JOINT OPTIMIZATION OF LCMV BEAMFORMING AND ACOUSTIC ECHO CANCELLATION FOR AUTOMATIC SPEECH RECOGNITION

W. Herbordt[†]*, S. Nakamura*[†]*, and W. Kellermann*[‡]

email: {wolfgang.herbordt, satoshi.nakamura}@atr.jp, wk@LNT.de
[†]ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seikacho Soraku-gun, Kyoto 619-0288, Japan
[‡]Telecommunications Laboratory, University Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany

ABSTRACT

For full-duplex hands-free acoustic human/machine interfaces, often a combination of acoustic echo cancellation and speech enhancement in order to suppress acoustic echoes, local interference, and noise is required. In order to optimally exploit positive synergies between acoustic echo cancellation and speech enhancement, we presented in an earlier work a combined least-squares (LS) optimization criterion for the integration of acoustic echo cancellation and adaptive linearly-constrained minimum variance (LCMV) beamforming [1]. In this contribution, we illustrate the efficiency of the proposed solution in situations with high levels of background noise and with time-varying echo paths and frequent double-talk by speech recognition experiments.

1. INTRODUCTION

For audio signal acquisition in hands-free human/machine interfaces, adaptive beamforming microphone arrays can be efficiently used for enhancing a desired signal while suppressing interference and noise [2]. For full-duplex communication systems, not only local interferers and noise corrupt the desired signal, but also acoustic echoes of loudspeakers. For suppressing acoustic echoes, acoustic echo cancelers (AECs) [3] are the optimum choice since they exploit the available loudspeaker signals as reference information.

For maximally suppressing local interference and echo signals, it is thus desirable to combine acoustic echo cancellation with adaptive beamforming in the acoustic human/machine interface. For optimum performance, a maximum number of degrees of freedom should be available for echo and interference suppression for the acoustic conditions which are met in practical applications. The statistics of the signals and of the wave propagation range from slowly time-varying to fast time-varying with wide ranges of signal-to-noise (power) ratios (SNRs) and of signal-to-acoustic-echoes (power) ratios (SERs).

In [4], a structure with the acoustic echo cancelers (AECs) in the sensor channels (see Fig. 1) in front of the adaptive beamformer ('AEC first') was identified as the optimum solution in terms of maximum interference and echo suppression, since, after convergence of the AECs, (a) all degrees of freedom of the adaptive beamformer are available for interference suppression, and (b) the AECs generally provide higher echo suppression than the adaptive beamformer. However, in practical situations, we often have to deal with frequent double-talk between acoustic echoes and the desired speaker and interference, so that the adaptation of the AEC is difficult. In combination with highly time-varying echo paths, this leads to reduced performance of the AEC and to reduced echo and interference suppression of the combined system in turn. Moreover, one AEC is necessary for each sensor channel so that the M-fold computational power, where M is the number of microphones, is required at least for the filtering and for the filter update compared to a single AEC. Even with moderate numbers of microphones ($4 \leq M \leq 8$), this is a limiting factor for the usage of 'AEC first' in cost-sensitive systems [4].

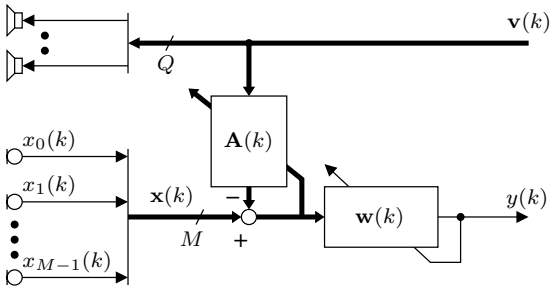

Fig. 1. Combination of AEC and beamforming ('AEC first') [4, 5, 6].

For resolving these problems, combined optimization of the AEC and of the adaptive beamformer would be the better solution. Ideally, this would allow to adapt the AEC during activity of the desired speaker and interference so that transient echo paths can be optimally tracked and echo and interference suppression are maximized. Therefore, a combined optimization criterion of linearly-constrained minimum-variance (LCMV) beamforming and multi-channel acoustic echo cancellation was presented in [1]. For studying the behavior, the proposed optimization criterion was realized in combination with a computationally efficient generalized sidelobe canceler (GSC) [7] ('generalized echo and interference canceler' (GEIC)). In this contribution, we study the behavior of the system in more detail, and, especially, apply the proposed system as a front-end for an automatic speech recognizer for time-invariant and time-varying acoustic conditions.

