ON THE COMBINATION OF VOICE PROMPT SUPPRESSION WITH MAXIMUM KURTOSIS BEAMFORMING

John McDonough, Bhiksha Raj

Carnegie Mellon University
Gates-Hillman Complex, 5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{johnmcd,bhiksha}@cs.cmu.edu

Kenichi Kumatani

Disney Research, Pittsburgh
4720 Forbes Avenue, Lower Level, Suite 110
Pittsburgh, PA 15213, USA
kenichi.kumatani@disneyresearch.com

ABSTRACT

In earlier work, we proposed a *voice prompt suppression* (VPS) algorithm based on a Kalman filter, in which the *temporal update* or *correction step* is performed in *information space*. The advantage of this approach is that the *information matrix* can be diagonally loaded in order to control the magnitude of the subband filter coefficients, which provides for better robustness. In this work, we extend that earlier work by proposing a square root implementation of the information filter VPS algorithm, as well as a technique for diagonally loading the Cholesky factor of the error covariance matrix used in this implementation. We also investigate the effectiveness of cascading VPS after *maximum kurtosis* beamforming, which has been shown to provide performance superior to all conventional beamforming techniques. In a set of distant speech recognition experiments we demonstrate that VPS can reduce word error rate from 19.9% to 16.1% for an adult speaker, and from 44.4% to 40.0% for a child.

Index Terms— acoustic echo cancellation, speech recognition, beamforming, information filter

1. INTRODUCTION

Modern speech enabled applications provide for dialog between a machine and one or more human users. The machine prompts the user with queries that are either prerecorded or synthesized on the fly. The human users respond with their own voices, and their speech is then recognized and understood by a human language understanding module. In order to achieve as natural an interaction as possible, the human user(s) must be allowed to interrupt the machine during a voice prompt. This implies that the recognition engine must be running even during the voice prompt; hence, the capacity to suppress the voice prompt in the signals captured by one or more far-field microphones is essential. The task of *voice prompt suppression* (VPS) is similar to that of *acoustic echo cancellation* (AEC).

In McDonough *et al.* [1], we proposed a VPS algorithm based on a Kalman filter, in which the *temporal update* or *correction* is performed in *information space*. The advantage of this approach is that the *information matrix* can be diagonally loaded in order to control the magnitude of the subband filter coefficients, which provides for better robustness. In this work, we extend that earlier work by proposing a square root implementation of the information filter VPS algorithm. It is well known that square root implementations—both of covariance and information forms of the Kalman filter—effectively double the numerical precision of the respective algorithms [2, §6.3–6.4]. Moreover, the square root implementations elminate the *explosive divergence* phenomenon to which the direct form implementations are prone [3, §11]. The latter occurs when the error covariance matrices—which by definition must be *positive definite*—that are propagated forward in time during the state estimate update inhherent in a Kalman filter can become indefinite due to finite precision effects. As the Cholesky factorization exists solely for positive definite matrices, stipulating that the Cholesky decomposition exists is equivalent to requiring that the error covariance matrix remains positive definite.

Kumatani [4] proposes an adaptive beamforming algorithm based on a *maximum kurtosis* (MK) optimization criterion, and demonstrates that this algorithm provides performance superior to the conventional *super-directive* (SD) beamformer [5, §13.3.4] in a series of *distant speech recognition* (DSR) experiments. Here, we investigate the effectiveness of cascading VPS after MK beamforming. It is also possible to perform VPS on the signal from each sensor in a microphone array *prior* to beamforming [6, §13.4.4], although we do not investigate that approach in the present work.

The balance of this work is organized as follows. In Section 2, we review the conventional covariance Kalman filter techniques for *voice prompt suppression* (VPS). We also present the VPS algorithm based on the information Kalman filter proposed here and discuss its similarities and differences with the covariance form of the filter. We then present a square root implementation of the information filter, as well as a novel technique for performing diagonal loading on the Cholesky factor of the information covariance matrix. Finally, our initial DSR results comparing the performance of beamforming when followed by both the direct and square root forms of the information filter are presented and discussed in Section 3.

