A MINIMUM SPEECH DISTORTION MULTICHANNEL ALGORITHM FOR NOISE REDUCTION

Jacob Benesty[1]*, Jingdong Chen*[2]*, and Yiteng (Arden) Huang*[2]

[1]: INRS-EMT, University of Quebec
800 de la Gauchetière Ouest, Suite 6900
Montreal, Quebec, H5A 1K6, Canada
e-mail: benesty@emt.inrs.ca

[2]: Bell Labs, Alcatel-Lucent
600 Mountain Avenue
Murray Hill, NJ, 07974, USA
e-mail:{jingdong,arden}@research.bell-labs.com

ABSTRACT

Noise reduction using multiple microphones remains a challenging and crucial research problem. This paper presents a new multichannel noise-reduction algorithm based on spatio-temporal prediction. Unlike many multichannel techniques that attempt to achieve both speech dereverberation and noise reduction at the same time, this new approach puts aside speech dereverberation and formulates the problem as one of estimating the speech component received at one microphone using the observations from all the available microphones. In comparison with the existing techniques such as beamforming, this new multichannel approach has many appealing properties: it does not require the knowledge of the source location or the channel impulse responses; the multiple microphones do not have to be arranged into a specific array geometry; it works the same for both the far-field and near-field cases; and most importantly, it can produce very good noise reduction with minimum speech distortion in real acoustic environments.

Index Terms— Microphone arrays, noise reduction, speech enhancement, beamforming.

1. INTRODUCTION

The problem of noise reduction using multiple microphones has attracted a considerable amount of research attention. This is due to the fact that, in theory, it is possible to develop multichannel algorithms that can achieve significant noise reduction without distorting the desired speech signal. The multichannel noise-reduction problem is illustrated in Fig. 1, where we have a speech source in the sound field and use N microphones to collect the signals from their field of view. The output of the nth microphone is given by

$$\begin{aligned} y_n(k) &= s(k) * g_n + v_n(k) \\ &= x_n(k) + v_n(k), \quad n = 1, 2, \ldots, N, \end{aligned} \tag{1}$$

where $*$ denotes convolution, $s(k)$ is the source signal, g_n represents the acoustic channel impulse response from the source to microphone n, and $x_n(k)$, $v_n(k)$, and $y_n(k)$ are, respectively, the speech, the noise, and the noisy signals at the nth microphone. It is assumed that $v_n(k)$ is a zero-mean random process that is uncorrelated with $x_n(k)$. The objective of noise reduction is to mitigate the effect due to the noise signals, $v_n(k)$, $n = 1, 2, \ldots, N$.

In most of the traditional multichannel techniques such as beamforming, the speech enhancement problem consists in the estimation of the source signal, $s(k)$, from the observed noisy signals, $y_n(k)$ [1]–[4]. This indeed involves two simultaneous subtasks, i.e., speech dereverberation and noise reduction. However, speech dereverberation alone is a very difficult problem and there have not been any practical solutions thus far. Therefore, it seems more reasonable that we take speech dereverberation and noise reduction apart, and tackle the two problems one at a time. This philosophy, also used in [5], is embraced in this study. Here, we put aside speech dereverberation and focus exclusively on noise reduction. So, the problem considered in this paper can be described as one of estimating the speech signal observed at one microphone from the noisy signals received at all the N microphones. Let us assume that we want to estimate the speech signal at the mth microphone. Then, the objective of this paper is to estimate $x_m(k)$, given $y_n(k)$, $n = 1, 2, \ldots, N$.

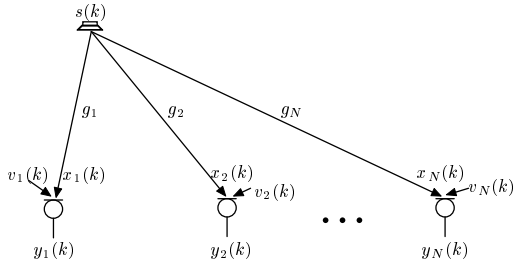

Fig. 1. Illustration of a multichannel system.

