Pergamon

Neural Networks, Vol. 7, No. 3, pp. 441–447, 1994
Copyright © 1994 Elsevier Science Ltd
Printed in the USA. All rights reserved
0893-6080/94 $6.00 + .00

0893-6080(93)E0010-5

CONTRIBUTED ARTICLE

Acoustic Binaural Correspondence Used for Localization of Natural Acoustic Signals

NUR ARAD,[1] ERIC L. SCHWARTZ,[2] ZVI WOLLBERG,[1] AND YEHEZKEL YESHURUN[1]

[1]Tel-Aviv University and [2]Boston University

(*Received 19 October 1992; revised and accepted 10 November 1993*)

Abstract—*The left–right signal correspondence problem, that is considered as one of the most prominent problems by visual stereoscopic computational models, is much ignored by computational auditory stereophonic models. The correspondence problem, which is trivial if only one acoustic source is present, is highly complicated for a multiple sources environment. We present a computational model able to perform localization of natural complex acoustic signals (one or two human speakers). The model relies mainly on computing the cross-correlation functions of selected frequency channels arriving at the two ears, and performing a weighted integration on these functions. Thus, first attempts are made to establish a correspondence between acoustic features of the two channels. Preliminary results show that this model, which might be compared to "early vision" models in computational vision research, can serve as a first step in analyzing the acoustic scene.*

Keywords—Localization, Stereo audition, Auditory scene analysis.

1. INTRODUCTION

Complex auditory signals, such as speech generated by one or more individuals, can be fairly well separated and localized by human subjects with normal hearing. To date, a vast number of computational models performing localization have been proposed, but these either tackle the case of artificial stimuli (Lyon, 1982; Stern, Zeiberg, & Trahiotis, 1988; Grantham & E.Luethke, 1988), deal with a single sound source (Trahiotis & Bernstein, 1986; Stern & Colburn, 1989), or specifically deal with the technical calculation of a disparity parameter (Shamma, Shen, & Gopalaswamy, 1989) (usually correlation or some variant). Many studies have been directed towards lateralization, which, although related to localization, usually deals with stimuli of very specific acoustic properties that can be easily formulated mathematically. The model we propose is an attempt to simulate localization of two human speakers with errors in the range of human subjects.

It is widely agreed that interaural intensity differences (IIDs) and interaural temporal differences (ITDs) are the two major cues used by biological auditory systems in computing the directional location of a sound source. Low-frequency signals are localized mainly by ITDs, whereas high-frequency sounds are localized by IIDs (Coren, Porac, & Ward, 1984). It is therefore reasonable to assume that the two cues should be used in unison to achieve accurate localization. A major difficulty in constructing a computational model of natural acoustic stimuli localization lies in the fact that in a multiple source situation, each ear receives a signal that is the superposition of the original signals, but because these originate from different directions, they feed different ITD and IID into the system. Thus, setting up a correspondence between acoustic features of the left and right input channels is a necessity in any attempt to achieve localization. The setting up of a similar correspondence (between left and right visual channels) is cardinal in performing stereopsis (Marr, 1982), but is usually neglected in auditory research because the problem is trivial for a single acoustical source and synthetic stimuli.

The problem of identifying different features as belonging to the same acoustic source is related to other auditory phenomena [e.g., the cocktail party effect (Strube, 1981)], and a computationally simple model of one of these phenomena may contribute to the understanding of the other.

A number of psychoacoustic parameters have been proposed to measure auditory spatial acuity; the best known is the *minimum audible angle* (m.a.a.) devel-

Acknowledgements: This work was supported by the US–Israel Bi-National Science Foundation (BSF), Grant No. 8900078. We thank Yair Nisimov of Kibbutz Ha'ogen for the valuable knowledge and studio time needed to produce the recordings used in this research.

Requests for reprints should be sent to Nur Arad, Department of Mathematics, Tel-Aviv University, 69978 Tel-Aviv, Israel.

oped by Mills (Mills, 1958). However, m.a.a. is measured by a serial presentation of definite auditory stimuli, and we are interested in the ability to localize spatially segregated but concurrent events. Thus, a more suitable measure for our purposes is *concurrent minimum audible angle* (c.m.a.a.) (Perrott, 1994).