2. THE INFORMATION FILTER

In this section, we describe the components of a VPS system. We then briefly present the operational details of the covariance form of the Kalman filter, as well as those of both the direct and square root implementations of the information filter.

2.1. Voice Prompt Suppression

Let us define the following components of our voice prompt suppression system:

- $V(z)$ denotes the transform of the known voice prompt;
- $S(z)$ denotes the transform of the unknown desired speech;

- $R(z) \triangleq \sum_{n=0}^{L-1} r[n]z^{-n}$ denotes the transform of FIR filter simulating the room impulse response;
- $G(z)$ is the transform of the actual, unknown *room impulse response* (RIR) for the voice prompt;
- $H(z)$ is the transform of the actual, unknown RIR for the desired speech;
- $A(z) \triangleq G(z)V(z) + H(z)S(z)$ is the combined signal reaching a single channel of the microphone array;
- $E(z) \triangleq A(z) - R(z)V(z)$ is the residual signal after removal of the voice prompt.

We will assume that the desired speech is a stochastic zero-mean process such that $\mathcal{E}\{S(e^{j\omega})\} = 0$, for all $-\pi \leq \omega \leq \pi$. The voice prompt $V(z)$, on the other hand, is assumed to be a *known* signal. The complete system for voice prompt suppression is shown schematically in Figure 1.

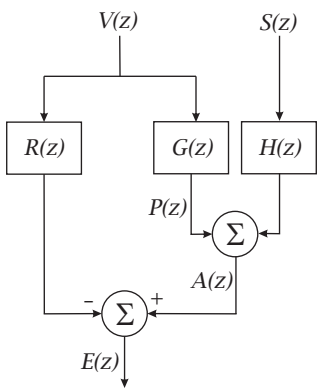

Figure 1: Block diagram for voice prompt suppression.

Based on the definitions above, we can formulate the VPS problem as one of minimizing the spectral energy of $E(z) = A(z) - V(z)\,R(z)$ whenever only the voice prompt is active. Posed as such, the VPS problem is tantamount to determining the *minimum mean square error* (MMSE) solution for the subband filter coefficients $R(z)$. It is well known that the MMSE solution is equivalent to the estimate of the *conditional mean* provided by the Kalman filter [5, §4.1]. Moreover, the Kalman filter formulation is attractive because it subsumes the *recursive least squares* (RLS) formulation, and—as we discuss in the next section—provides for finer control of the evolution of the subband filter coefficients through the inclusion of both a transition matrix and a process noise.

2.2. Kalman Filter Formulation

Given the *state* $\mathbf{x}_k$ and observation $\mathbf{y}_k$ at time k, the *state model* of the Kalman filter can be expressed as

$$\mathbf{x}_k = \mathbf{F}_{k|k-1}\,\mathbf{x}_{k-1} + \mathbf{u}_{k-1}, \tag{1}$$

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k, \tag{2}$$

where $\mathbf{F}_{k|k-1}$ and $\mathbf{H}_k$ are the known *transition* and *observation matrices*. The noise terms $\mathbf{u}_{k-1}$ and $\mathbf{v}_k$ in (1–2) are by assumption zero mean, white Gaussian random vector processes with covariance matrices $\mathbf{U}_k \triangleq \mathcal{E}\{\mathbf{u}_k\mathbf{u}_k^H\}$ and $\mathbf{V}_k = \mathcal{E}\{\mathbf{v}_k\mathbf{v}_k^H\}$, respectively. Moreover, by assumption $\mathbf{u}_k$ and $\mathbf{v}_k$ are statistically independent.