In Sect. 2, we review the combined optimization criterion. Section 3 describes the practical realization of the combined system

This research was supported in part in a program with the National Institute of Information and Communications Technology (NICT), Japan, entitled "JAPAN TRUST International Research Coorporation Program".

based on the GSC structure. Section 4 illustrates the performance by experimental results for practically relevant scenarios.

2. OPTIMIZATION CRITERION

In contrast to 'AEC first' in Fig. 1, where the AEC is optimized independently of the beamformer, we propose to use the output signal $y(k)$ for the optimization of both the AEC and the LCMV beamformer as shown in Fig. 2. The reference loudspeaker signals $\mathbf{v}(k)$ can thus be interpreted as additional input signals for the adaptive beamformer.[1]

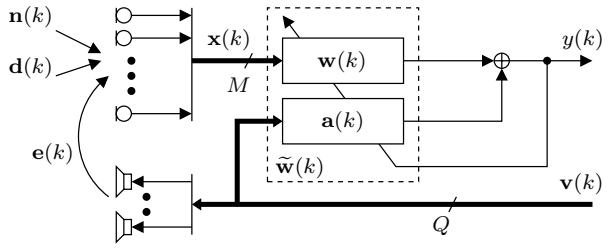

Fig. 2. Joint optimization of LCMV beamforming and acoustic echo cancellation.

We assume that the sensor signals $\mathbf{x}(k)$ are given by the superposition of the desired signal $\mathbf{d}(k)$, local interference $\mathbf{n}(k)$, and acoustic echoes $\mathbf{e}(k)$,

$$\mathbf{x}(k) = \mathbf{d}(k) + \mathbf{n}(k) + \mathbf{e}(k)\,, \tag{1}$$

where $\mathbf{d}(k)$, $\mathbf{n}(k)$, and $\mathbf{e}(k)$ are zero-mean and mutually uncorrelated. The output signal $y(k)$ of the combined system can be written as a function of the sensor signals $\mathbf{x}(k)$, the loudspeaker signals $\mathbf{v}(k)$, the stacked beamformer weight vector $\mathbf{w}(k)$, and the stacked AEC weight vector $\mathbf{a}(k)$ as

$$y(k) = \mathbf{w}^T(k)\mathbf{x}(k) + \mathbf{a}^T(k)\mathbf{v}(k)\,, \tag{2}$$

where

$$\begin{aligned}
\mathbf{x}(k) &= (\mathbf{x}_0(k),\, \mathbf{x}_1(k),\, \ldots,\, \mathbf{x}_{M-1}(k))^T\,, && (3)\\
\mathbf{x}_m(k) &= (x_m(k),\, x_m(k-1),\, \ldots,\, x_m(k-N_{\mathbf{w}}+1))^T, && (4)\\
\mathbf{v}(k) &= (\mathbf{v}_0(k),\, \mathbf{v}_1(k),\, \ldots,\, \mathbf{v}_{Q-1}(k))^T\,, && (5)\\
\mathbf{v}_q(k) &= (v_q(k),\, v_q(k-1),\, \ldots,\, v_q(k-N_{\mathbf{a}}+1))^T\,, && (6)\\
\mathbf{w}(k) &= (\mathbf{w}_0(k),\, \mathbf{w}_1(k),\, \ldots,\, \mathbf{w}_{M-1}(k))^T\,, && (7)\\
\mathbf{w}_m(k) &= (w_{0,m}(k),\, w_{1,m}(k),\, \ldots,\, w_{N_{\mathbf{w}}-1,m}(k))^T\,, && (8)\\
\mathbf{a}(k) &= (\mathbf{a}_0(k),\, \mathbf{a}_1(k),\, \ldots,\, \mathbf{a}_{Q-1}(k))^T\,, && (9)\\
\mathbf{a}_q(k) &= (a_{0,q}(k),\, a_{1,q}(k),\, \ldots,\, a_{N_{\mathbf{a}}-1,q}(k))^T\,. && (10)
\end{aligned}$$