Cross-correlation is widely used in computing ITDs of two signals, and a rudimentary localization model that cross-correlates the left–right signals can achieve moderate results when dealing with a single artificial sound source. However, such a model yields poor results when applied to complex natural signals. In the model presented here, the cross-correlation functions are computed for different frequency channels separately (see Figure 1 for a schematic illustration of the model). This approach agrees with the finding that the various nuclei along the mammalian auditory pathway are tonotopically organized, that is, neurons along a given direction respond best when the system is stimulated by sounds of a certain frequency, and an increase in the stimulating frequency induces a monotone shift in the place of maximal activity (Knudsen, du Lac, & Esterly, 1987; Knudsen, 1982). In addition, this technique is chosen as a first step in setting up the correspondence between left–right channels, because any correspondence scheme should first try to match the corresponding frequencies of the two channels. Cross-correlation is performed to preserve the fine temporal resolution achieved during sampling of the signals. Cross-correlation functions are integrated with weights corresponding to the relative intensity of the channels over frequency ranges to obtain an omni-frequency cross-correlation function related to a specific temporal window of the input. Local maxima of this function are then translated to directions. In the single sound source case, the use of one temporal window (3–10 ms) is sufficient to accomplish localization with errors similar to that of human subjects (Perrott, 1994). With two complex sound sources located at distinct directions a single temporal window yields nonsatisfactory results, and longer time intervals should be taken. This is achieved either by integrating the functions over several consecutive temporal windows, or by displaying them serially on a two-dimensional surface, with magnitude translated to gray-level. Given such a representation, localization breaks down to identifying straight lines (parallel to the time axis) in the image. This task can be readily achieved using numerous computer vision algorithms for edge detection, edge linking, and detection of line segments.

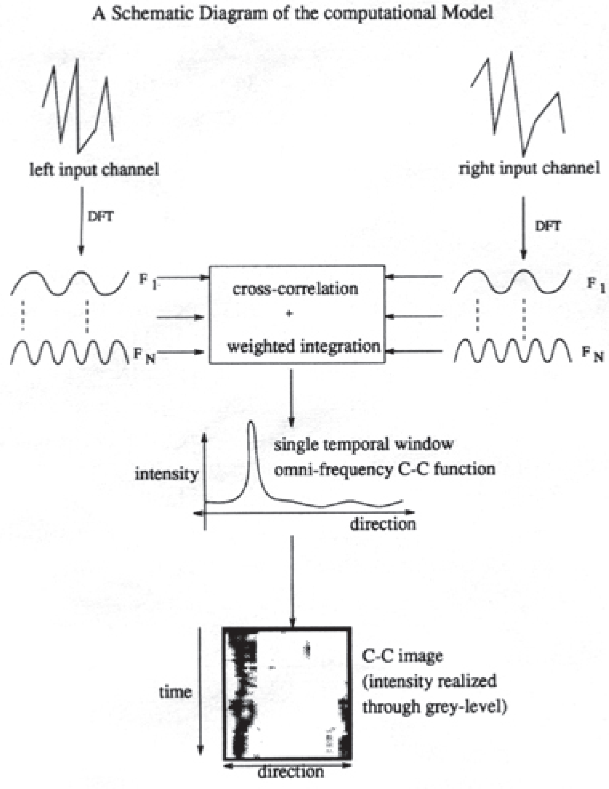

FIGURE 1. A schematic diagram of the computational model. See text for explanation.

The computational model described here can be seen as an operator that maps pairs of input signals (actually one signal interpreted by two slightly different sensors) to a one-dimensional function with domain $[-\pi/2, \pi/2]$ and whose value at direction α is the relative expectancy that a sound source is present at α. The output of this operator can be used as a front end in a system that ultimately gives a quantitative representation of the auditory scene, in quite the same manner that low-level computer vision algorithms (such as edge-detection) are used in creating a visual scene.

