SPEECH ACQUISITION AND ENHANCEMENT IN A REVERBERANT, COCKTAIL-PARTY-LIKE ENVIRONMENT

Yiteng (Arden) Huang[†]*, Jacob Benesty*[‡]*, and Jingdong Chen*[†]

[†] Bell Laboratories, Lucent Technologies
600 Mountain Avenue
Murray Hill, New Jersey 07974, USA
{arden, jingdong}@research.bell-labs.com

[‡] Université du Québec, INRS-EMT
800 de la Gauchetière Ouest, Suite 6900
Montréal, Québec, H5A 1K6, Canada
benesty@emt.inrs.ca

ABSTRACT

Developing a successful multi-microphone speech acquisition system in a reverberant, cocktail-party-like environment is a very challenging problem since both interfering sources and reverberation need to be well controlled. In this paper, we propose an algorithm based on blind SIMO identification. We first blindly identify the channels from the interfering sources to all the microphones. Then we extract the speech signal of interest. Finally speech dereverberation is performed using the MINT method. Simulations with acoustic impulse responses measured in the varechoic chamber at Bell Labs are carried out to verify the proposed algorithm.

1. INTRODUCTION

Extracting a desired speech signal from its corrupted observations is essential for tremendous applications of speech processing and speech communication since we are living in a natural environment where noise or disturbance is perpetual and ubiquitous. Speech signal can seldom be recorded in pure form. In most cases, they are immersed in acoustic ambient noise and interference, and are distorted by reverberation.

A human has the ability of choosing to focus on a specific speaker in a room where several people are talking concurrently and where noise sources might meanwhile exist. This phenomenon is referred to as the *cocktail party effect* or *attentional selectivity* [1]. This effect is mainly attributed to the fact that we have two ears and our perception of speech is based on binaural hearing, which can be easily demonstrated by observing the difference in understanding between using both ears and with either ear covered when listening in a cocktail-party-like environment. This suggests the use of multiple microphones in the development of practical speech acquisition systems.

In this paper, we consider a cocktail-party-like environment where there are one speech source of interest, $M-1$ other sound sources, and N microphones with $M < N$. The speech source and $M-1$ other sound sources are mutually independent. The interfering sound sources can be speech or noise, and are regarded as interference. Without loss of generality, we label the speech source of interest as the first. Such an acoustic environment is mathematically modeled as an $M \times N$ MIMO FIR system. At the nth microphone and at the kth sample time, we have:

$$x_n(k) = \sum_{m=1}^{M} \mathbf{h}_{nm}^T \mathbf{s}_m(k) + b_n(k), \quad n = 1, 2, \cdots, N, \qquad (1)$$

where $(\cdot)^T$ denotes the transpose of a matrix or a vector,

$$\mathbf{h}_{nm} = \left[\begin{array}{cccc} h_{nm,0} & h_{nm,1} & \cdots & h_{nm,L_h-1} \end{array} \right]^T$$

is the impulse response (of length L_h, $\forall m, n$) between source m and microphone n,

$$\mathbf{s}_m(k) = \left[\begin{array}{cccc} s_m(k) & s_m(k-1) & \cdots & s_m(k-L_h+1) \end{array} \right]^T$$

contains the last L_h samples of the mth source signal s_m, and $b_n(k)$ is zero-mean additive white Gaussian noise (AWGN). Since dominant noise sources have been modeled as inputs, additive noise in this environment is deemed significantly weak. We have no *a priori* knowledge about the speech source $s_1(k)$, its interfering signals $s_m(k)$ $(m = 2, \cdots, M)$, or the channel impulse responses h_{nm}. But we intend to blindly estimate $s_1(k)$ using only the second-order statistics of $x_n(k)$ $(n = 1, 2, \cdots, N)$.