Let $\hat{\mathbf{x}}_{k|k-1}$ denote the *predicted state estimate* at time k using all observations up to time $k-1$. Moreover, let $\mathbf{y}_{1:k-1}$ denote all past observations up to time $k-1$, and let $\hat{\mathbf{y}}_{k|k-1}$ denote the MMSE estimate of the next observation $\mathbf{y}_k$ given all prior observations, such that, $\hat{\mathbf{y}}_{k|k-1} = \mathcal{E}\{\mathbf{y}_k|\mathbf{y}_{1:k-1}\}$. By definition, the *innovation* is the difference

$$\mathbf{s}_k \triangleq \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1} \tag{3}$$

between the actual and the predicted observations. This quantity is given the name innovation, because it contains all the "new information" required for sequentially updating the filtering density $p(\mathbf{x}_{0:k}|\mathbf{y}_{1:k-1})$; i.e., the innovation contains that information about the time evolution of the system that cannot be predicted from the state space model (1–2).

Let us begin our exposition of the Kalman filter by stating how the predicted observation may be calculated based on the current state estimate, according to

$$\hat{\mathbf{y}}_{k|k-1} = \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}. \tag{4}$$

In light of (3) and (4), we may write

$$\mathbf{s}_k = \mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}. \tag{5}$$

Substituting (2) into (5), we find

$$\mathbf{s}_k = \mathbf{H}_k\boldsymbol{\epsilon}_{k|k-1} + \mathbf{v}_k, \tag{6}$$

where $\boldsymbol{\epsilon}_{k|k-1} \triangleq \mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}$ is the *predicted state estimation error* at time k, using all data up to time $k-1$. It can be readily shown that $\boldsymbol{\epsilon}_{k|k-1}$ is orthogonal to $\mathbf{u}_k$ and $\mathbf{v}_k$ [3, §10.1]. Using (6) and exploiting the statistical independence of $\mathbf{u}_k$ and $\mathbf{v}_k$, the covariance matrix of the innovations sequence can be expressed as

$$\mathbf{S}_k \triangleq \mathcal{E}\left\{\mathbf{s}_k\mathbf{s}_k^H\right\} = \mathbf{H}_k\mathbf{K}_{k|k-1}\mathbf{H}_k^H + \mathbf{V}_k, \tag{7}$$

where the *predicted state estimation error covariance matrix* is defined as

$$\mathbf{K}_{k|k-1} \triangleq \mathcal{E}\left\{\boldsymbol{\epsilon}_{k|k-1}\boldsymbol{\epsilon}_{k|k-1}^H\right\}. \tag{8}$$

The sequential update of the Kalman filter can be partitioned into two steps:

- First, there is a *prediction*, which can be expressed as

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_{k|k-1}\hat{\mathbf{x}}_{k-1|k-1}. \tag{9}$$

  Clearly the prediction is so-called because it is made without the advantage of any information derived from the current observation $\mathbf{y}_k$.
- The latter information is instead folded into the current estimate through the *update* or *correction*, according to

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{G}_k\mathbf{s}_k, \tag{10}$$

  where the *Kalman gain* is defined as

$$\mathbf{G}_k \triangleq \mathcal{E}\{\mathbf{x}_k\mathbf{s}_k^H\}\mathbf{S}_k^{-1}, \tag{11}$$

$\mathbf{s}_k$ and $\mathbf{S}_k$ are given by (5), and (7), respectively, and $\hat{\mathbf{x}}_{k|k}$ denotes the *filtered state estimate* using all observations $\mathbf{y}_{1:k}$. Note that (10) is of paramount importance, as it shows how the MMSE or Bayesian state estimate can be recursively updated. To wit, it is only necessary to premultiply the prior estimate $\hat{\mathbf{x}}_{k|k-1}$ by the transition matrix $\mathbf{F}_{k|k-1}$, then to add a correction factor consisting of

the Kalman gain $\mathbf{G}_k$ multiplied by the innovation $\mathbf{s}_k$. Hence, the entire problem of recursive MMSE estimation under the assumptions of linearity and Gaussianity reduces to the calculation of the Kalman gain (11), whereupon the state estimate can be updated according to (10). From (9) and (10), we deduce that the KF has the predictor-corrector structure shown in Figure 2. The Kalman