2. THE SETUP

Different human speech signals were recorded in a setup simulating stereophonic audition: two microphones were placed at opposite sides of a dummy head. The human speakers were situated approximately 2 m from the center of the head at different angles relative to the midline passing through the two microphones. In the single-speaker case one of four positions (0°, 30°, 60°, and 90°) was used in each recording, and in the two-speaker case the speakers used two out of five positions (−90°, −45°, 0°, 45°, and 90°) in each recording. Altogether eight different recordings were made in the single-speaker case (differing in either direction or speaker) and five different recordings were made in the two-speaker case. The signals themselves were digits spoken in Hebrew. The recordings were made in a commercial studio, and were digitally sampled at 22 kHz. No filtering was applied to the raw signals before application of the direction operator. A schematic illustration of the recording scheme is depicted in Figure 2.

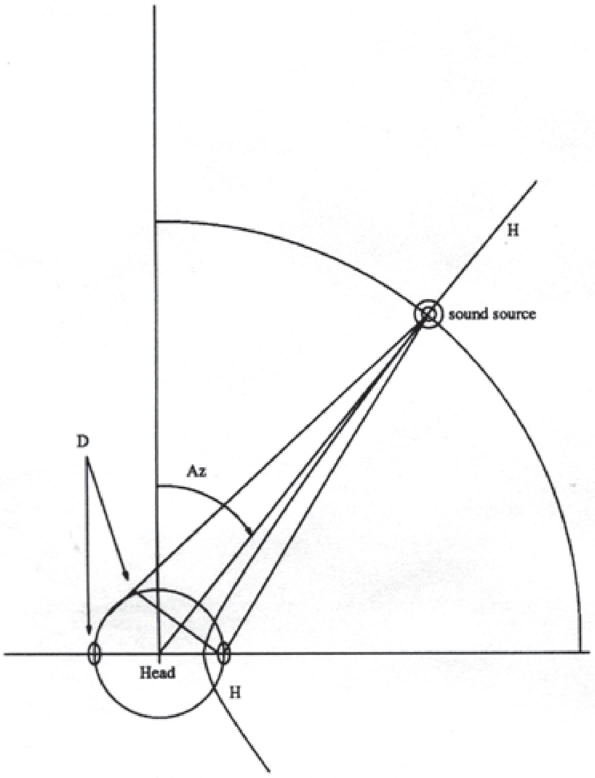

FIGURE 2. A top view of the recordings setup. Az = azimuth of sound source, D = difference in distances from sound source to ears, H = hyperbola for which D is constant.

A main step in achieving localization is the translation of the temporal shift between the input arriving at the ears to a direction. Disregarding bone conduction, the shape of the head and ears, and their acoustic characteristics, the ears can be seen as two microphones in free field placed about 20 cm apart. Because points for which difference in the distance to the ears is constant lie on hyperbolas centered at the midpoint between the ears, and these hyperbolas have linear asymptotic behavior, it follows that the direction α depends, even for moderate distances, solely on the difference in the distance from the source to the two ears. It is easily shown that in such a situation the direction α of the sound source located at $\bar{x}$ is given by

$$\alpha = \arcsin[K(|\bar{x} - \bar{e}_l| - |\bar{x} - \bar{e}_r|] \qquad (1)$$

where $\bar{e}_l$ and $\bar{e}_r$ are the positions of the left and right ears and K is an appropriate constant.

These calculations do not take into consideration factors such as the shape of the head and ears. However, experiments with analogue equipment have shown that the approximation we use is quite accurate across a wide range of frequencies (Mills, 1958), certainly within the resolution of the system we are simulating.

Equation (1) gives the theoretical foundation to the fact that the resolution in localization is much better in the frontal areas that in the lateral ones. Furthermore, assuming resolution in the frontal areas is 3° (resolution that is of the order of the resolution exhibited in ideal situations), this translates to a 50-μs shift in the arrival of the signal at the ears; thus, a sampling rate of 20 kHz is the minimum rate needed to achieve accurate localization (by biological standards).