The z-transform of the MIMO signal model (1) is expressed (in the vector/matrix form) as

$$\vec{\mathbf{X}}(z) = \mathbf{H}(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{B}}(z), \qquad (2)$$

where

$$\begin{aligned} \vec{\mathbf{X}}(z) &= \left[\begin{array}{cccc} X_1(z) & X_2(z) & \cdots & X_N(z) \end{array} \right]^T, \\ \mathbf{H}(z) &= \left[\begin{array}{cccc} H_{11}(z) & H_{12}(z) & \cdots & H_{1M}(z) \\ H_{21}(z) & H_{22}(z) & \cdots & H_{2M}(z) \\ \vdots & \vdots & \vdots & \vdots \\ H_{N1}(z) & H_{N2}(z) & \cdots & H_{NM}(z) \end{array} \right], \\ \vec{\mathbf{S}}(z) &= \left[\begin{array}{cccc} S_1(z) & S_2(z) & \cdots & S_M(z) \end{array} \right]^T, \\ \vec{\mathbf{B}}(z) &= \left[\begin{array}{cccc} B_1(z) & B_2(z) & \cdots & B_N(z) \end{array} \right]^T. \end{aligned}$$

In addition, $X_n(z)$, $H_{nm}(z)$, $S_m(z)$, and $B_n(z)$ $(m = 1, 2, \cdots, M,\ n = 1, 2, \cdots, N)$ are the z-transforms of $x_n(k)$, h_{nm}, $s_m(k)$, and $b_n(k)$, respectively.

2. ASSUMPTIONS AND BLIND SIMO IDENTIFICATION

In a cocktail-party-like environment, sound sources are presumably motionless or move very slowly. Consequently their corresponding channel impulse responses change very slowly in time. It is further assumed that from time to time each interfering source occupies at least one exclusive interval alone. Then during every single-talk interval, the SIMO systems that correspond to the interfering sources are blindly identified and their channel impulse

responses are saved for later extraction of the interested speech source when all sources voice out simultaneously. Although these assumptions make the developed algorithm less flexible, they still are reasonable and stand in many practical scenarios.

In addition, we assume that the SIMO system corresponding to each sound source is irreducible such that it can be blindly identified using only second order statistics (SOS). In other words, $H_{1m}(z), H_{2m}(z), \cdots, H_{Nm}(z)$ $(\forall m)$ do not share any common zeros. The idea of blind SIMO identification using only SOS was first proposed by Tong et al. [2]. So far there have been a number of adaptive algorithms for blind SIMO identification developed by the authors, among which the so-called frequency-domain normalized multichannel LMS (FN-MCLMS) algorithm performs well with acoustic systems [3] and will be used in this study.

3. SPEECH EXTRACTION BY CANCELING INTERFERING SOUND SOURCES

In this section, we explain how to extract the speech signal of interest from concurrent interfering sound sources. Although the fundamental principle is similar to what was used in [4], the two algorithms are significantly distinct. In [4], all acoustic sources to be separated are equally important. But here, our focus is exclusively on one speech source of interest. This makes the developed algorithm less complicated. In addition, the developed algorithm does not require that the speech source of interest would have to occupy a period of time with no other sound sources. As a result, the algorithm developed in this paper is more useful in practice.

It is supposed that so far the SIMO systems corresponding to the interfering sources have been blindly identified. Then the estimates of their channel impulse responses will be used to convert the $M \times N$ MIMO system into a SIMO system with the speech source of interest as the sole input.

From the N microphone outputs, let's choose M at a time. We have $P = C_N^M$ different ways of doing so. For the pth $(p = 1, 2, \cdots, P)$ combination, we denote the index of the M selected microphone signals as p_m $(m = 1, 2, \cdots, M)$, and get an $M \times M$ MIMO subsystem. For this subsystem, let $\mathbf{H}_p(z)$ be the $M \times M$ matrix obtained from the system's channel matrix $\mathbf{H}(z)$ by keeping its rows corresponding to the M selected microphone signals. Then similar to (2), we have

$$\vec{\mathbf{X}}_p(z) = \mathbf{H}_p(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{B}}_p(z), \tag{3}$$

where

$$\begin{aligned} \vec{\mathbf{X}}_p(z) &= \left[\, X_{p_1}(z) \;\; X_{p_2}(z) \;\; \cdots \;\; X_{p_M}(z) \,\right]^T, \\ \vec{\mathbf{B}}_p(z) &= \left[\, B_{p_1}(z) \;\; B_{p_2}(z) \;\; \cdots \;\; B_{p_M}(z) \,\right]^T. \end{aligned}$$

