4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A Database for Microphone Array Experimentation*

Ea-Ee Jan[1], Piergiorgio Svaizer[2†] and James L. Flanagan[1]

[1] CAIP Center, Rutgers University, Piscataway, New Jersey 08855

[2] IRST - Istituto per la Ricerca Scientifica e Tecnologica, 38050 Povo di Trento, Italy

ABSTRACT

An extensive speech database was collected in two experimental enclosures for research on microphone arrays. The first enclosure is a digitally controlled variable acoustics facility where the reverberation time can be adjusted from 1.7 s to 0.1 s, and the second is a regular hard-walled laboratory. Two different sets of data were collected to address source location and spatial volume selectivity issues. Preliminary results on Matched-filter processing and on Source location are reported for the new database.

I. Introduction

Performance of microphone sound pick up is degraded by deleterious properties of the acoustic environment, such as multipath distortion (reverberation) and ambient noise. The degradation becomes more prominent in a teleconferencing environment in which the microphone is positioned far away from the speaker. Besides, the ideal teleconference should feel as easy and natural as face-to-face communication with another person. This suggests hands-free sound capture with no tether or encumbrance by hand-held or body-worn sound equipment. Microphone arrays represent an appropriate approach for this application .

Microphone arrays have been employed to mitigate room reverberation and ambient noise by taking advantage of their spatial selectivity in picking-up distant acoustic sources. Major issues addressed in microphone array research include: Delay-and-sum beamformers, Adaptive beamformers, Matched-filter and multiple beamforming, and Source location. Simulation results have shown the possibility of remarkable improvement in sound capture quality compared with a single microphone. However, due to the lack of reliable multi-channel data acquisition systems and a controllable acoustical environment, limited results of microphone array research have been reported using real room data. This paper describes a database collected for microphone array research in two experimental rooms. The first one is a controllable acoustics enclosure, the second one is a regular hard-walled laboratory. Preliminary results on Matched-filter processing and on Source location are reported for the new database.

II. Data Collection

Two sections are included in the database. The first section was collected in a digitally controllable variable acoustics facility, i.e. the Varechoic Chamber at AT&T Bell Laboratories in Murray Hill, NJ [7]. The second section was collected in a regular hard-walled laboratory at CAIP Center, Rutgers University. This room is more representative of a real room environment.

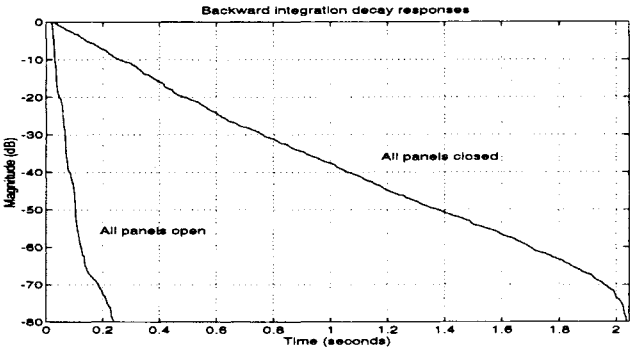

Figure 1: Backward-integrated impulse responses for the Varechoic Chamber, with polynomial fits, in fully-opened and fully-closed states. (after Ward et al.)

A. Description of the Experimental Rooms

The Varechoic Chamber is a digitally controlled variable acoustics facility. The walls of the room are constructed by two layered sheets of perforated stainless steel. The chamber is of dimension 6.7m x 6.1 m x 2.9 m and consist of 368 independent shutters with area of about $0.38m^2$ each. The holes on each shutter can be controlled either open or closed. Ten cm depth of fiberglass is filled at the back of the panel to absorb the acoustic signals when the holes are opened. Different acoustic conditions are achieved by opening and closing different sets of shutters. A total of 2^{368} conditions can be provided. The reverberation times, corresponding to 60 dB attenuation of the impulse response, can be changed between 0.1 and 1.7 seconds by controlling all the shutters from fully open to fully closed. Figure 1 shows the reverberation decay computed from the backward-integrated impulse response with polynomial fits for the Varechoic Chamber when the panels are all closed or opened. A climate control system exists in the chamber to control room temperature and humidity. The room was sensibly empty during the data collection. The measured ambient noise was 21 dB spl on the A scale and 49 dB spl on the C scale.