3. THE DIRECTION OPERATOR

The input signals were DFT'd with a window of $N = 64$ to $N = 512$ samples (3.2 ms to 12.8 ms). The signals were then represented as:

$$X_i(j) = \sum_{k=1}^{N/2} a_i(k)\sin(2\pi k/N + \varphi_{i,k}),$$
$$i = 1, 2, \quad 0 \le k \le N - 1,$$

where $i = 1, 2$ represents the left and right signals, respectively. The coefficients $a_i(0)$ were discarded because they only add a constant factor to each of the signals. The use of a Hamming window did not have a major effect on the results.

The corresponding frequency channels of the two signals were logarithmically scaled, and then their cross-correlation was computed to obtain the frequency-dependent correlation functions (C-C functions) $\tau_j(t)$ for $1 \le j \le N/2 - 1$, $-T \le t \le T$, where $T = 17$ is the maximal sample shift possible (when the signal originates from an azimuth of 90°). The functions τ_j are themselves sine functions with period N/j and are thus completely determined by their phase and amplitude.

Actually not all cross-correlation functions were computed: because the input signals are of human speech, it is argued that most of the acoustic energy is contained in relatively few frequencies. Thus, cross-correlation functions were computed for frequencies j satisfying the relations:

$$|b_i(j)| > |b_i(j-1)| \quad \text{or}$$
$$|b_i(j)| > |b_i(j+1)|, \quad j = 1, 2,$$

where the coefficients $b_i(j)$ were smoothed-out versions of the $a_i(j)$'s defined by:

$$b_i(j) = 0.25a_i(j-1) + 0.5a_i(j) + 0.25a_i(j+1).$$

The criterion used here is a compromise chosen to minimize the computations involved while using enough information to enable separation by peak detection methods between two distinct speakers. In addition, the number of frequencies that satisfy this criterion was, in practice, independent of the speakers, their number, or position.

The cross-correlation functions of the frequencies used are summed, obtaining the omni-frequency cor-

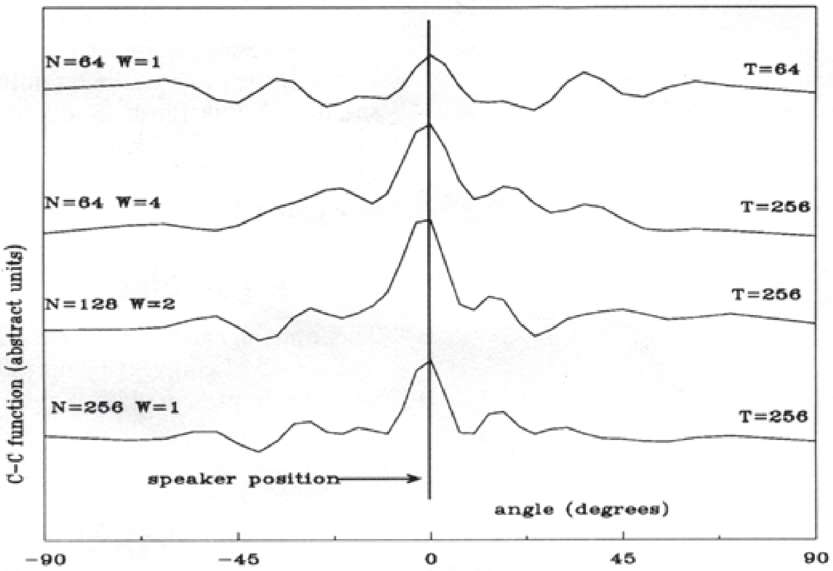

FIGURE 3. The C-C functions in a single-speaker case with varying window size and number of windows. N = window size, W = number of windows used, T = total number of samples used.

relation function associated with the signal involved and the particular temporal window used:

$$\tau_W(t) = \sum_{j\epsilon F} \tau_j(t), \quad F = \{\text{frequencies used}\},$$

$$W = \text{temporal window.}$$

To switch the temporal parameterization of $\tau_W(t)$ to directional parameterization, the transformation $\alpha = \arcsin(Kt)$ is used, and thus we represent the omni-frequency cross-correlation function as a function of the direction α. The significant peaks of this function correspond to directions where acoustic energy is present at the temporal window W.