2. PROPOSED MMSE ESTIMATOR WITH MULTIPLE MICROPHONES

With the signal model given in (1), an estimate of the speech component $x_m(k)$ can be obtained by passing the N observed signals through N temporal filters, i.e.,

$$\begin{aligned} \hat{x}_m(k) &= \mathbf{h}_{1m}^T \mathbf{y}_1(k) + \mathbf{h}_{2m}^T \mathbf{y}_2(k) + \cdots + \mathbf{h}_{Nm}^T \mathbf{y}_N(k) \\ &= \sum_{n=1}^{N} \mathbf{h}_{nm}^T \mathbf{y}_n(k), \end{aligned} \tag{2}$$

where

$$\mathbf{y}_n(k) = \left[\begin{array}{cccc} y_n(k) & y_n(k-1) & \cdots & y_n(k-L+1) \end{array} \right]^T,$$

and

$$\mathbf{h}_{nm} = \left[\begin{array}{cccc} h_{nm,0} & h_{nm,1} & \cdots & h_{nm,L-1} \end{array} \right]^T, \quad n = 1, 2, \ldots, N,$$

are, respectively, the observation signal vectors and the N FIR filters of length L. The corresponding error signal obtained by this

estimation is written as

$$\begin{aligned} e_m(k) &= \hat{x}_m(k) - x_m(k) \\ &= \sum_{n=1}^{N} \mathbf{h}_{nm}^T \mathbf{y}_n(k) - x_m(k). \end{aligned} \quad (3)$$

From the signal model given in (1), we have

$$\mathbf{y}_n(k) = \mathbf{x}_n(k) + \mathbf{v}_n(k), \quad n = 1, 2, \ldots, N, \quad (4)$$

where $\mathbf{x}_n(k)$ and $\mathbf{v}_n(k)$ are the speech and noise signal vectors, defined similarly to $\mathbf{y}_n(k)$. Substituting (4) into (3), we can decompose the error signal into the following form:

$$e_m(k) = e_{x,m}(k) + e_{v,m}(k), \quad (5)$$

where

$$e_{x,m}(k) = \sum_{n=1}^{N} \mathbf{h}_{nm}^T \mathbf{x}_n(k) - x_m(k) \quad (6)$$

and

$$e_{v,m}(k) = \sum_{n=1}^{N} \mathbf{h}_{nm}^T \mathbf{v}_n(k). \quad (7)$$

The term $e_{x,m}(k)$ quantifies how much the speech sample $x_m(k)$ is distorted due to the filtering operation. The larger the mean-square value of $e_{x,m}(k)$, the higher the speech is distorted. In comparison, the term $e_{v,m}(k)$ tells how much the noise is reduced. The smaller the mean-square value of $e_{v,m}(k)$, the more the noise is reduced. So, ideally, noise reduction is a problem of finding an optimal set of the filters $\mathbf{h}_{nm}$ $(n = 1, 2, \ldots, N)$ such that the mean-square error (MSE) corresponding to the residual noise is minimized while keeping the speech distortion $e_{x,m}(k)$ close to 0.

From (7), we can write the MSE associated with the residual noise as

$$\begin{aligned} J_{v,m}(\mathbf{h}_m) &= E\left[e_{v,m}^2(k)\right] \\ &= \mathbf{h}_m^T \mathbf{R}_{vv} \mathbf{h}_m, \end{aligned} \quad (8)$$

where

$$\mathbf{h}_m = \left[\begin{array}{cccc} \mathbf{h}_{1m}^T & \mathbf{h}_{2m}^T & \cdots & \mathbf{h}_{Nm}^T \end{array} \right]^T, \quad (9)$$

$\mathbf{R}_{vv} = E[\mathbf{v}(k)\mathbf{v}^T(k)]$ is the noise correlation matrix, and

$$\mathbf{v}(k) = \left[\begin{array}{cccc} \mathbf{v}_1^T(k) & \mathbf{v}_2^T(k) & \cdots & \mathbf{v}_N^T(k) \end{array} \right]^T. \quad (10)$$

Now, the noise-reduction problem is equivalent to finding the optimal filter as follows:

$$\mathbf{h}_{m,\mathrm{o}} = \arg\min_{\mathbf{h}_m} J_{v,m}(\mathbf{h}_m) \quad \text{subject to} \quad e_{x,m}(k) = 0. \quad (11)$$

The solution to (11) depends on the number of microphones and how the spatio-temporal information is exploited. We have two cases: $N = 1$ and $N \geq 2$.

Case 1: $N = 1$.

In this case, we have $m = N = 1$. If the current speech sample $x_1(k)$ cannot be completely predicted from its past samples (which is generally true in practice), we can easily check that the solution to (11) is

$$\mathbf{h}_{1,\mathrm{o}} = \mathbf{u}_1, \quad (12)$$

where

$$\mathbf{u}_1 = \left[\begin{array}{cccc} 1 & 0 & \cdots & 0 \end{array} \right]^T \quad (13)$$

is a unit vector of length L. With this degenerate filter, there will be no noise reduction. So, in the single-channel scenario, if we want to keep the speech undistorted, there will be no noise reduction. But if we still want to achieve some noise reduction, we need to loosen the constraint to allow some speech distortion. Indeed, this is almost the *de facto* standardized practice in the existing single-channel noise-reduction techniques, where noise reduction is achieved at the expense of speech distortion [6], [7].

