 130th Convention
2011 May 13–16 London, UK

Detection of 'solo intervals' in multiple microphone multiple source audio applications

Elias K. Kokkinis[1], Joshua Reiss[2], and John Mourjopoulos[1]

[1] *Audio and Acoustic Technology Group, Department of Electrical and Computer Engineering, University of Patras, 26504, Patras, Greece*

[2] *Center for Digital Music, Department of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, London, UK*

Correspondence should be addressed to Elias K. Kokkinis (ekokkinis@upatras.gr)

ABSTRACT

In this work a simple and effective method is proposed to detect time intervals where only a single source is active (solo intervals) for multiple microphone, multiple source settings commonly encountered in audio applications, such as live sound reinforcement. The proposed method is based on the short term energy ratios between all available microphone signals and a single threshold value is used to determine if and which source is solely active. The method is computationally efficient and results indicate that it is accurate and fairly robust with respect to reverberation time and amount of source interference.

1. INTRODUCTION

Many audio applications, especially live sound reinforcement, involve multiple sound sources that are simultaneously active and multiple microphones set to capture the sound of each source. The interaction between the various sources (and moreover between sources and room acoustics) results in the well-known problem of microphone leakage. In most modern audio setups, the close-microphone technique is used in order to control the amount of microphone leakage, having limited success. In general, the presence of interfering energy in the primary signal picked up by the microphone makes the subsequent processing of the microphone signal by advanced algorithms and techniques (feature extraction for adaptive audio effects [1], automatic instrument recognition [2], etc) difficult and error prone. However in music there are several time intervals where only a single instrument is active. During these periods, termed solo intervals, the clean source

signal is present at the microphone, allowing further processing.

A related problem is often present in automatic transcription of meeting recordings. In such applications, the distinction between single speaker, multiple speaker (crosstalk) and non-speech periods is critical for the successful transcription of speech to text. Most approaches proposed to address this problem consist of machine learning methods using large sets of features and trained with appropriate data sets [3, 4, 5]. Another related problem is encountered in the framework of array processing, where the number of active sources needs to be estimated, typically using information theoretic criteria [6, 7]. The solutions proposed for the problems mentioned above are largely offline and computationally expensive methods that are not readily applicable to audio applications, where real-time, simple and efficient solutions are desired. While the framework of voice activity detection (VAD) could be employed [8], most methods make use of speech specific features [9] while multichannel formulations assume only one active speech source [10, 11].

In this work, an energy based method is proposed for the detection of solo intervals, based on the identification of solo audio frames by examining the energy present at each microphone with respect to the energy of all other microphones with the help of a single measure. In Section 2 the proposed method will be described in detail while in Section 3 results for simulated and real cases are presented, indicating that the proposed method is successful in a number of different scenarios.

2. METHOD DESCRIPTION

Consider a setup with M source signals $s_m(k)$ and the respective close microphone signals $x_m(k)$ given by

$$x_m(k) = \sum_{i=1}^{M} s_i(k) * h_{mi}(k) \tag{1}$$

where $h_{mi}(k)$ is the FIR filter that models the response of the acoustic path (namely the room impulse response) between the ith source and the mth microphone including microphone properties. For the purpose of this work, the acoustic path can be reduced to a single scalar a_{mi} which represents a gain that controls the amount of leakage from the ith source to the mth microphone. Thus eq. 1 becomes

$$x_m(k) = a_{mm}s_m(k) + \sum_{\substack{i=1 \\ i\neq m}}^{M} a_{mi}s_i(k) \tag{2}$$

where $a_{mm}s_m(k)$ is the direct source and $\sum_{\substack{i=1 \\ i\neq m}}^{M} a_{mi}s_i(k)$ is the leakage present at the mth microphone.

The microphone signals are subdivided into non-overlapping, consecutive frames, of length L_b

$$\mathbf{x}_m(\kappa) = [x_m(\kappa L_b) \dots x_m(\kappa L_b + L_b - 1)]^T \tag{3}$$

where κ is the discrete frame index. The detection of solo intervals is equivalent to the detection of solo frames, that is frames during which only one source is active. In order to detect such frames, the proposed method examines the energy of each microphone signal with respect to all other signals for each frame and labels that frame accordingly as solo or non-solo.