In the sequel we compose different operations on the output of the direction operator to achieve localization in different situations. In the single-speaker case, the use of one window of size 3–12 ms was enough to obtain an omni-frequency cross-correlation function with a significant peak. In the two-speaker case, it is unreasonable to believe that such a short temporal window will suffice, because the speakers were producing natural speech signals, and, as such, their acoustic energy was not necessarily concentrated at the same temporal windows. In these cases, several consecutive windows were averaged, usually with a shift of $N/2$ (N = sample size of one window). The number of windows was of the order of 4 to 12. As a whole, the results improved with the increase in the number of windows used. An alternative approach used was creating a *directional image* by translating the values of each C-C function to gray-level and then exhibiting consecutive C-C functions as an image.

4. RESULTS

4.1. Single Speaker

As mentioned in Section 3, in the single-speaker case, the use of one window is sufficient to establish localization, with an error of no more than two samples. The window size and the number of windows used had

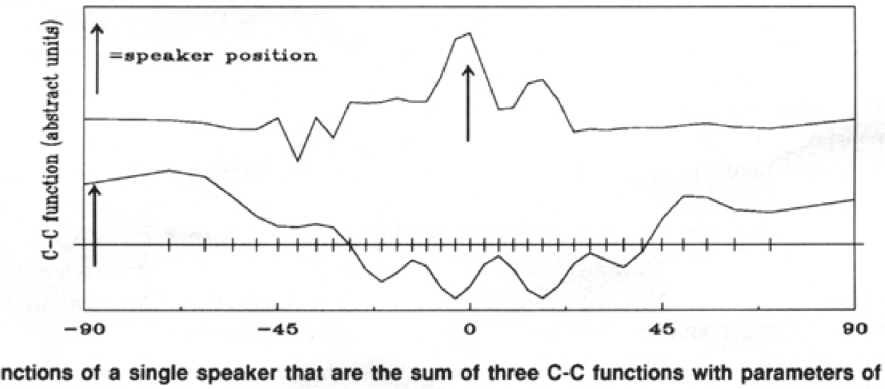

FIGURE 4. C-C functions of a single speaker that are the sum of three C-C functions with parameters of the lower functions of Figure 4. Note that in the lower function the absolute error is quite large, although it is only of one sample in temporal units.

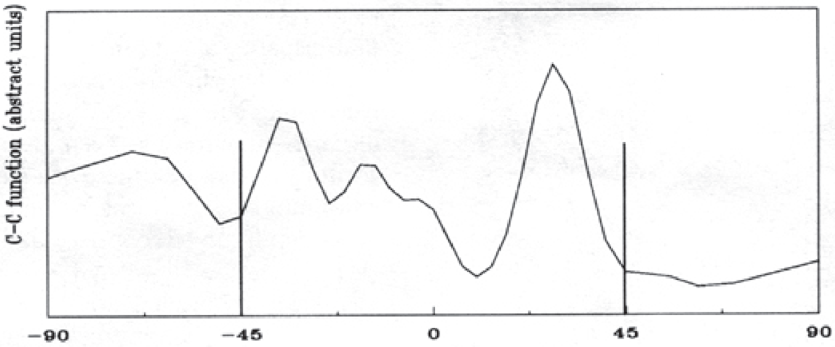

FIGURE 5. The result of the integration of 16 C-C functions in a two-speaker case. Each window is of N = 128 samples, and the shift between windows is $N/2$. Total number of samples is 1088.

a smoothing effect on the cross-correlation functions: a short window (64 samples) had the effect of C-C functions with relatively many sharp peaks. Increasing the number of windows used or enlarging the window size admits much better results. The effect of window size and number of windows is depicted in Figure 3. Integration in time plays an important role in the two-speaker case to be discussed.

