Cleaning and Quality Classification of Optically Recorded Voice Signals

Yevgeny Beiderman[1], Yaniv Azani[2], Yoni Cohen[2], Chen Nisankoren[2], Mina Teicher[1], Vicente Mico[3], Javier Garcia[3] and Zeev Zalevsky[*,2]

[1]*Department of Mathematics, Bar-Ilan University, Ramat-Gan 52900, Israel*

[2]*School of Engineering, Bar Ilan University, Ramat Gan, 52900, Israel*

[3]*Departamento de Óptica, Universitat de València, c/Dr. Moliner, 50, 46100 Burjassot, Spain*

Abstract: A newly developed optical technology for remote recording of voice signal was recently demonstrated. In this paper we present a signal processing approach for improving the quality of the recording and then for classifying the characteristics of the recording done using this system. In both cases the proposed signal processing operations are applied over the spectrogram of the optically recorded signals.

Keywords: Optical microphone, spectrogram, corner detection, classification.

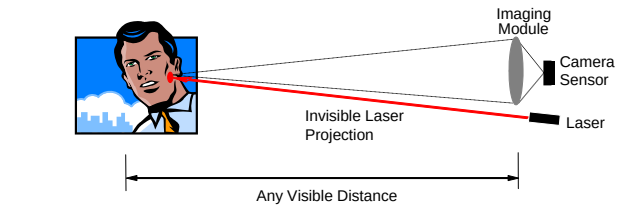
1. INTRODUCTION

The ability of dynamic extraction of remote sound signals is very appealing. Modern techniques being used for processing of optically recorded voice signals are based on either intensity, phase or polarization modulation [1]. A new technology that is based upon fast camera and a small laser light source allows the extraction and the separation of remote sound sources from distances of up to few hundreds of meters was recently developed [2,3]. The operation principle of this optical microphone system involves special type of tracking of self-interference patterns that are called speckles [4-6] and which are generated inside the illuminating laser spot when it is illuminating a vibrating object. The approach is very modular and it does not apply any constraints regarding the orientation of the speaker or the relative positions of the sound sources in respect to the detection device. The optical setup performing the detection is very simple and versatile [3].

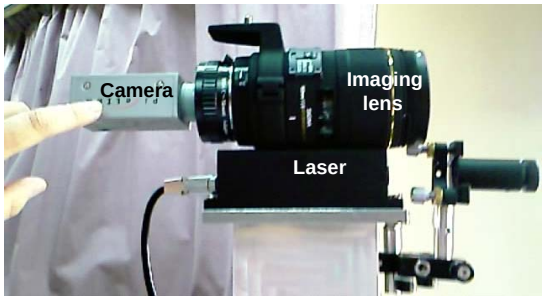
The approach was successfully demonstrated in the detection of various sound sources from several hundreds of meters while the sound sources were fully separated since every spatially separated source was imaged by the camera and therefore decoded by different pixel of the camera (each group of camera's pixels tracked the movement of the speckles and accordingly extracted the sound source that was corresponding to imaging of that group of pixels). Thus, the system is capable of being a "blind source separation" filter as well as a filter for eliminating the surrounding noise. However, electronic and optical noises are not automatically filtered out in this setup. The purpose of this work is to present a way of suppressing and eliminating these type of noises.

*Address correspondence to this author at the School of Engineering, Bar Ilan University, Ramat Gan, 52900, Israel; Tel: 972-3-5317055; Fax: 972-3-7384051; E-mail: z_zalevsky@yahoo.com

The picture of the developed system itself is shown in Fig. (**1**). The setup basically includes a fast camera and a laser illuminating the region of interest. The required



(a)



(b)

Fig. (1). (**a**) Scheme of the optical microphone system showing a fast camera with its optics as well as a laser. (**b**) An image of the system.

temporal sampling rate of the camera corresponds to Nyquist rule of sampling (twice as the relevant bandwidth of the sampled signal). The bandwidth of speech signals is approximately 4 kHz [7] and thus sampling at a rate of 8000

1877-6124/10 2010 Bentham Open

fps (frames per second) is enough for the reconstruction. Nowadays digital cameras can allow even higher sampling frame rates at predefined spatial regions of interest, e.g., 256 x 256 pixels (this region for instance can potentially separate 256 x 256 different sound sources) [8]. We saw that sampling at rate of about 2.4 kHz provides sufficiently good recording quality (corresponds to signals with bandwidth of up to 1.2 kHz).