The energy of each frame is given by the energy operator $\mathcal{E}(\cdot)$ defined as

$$\mathcal{E}\left[\mathbf{x}_m(\kappa)\right] = \frac{1}{L_b}\|\mathbf{x}_m(\kappa)\|^2 \tag{4}$$

and the energy ratio (ER) for the mth microphone during κth frame is defined as

$$ER(m,\kappa) = \frac{\mathcal{E}\left[\mathbf{x}_m(\kappa)\right]}{\sum_{\substack{i=1 \\ i\neq m}}^{M} \mathcal{E}\left[\mathbf{x}_i(\kappa)\right]} \tag{5}$$

During a solo interval, when only the mth source is active, eq. 5 simplifies to

$$ER(m,\kappa) = \frac{a_{mm}^2}{\sum_{\substack{i=1 \\ i\neq m}}^{M} a_{mi}^2} \tag{6}$$

Since in general the direct signal reaching a microphone placed in close proximity to the source is much larger than leakage (i.e. $a_{mm} \gg a_{mi}$) then during a solo of the m_s source, $ER(m_s,\kappa) \gg 1$. In fact if no leakage was present (that is $a_{mi} = 0\ \forall i$) then $ER(m_s,\kappa)$ would be infinite. However, in practice leakage is always present and hence the value of the ER will depend on the direct gain a_{mm} and

the leakage gains a_{mi}. The problem now is that despite knowning that for a solo frame $ER(m_s, \kappa) \gg 1$ there is no clear indication of how large the energy ratio will be and thus it is not easy to set a single threshold T_s applicable to all cases, such that when $ER(m_s, \kappa) > T_s$ the frame will be labelled as solo. To overcome the problem of choosing a different threshold value depending on the application, a bounding function $f(\cdot)$ is used to limit the value of the energy ratio in $[0, 1]$.

$$ER(m, \kappa) = f\left(\frac{\mathcal{E}\left[\mathbf{x}_m(\kappa)\right]}{\sum_{\substack{i=1 \\ i \neq m}}^{M} \mathcal{E}\left[\mathbf{x}_i(\kappa)\right]}\right) \tag{7}$$

The parametric version of the sigmoid function is used here as a bounding function, expressed as

$$f(x) = \frac{2}{1 + e^{-\alpha x}} - 1 \tag{8}$$

where α is a 'steepness' control parameter which can be varied to obtain different shapes of the sigmoid function (Figure 1). Since, we know that the energy

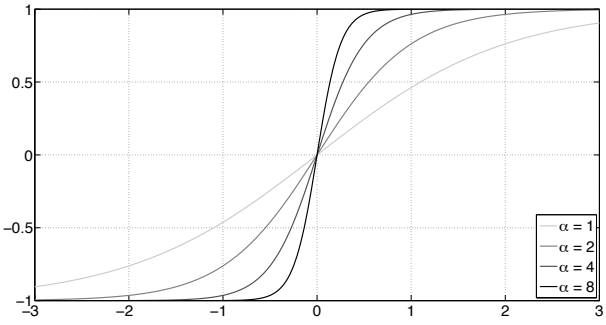


Fig. 1: The shape of the sigmoid function (eq. 8) for various values of the steepness control parameter α.

ratio will take very large values during solo frames, then the use of the bounding function enables to set a constant threshold value $T_s = 1$. For low interference settings, *ER* will take large enough values so that the bounded *ER* will take values equal to unity. However for higher interference settings the leakage gains a_{mi} result in lower *ER* values with the bounded ratio being less than unity. The steepness control parameter can address this issue effectively as shown in Figure 2, restoring the bounded *ER* to

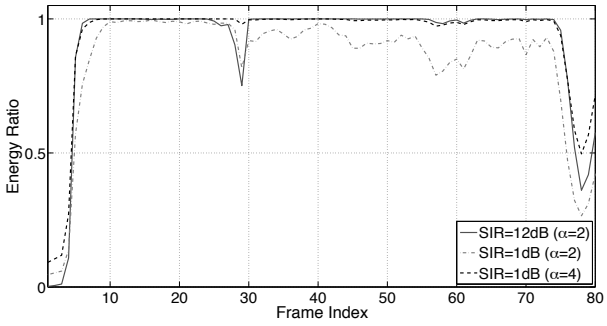


Fig. 2: The bounded energy ratio for a single microphone over several solo frames. For high Signal to Interference Ratio (SIR) the bounded ratio reaches unity while for lower SIR its values are well below unity. A more steep bounding function addresses this issue effectively.

unity and hence enabling the use of the same decision process for all cases.