Q is the number of loudspeaker channels, and $N_{\mathbf{w}}$ and $N_{\mathbf{a}}$ are the number of filter coefficients of the beamformer weight vectors $\mathbf{w}_m(k)$ and of the AEC filters $\mathbf{a}_q(k)$, respectively. With stacked vectors

$$\widetilde{\mathbf{w}}(k) = \left(\mathbf{w}^T(k),\, \mathbf{a}^T(k)\right)^T\,, \tag{11}$$

$$\widetilde{\mathbf{x}}(k) = \left(\mathbf{x}^T(k),\, \mathbf{v}^T(k)\right)^T\,, \tag{12}$$

we can write $y(k)$ as

$$y(k) = \widetilde{\mathbf{w}}^T(k)\widetilde{\mathbf{x}}(k)\,. \tag{13}$$

A LS optimization criterion is obtained by minimizing the windowed sum of squared output signal samples $y(k)$ subject to constraints which assure that the desired signal is not distorted by $\widetilde{\mathbf{w}}(k)$. That is,

$$\min_{\widetilde{\mathbf{w}}(k)} \sum_{i=0}^{k} w_i(k) y^2(i) \quad \text{subject to} \quad \widetilde{\mathbf{C}}^T(k)\widetilde{\mathbf{w}}(k) = \mathbf{c}(k)\,. \tag{14}$$

The windowing function $w_i(k)$ extracts desired samples from the output signal $y(k)$ which should be included into the optimization. For example, infinite memory with exponential averaging is obtained with $w_i(k) = \lambda^{k-i}$ [9]. The constraint matrix $\widetilde{\mathbf{C}}(k)$ of size $(MN_{\mathbf{w}} + QN_{\mathbf{a}}) \times C$ and the constraint column vector $\mathbf{c}(k)$ of length C put C spatial constraints onto $\widetilde{\mathbf{w}}(k)$ in order to assure unity beamformer response for the direction-of-arrival of the desired signal [10]. Since the Q loudspeaker signals $\mathbf{v}(k)$ are assumed to be uncorrelated with the desired signal, the constraints are only required for the microphone signals, just as for conventional LCMV beamformers [10]. We can thus write $\widetilde{\mathbf{C}}(k)$ as

$$\widetilde{\mathbf{C}}(k) = \left(\mathbf{C}^T(k),\, \mathbf{0}_{C\times QN_{\mathbf{a}}}\right)^T\,, \tag{15}$$

where $\mathbf{C}(k)$ of size $MN_{\mathbf{w}} \times C$ is a conventional constraint matrix known from LCMV beamforming [10]. We thus obtain with (14) a formally simple optimization criterion, where only one single error signal needs to be minimized for an arbitrary number of microphones. This combined optimization allows to update the beamformer and the AEC simultaneously – in contrast to 'AEC first', where the AEC can only be updated if local interference and desired signal are not active. As a consequence, it is assured that acoustic echoes and interference are maximally suppressed even in highly non-stationary acoustic conditions with frequent double-talk and/or high background noise levels.[2] The number of spatial degrees of freedom for interference suppression and for echo cancellation are increased by the number of loudspeakers Q relative to a beamformer alone.

3. REALIZATION AS A GENERALIZED SIDELOBE CANCELER

A direct solution of (14) can be determined using Lagrange multipliers [10]. However, with regard to an efficient realization of this combined system, we transform the constrained optimization problem into an unconstrained one using the structure of the GSC [7, 11]. Applying this transformation to (14) [1], we obtain the generalized echo and interference canceler as depicted in Fig. 3, where the AEC is combined with the interference canceler $\mathbf{w}_{\mathrm{a}}(k)$. That is,

$$\widetilde{\mathbf{w}}_{\mathrm{a}}(k) = (\mathbf{w}_{\mathrm{a}}^T(k),\, \mathbf{a}^T(k))^T\,. \tag{16}$$

The weight vector $\mathbf{w}_{\mathrm{a}}(k)$ of length $N_{\mathbf{w}}$ is defined according to $\mathbf{w}(k)$ with the filter coefficients of the adaptive beamformer replaced by the filter coefficients of the interference canceler.