Consider the following equation:

$$Y_p(z) = \vec{\mathbf{W}}_p^T(z)\vec{\mathbf{X}}_p(z), \;\; p = 1, 2, \cdots, P, \tag{4}$$

where

$$\vec{\mathbf{W}}_p(z) = \left[\, W_{p,1}(z) \;\; W_{p,2}(z) \;\; \cdots \;\; W_{p,M}(z) \,\right]^T.$$

Substituting (3) into (4) produces

$$Y_p(z) = \vec{\mathbf{W}}_p^T(z)\mathbf{H}_p(z)\vec{\mathbf{S}}(z) + \vec{\mathbf{W}}_p^T(z)\vec{\mathbf{B}}_p(z). \tag{5}$$

In order to remove the interference from other competing sources, the objective here is to find the vector $\vec{\mathbf{W}}_p(z)$ whose components are linear combinations of $H_{nm}(z)$ $(m = 2, 3, \cdots, M,$ $n = 1, 2, \cdots N)$ such that

$$\vec{\mathbf{\Phi}}_p^T(z) \triangleq \vec{\mathbf{W}}_p^T(z)\mathbf{H}_p(z) = \left[\, F_p(z) \;\; 0 \;\; \cdots \;\; 0 \,\right]. \tag{6}$$

Consequently, we have

$$Y_p(z) = F_p(z)S_1(z) + U_p(z), \tag{7}$$

where $U_p(z) = \vec{\mathbf{W}}_p^T(z)\vec{\mathbf{B}}_p(z)$.

Obviously a good choice is to let the ith element of $\vec{\mathbf{W}}_p(z)$ be the $(i,1)$th cofactor[1] of $\mathbf{H}_p(z)$. Consequently, the polynomial $F_p(z)$ would be the determinant of $\mathbf{H}_p(z)$, which is not trivial since $\mathbf{H}_p(z)$ has full column normal rank[2] in acoustic environments as assumed in this paper. Note that the $(i,1)$th cofactor of $\mathbf{H}_p(z)$ is only a linear combination of $H_{nm}(z)$ $(m = 2, 3, \cdots, M, n = 1, 2, \cdots N)$. Therefore even though the channel impulse responses corresponding to the speech signal of interest $s_1(k)$ are not known or at least have not yet been blindly identified, we still are able to extract $s_1(k)$ from the interfering sources.

We know that $H_{1m}(z), H_{2m}(z), \cdots, H_{Nm}(z)$ $(\forall m)$ share no common zeros as assumed above. Consequently, $W_{p,1}(z), W_{p,2}(z), \cdots, W_{p,M}(z)$ share no common zeros either. Note that

$$F_p(z) = \vec{\mathbf{W}}_p^T(z)\vec{\mathbf{H}}_{p,:1}(z) = \sum_{q=1}^{M} W_{p,q}(z)H_{p,q1}(z), \tag{8}$$

where $\vec{\mathbf{H}}_{p,:1}(z)$ is the first column vector of $\mathbf{H}_p(z)$. Then it becomes clear that the polynomials $F_p(z)$ $(p = 1, 2, \cdots, P)$ do not share any common zeros. As a result, we obtain an *irreducible* $1 \times P$ SIMO system with $s_1(k)$ as the input. In addition, it can easily be checked that the length of the channel impulse responses of this SIMO system L_f would be less than or equal to $M(L_h - 1) + 1$.

4. RECOVERING THE SPEECH OF INTEREST BY SIMO DEREVERBERATION

In the previous section, we showed how to extract the speech signal of interest from the mixtures with interfering sources. Although the interference has been mitigated, the obtained signals may sound more reverberant due to the prolonged impulse responses of the equivalent channels. Therefore we need to suppress the distortion caused by reverberation.

For the SIMO system with $s_1(k)$ as the input, we do not know the channel impulse responses f_p $(p = 1, 2, \cdots, P)$ since $\vec{\mathbf{H}}_{p,:1}$ in (8) has not yet been identified. But we know that this SIMO system is irreducible and it can be blindly identified with the technique described in Section 2. Thereafter, speech dereverberation

[1] The (i,j)th cofactor c_{ij} of a matrix $\mathbf{A}$ is a signed version of $\mathbf{A}$'s minor d_{ij}: $c_{ij} \triangleq (-1)^{i+j} d_{ij}$, where the minor d_{ij} is the determinant of a reduced matrix that is formed by omitting the ith row and jth column of the matrix A.