Having calculated the energy ratios for all microphones during the κth frame, we look for an indication of a solo frame by calculating

$$n_s = \{m \in \Omega_M : ER(m, \kappa) = 1\} \tag{9}$$

where $\Omega_M = \{1, 2, \ldots, M\}$. Due to the highly non-stationary properties of audio signals there can be frames during which the ratio of one or more microphones is equal to one, especially if percussive instruments are present. However, a frame is labelled as solo, only if $|n_s| = 1$ and the solo microphone is of course that with an *ER* equal to unity. The process for detecting a solo frame is summarized in the flowchart of Figure 3.

One of the main points to note here is that the method does not make use of any signal specific features (such as periodicity or harmonicity) and only assumes the use of the close-microphone technique. The method involves the calculation of the energy at each microphone and the formulation of the energy ratios, merging the information about the relative microphone activity into a single measure and by the use of the bounding function, a single decision process with a fixed threshold can be applied.

3. TESTS AND RESULTS

Six datasets consisting of four channels of single instrument and vocal recordings were constructed

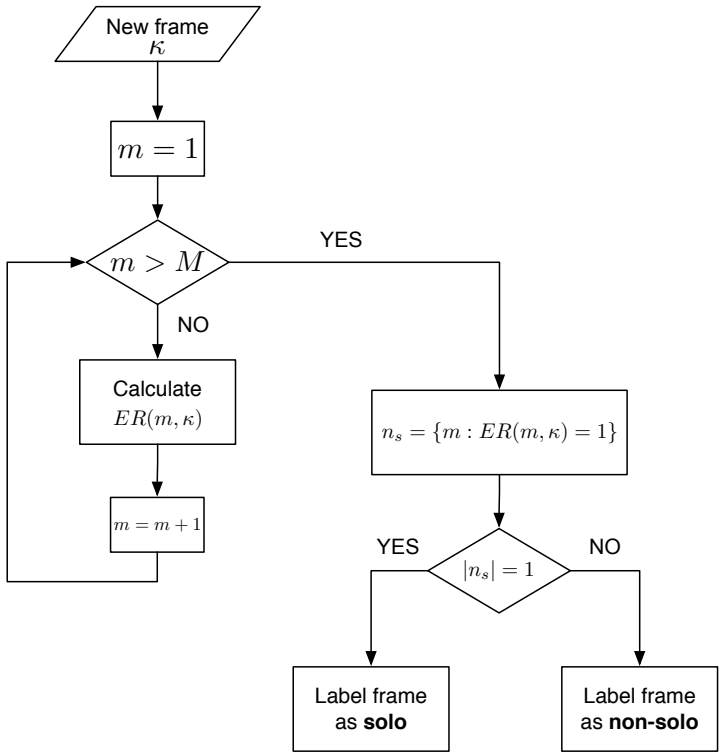


Fig. 3: The flowchart describing the process of detecting a solo frame.

and each channel was assigned a solo time interval during which all other channels were muted. Subsequently the channels were convolved with simulated room impulse responses inside a room with dimensions $12 \times 8 \times 4.5$m and variable reverberation time (Figure 4).

The performance assessment of the proposed method is based on (a) the successful detection of solo frames and (b) of non-solo frames. Hence, two performance metrics are defined:

- The *solo detection rate* (SDR) which is the ratio of the correctly labelled solo frames to the total number of solo frames expressed as percentage.
- The *solo misdetection rate* (SMR) which is the ratio of the non-solo frames incorrectly labelled as solo to the total number of non-solo frames, again expressed as percentage.

In Figure 5 the performance of the proposed method is shown for various levels of interference (indicated with Signal to Interference Ratio - SIR [12]) and values of the steepness control parameter α, averaged over all datasets. For high SIRs the proposed method achieves a fairly consistent performance with a detection rate above 80%, while when more interference is present the performance drops significantly. In such cases the steepness of the bounding function plays an important role since a more steep function can increase performance by almost 30%. Note, that the performance of the misdetection rate follows a different trend. While it remains relatively low in all cases, high values of α result in increased misdetection rate. In general, the steepness of the bounding function provides a trade-off between the accuracy of solo frame detection and non-solo frame misdetection.

Next, the effect of the analysis frame length on the performance of the method will be examined (see Figure 6). It can be easily seen that longer frames result in lower performance in terms of both metrics. This indicates that a high resolution in the time

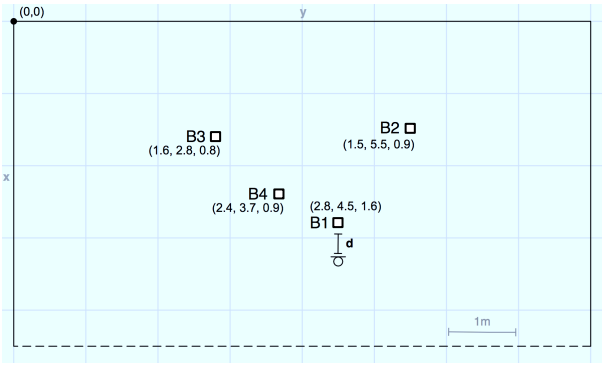


Fig. 4: The positions of the sources for the room impulse responses. In front of each source a microphone is placed at a distance of 10cm. Note that only the area around the "stage" is shown.