Case 2: $N \geq 2$.

In the single-channel situation, there is a fundamental compromise between noise reduction and speech distortion. But if we use multiple microphones, we can take advantage of the redundancy among the microphones (or in other words, the spatio-temporal information) to achieve noise reduction without introducing any speech distortion.

Let us assume that we can find N filter matrices, $\mathbf{W}_{nm}$ $(n = 1, 2, \ldots, N)$, such that

$$\mathbf{x}_n(k) = \mathbf{W}_{nm} \mathbf{x}_m(k), \; n = 1, 2, \ldots, N. \quad (14)$$

For $n = m$, we have $\mathbf{W}_{mm} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. We will discuss later on how to determine an optimal estimate of the matrix $\mathbf{W}_{nm}$ for $n \neq m$; but for now, we assume that $\mathbf{W}_{nm}$ are known. Substituting (14) into (6), we obtain

$$e_{x,m}(k) = \mathbf{x}_m^T(k) \left[\mathbf{W}_m \mathbf{h}_m - \mathbf{u}_1 \right], \quad (15)$$

where

$$\mathbf{W}_m = \left[\begin{array}{cccc} \mathbf{W}_{1m}^T & \mathbf{W}_{2m}^T & \cdots & \mathbf{W}_{Nm}^T \end{array} \right]. \quad (16)$$

With this expression of the speech distortion, we can rewrite the constrained estimation problem in (11) into the following form:

$$\mathbf{h}_{m,\mathrm{o}} = \min_{\mathbf{h}_m} J_{v,m}(\mathbf{h}_m) \quad \text{subject to} \quad \mathbf{W}_m \mathbf{h}_m = \mathbf{u}_1. \quad (17)$$

If we use a Lagrange multiplier to adjoin the constraint to the cost function, the estimation problem in (17) can be written as

$$\mathbf{h}_{m,\mathrm{o}} = \arg\min_{\mathbf{h}_m} \mathcal{L}\left(\mathbf{h}_m, \boldsymbol{\lambda}\right), \quad (18)$$

where

$$\begin{aligned} \mathcal{L}\left(\mathbf{h}_m, \boldsymbol{\lambda}\right) &= J_{v,m}(\mathbf{h}_m) + \boldsymbol{\lambda}^T \left(\mathbf{W}_m \mathbf{h}_m - \mathbf{u}_1\right) \\ &= \mathbf{h}_m^T \mathbf{R}_{vv} \mathbf{h}_m + \boldsymbol{\lambda}^T \left(\mathbf{W}_m \mathbf{h}_m - \mathbf{u}_1\right), \end{aligned}$$

and vector $\boldsymbol{\lambda}$ is the Lagrange multiplier. Evaluating the gradient of $\mathcal{L}\left(\mathbf{h}_m, \boldsymbol{\lambda}\right)$ with respect to $\mathbf{h}_m$ and equating the result to zero gives

$$\frac{\partial}{\partial \mathbf{h}_m} \mathcal{L}(\mathbf{h}_m, \boldsymbol{\lambda}) = 2\mathbf{R}_{vv}\mathbf{h}_m + \mathbf{W}_m^T \boldsymbol{\lambda} = \mathbf{0}. \quad (19)$$

Now if we assume that the noise signals at the different microphones are not completely coherent so that the noise covariance matrix $\mathbf{R}_{vv}$

is full rank, from (19) and using the constraint, we easily find the solution to (18):

$$\mathbf{h}_{m,\mathrm{o}} = \mathbf{R}_{vv}^{-1}\mathbf{W}_m^T \left[\mathbf{W}_m\mathbf{R}_{vv}^{-1}\mathbf{W}_m^T\right]^{-1}\mathbf{u}_1. \quad (20)$$

We see that, in order to compute the optimal filter $\mathbf{h}_{m,\mathrm{o}}$, we need to know the two matrices $\mathbf{R}_{vv}$ and $\mathbf{W}_m$. The noise correlation matrix $\mathbf{R}_{vv}$ can be estimated during the absence of speech if we have a voice activity detector (VAD). In the next section, we will discuss how to determine the $\mathbf{W}_m$ matrix.