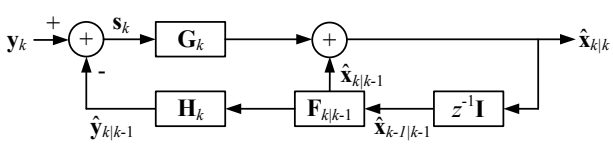

Figure 2: Predictor-corrector structure of the Kalman filter.

gain (11) can be efficiently calculated according to

$$\mathbf{G}_k = \mathbf{K}_{k|k-1}\mathbf{H}_k^H\mathbf{S}_k^{-1}, \tag{12}$$

where the covariance matrix $\mathbf{S}_k$ of the innovations sequence is defined in (7). The *Riccati equation* then specifies how $\mathbf{K}_{k|k-1}$ can be sequentially updated, namely as,

$$\mathbf{K}_{k|k-1} = \mathbf{F}_{k|k-1}\,\mathbf{K}_{k-1}\,\mathbf{F}_{k|k-1}^H + \mathbf{U}_{k-1}. \tag{13}$$

The matrix $\mathbf{K}_k$ in (13) is, in turn, obtained through the recursion,

$$\mathbf{K}_k = (\mathbf{I} - \mathbf{G}_k\mathbf{H}_k)\mathbf{K}_{k|k-1}. \tag{14}$$

This matrix $\mathbf{K}_k$ can be interpreted as the covariance matrix of the *filtered state estimation error* [3, §10], such that, $\mathbf{K}_k \triangleq \left\{\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^H\right\}$, where $\boldsymbol{\epsilon}_k \triangleq \mathbf{x}_k - \hat{\mathbf{x}}_{k|k}$. Note the critical difference between $\boldsymbol{\epsilon}_{k|k-1}$ and $\boldsymbol{\epsilon}_k$, namely, $\boldsymbol{\epsilon}_{k|k-1}$ is the error in the state estimate made *without* knowledge of the current observation $\mathbf{y}_k$, while $\boldsymbol{\epsilon}_k$ is the error in the state estimate made *with* knowledge of $\mathbf{y}_k$.

In order to formulate the voice prompt suppression system as a Kalman filter, we associate the state $\mathbf{x}_k$ with the subband coefficients of $R(z)$, and the observation matrix (vector) $\mathbf{H}_k$ with the current and delayed subband samples of $V(z)$. Moreover, the (scalar) observation $\mathbf{y}_k$ is associated with the signal $A(z)$ arriving at the microphone. The scalar observation noise $\mathbf{v}_k$ is associated with the term $H(z)\,S(z)$, the variance of which can be estimated with a sliding exponential window

$$\hat{\sigma}_{u,m}^2(k) = (1-\lambda)\hat{\sigma}_{u,m}^2(k-1) - \lambda\,|E_m(k)|^2, \tag{15}$$

where $0 < \lambda < 1$ is a *forgetting factor* that controls how quickly past observations are discounted. Note that the update in (15) should be performed exclusively when the voice prompt is *not* active. The innovation $\mathbf{s}_k$ is then associated with the error term $E(z)$. The transition matrix $\mathbf{F}_{k|k-1}$, the covariance matrix $\mathbf{U}_{k-1}$ of the process noise, and the initial value assigned to the error covariance matrix $\mathbf{K}_k$ can be treated as system parameters to be tuned for optimal performance.