	RT_{60} (sec)	SDR (%)	SMR (%)
Live	1.47	84.85	4.69
Studio	≈ 0.4	75.49	2.82

Table 1: Performance of the solo detection method for real recordings with $L_b = 2048$ and $\alpha = 8$.

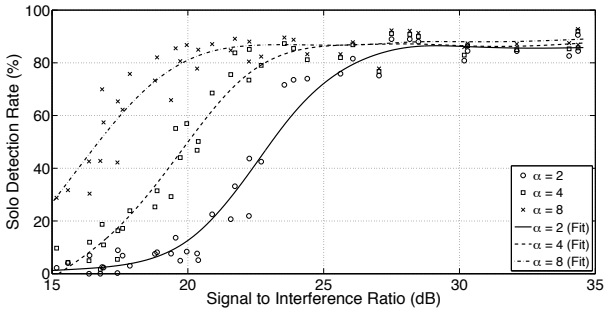
domain is desired in order to cope with the highly non-stationary energy profiles of audio signals.

As discussed in previous work [13], the interference that results in microphone leakage does not depend only on the sound level produced by the sources but also on room acoustics. Hence, the performance of the proposed method will also be examined for various reverberation times. As shown in Figure 7, the proposed method performs similarly for all reverberation times examined here, exhibiting a small decrease in performance for longer reverberation.

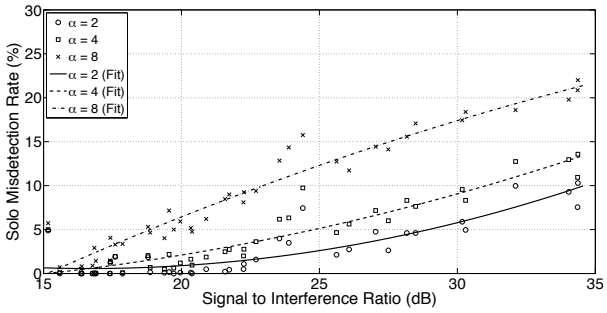
Finally the proposed method was evaluated for real scenarios including an actual live performance (four channels) and a studio recording session (eight channels), where it performed well, with high solo detection rate and rather low misdetection rates (see Table 1).

4. CONCLUSIONS

A simple and efficient method for the detection of solo intervals was presented, based on the energy ratio between all microphone signals. The method was evaluated for simulated multichannel setups in



(a)



(b)

Fig. 5: Average performance of the proposed method in a simulated room with $RT_{60} = 0.5sec$ for various values of the steepness control parameter ($L_b = 2048$). Fitted curves with (a) Smoothing spline and (b) 2nd order polynomial.

a room with varying reverberation time and SIRs ranging from 15 to 35dB. A parametric sigmoid function was used as a bounding function for the energy ratio, with a steepness control parameter providing a flexible way to control the sensitivity of the detection method.

The performance of the method was shown to be relatively consistent for SIR above 25dB in terms of solo detection rate for the cases examined here. It was also shown that short analysis frames provide a better time resolution, which is required for the accurate detection of solo frames, and that reverberation time affects performance to a certain extent.

ACKNOWLEDGEMENT

The research activities that led to these results, were co-financed by Hellenic Funds and by the European Regional Development Fund (ERDF) under the Hellenic National Strategic Reference Frame-