3. ESTIMATION OF THE W MATRIX

From (14), we can construct the following MSE cost function:

$$J(\mathbf{W}_{nm}) = E\left\{[\mathbf{x}_n(k) - \mathbf{W}_{nm}\mathbf{x}_m(k)]^T \times [\mathbf{x}_n(k) - \mathbf{W}_{nm}\mathbf{x}_m(k)]\right\}. \quad (21)$$

Differentiating $J(\mathbf{W}_{nm})$ with respect to $\mathbf{W}_{nm}$ and equating the result to zero, we obtain an optimal estimate of the $\mathbf{W}_{nm}$ matrix:

$$\mathbf{W}_{nm,\mathrm{o}} = \mathbf{R}_{x_nx_m}\mathbf{R}_{x_mx_m}^{-1}, \quad (22)$$

where

$$\mathbf{R}_{x_nx_m} = E\left[\mathbf{x}_n(k)\mathbf{x}_m^T(k)\right]$$

and

$$\mathbf{R}_{x_mx_m} = E\left[\mathbf{x}_m(k)\mathbf{x}_m^T(k)\right]$$

are, respectively, the cross-correlation and correlation matrices of the speech signals. However, the signals $\mathbf{x}_n(k)$ and $\mathbf{x}_m(k)$ are not observable so the direct computation of $\mathbf{W}_{nm,\mathrm{o}}$ seems difficult. But using the relation $\mathbf{x}_n(k) = \mathbf{y}_n(k) - \mathbf{v}_n(k)$ and the fact that noise and speech are uncorrelated, we can verify that

$$\mathbf{R}_{x_nx_m} = \mathbf{R}_{y_ny_m} - \mathbf{R}_{v_nv_m}, \quad (23)$$

$$\mathbf{R}_{x_mx_m} = \mathbf{R}_{y_my_m} - \mathbf{R}_{v_mv_m}, \quad (24)$$

where $\mathbf{R}_{y_ny_m}$ and $\mathbf{R}_{v_nv_m}$ are defined similarly to $\mathbf{R}_{x_nx_m}$, and $\mathbf{R}_{y_my_m}$ and $\mathbf{R}_{v_mv_m}$ are defined similarly to $\mathbf{R}_{x_mx_m}$. As a result

$$\mathbf{W}_{nm,\mathrm{o}} = \left(\mathbf{R}_{y_ny_m} - \mathbf{R}_{v_nv_m}\right)\left(\mathbf{R}_{y_my_m} - \mathbf{R}_{v_mv_m}\right)^{-1}. \quad (25)$$

Now the optimal filter matrix depends only on the second-order statistics of the noisy and noise signals. The statistics of the noisy signals can be directly computed from the observed signals. If we assume that the noise is stationary or at least slowly-varying so that its characteristics stay the same from a silence period [i.e., when $x_n(k) = 0$] to the following period when speech is active and with the help of a VAD, the noise characteristics can be estimated during silence periods.

Using either (22) or (25), we can obtain an optimal estimate of the $\mathbf{W}_m$ matrix, i.e., $\mathbf{W}_{m,\mathrm{o}}$. Substituting $\mathbf{W}_{m,\mathrm{o}}$ into (20), the optimal filter $\mathbf{h}_{m,\mathrm{o}}$ can be rewritten as:

$$\mathbf{h}_{m,\mathrm{o}} = \mathbf{R}_{vv}^{-1}\mathbf{W}_{m,\mathrm{o}}^T \left[\mathbf{W}_{m,\mathrm{o}}\mathbf{R}_{vv}^{-1}\mathbf{W}_{m,\mathrm{o}}^T\right]^{-1}\mathbf{u}_1. \quad (26)$$

If $\mathbf{x}_n(k) = \mathbf{W}_{nm,\mathrm{o}}\mathbf{x}_m(k)$, applying $\mathbf{h}_{m,\mathrm{o}}$ to filter the observed signals can reduce noise without introducing any speech distortion. In practice, however, we do not have exactly $\mathbf{x}_n(k) = \mathbf{W}_{nm,\mathrm{o}}\mathbf{x}_m(k)$ so that some speech distortion is expected. But for long filters, we can approach this equality so that the distortion can be kept very low.