[1] This idea was first used in [8] for a combination of acoustic echo cancellation and multi-channel noise-reduction based on generalized singular value decomposition (GSVD).

[2] This statement is subject to the condition that the realization of the beamformer allows for sufficient tracking capability in such situations. For our realization based on a generalized sidelobe canceler we show in Section 4 that this condition is fulfilled.

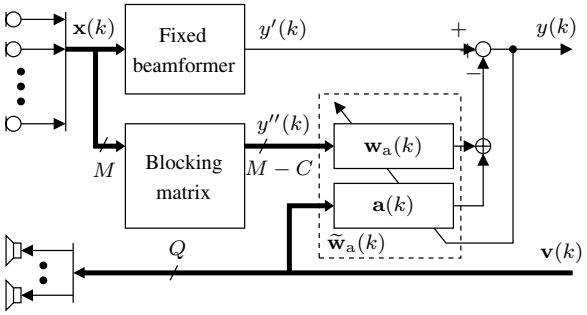

Fig. 3. Generalized echo and interference canceler (GEIC).

For realizing the GEIC for practical applications, we use the implementation of the GSC in the discrete Fourier transform (DFT) domain after [6]. This implementation uses an adaptive blocking matrix for tracking movements of the desired source and for robustness against distortion of the desired signal in reverberant environments. Blocking matrix and $\widetilde{\mathbf{w}}_\mathrm{a}(k)$ are realized using DFT-domain adaptive filtering. The blocking matrix is adapted for presence of only desired signal, $\widetilde{\mathbf{w}}_\mathrm{a}(k)$ is adapted for presence of local interference and/or acoustic echoes. A separate adaptation control for the AEC as for 'AEC first' is thus not required. The fixed beamformer is realized as a uniformly weighted delay&sum beamformer.

Among others [1, 6], it should be especially considered for the realization of GEIC that the AEC $\mathbf{a}(k)$ and the interference canceler $\mathbf{w}_\mathrm{a}(k)$ should have the same filter length $N_\mathbf{a} = N_\mathbf{w}$ in order to assure the same convergence speed. Therefore, $\mathbf{a}(k)$ should not be viewed as a conventional AEC but as additional degrees of freedom for the interference canceler. As a result, the echo suppression of GEIC will be smaller than the echo suppression of 'AEC first' for stationary conditions especially for reverberant environments, where generally $N_\mathbf{a} > N_\mathbf{w}$. For environments with low reverberation the performance of GEIC for stationary conditions approaches that of 'AEC first'.

4. EXPERIMENTAL RESULTS

We study the performance of GEIC relative to the GSC alone and relative to 'AEC first' as a front-end for a speaker-independent connected digit speech recognizer.

4.1. Experimental setup

The speech recognizer is based on the HTK software [12]. The sampling rate is 8 kHz, the frequency range is 0.2–4 kHz. 13 Mel-frequency cepstral coefficients (including the coefficient of order 0) plus the corresponding delta coefficients are calculated from the pre-emphasized input signals using cepstral mean normalization. The digits are modeled as whole word Hidden Markov Models with 18 states per word including entry and exit state without skips over states using mixtures of 3 Gaussians with mean and diagonal covariance matrices for modeling the output probabilities. A voice activity detector is not used. The recognizer is trained using the clean training set of the TIDigits data base [13]. The baseline word accuracy for the clean test set of the TIDigits database is 98.29%.

The acoustic environment is the passenger cabin of a car with presence of slowly-time varying car noise and acoustic echo signals. The room impulse responses between the loudspeakers and the microphones and between the desired source and the microphones are simulated using the image method [14] with a simulated reverberation time $T_{60} = 50$ ms. The desired signal is the test set of the TIDigits database, the loudspeaker signals are stereophonic music. The microphone signals are obtained by convolution of the clean source signals with the room impulse responses and superposition with variable SNR and fixed $SER = 7$ dB. The microphone array consists of $M = 4$ sensors with spacing 4 cm, the array aperture is 12 cm ('AEC first': $N_\mathbf{a} = 512$, $N_\mathbf{w} = 256$; GEIC, GSC: $N_\mathbf{a} = N_\mathbf{w} = 256$).