In this paper we present two types of signal processing approaches related to the optical microphone system. The first is an algorithm used to improve the quality of recording by cleaning the recorded signal. This algorithm is based upon applying Harris' corner detector algorithm [9] over the spectrogram of the recorded signal. The second algorithm includes the construction of a set of five rules extracted from the spectrogram of the recorded signal, fusing them and obtaining an overall grade that is rating the quality of recording. By using the proposed algorithm, the operator of the system may have a real- time evaluation tool allowing him to know if the optical system is properly aligned and whether or not the recording process should be repeated.

In Section 2 we present the cleaning algorithm. In Section 3 we show our characterization process. The paper is concluded in Section 4.

2. SIGNALS' CLEANING

The spectrogram [10] of a recorded signal can be represented as follows:

$$Spectrogram(t,\omega) = |STFT(t,\omega)|^2 \tag{1}$$

where STFT stands for short time Fourier transform which is defined as:

$$STFT(t,\omega) = \int_{-\infty}^{\infty} s(\tau)W(\tau - t)\exp(-i\omega\tau)\,d\tau \tag{2}$$

s(t) denotes our temporal signal and *W(t)* is a window function. For the discrete case the definition is:

$$STFT(m,\omega) = \sum_{n=-\infty}^{\infty} s[n]W[n-m]\exp(-i\omega n) \tag{3}$$

The inverse STFT which is required to obtain the inverse spectrogram can be obtained as follows:

$$s(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} STFT(t,\omega)\exp(i\omega\tau)\,dt\,d\omega \tag{4}$$

Harris corner detection algorithm is an image processing algorithm allowing detecting corners in an image. Mathematically if we denote our image by *I(x,y),* we obtain:

$$E(u,v) = \sum_{x,y} w(x,y)\left[I(x+u,y+v) - I(x,y)\right]^2 \tag{5}$$

where *w(x,y)* is a weighting function and *E(u,v)* is the output of the processing. For small shifts of *u,v* the following approximation can be used:

$$E(u,v) \cong \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \tag{6}$$

where the matrix *M* is defined as:

$$M = \sum_{x,y} w(x,y)\begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \tag{7}$$

and I_x, I_y are the partial spatial derivatives according to *x* and *y* axes, respectively. The corner response measure is defined according to:

$$R = \det(M) - k\left[trace(M)\right]^2 \tag{8}$$

where *k* is a constant determining the sensitivity of the algorithm. From this definition one may see that

- *R* depends only on eigen-values of *M*.
- *R* is large for a corner.
- *R* is negative with large magnitude for an edge.
- |*R*| is small for spatially flat region.

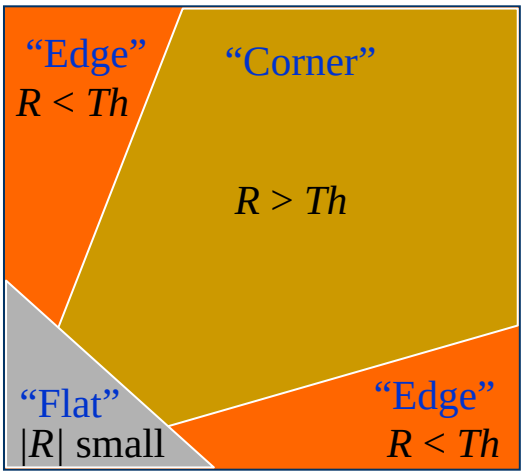


Fig. (2). Harris corner detector algorithm applied over the recorded spectrogram.