2.3. Information Filter Formulation

The *Fisher information matrix* and *information vector* are defined as

$$\mathbf{Z}_k \triangleq \mathbf{K}_k^{-1}, \tag{16}$$

$$\hat{\mathbf{d}}_{k|k-1} \triangleq \mathbf{Z}_{k|k-1}\hat{\mathbf{x}}_{k|k-1}, \tag{17}$$

respectively. The *temporal update* or *prediction* in information space is performed according to [7, §6.8]

$$\mathbf{A}_k \triangleq \mathbf{F}_k^{-H}\mathbf{Z}_{k-1}\mathbf{F}_k^{-1}, \tag{18}$$

$$\mathbf{Z}_{k|k-1} = \left[\mathbf{I} - \mathbf{A}_k\left(\mathbf{A}_k + \mathbf{U}_{k-1}^{-1}\right)^{-1}\right]\mathbf{A}_k, \tag{19}$$

$$\hat{\mathbf{d}}_{k|k-1} = \left[\mathbf{I} - \mathbf{A}_k\left(\mathbf{A}_k + \mathbf{U}_{k-1}^{-1}\right)^{-1}\right]\mathbf{F}_k^{-H}\hat{\mathbf{d}}_{k-1|k-1}. \tag{20}$$

Alternatively, it is possible to simply calculate $\mathbf{K}_k = \mathbf{Z}_k^{-1}$, perform the temporal update as in (13), then invert $\mathbf{K}_{k|k-1}$; we chose to implement our direct form information filter using this latter more "naive" approach, taking care to calculate numerically well-conditioned inverses. The *observational update* or *correction* in the information filter is then performed according to [7, §6.8]

$$\mathbf{Z}_k = \mathbf{Z}_{k|k-1} + \mathbf{H}_k^H\,\mathbf{V}_k^{-1}\,\mathbf{H}_k, \tag{21}$$

$$\hat{\mathbf{d}}_{k|k} = \hat{\mathbf{d}}_{k|k-1} + \mathbf{H}_k^H\,\mathbf{V}_k^{-1}\,\mathbf{y}_k. \tag{22}$$

Once $\mathbf{Z}_k$ has been calculated from (21) it can be diagonally loaded according to $\mathbf{Z}_k' = \mathbf{Z}_k + \sigma_{\mathrm{D}}^2\mathbf{I}$, whereupon the updated state vector (i.e., the subband filter coefficients) can be calculated according to

$$\mathbf{K}_k = \left(\mathbf{Z}_k'\right)^{-1}, \tag{23}$$

$$\hat{\mathbf{x}}_{k|k} = \mathbf{K}_k\,\hat{\mathbf{d}}_{k|k}. \tag{24}$$

At this point, all is ready for the next time step. The larger the diagonal loading term σ_{D}^2, the smaller the final subband filter coefficients, which is apparent from (23–24). The magnitude σ_{D}^2 of the diagonal loading can be treated as yet another system parameter to be tuned for optimimum performance.

2.4. Square Root Information Filter Formulation

In the square-root implementation, it is not $\mathbf{Z}_k$ that is propagated forward in time, but rather the lower triangular *Cholesky factor* $\mathbf{Z}_k^{1/2}$, which achieves [7, §6.8]

$$\mathbf{Z}_k = \mathbf{Z}_k^{1/2}\mathbf{Z}_k^{H/2}. \tag{25}$$

Moreover, the information vector (17) is replaced with the *square root information state*

$$\mathbf{z}_k \triangleq \mathbf{Z}^{H/2}\mathbf{x}_k. \tag{26}$$

Prediction then proceeds by calculating an orthogonal matrix $\boldsymbol{\Theta}_{\text{pred}}$ that imposes a lower triangular structure on the *prearray* such that

$$\begin{bmatrix} \mathbf{U}_{k-1}^{-1/2} & -\mathbf{F}_{k|k-1}^{-H}\mathbf{Z}_{k-1}^{1/2} \\ \mathbf{0} & \mathbf{F}_{k|k-1}^{-H}\mathbf{Z}_{k-1}^{1/2} \\ \mathbf{0} & \hat{\mathbf{z}}_{k-1|k-1}^H \end{bmatrix}\boldsymbol{\Theta}_{\text{pred}} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{B}_{21} & \mathbf{Z}_{k|k-1}^{1/2} \\ \mathbf{b}_{31}^H & \hat{\mathbf{z}}_{k|k-1}^H \end{bmatrix}, \tag{27}$$