4. EXPERIMENTS

In this section we evaluate the performance of the developed multichannel noise-reduction algorithm in real acoustic environments. We set up a multiple-microphone system in the varechoic chamber at Bell Labs [which is a room that measures 6.7 m long by 6.1 m wide by 2.9 m high ($x \times y \times z$)]. A total of ten microphones are used and their locations are, respectively, at (2.437, 5.600, 1.400), (2.537, 5.600, 1.400), (2.637, 5.600, 1.400), (2.737, 5.600, 1.400), (2.837, 5.600, 1.400), (2.937, 5.600, 1.400), (3.037, 5.600, 1.400), (3.137, 5.600, 1.400), (3.237, 5.600, 1.400), and (3.337, 5.600, 1.400). To simulate a sound source, we place a loudspeaker at (1.337, 3.162, 1.600), playing back a speech signal prerecorded from a female speaker. To make the experiments repeatable, we first measured the acoustic channel impulse responses from the source to the ten microphones (each impulse response is first measured at 48 kHz and then downsampled to 8 kHz). These measured impulse responses are then treated as the true ones. During experiments, the microphone outputs are generated by convolving the source signal with the corresponding measured impulse responses. Noise is then added to the convolved results to control the (input) SNR level.

We choose the first microphone as the reference microphone. Substituting the optimal filter into (2) and set $m = 1$, we obtain the optimal speech estimate as

$$\hat{x}_1(k) = \sum_{n=1}^{N}\mathbf{h}_{n1,\mathrm{o}}^T\mathbf{y}_n(k) = x_{1,\mathrm{nr}}(k) + v_{1,\mathrm{nr}}(k),$$

where $x_{1,\mathrm{nr}}(k) = \sum_{n=1}^{N}\mathbf{h}_{n1,\mathrm{o}}^T\mathbf{x}_n(k)$ and $v_{1,\mathrm{nr}}(k) = \sum_{n=1}^{N}\mathbf{h}_{n1,\mathrm{o}}^T\mathbf{v}_n(k)$ are, respectively, the speech filtered by the optimal filter and the residual noise. To assess the performance, we evaluate two criteria, namely the output SNR and the Itakura-Saito (IS) distance. The output SNR is defined as

$$\mathrm{SNR}_{\mathrm{o}} = \frac{E\left[x_{1,\mathrm{nr}}^2(k)\right]}{E\left[v_{1,\mathrm{nr}}^2(k)\right]}.$$

This measurement, when compared with the input SNR, tells us how much noise is reduced. The IS distance is a speech-distortion measure. For a detailed description of the IS distance, we refer to [8]. Many studies have shown that the IS measure is highly correlated with subjective quality judgements and two speech signals would be perceptually nearly identical if the IS distance between them is less than 0.1. In this experiment, we compute the IS distance between $x_1(k)$ and $x_{1,\mathrm{nr}}(k)$, which measures the degree of speech distortion due to the optimal filter.

In order to estimate and use the optimal filter given in (20), we need to specify the filter length L. If there is no reverberation, it is relatively easy to determine L, i.e., it needs only to be long enough to cover the maximal TDOA between the reference and the other microphones. In presence of reverberation, however, the determination of L would become more difficult and its value should, in theory, depend on the reverberation condition. Generally speaking, a longer filter has to be used if the environment is more reverberant. This experiment investigates the impact of the filter length on the algorithm performance. To eliminate the effect due to noise estimation, here we assume that the statistics of the noise signals are known *a priori*. The input SNR is 10 dB and the reverberation condition is controlled such that the reverberation time T_{60} is approximately 240 ms. The results are plotted in Fig. 2. One can see from Fig. 2(a) that the output SNR increases with L. So, the longer is the filter, the more the noise is reduced. Compared with $\mathrm{SNR}_{\mathrm{o}}$, the IS distance

decreases with L. This is understandable. As L increases, we will get a better spatio-temporal prediction of $\mathbf{x}_1(k)$ from $\mathbf{x}_n(k)$. Consequently, the algorithm achieves more noise reduction and meanwhile causes less speech distortion. We also see from Fig. 2 that the output SNR increases almost linearly with L. Unlike the SNR curve, the relationship between the IS distance and the filter length L is not linear. Instead, the curve first decreases quickly as the filter length increases, and then continues to decrease but with a slower rate. After $L = 250$, continuing to increase L does not seem to further decrease the IS distance. So, from speech-distortion point of view, $L = 250$ is long enough for reasonably good performance.