4.2. Fixed source positions

In this experiment, we consider fixed positions of the desired source and of the loudspeakers. The desired source and the loudspeakers are located in broadside direction ($\theta = 90$ degrees) and in the two endfire directions ($\theta = 0,\ 180$ degrees), respectively, at a distance of 60 cm from the array center. The echo suppression $ERLE$ and the noise reduction NR averaged over the whole test set are given in Fig. 4 as a function of the SNR at the sensors. The word accuracies are given in Fig. 5a[3].

For high SNR (equivalent to high echo-to-noise ratio (ENR), since SER is fixed), the AECs of 'AEC first' converge in speech pauses and provide high echo suppression, which translates to a greater $ERLE$ and NR of 'AEC first' relative to GSC and GEIC (Fig. 4). With decreasing ENR, the echo suppression of the AECs of 'AEC first' decreases until the AECs are inefficient and $ERLE$ and NR of 'AEC first' are equivalent to the GSC. Here, the GEIC outperforms 'AEC first', since the number of degrees of freedom does not depend on the ENR. Nevertheless, $ERLE$ of GEIC decreases with decreasing ENR, since the system concentrates on the suppression of the stronger car noise. These characteristics are directly reflected by the word accuracies of the speech recognizer (Fig. 5a). For comparison, the averaged word accuracies of the unprocessed microphone signals and the word accuracy of a simple delay&sum beamfomer are depicted, too.

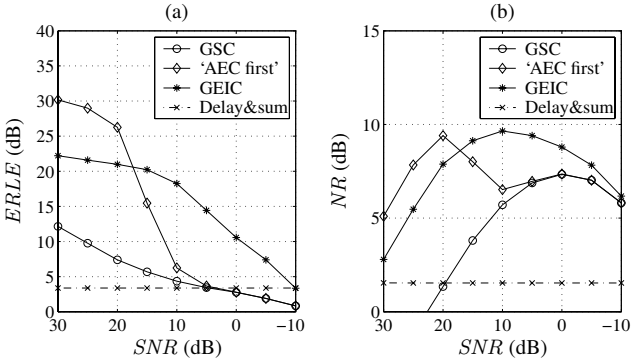

Fig. 4. Noise reduction NR and echo suppression $ERLE$ for GSC alone, 'AEC first', and GEIC for fixed echo paths and fixed source position for $SER = 7$ dB.

4.3. Time-varying echo path and moving desired source

In this experiment, we consider a time-varying echo path and a moving desired source. The desired source position is switched randomly for each file of the TIDigits test set in the interval $\theta =$

[3]Experimental results for the same recognition task without presence of acoustic echoes and with varying numbers of sensors can be found in [6].

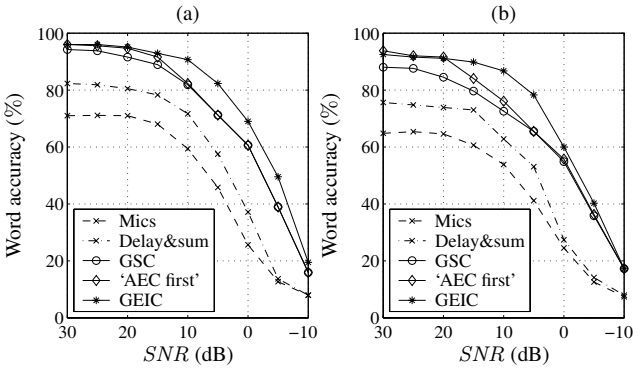

Fig. 5. Word accuracy for GSC alone, 'AEC first', and GEIC for **(a)** fixed source position and fixed echo paths and **(b)** time-varying source position and time-varying echo paths for $SER = 7\,\text{dB}$.

$80 \ldots 100$ degrees in steps of 2 degrees[4] with equal probability for all directions.