In order to detect the corners we will compare the value of *R* to a threshold (denoted as *Th*) as seen in Fig. (**2**).

Obviously the value of *Th* significantly affects the obtained results. Fig. (**3**) presents how the value of the threshold affects the final result. In the left side of Fig. (**3**) we present the recorded spectrogram. In the middle one may see the Harris corner detector algorithm applied with *Th* of 0.2. In the right side we present the Harris corner detector algorithm applied with *Th* of 1000. Based on this analysis we made various recordings using the proposed optical microphone system. We found that value of 2 for the threshold provides good results.

This threshold value was defined by an optimization process which was a tradeoff between two parameters: signal to noise ratio (SNR) of the image versus the informative content of the image. If *Th* is high then the SNR is high but the informative content is low and vise versa. In Fig. (**4**) one may see another spectrogram of a different optical recording with higher SNR than the one used in Fig. (**3**), and which also was obtained using the aforementioned optical microphone system. On the left side of the figure one may see the originally recorded spectrogram and on right side the same spectrogram but after applying the Harris corner detection algorithm.

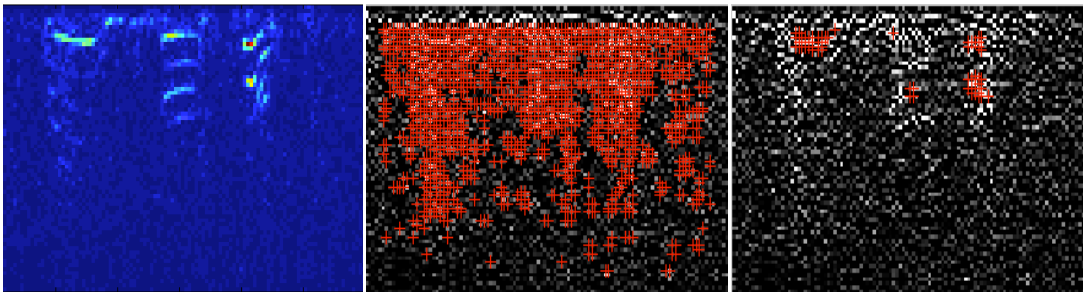


Fig. (3). Left: the recorded spectrogram. **Middle**: the Harris corner detector algorithm applied with *Th* of 0.2. **Right**: the Harris corner detector algorithm applied with *Th* of 1000.

The algorithm that we propose to clean the optically recorded signal is as follows: The corners are detected using Harris corner detection algorithm, after the recording the spectrogram is computed.

The original values of the recorded spectrogram at the positions where the corners were detected have been left unchanged, while all the rest of the regions in the spectrogram were zeroed. Then, an inverse spectrogram operation was applied in order to return to the time domain (see Eq. 4).

In Fig. (**5**) we present an example for the proposed processing. In this example one may see how the noisy signals that were recorded using the optical microphone system were cleaned using the proposed spectrogram-based algorithm. The optical recording resulted from counting: "one", "two", and "three". Sampling frequency in this example is 4666 Hz.

3. ASSESSMENT OF RECORDING QUALITY

In this section we present how we can classify the recording quality and thus to know in real-time if the optical recording was good enough or whether it should be repeated. The classification process is applied over the spectrogram of the recorded signal. There are five parameters that we used in order to perform the classification: number of spectral strips, the highest frequency in the spectrogram, average gap width between strips, average strip width and the SNR of the spectrogram of the recorded signal. The highest frequency was determined as the highest frequency strip that remained after the cleaning process. The SNR was computed as following: a cleaned image was considered as a signal and the original image as a signal plus noise. A subtraction of both produced a noise image. Energy ratio of signal to noise images yields the SNR. Four out of the specified five parameters are presented on top of one recorded spectrogram in Fig. (**6**) (SNR is not presented).