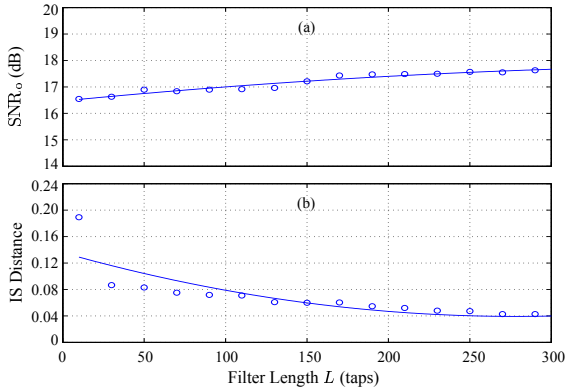

Fig. 2. The output SNR and IS distance, both as a function of the filter length L: (a) SNR$_o$ and (b) IS distance. The source is a speech signal from a female speaker; the background noise at each microphone is a computer-generated white Gaussian process; input SNR $= 10$ dB; and $T_{60} = 240$ ms. The fitting curve is a second-order polynomial.

The second experiment is to test the robustness of the multichannel algorithm to reverberation. The parameters used are: $L = 250$, $N = 10$, and input SNR $= 10$ dB. Compared with the previous experiments, this one does not assume to know the noise statistics. Instead, we developed a short-term energy based VAD to distinguish speech-plus-noise from noise-only segments. The noise covariance matrix is then computed from the noise-only segments using a batch method and the optimal filter is subsequently estimated according to (26). We tested the algorithm in two noise conditions: computed generated white Gaussian noise and a noise signal recorded in a New York Stock Exchange (NYSE) room. The results are depicted in Fig. 3. We see that the output SNR in both situations does not vary much when reverberation time is changed. This indeed demonstrates that the developed multichannel algorithm is very robust to reverberation. In comparison with the output SNR, we see that the IS distance grows with the reverberation time. This result should not come as a surprise. As the reverberation time T_{60} increases, it becomes more difficult to predict the speech observed at one microphone from that received at another microphone. As a result, more speech distortion is unavoidable but it is still perceptually almost negligible.

5. CONCLUSIONS

The existing multichannel noise-reduction techniques, such as beamforming, formulate the problem as one of recovering the desired source signal from the outputs of an array of microphones. Since

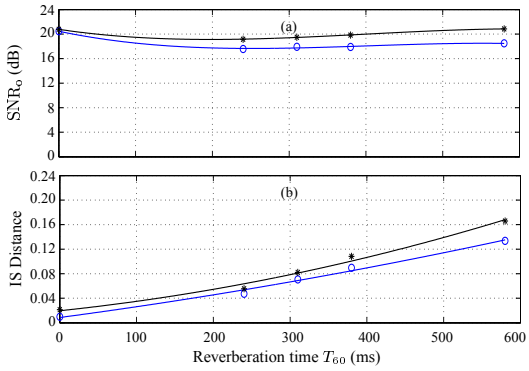

Fig. 3. Noise-reduction performance versus T_{60}. $*$: in white Gaussian noise; $\circ$: in NYSE noise; $L = 250$; input SNR $= 10$ dB. The fitting curve is a second-order polynomial.

they have to deal with both speech dereverberation and noise reduction at the same time, these techniques in general lack robustness and often produce very limited performance in real environments. In addition, they normally require information such as the room impulse responses, which is difficult to acquire reliably in practice. To overcome the drawbacks of the existing techniques, this paper reformulated the multichannel noise-reduction problem. Instead of estimating the desired source signal, this formulation puts aside the speech-dereverberation part and formulated the problem as one of estimating the speech component received at one of the multiple microphones. We then developed an MMSE estimator based on spatio-temporal prediction. Experiments demonstrated that the developed technique can achieve significant noise reduction and the resulting speech distortion is perceptually almost negligible. Compared with the traditional beamforming techniques, the developed algorithm has many appealing properties: it does not require the knowledge of the source location or the channel impulse responses; the multiple microphones do not have to be arranged into a specific array geometry; it works the same for both the far-field and near-field cases; and it can produce very good and robust noise reduction with minimum speech distortion in real acoustic environments.

6. REFERENCES

[1] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

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

[3] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive Signal Processing: Applications to Real-World Problems*, J. Benesty and Y. Huang, eds., Berlin, Germany: Springer, 2003.

[4] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 1053–1065, Mar. 2007.

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.*, vol. 49, pp. 1614–1626, Aug. 2001.

[6] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Process.*, vol. 3, pp. 251–266, July 1995.

[7] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 1218–1234, July 2006.

[8] F. Itakura and S. Saito, "A statistical method for esimation fo speech spectral density and formant frequencies," *Electron. Commun. Japan*, vol. 53A, pp. 36–43, 1970.