The second room is a hard-walled laboratory of dimension of 6x6x2.7 meters. The reverberation time in this case was about 0.5 sec. Several computers generating fan noise were inside the room during data acquisition. Several computer desks were also inside the enclosure. The measured ambient noise was 45 dB spl on the A scale and 60 dB spl on the C scale.

*Special thanks to AT&T Bell Labs, and research scientists Dr. G. Elko, Dr. J. West and Dr. D. Berkley for arranging our use of the Varechoic Chamber at AT&T Bell Labs.

† Visited CAIP in Jul.- Dec., 1994, currently at IRST, Trento - Italy

B. Experimental Setup and Data Collection Procedures

Two loudspeakers were employed as acoustic sources: the first one produced the target speech signal, while the second one acted as a competing source and was adjusted to produce different conditions of interference. Four equally powered excitation signals were used during the data collection. They included a Maximum Length (ML) pseudo random sequence, used to calculate the room impulse responses [4], a Gaussian noise representing a competing noise and two similar length utterances, produced by a male and a female, representing the target speech and interfering signals. Omni-directional electret microphones (Panasonic MW 54) were employed in data acquisition at CAIP's laboratory. High quality B&K condenser microphones were employed at the Varechoic chamber. Measurements were made at 8 kHz and 16 kHZ sampling rates, corresponding to 4 and 8 kHz bandwidths, using an Ariel ProPort for the D/A and A/D conversions. A DSP driver was implemented to acquire either one or two A/D channels while synchronously sending out either one or two D/A channels.

Four speech files, associated with the following excitation signals, were recorded at each microphone position.

1. A ML pseudo random sequence at the focal point.
2. A male utterance at the same focal point.
3. A male utterance at the focal point and an interfering 0 dB Gaussian white noise at the off-focus position.
4. A male utterance and a 0 dB interfering female utterance at the off-focus point.

C. Varechoic Chamber Data Collection

Two sets of data were collected at AT&T Bell Labs. Each set of data was recorded in three different acoustic conditions: highly reverberant, moderately reverberant, and slightly reverberant. The corresponding reverberation times were approximately 1.7, 0.6, and 0.1 seconds, respectively. The first data set addresses sound capture experiments. Two orthogonal line arrays with 5 octave harmonic spacing were configured to cover a frequency bandwidth from 250 Hz to 8000 Hz. The microphone spacing, from lowest frequency octave to highest frequency octave, was 32, 16, 8, 4, and 2 cm. Thirteen microphones were included in each octave. The total number of microphones in one line array was 37 with some sensors being shared across octaves. The center of these line arrays were (1,297.5,151.1)cm and (389.9,1,151.1) cm, respectively.

To analyze the spatial volume selectivity of a given array, a focus position was selected in the room and the target signals were generated at the "in focus" position and at various distances from the focus. Competing noise or interfering utterance was also generated at different positions and interposed between the microphones and the desired source. The noise was

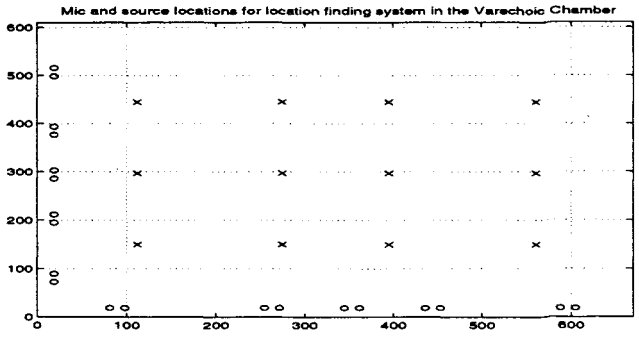

Figure 2: Configuration of source location finding in the Varechoic Chamber.

located at (226,536.2,151.1) cm, and the major focal point was at (376.6,297.5,151) cm. The focal points were moved by ±15 cm in the X axis, and ±15cm, ±30 cm, -60 cm and -100 cm in the Y axis. Thus, a total of nine different target locations were included. These data were collected when the shutters of the Varechoic Chamber were all closed.