[2] For a square matrix $(M \times M)$, the normal rank is full if and only if the determinant, which is a polynomial in z, is not identically zero for all z. In this case, the rank is less than M only at a finite number of points in the z plane.

can be performed using the MINT method [5], again because the SIMO system is irreducible. Due to space limitations, we cannot elaborate the development of the MINT method. But the reader can refer to [5] for the principle and [4] for a practical implementation.

5. SIMULATIONS

5.1. Performance Measures

Similar to what was adopted in our earlier study [3], we will use the normalized projection misalignment (NPM) [6] to evaluate the performance of a BCI algorithm.

To assess the performance of speech extraction and dereverberation, two measures, namely signal-to-interference ratio (SIR) and speech spectral distortion, are used in the simulations. For speech spectral distortion, we employed the Itakura-Saito (IS) distortion measure [7]: $d_{\rm IS}$. The SIR would be defined in a way where a component contributed by $s_1(k)$ is treated as the signal and the rest as the interference since only the first speech source is what we are interested in. We first define the input SIR at microphone n as:

$$\mathrm{SIR}_n^{\mathrm{in}} \triangleq \frac{E\left\{[h_{n1} * s_1(k)]^2\right\}}{\sum_{m=2}^{M} E\left\{[h_{nm} * s_m(k)]^2\right\}},\ n = 1, 2, \cdots, N, \quad (9)$$

where $E\{\cdot\}$ and $*$ denote mathematical expectation and linear convolution, respectively. Then the overall average input SIR is computed as $\mathrm{SIR}^{\mathrm{in}} \triangleq \left[\sum_{n=1}^{N} \mathrm{SIR}_n^{\mathrm{in}}\right]/N$.

The output SIR is defined using the same principle but the expression is more complicated. For a concise presentation, we denote ϕ_{p,s_m} ($p = 1, 2, \cdots, P, m = 1, 2, \cdots, M$) as the impulse response of the equivalent channel from the mth source $s_m(k)$ to the output $y_p(k)$ for the pth $M \times M$ separation subsystem. From (5) and (6), we know that ϕ_{p,s_m} corresponds to the mth element of $\vec{\mathbf{\Phi}}_p(z)$ and $\phi_{p,s_1} = f_p$. Then the average output SIR for the pth subsystem is:

$$\mathrm{SIR}_p^{\mathrm{out}} \triangleq \frac{E\left\{[f_p * s_1(k)]^2\right\}}{\sum_{m=2}^{M} E\left\{[\phi_{p,s_m} * s_m(k)]^2\right\}},\quad p = 1, 2, \cdots, P. \quad (10)$$

Finally, the overall average output SIR is found as $\mathrm{SIR}^{\mathrm{out}} \triangleq \left[\sum_{p=1}^{P} \mathrm{SIR}_p^{\mathrm{out}}\right]/P$.

5.2. Experimental Setup and Results

The simulations were conducted with the impulse responses measured in the varechoic chamber at Bell Labs [8]. A diagram of the floor plan layout is given in Fig. 1, which shows the positions of the four microphones and three sound sources. The first female speech is the target for extraction. The other two sources include one male speaker and one noise source. The two speech sources are equally loud in volume while the noise source is 5 dB weaker than the speech sources. The noise signal was babbling noise recorded in the New York Stock Exchange (NYSE). The sampling rate was 8 kHz. The time sequence and spectrogram (30 Hz bandwidth) of the NYSE babbling noise are shown in Fig. 2. It is clear that the NYSE babbling noise is a pretty wide-band signal. The varechoic chamber is a rectangular room which measures 6.7 m wide