Quite a sharp peak in the C-C function can be obtained when several C-C functions of different window sizes are integrated to one function. Such functions are depicted in Figure 4.

4.2. Two Speakers

The use of short temporal windows (total number of samples used is small) was found to be insufficient for accurate localization of both speakers. If the C-C functions are integrated over several consecutive windows (with overlapping), the output of the operator gives a much more accurate description of the location of the speakers. In a two-speaker typical situation we used integration over a period of the order of 1000 samples (50 ms). The output of such an operation can be seen in Figure 5.

A different representation that facilitates localization in a complex situation is one in which the distinct C-C functions are displayed serially on a video monitor with values translated to gray-level (C-C image). The dimensions of the image are direction vs. time. Figure 6 shows such images in single-speaker cases.

Figure 7 shows the resulting images in the case of two speakers. In both images, 50 windows of 256 samples were used, with a shift of 10 samples between windows. We note that the lower images of Figures 6 and 7 were all obtained from the upper ones by the same standard image processing operation: the intensity levels of the image were remapped (gamma function application) in a manner that is a continuous analogue of threshold application. The speaker positions in the examples shown clearly correspond to dark vertical lines in the images.

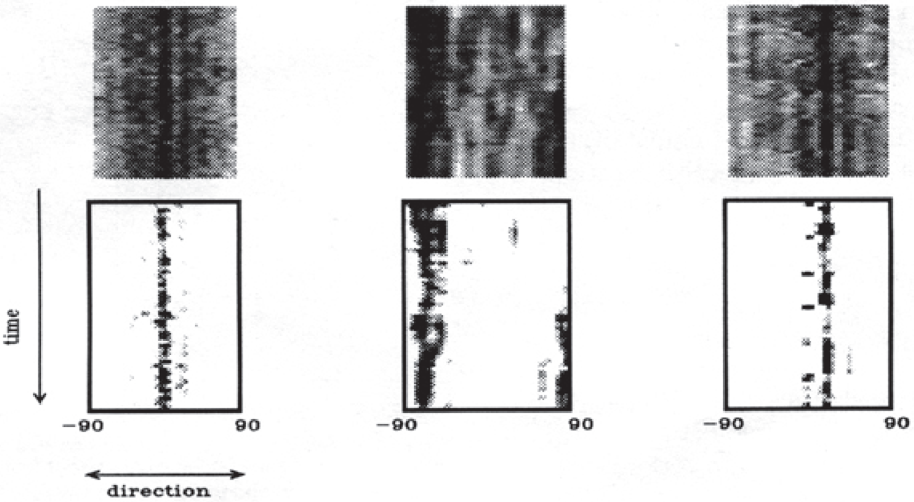

FIGURE 6. Typical C-C images in a single-speaker case. Fifty windows of 258 samples were used, with a shift of 10 samples. Total number of samples is around 750 samples. Top: original C-C images. Bottom: images after applying a threshold filter. The position of the sound source is clear from the processed images. In the middle example a Hamming window was used, with results similar to the other examples.

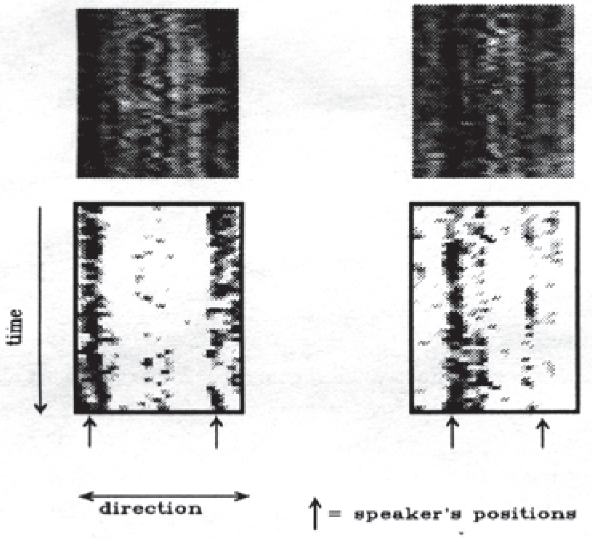

FIGURE 7. C-C images in typical two-speaker cases. Top: original C-C images. Bottom: images after applying threshold filter (identical filter in all cases). Dark areas correspond to high values of the C-C functions.