where the terms $\mathbf{B}_{11}$, $\mathbf{B}_{12}$, and $\mathbf{b}_{31}$ need not be retained. Demonstrating that (27) is equivalent to (18–20) involves applying the *matrix factorization lemma* [3, §11.1] to the former, and then showing that

$$\mathbf{B}_{11} = \left(\mathbf{A}_k + \mathbf{U}_{k-1}^{-1}\right)^{1/2},$$

$$\mathbf{B}_{21} = -\mathbf{A}_k\left(\mathbf{A}_k + \mathbf{U}_{k-1}^{-1}\right)^{-H/2}, \text{ and}$$

$$\mathbf{b}_{31} = -\left(\mathbf{A}_k + \mathbf{U}_{k-1}^{-1}\right)^{-1/2}\mathbf{F}_{k|k-1}^{-1}\hat{\mathbf{d}}_{k-1|k-1}.$$

Correction is performed by calculating a second orthogonal matrix $\mathbf{\Theta}_{\text{corr}}$ achieving

$$\begin{bmatrix} \mathbf{Z}_{k|k-1}^{1/2} & \mathbf{H}_k^H \mathbf{V}_k^{-1/2} \\ \mathbf{z}_{k|k-1}^H & \mathbf{y}_k^H \mathbf{V}_k^{-1/2} \end{bmatrix} \mathbf{\Theta}_{\text{corr}} = \begin{bmatrix} \mathbf{Z}_k^{1/2} & \mathbf{0} \\ \mathbf{z}_{k|k}^H & \beta \end{bmatrix}, \tag{28}$$

where the scalar β is not required. Diagonal loading is performed in square root information space by calculating yet another orthogonal matrix $\mathbf{\Theta}_{\text{diag}}$ that achieves

$$\begin{bmatrix} \mathbf{Z}_k^{1/2} & \sigma_{\text{D}}\, \mathbf{e}_n \end{bmatrix} \mathbf{\Theta}_{\text{diag}} = \begin{bmatrix} (\mathbf{Z}_k')^{1/2} & \mathbf{0} \end{bmatrix}, \tag{29}$$

where $\mathbf{e}_n$ is the nth unit vector. This loading (29) must be repeated for each diagonal element of $\mathbf{Z}_k^{1/2}$. Performing diagonal loading in this manner was described in [5, §13.4.4] for RLS beamforming, but—to the authors' knowledge—this is the first instance wherein this algorithm has been proposed for subband adaptive filtering or voice prompt suppression. In our experiments, we constructed $\mathbf{\Theta}_{\text{pred}}$, $\mathbf{\Theta}_{\text{corr}}$, and $\mathbf{\Theta}_{\text{diag}}$ from series of Givens rotations, but Householder transformations could be used as well [7, §6.8].

Once $\mathbf{z}_{k|k}^H$ has been determined, the new filter coefficients can be calculated through forward substitution on $\hat{\mathbf{z}}_{k|k} = \mathbf{Z}^{H/2}\hat{\mathbf{x}}_{k|k}$.

3. DISTANT SPEECH RECOGNITION EXPERIMENTS

The data collection scenario used for the DSR experiments described here was a simple listen-and-repeat task known as *Copycat*, in which children were shown an illustration of an object and asked to repeat the referring phrase spoken by the experimenter (e.g., "I want the dragon's tail," or "Give her the crown"). To obtain a large number of segments of high overlap between a voice prompt and speech of the subjects, the former was artificially mixed with the latter after capture with far-field microphones. All far-field data capture was conducted with a 64 channel linear microphone array with an intersensor spacing of 2 cm. Further details of the sensor configuration used to capture the far-field data are given in [8].

Our basic DSR system was trained on three corpora of children's speech as described in [8]. The conventional model had 1,200 states and a total of 25,702 Gaussian components. Conventional training was followed by *speaker-adapted training* (SAT) as described in [5, §8.1.3]. Details of the front end used for feature extraction in our system are given in [8].