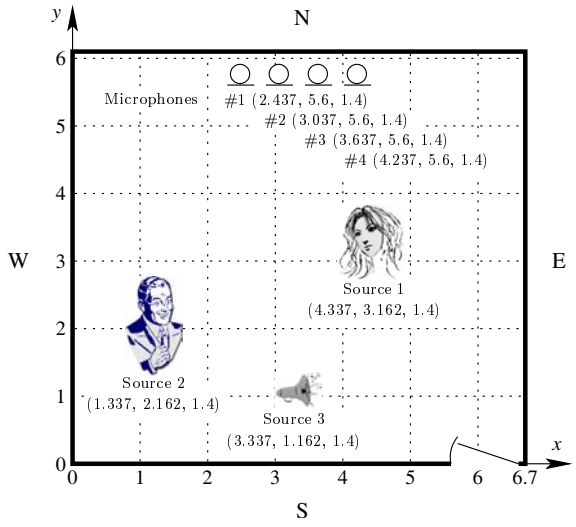

Figure 1: Floor plan of the varechoic chamber at Bell Labs (coordinate values measured in meters).

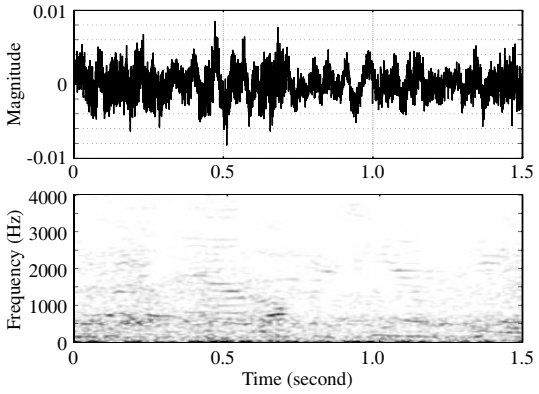

Figure 2: Time sequence and spectrogram (30 Hz bandwidth) of the babbling noise recorded in the NYSE for the first 1.5 seconds.

by 6.1 m deep by 2.9 m high. Its surfaces (walls, floor, and ceiling) are covered with 368 electronically controlled panels that vary the acoustic absorption of the surfaces. The panels are individually controlled and can be either fully opened (absorbing state) or fully closed (reflective state). Therefore, by varying the binary state of each panel in any combination, 2^{238} different room acoustics can be simulated. In the database of [8], there are four panel configurations with 89%, 75%, 30%, and 0% of panels open, which were all used in this paper for assessing performance of the proposed algorithm. The original impulse response measurements have 4096 samples under the 8 kHz sampling rate. When used, they were truncated according to the specified L_h.

Table 1 summarizes the results of the four experiments. Figure 3 visualizes the procedure of speech extraction and dereverberation for the experiment with all panels closed. These results verify the effectiveness of the proposed algorithm.