These images clearly show the advantages of using such a representation: while in a one-dimensional representation, such as shown in Figure 4, we are looking for relative maxima, and these can be masked easily by noise; in the two-dimensional representation we are looking for vertical edges, and these are much less sensitive to noise. Moreover, if ultimately a human is to perform localization with the aid of the output of the direction operator, this representation is tailored for our own visual system. Finally, if a localization algorithm is wanted at this stage, it can use a variety of computer-vision algorithms designed for the detection of straight lines in an image (edge detector filters, or the Hough Transform). A glance at Figure 4 shows that even in a single-speaker case, spurious peaks in a single C-C function may be interpreted as an additional sound source at distinct direction. Lining up the C-C functions serially and applying a line/edge detector vastly reduces the effect of such peaks.

Figure 8 shows the results of the output of the direction operator for all the single-speaker experiments conducted.

5. DISCUSSION

We have exhibited an operator that maps two acoustic signals arriving at two sensors to what we called C-C functions. In a natural scene sequences of these functions are presented together in a *direction map*—a representation of the acoustic stimuli that may be fed as input into higher-order operators to perform higher-order recognition.

In stereo audition a large effort has been aimed at the localization of a single sound source under varying conditions. The visual counterpart of auditory localization is depth perception. It has been demonstrated that disparity alone can be used in the calculation of depth (although performance improves considerably when other cues are used); thus, the evaluation of the disparity function has been an objective of many algorithms. Moreover, there is a clear-cut geometric relationship between disparity and depth. On the other hand, depth can be conceived quite well without stereo vision.

In auditory localization the situation is quite different: on the one hand, it is known that binaural localization ability is of the order of 20 times better than monaural localization; thus, binaural audition is essential to localization. On the other hand, acoustic disparity (the differences in the input to the two ears) is in many cases insufficient input for localization in a multiple source situation (Perrott, 1994); thus, monaural cues may enhance localization in complex acoustic scenes. Returning back to the vision/depth analogy, accurate localization can be enhanced by segmentation of the input stimuli into distinct acoustic events, setting up a correspondence between the events in the two channels, and finally performing localization. Such a localization procedure can run in parallel to a procedure that performs localization on a much coarser scale without segmentation. On the other hand, the output of any direction operator can be fed into higher-order systems that perform segregation and ultimately recognition of acoustic patterns. Natural candidates for such segmentation cues are onset/offset times, spectral properties of stimuli, and frequency/amplitude variation. All these cues have been found to be significant in the formation of the *auditory scene* (Bregman, 1990). Of these cues, the most important is the spectral

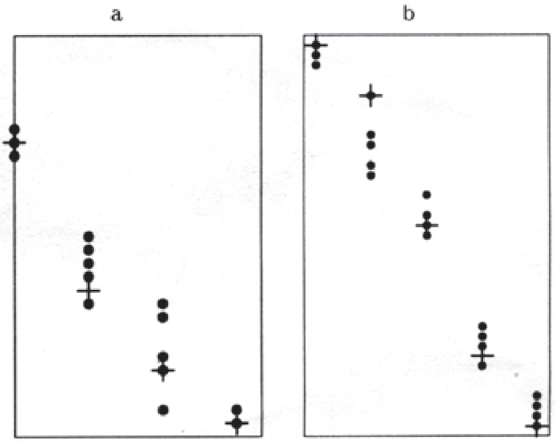

FIGURE 8. Results of applying the direction operator on acoustic scenes. In each instance the location of the sound sources was computed as the position of the brightest vertical line segments in the corresponding direction images, and are marked by dots. (a) single speaker. (b) two speakers. The crosses indicate the actual position of the speakers. In some of the two-speaker cases the operator was unable to detect the second sound source, and only one line segment was present.

properties of the stimuli, and it is this cue that we have used in our model. Thus, we see it as a localization model that uses segregated input information. Although our direction operator is of quite rudimentary nature, its performance coupled with the direction map representation scheme implicitly shows that setting up a correspondence between acoustic cues in the two input channels is a must in any natural localization task. We believe that the use of more segregation cues will improve the performance of such models in the sense that a better correspondence will be achieved. Alternately, we expect that the use of cochlear models such as those of Lyon (1983), Shamma et al. (1986), or Deng (1991) will improve the performance of the operator.