A set of 20 recordings was obtained and used as a test group. Those recordings were used in order to determine the range of values for the five parameters. All parameters are equally weighted and averaged together into a total rating grade. All the sentences were based on the same wording: “one”, “two” and “three”. However, the speakers could be changed from one recording to the next.

On the left part of Fig. (**7**) one may see the original spectrogram (up) and the spectrogram after applying the cleaning algorithm described in the previous section (down). On the right part of the figure one may see an example for the parameters rating obtained for the recording shown on the left part of the figure. From the test group we found the proper range of parameters that provided good quality of recording while the quality of recording was graded perceptually by a group of people participating in the test group (3 persons). Reference points for the grading were made by the following two conditions: good recording

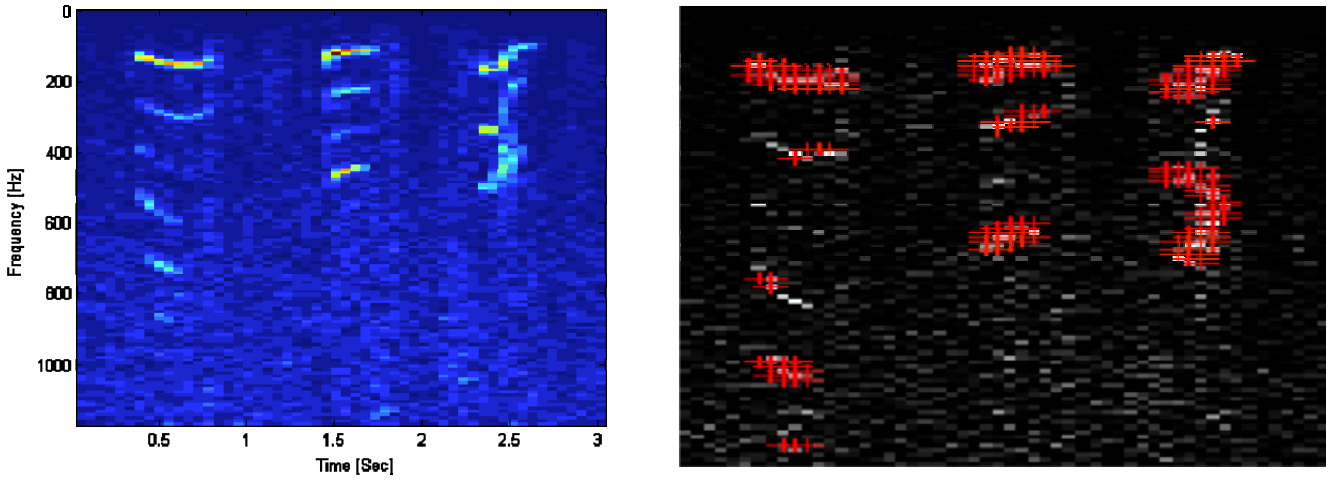
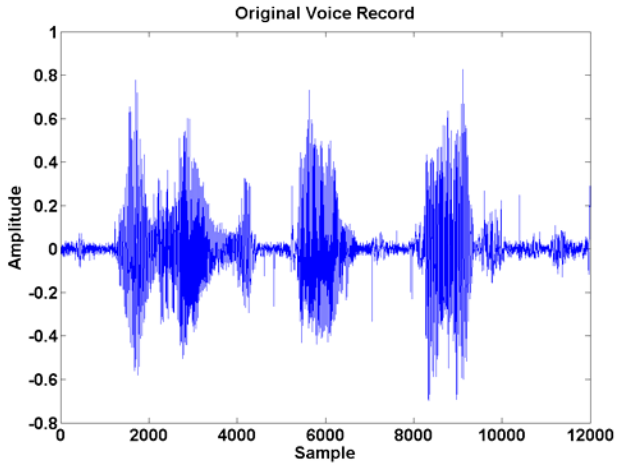


Fig. (4). The recorded spectrogram (**left**) and the spectrogram after applying the Harris corner detector algorithm (**right**).

quality as assessed by human perception and high quality cleaning results visually checked on the recordings plots. In the next step we collected another set of 20 recordings and extracted from them the same five parameters as before.