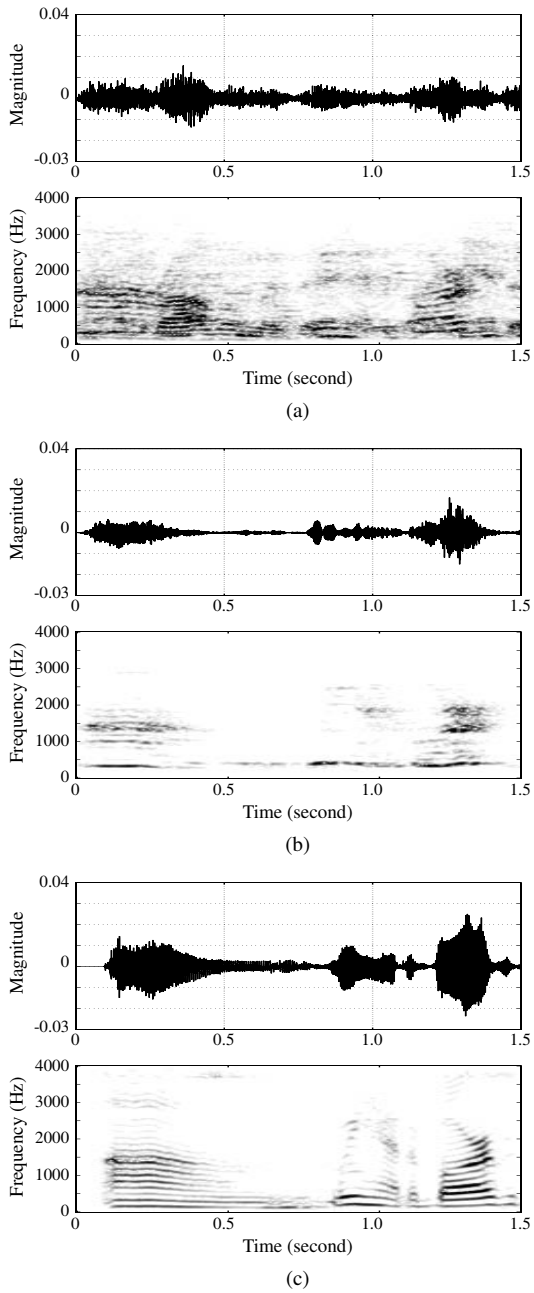

Figure 3: Time sequence and spectrogram (30 Hz bandwidth) of (a) $x_1(k)$, (b) $y_1(k)$, and (c) $\hat{s}_1(k)$ for the experiment carried out in the varechoic chamber with all panels closed.

6. CONCLUSIONS

Capturing a speech signal of interest among a number of interfering sound sources in reverberant, cocktail-party-like environments is difficult but essential to practical speech communication and processing systems. A close-talking microphone is a common engineering solution to this problem. But it has been well received that users of communication and information systems would benefit from an acoustic interface that allows freedom of movement without a body-worn or tethered microphone. In this paper, we proposed an algorithm for speech extraction and enhancement using multiple microphones. The algorithm is based on the technique of blind SIMO identification. The channel impulse responses from an interfering source to the microphones were blindly estimated in the period that the source exclusively occupies. Then the knowledge is used to extract the speech source of interest when all sources voice out by converting a MIMO to a SIMO system with the speech source of interest as the input. The SIMO system is further processed with the MINT method to reduce reverberation in the captured speech signal. Experiments using real impulse responses measured in the varechoic chamber at Bell Labs were conducted and the results demonstrated the promise of the proposed algorithm.

Table 1: Performance of the speech acquisition and enhancement algorithm in reverberant, cocktail-party-like environments simulated with data measured in the varechoic chamber at Bell Labs with different panel configurations.

Open	T_{60}	L_h	NPM (dB)			$\mathrm{SIR}^{\mathrm{in}}$	$\mathrm{SIR}^{\mathrm{out}}$	$d_{\mathrm{IS}}^{\mathrm{SE}}$	$d_{\mathrm{IS}}^{\mathrm{SD}}$
Panels	(ms)		H_{s_2}	H_{s_3}	H_{s_1}	(dB)	(dB)		
89%	240	256	-16.82	-20.75	-18.64	0.945	44.755	4.41	0.07
75%	310	256	-18.93	-23.72	-17.92	1.870	45.163	5.48	0.19
30%	380	512	-13.04	-12.55	-12.13	0.836	40.274	5.65	0.32
0%	580	512	-13.51	-16.90	-12.56	1.725	42.275	9.54	0.14

NOTES:

H_{s_m} represents the SIMO system corresponding to source s_m.
T_{60} denotes 60-dB reverberation time in the 20-4000 Hz band.
$d_{\mathrm{IS}}^{\mathrm{SE}}$ and $d_{\mathrm{IS}}^{\mathrm{SD}}$ stand for the Itakura-Saito measures after speech extraction (SE) and speech dereverberation (SD).

7. REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sept. 1953.

[2] L. Tong, G. Xu, and T. Kailath, "A new approach to blind identification and equalization of multipath channels," in *Proc. 25th Asilomar Conf. on Signals, Systems, and Computers*, 1991, vol. 2, pp. 856–860.

[3] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multi-channel identification," *IEEE Trans. Signal Processing*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[4] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 882–896, Sept. 2005.

[5] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 145–152, Feb. 1988.

[6] D. R. Morgan, J. Benesty, and M. M. Sondhi, "On the evaluation of estimated impulse responses," *IEEE Signal Processing Lett.*, vol. 5, no. 7, pp. 174–176, Jul. 1998.

[7] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[8] A. Härmä, "Acoustic measurement data from the varechoic chamber," Technical Memorandum, Agere Systems, Nov. 2001.