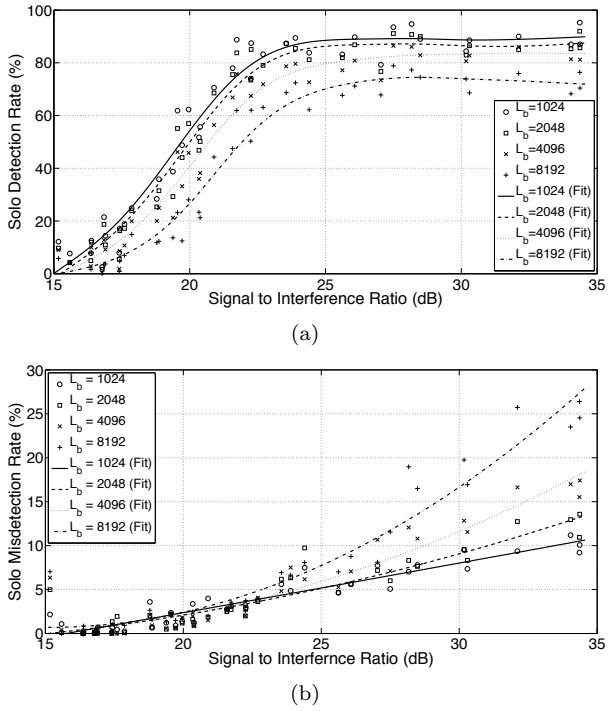


Fig. 6: Average performance of the proposed method in a simulated room with $RT_{60} = 0.5sec$ for various block sizes ($\alpha = 4$). Fitted curves with (a) Smoothing spline and (b) 2nd order polynomial.

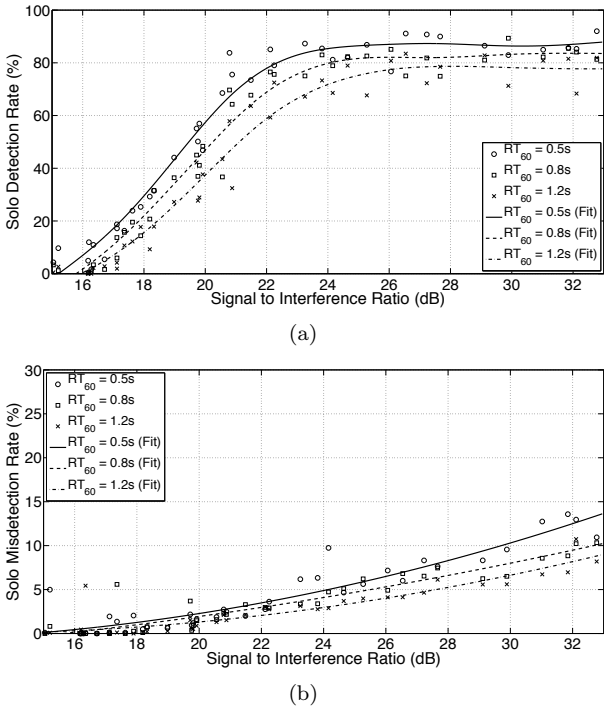


Fig. 7: Average performance of the proposed method in a simulated room with variable reverberation time ($L_b = 2048, \alpha = 4$). Fitted curves with (a) Smoothing spline and (b) 2nd order polynomial.

work (ESPA) 2007-2013, according to Contract no. MICRO2-07/E-II-A.

5. REFERENCES

[1] E. P. Gonzalez and J. D. Reiss. Automatic mixing. In U. Zoelzer, editor, *Digital Audio Effects*. John Wiley and Sons, 2011.

[2] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. 10th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2009.

[3] T. Pfau, D. P. W. Ellis, and A. Stolcke. Multispeaker speech activity detection for the icsi meeting recorder. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2001.

[4] 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.*, 13(1):84 – 91, January 2005.

[5] K. Laskowski and T. Schultz. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. (ICASSP)*, 2006.

[6] Y Q Yin and P R Krishnaiah. On some nonparametric methods for detection of the number of signals. *IEEE Trans. on Acoustics, Speech and Sig. Proc.*, 35(11):1533–1538, 1987.

[7] S Valaee and P Kabal. An information theoretic approach to source enumeration in array signal processing. *IEEE Trans. on Sig. Proc.*, 52(5):1171–1178, 2004.

[8] S. G. Tanyer and H. Özer. Voice activity detection in nonstationary noise. *IEEE Trans. on Speech and Audio Proc.*, 8(4):478–482, July 2000.

[9] E Nemer, R Goubran, and S Mahmoud. Robust voice activity detection using higher-order statistics in the LPC residual domain. *IEEE Trans. On Speech And Audio Proc.*, 9(3):217–231, 2001.

[10] J. Cho and A. Krishnamurthy. Voice activity detection using microphone array. In *AES 32nd Int. Conf.: DSP For Loudspeakers*, 2007.

[11] F. Talantzis and A. G. Constantinides. A multimicrophone voice activity detection system based on mutual information. *J. Audio Eng. Soc*, 57(11):937–950, 2009.

[12] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 14(4):1462 –1469, July 2006.

[13] E. K. Kokkinis and J. Mourjopoulos. Unmixing acoustic sources in real reverberant environments for close-microphone applications. *J. Audio Eng. Soc.*, 58(11):1–10, November 2010.