(**a**)



(**b**)

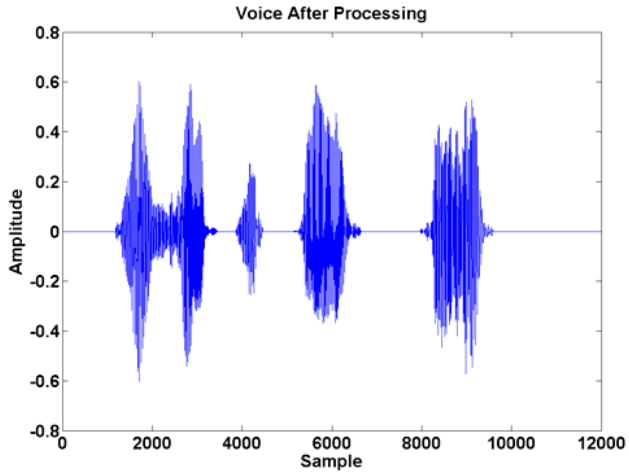


Fig. (5). (**a**). The temporal distribution of the optically noisy recorded signal, which is containing a counting of "one", "two", "three". (**b**) The temporal signal after it was cleaned using the proposed algorithm.

The distribution range of the five parameters in the experimental group was compared to the range obtained from the test group. Test group representative is shown in Fig. (**8**). The quality of recording as was tagged by the members of the experimental group was compared to the automatic grading obtained according to the processing parameters (features) learned from the test group. Grades are linearly dependant on the actual features' values. Table **1** summarizes the data values of the features extracted from the recordings of the test group as shown in Fig. (**8**). The overall sound quality grade is given in the last column of Table **1**.

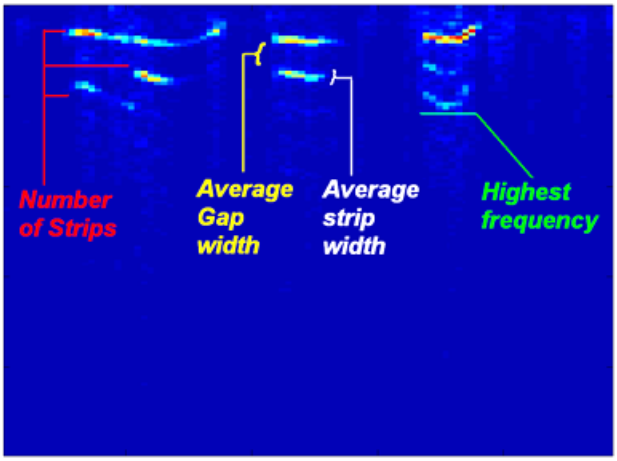


Fig. (6). The spectrogram of a recorded signal with the designation of four of the parameters that we used in order to characterize the quality of the recording.

In Fig. (**9**) one may see an example of a plot and a data table for grading that was given to 4 representative recordings. Those recordings were graded as good ones by the experimental group. One may see that good matching is obtained between the two types of grading (test and representative recordings) and thus the automatic classification based upon the parameters extracted from the test group can perform well on classifying the recording quality of the experimental recording done in real time.

In Figs. (**8**, **9**) one may also see, for comparison reasons, optimal features values range marked by dotted lines obtained from the test group. Optimal features values range was defined as a gap between minima and maxima values in the test group.

Note that not every recording succeeded to obtain good total rating. For instance, two representatives that are shown in Fig. (**10**) have two features values outside the range. The total grating is also outside the range. Data table of these recording marked as B1 and B2 is shown right below the plot. Again, optimal features values range is marked by dotted lines.

4. CONCLUSIONS

In this paper we have presented two types of signal processing algorithms. The first one was applied on the spectrogram and allowed improving the clarity of speech signals that were recorded using special optical microphone system. Then, in the second step we have experimentally developed a data fusion processing allowing classifying the quality of recording that was obtained using this system.