For comparison, the data were also collected once by placing the source at the major target position with the shutters fully open. Additional data were collected with the noise source interposed between microphones and the target point. This condition is particularly critical for a delay-and-sum beamformer. The shutters were set at fully closed and 50% closed.

The second set of data addresses the location finding problem. Two orthogonal linear arrays, each consisting of 10 microphones arranged into 5 pairs, were employed in the Varechoic chamber (intra-pair distance was set to 17 cm). Different configurations of equally or unequally spaced microphone pairs can be explored by choosing different sets of microphones. Stationary acoustic sources were placed at 12 positions producing target male utterances and interfering signals, either white noise or female utterance. The pseudo random sequence was used as the impulse excitation. Two loudspeakers, facing toward the array in the X axis, were positioned in pairs, e.g., positions 1 and 7, 2 and 8, ..., and 6 and 12 were pairs. One of the positions was used as the target position, the other was the interfering source location. A total of 12 target positions were included under three different reverberation conditions. The shutters were 100%, 50% and 0% closed.

The data were acquired on two channels simultaneously from each pair of microphones. Figure 2 demonstrates the configuration of the microphone locations and the loudspeaker locations. The "o" symbols denote the microphone positions and the "x" symbols denote the loudspeaker positions.

D. Hard-walled Room Data Collection

One section of data was collected at a CAIP Center laboratory where the microphones were configured as four line arrays. The lines were placed one above the other with a separation of 15 cm. A total of 23 microphones, with a spacing of 15 cm, were included

in each line array. The microphones in lines 1 and 2 were collected in pairs, similarly for lines 3 and 4. Thus, the n^{th} microphone in lines 1 and 2 were collected simultaneously. The target source for this data was similar to that collected at Bell Labs, except 16 target positions were included. The target location was set at (281.6, 304.8) cm and the interfering source was at (66, 246)cm. The center of the line array was at (281.6, 26, 110) cm. The target source was moved ±10 cm, ±20 cm, 30 cm, 45 cm, 60 cm, 80 cm, and 100 cm in the X axis and ±15 cm, -30 cm, -50 cm, -80 cm, and -110 cm in the Y axis. An additional set of data was collected with the interfering source interposed between microphones and target location.

E. Summary of the Database

Considering all microphones, all acoustic conditions and all source and noise placements, a total of 8696 signal files were collected concerning the sound capture issue. Additionally 2880 signal files were recorded concerning the source location application, yielding a total of more than 11,500 signal files (3.5 GB of memory space).

III. Data Analysis

The database has been employed in preliminary experiments by applying the Matched-Filter Array (MFA) algorithm [1, 2, 3] and a Source location algorithm based on time delay estimation using the Cross-power Spectrum Phase (CSP) technique [5].

In the MFA, the output of each microphone is processed by a Matched-filter which is the time reverse of the impulse response from the focal point to that microphone. The array output is the summation of outputs from each Matched-filter. Because the time-reverse impulse response is typically non-causal, truncation and fixed delay are required to realize a causal filter which approximates the desired response. For a source located at the focal point emitting a signal $s(t)$, the output of the MFA is

$$O(t) = \sum_{n=1}^{N} s(t) * h_{nf}(t) * h_{nf}(-t) = s(t) * \sum_{n=1}^{N} h_{nf}(t) * h_{nf}(-t) \quad (1)$$

where $h_{nf}(t)$ is the impulse response from the focal point to the nth sensor, N is the total number of sensors, and $*$ denotes convolution. Notice the latter term of Eq. (1) is the autocorrelation of $h_{nf}(t)$.