One of the loudspeakers is located at $\theta = 180$ degrees. The second loudspeaker position is switched every 20000 samples between $\theta = 0$ and $\theta = 60$ degrees. The distance between the sources and the array center is fixed at 60 cm. The results are given in Fig. 6 and in Fig. 5b. Compared to fixed source positions, we notice that $ERLE$ and NR for 'AEC first' are reduced relative to GEIC and GSC for $SNR \geq 20\,\text{dB}$. This effect can be explained by the reduced efficiency of the AECs of 'AEC first' and the missing capability to adapt the AEC during double-talk of interference and acoustic echoes. The performance loss is mostly related to the time-variance of the acoustic echo paths, since experiments show that the performance for fixed echo paths and time-varying source position ($\theta = 80 \ldots 100$ degrees) is almost equivalent to fixed echo paths and fixed source position in Fig. 5a.

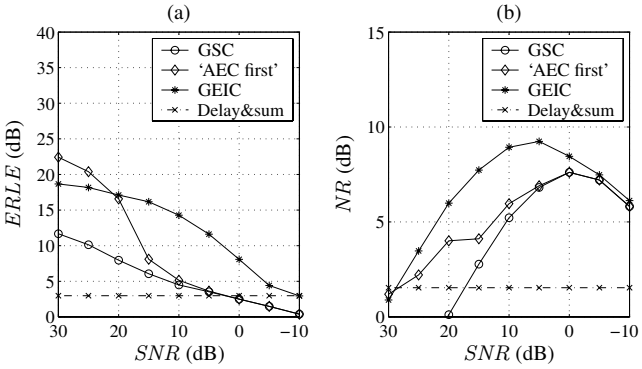

Fig. 6. Noise rejection NR and echo suppression $ERLE$ for GSC alone, 'AEC first', and GEIC for time-varying echo paths and fixed source position for $SER = 7\,\text{dB}$.

5. CONCLUSIONS

We studied a technique for joint optimization of acoustic echo cancellation and adaptive LCMV beamforming. With a realization example based on a robust GSC, we showed that this structure is especially efficient for (a) transient echo paths if frequent double-talk between acoustic echoes, local interference, and desired speakers is to be expected and (b) high levels of background noise. For stationary conditions and low levels of background noise, the performance of GEIC is reduced relative to 'AEC first' due to the limited number of filter taps of the beamformer weight vector. The proposed solution requires only one AEC for an arbitrary number of microphones and no separate adaptation control for the AEC.

[4]This range corresponds to the 5 dB-width of the mainlobe at 4 kHz.

6. REFERENCES

[1] W. Herbordt, W. Kellermann, and S. Nakamura, "Combined optimization of lcmv beamforming and accoustic echo cancellation," *Proc. EURASIP European Signal Processing Conference*, 2004.

[2] M.S. Brandstein and D.B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin, 2001.

[3] C. Breining et al., "Acoustic echo control - an application of very-high-order adaptive filters," *IEEE Signal Processing Magazine*, vol. 16, no. 4, pp. 42–69, July 1999.

[4] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M.S. Brandstein and D.B. Ward, Eds., chapter 13, pp. 281–306. Springer, Berlin, 2001.

[5] R. Martin, *Freisprecheinrichtungen mit mehrkanaliger Echokompensation und Störgeräuschreduktion*, Ph.D. thesis, Aachener Institut für Nachrichtengeräte und Datenkommunikation, 1995.

[6] W. Herbordt, *Sound capture for human/machine interfaces: Practical aspects of microphone array signal processing*, Springer, Heidelberg, Germany, 2005.

[7] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, January 1982.

[8] S. Doclo, M. Moonen, and E. De Clippel, "Combined acoustic echo and noise reduction using GSVD-based optimal filtering," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1051–1054, June 2000.

[9] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, New Jersey, 3rd edition, 1996.

[10] H.L. Van Trees, *Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory*, John Wiley & Sons, Inc., New York, 2002.

[11] K.M. Buckley, "Broad-band beamforming and the generalized sidelobe canceller," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1322–1323, October 1986.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*, Entropic Cambridge Research Labs, Cambridge, UK, 2000.

[13] R.G. Leonard, "A database for speaker independent digit recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 42.11.1–42.11.4, March 1984.

[14] J.B. Allen, "Multimicrophone signal-processing technique to remove room reverberation from speech signals," *Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, October 1977.