The proposed two algorithms were applied over signals that were experimentally extracted using the optical microphone system.

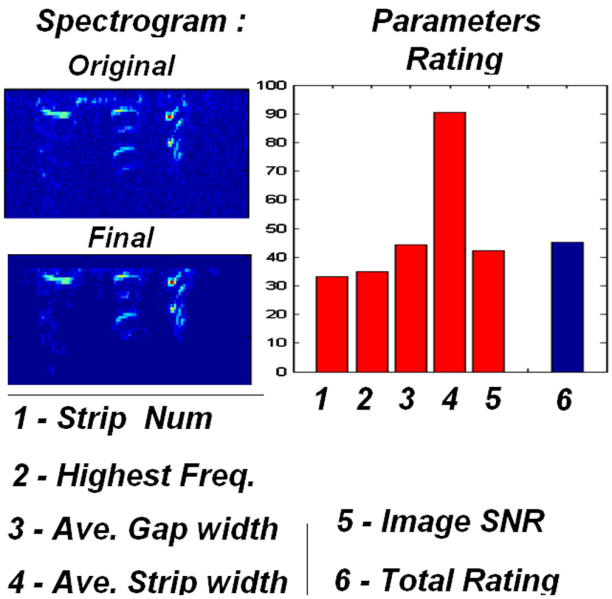


Fig. (7). On the left: The original recorded spectrogram (up) and the spectrogram after applying the presented cleaning algorithm (down). On the right: The parameters rating for the recording shown on the left. The rating is performed according to five different parameters and the overall rating is given after properly weighting all of them.

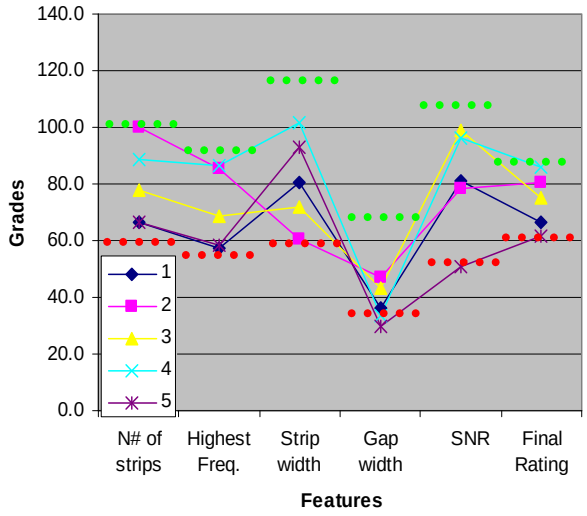


Fig. (8). The grading obtained on a set of 5 representative recordings (out of 20) in test group.

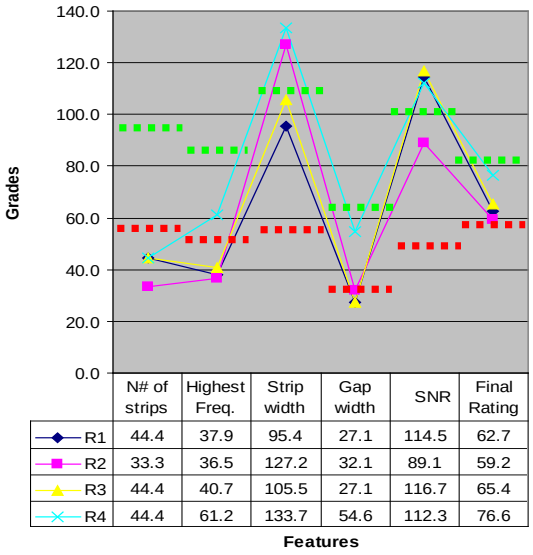
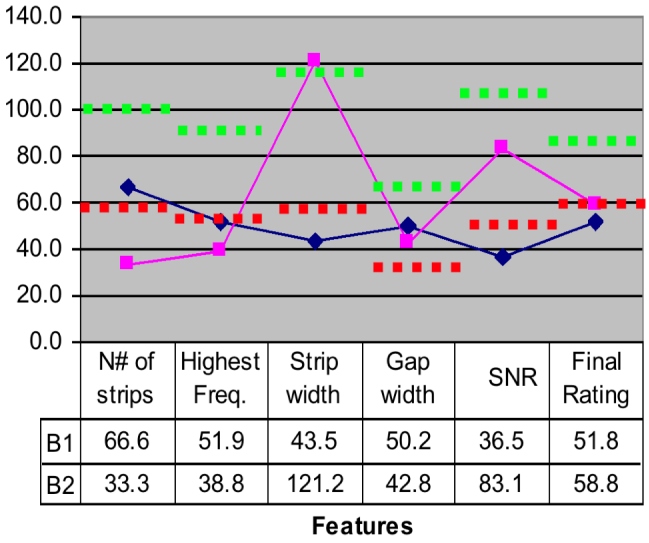