When the source is off the focal position, the temporal output of the array is

$$O(t) = s(t) * \sum_{n=1}^{N} h_{ns}(t) * h_{nf}(-t) \quad (2)$$

where $h_{ns}(t)$ is the impulse response from the source to the nth sensor. Again the latter term of Eq. (2) is the cross correlation of impulse responses from focus and from source. In the experiments, the impulse responses from the desired focal point to each receiver in the array were measured using pseudo random sequences [3, 4]. The quality of the reconstructed signal obtained as output of the MFA was perceptually estimated as being improved more than 10 dB when compared to a single microphone using data of 46 microphones collected at CAIP Center.

Behavior of the MFA using data collected in the Varechoic Chamber was also evaluated. A thirty-seven channel nested array is available in the database. However, to obtain better spatial resolution, an non-uniform spacing array with 20 microphones is constructed. The spacing of the 20 microphones is either as large as possible, or as few common dividers as possible in the consecutive microphone pairs. The result shows dramatic improvement of the speech quality over the MFA using all of the 37 microphones. Therefore, a better configuration of the microphone array can improve the performance of the MFA. In this case the real room data is consistent with the simulation results [1, 2, 3]. It also suggests that the MFA can be widely used to provide hands-free, high quality signals in multimedia and teleconferencing environments, without body-worn or hand-held equipment.

Source location experiments are performed using the Cross-power Spectrum Phase (CSP) technique in order to analyze the robustness of this technique in the presence of different amounts of reverberation. This approach is based on the use of phase difference information only to perform an interchannel Time Delay Estimation.

In a multipath environment, the output of a transducer, $x(t)$, associated with the source signal, $s(t)$, can be expressed as:

$$x(t) = \sum_{i=0}^{k} \alpha_i s(t - \tau_i) + n(t) \quad (3)$$

where k is the number of reflections considered, α_i is the amplitude of the i^{th} reflection, τ_i is the corresponding delay, and $n(t)$ is the noise. The direct path is associated with $i = 0$.

In a real system, Equation (3) cannot be employed without *a priori* information of $s(t)$. Instead, if the direct wavefront is predominant over reflections and two transducers are close enough, e.g., 10-20 cm, the output of those transducers, $x_1(t)$ and $x_2(t)$, will be similar except for a delay. This can then be expressed as:

$$x_2(t) = x_1(t - \tau) + n(t) \quad (4)$$

Note that this is also a major assumption for the delay-and-sum beamformer. Therefore, by ignoring the effects of noise (Eq.(4)) the delay can be estimated by the cross-power spectrum [5]:

$$f(t) = \mathcal{F}^{-1} \frac{X_1(w)\, X_2(w)^*}{|X_1(w)|\,|X_2(w)|} \quad (5)$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform, and $*$ denotes complex conjugate. The delay τ is the time

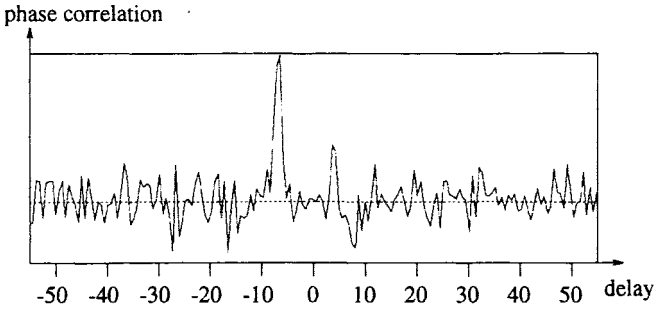

Figure 3: Example of phase correlation between two microphones. The peak of this function indicates the inter-channel delay.

index associated with peak value of $f(t)$. This delay estimator is computationally convenient and more robust to noise and reverberation than other approaches based on cross-correlation or adaptive filtering.