A natural question that arises is, biological considerations notwithstanding, what are the merits of using the C-C image approach we have described over simply cross-correlating the raw data of the input channels? We have found that selecting specific channels to be cross-correlated at each temporal window yields better results in a multisource scene. Moreover, the algorithm chosen for the selection of channels to be correlated is independent of the operator as a whole, and we intend to use additional biologically plausible cues for the selection of channels in the future.

In this paper we have applied the operator to the cases of single speaker and two human speakers. Demonstrating its performance on these cases, it is possible to apply the operator to more complex acoustic scenes, such as natural speech masked by noise or scenes with more than two sources. We note that creating a direction image in a scene with moving targets should result in diagonal trajectories, and Doppler shift notwithstanding, these should be detected by our operator.

REFERENCES

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge: MIT Press.

Coren, S., Porac, C., & Ward, L. M. (1984). *Sensation and perception.* San Diego: Harcourt Brace Jovanovich.

Deng, L. (1991). Processing of acoustic signals in a cochlear model incorporating laterally coupled suppressive elements. *Neural Networks,* **5,** 19–34.

Grantham, D. W., & E. Luethke, L. (1988). Detectability of tonal signals with changing interaural phase differences in noise. *Journal of the Acoustic Society of America,* **83,** 514–523.

Knudsen, E. I. (1982). Auditory and visual maps of space in the optic tectum of the owl. *The Journal of Neuroscience,* **2**(9), 1177–1194.

Knudsen, E. I., du Lac, S., & Esterly, S. D. (1987). Computational maps in the brain. *Annual Review of Neuroscience,* **10,** 41–65.

Lyon, R. F. (1982). A computational model of filtering, detection and compression in the cochlea. *International conference on acoustics, speech and signal processing,* Paris.

Lyon, R. F. (1983). A computational model of binaural localization and separation. *International conference on acoustics, speech and signal processing,* Boston.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Mills, A. W. (1958). On the minimum audible angle. *Journal of the Acoustic Society of America,* **32,** 127–146.

Perrott, D. R. (1994). Concurrent minimum audible angle: A reexamination of the concept of auditory spatial acuity. *Journal of the Acoustic Society of America,* **75**(4), 1201–1226.

Shamma, S. A., Chadwick, R., Wilbur, J., Moorish, K., & Rinzel, J. (1986). A biophysical model of cochlear processing: Intensity dependence of pure tone responses. *Journal of the Acoustic Society of America,* **80,** 133–145.

Shamma, S. A., Shen, N., & Gopalaswamy, P. (1989). Stereausis: Binaural processing without neural delays. *Journal of the Acoustic Society of America,* **88,** 2159–2170.

Stern, R. M., & Colburn, S. H. (1989). Theory of binaural interaction on auditory-nerve data. IV. A model for subjective lateral position. *Journal of the Acoustic Society of America,* **64**(1), 127–140.

Stern, R. M., Zeiberg, A. S., & Trahiotis, C. (1988). Lateralization of complex binaural stimuli: A weighted-image model. *Journal of the Acoustic Society of America,* **84**(1), 156–165.

Strube, H. W. (1981). Separation of several speakers recorded by two microphones (cocktail party processing). *Proceedings of the European Signal Processing Conference,* Lausanne.

Trahiotis, C., & Bernstein, L. R. (1986). Lateralization of sinusoidally amplitude-modulated tones: Effects of spectral locus and temporal variation. *Journal of the Acoustic Society of America,* **78,** 514–523.