Fig. (9). The grading obtained on a set of 4 representative recordings (out of 20 recordings performed in the experiment). The grading was determined according to 20 recordings used in a test group.

Table 1. Test Group Features Values and Grades as are Observed in Fig. (8)

N#	N# of Strips		Highest Freq.		Strip Width		Gap Width		SNR		Final Value	Sound Quality
	Value	Grade	[Hz]	Grade	[Hz]	Grade	[Hz]	Grade	Value	Grade		
1	6	66.6	747.3	57.5	72.3	80.4	54.3	36.2	3.7	81.1	66.3	6
2	9	100.0	1111.8	85.5	54.3	60.3	70.3	46.9	3.5	78.5	80.8	9
3	7	77.7	893.1	68.7	64.6	71.8	64.6	43.1	4.5	98.9	75.3	8
4	8	88.8	1125.6	86.6	91.5	101.6	50.5	33.7	4.3	96.0	86.1	10
5	6	66.6	756.8	58.2	83.4	92.7	44.9	30.0	2.3	50.9	61.6	5



	N# of strips	Highest Freq.	Strip width	Gap width	SNR	Final Rating
B1	66.6	51.9	43.5	50.2	36.5	51.8
B2	33.3	38.8	121.2	42.8	83.1	58.8

Features

Fig. (10). Two representatives have two features values and the total grading outside the marked range.

Test and experimental groups containing overall numbers of 40 recordings were used to construct and right after to validate the data fusion features allowing classifying the quality of obtained recordings.

REFERENCES

[1] Bilaniuk N. Optical microphone transduction technique. Appl Acoust 1997; 50: 35-63.

[2] Zalevsky Z, Garcia J. WO/013738 (2009). International Application No PCT/IL2008/001008 July 2008.

[3] Zalevsky Z, Beiderman Y, Margalit I, *et al*. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. Opt Express 2009; 17: 21566-21580.

[4] Dainty JC. Laser speckle and related phenomena. 2nd ed. Berlin: Springer-Verlag 1989.

[5] Pedersen HM. Intensity correlation metrology: a comparative study. Opt Acta 1982; 29: 105-118.

[6] Leedertz JA. Interferometric displacement measurements on scattering surfaces utilizing speckle effects. J Phys E Sci Instrum 1970; 3: 214-218.

[7] Bansal D, Raj B, Smaragdis P. Bandwidth expansion of narrowband speech using non negative matrix factorization. *Paper TR2005-135, 9th European Conference on Speech Communication* (Eurospeech) 2005.

[8] High speed digital cameras: http://www.photron.com/

[9] Harris C, Stephens M. A combined corner and edge detector. *Proceedings of the 4th Alvey Vision Conference*. 1988; pp. 147-151.

[10] Smith JO. Spectral audio signal processing: http://www.dsprelated.com/dspbooks/sasp/

Received: August 17, 2009 Revised: September 29, 2009 Accepted: November 6, 2009