In ideal conditions, the output of Equation (5) is a delta function centered on the correct delay. In real applications with a wide band signal, e.g., a speech signal, the outcome is not a perfect delta function. Rather it resembles a correlation function of a random process. The time index associated with the maximum value of the output of Equation (5) provides an estimation of the delay. The system can produce wrong answers when two or more peaks of similar amplitude are present, *i.e.*, in highly reverberant conditions. The resolution in delay estimation is limited in discrete systems by the sampling frequency. In order to increase the accuracy, oversampling can be applied in the neighborhood of the peak, to achieve sub-sample precision. Fig. 3 demonstrates an example of the result of a cross-power spectrum time delay estimator.

Once the relative delays associated with all considered microphone pairs are known, the source position (x_s, y_s) is estimated as the point that would produce the most similar delay values to the observed ones. This optimization is performed by a downhill simplex algorithm [6] applied to minimize the Euclidean distance between M observed delays $\hat{\tau}_i$ and the corresponding M theoretical delays τ_i:

$$(x_s, y_s) = \arg\min_{(x,y)} \sum_{i=1}^{M} (\hat{\tau}_i - \tau_i(x,y))^2 \qquad (6)$$

An analysis of the impulse responses associated with all the microphones, given an acoustic source emitting at a specific position, has shown that constructive interference phenomena occur in the presence of significant reverberation. In some cases, the direct wavefront happens to be weaker than a coincidence of reflections, inducing a wrong estimation of the arrival direction and leading to an incorrect result.

Selecting only microphone pairs that show the highest peaks of phase correlation generally alleviates this problem. Location results obtained with this strategy show comparable performance (mean position error of about 0.3 m) at reverberation times of 0.1 s and 0.6 s. When the reverberation time is set to 1.7 s, 3 position estimates out of 24 are completely wrong (*i.e.*, error > 1 m) and the remaining estimates still have an average error about 0.3 m. Table 1 summarizes the average error using either 10 microphone pairs or only 4 selected microphone pairs.

Reverb. Time	*Average Error*	
	10 mic pairs	4 mic pairs
0.1sec	38.4 cm	29.8 cm
0.6sec	51.3 cm	32.1 cm
1.7sec	105.0 cm	46.4 cm

Table 1: Average location error using either all 10 pairs or 4 pairs of microphones. Three reverberation time conditions are considered.

A proper selection of suitable microphone pairs seems to be the only solution to the source location problem in highly reverberant environments.

IV. Conclusion

A database for microphone array research has been collected. The database provides data from various acoustic environments. Preliminary results by applying the Matched-filter processing and source location algorithms under various acoustic conditions are presented. The research continues to optimize the configuration for the matched-filter array and to properly select reliable time delay estimators from microphone pairs.

References

[1] J. L. Flanagan, A. C. Surendran and E. E. Jan "Spatially selective sound capture for speech and audio processing" *Speech Communication 13*, 1993, pp. 207-222

[2] E. E. Jan, "Parallel processing of large scale microphone arrays for sound capture", Ph.D thesis, Electrical Engineering, Rutgers University, NJ, May 1995.

[3] E. E. Jan, P. Svaizer and J. L. Flanagan "Matched-filter processing of microphone array for spatial volume selectivity" *Proc. IEEE ISCAS*, Seattle, 1995. pp. 1460-1463

[4] F.J. MacWilliams, N.J.A. Sloane, "Pseudo-Random Sequences and Arrays," *Proc.IEEE*, vol. 68, May 1980, pp. 593-619.

[5] M. Omologo, P. Svaizer "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", *Proc. ICASSP*, Adelaide 1994, pp. II273-II276.

[6] W. Press, B. Flannery, S. Teukolsky, W. Vetterling, "Numerical Recipes in C - The Art of Scientific Computing", Cambridge University Press 1988.

[7] W. C. Ward, G. W. Elko, R. A. Kubi, and W. C. McDougald "The new Varechoic Chamber at AT&T Bell Labs", *Proceedings of the Wallace Clement Sabine Centennial Symposium*, Acoustic Society of America, pp. 343-346, 1994.