Our experiments involved three passes of speech recognition:

1. Recognize with the unadapted conventionally-trained model;
2. Estimate *vocal tract length normalization* (VTLN) [5, §9.1.1], *maximum likelihood linear regression* (MLLR) [5, §9.2.1] and *constrained maximum likelihood linear regression* (CMLLR) [5, §9.1.2] parameters, then recognize once more with the adapted conventially-trained model;
3. Estimate VTLN, MLLR and CMLLR parameters for the SAT model, then recognize with same.

For all but the first unadapted pass, unsupervised speaker adaptation was performed on word lattices from the previous pass.

The test set consisted of four sessions of the Copycat scenario. There were a total of 354 utterances and 1,297 words spoken by the children subjects. A total of 356 utterances and 1,305 words were spoken by the experimenter.

For the experiments described below, the variable system parameters were set as $\mathbf{F} = \mathbf{I}$, $\sigma_{\text{D}}^2 = 10^{-4}$, $\mathbf{U}_k = 10^{-4} \cdot \mathbf{I}$, and $\mathbf{K}_0 = 5 \cdot \mathbf{I}$. In particular, we choose the diagonal elements of $\mathbf{K}_0$ to be much larger than those of $\mathbf{U}_k$ so that the filter coefficients would converge rapidly at the start of the utterances from a given speaker, but would not oscillate once they had reached a more or less steady state.

The results of our DSR experiments—reported in terms of *word error rate* (WER)—using both implementations of the information filter are presented in Table 1; for these comparisons, VPS was performed *after* beamforming. Also shown in the table are the results obtained with MK beamforming without VPS. These results reveal

		Pass					
		1		2		3	
Type	L_{filt}	Exp.	Child	Exp.	Child	Exp.	Child
None		37.7	60.0	19.9	44.4	18.4	41.9
Direct	4	30.5	57.3	15.9	41.2	16.9	41.7
	8	31.2	57.6	17.1	42.3	16.7	43.0
	16	31.1	58.1	17.5	40.0	17.1	43.2
S/R	4	30.3	55.9	16.3	40.9	15.2	44.0
	8	31.0	56.7	16.1	40.0	16.7	42.3
	16	31.4	56.8	17.3	40.9	17.3	43.4

Table 1: Word error rates (WERs) for several subband filter lengths using both the direct form and square root (S/R) implementation of VPS after MK beamforming; also shown are comparable results with no VPS.

that both implementations of VPS provide reductions in error rate, although the square root implementation is arguably more consistent. In particular, applying VPS based on the square root implementation with a filter length of $L_{\text{filt}} = 8$ reduces WER from 19.9% to 16.1% for the adult experimenter, and from 44.4% to 40.0% for the children subjects after the second pass. Oddly enough, the third pass of recognition is largely ineffective in reducing WER with respect to the second pass when VPS is applied.

4. REFERENCES

[1] J. McDonough, K. Kumatani, B. Raj, and J. F. Lehman, "An information filter for voice prompt suppression," in *Proc. Asilomar*, submitted for publication, 2011.

[2] D. Simon, *Optimal State Estimation: Kalman, H_∞, and Nonlinear Aproaches*. New York: Wiley, 2006.

[3] S. Haykin, *Adaptive Filter Theory*, 4th ed. New York: Prentice Hall, 2002.

[4] K. Kumatani, J. McDonough, and B. Raj, "Block-wise incremental adaptation algorithm for maximum kurtosis beamforming," in *Proc. WASPAA*, submitted for publication, 2011.

[5] M. Wölfel and J. McDonough, *Distant Speech Recognition*. London: Wiley, 2009.

[6] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Heidelberg, Germany: Springer, 2000, pp. 281–306.

[7] M. S. Grewal and A. P. Andrews, *Kalman Filtering: Theory and Practice*. Upper Saddle River, NJ: Prentice Hall, 1993.

[8] J. McDonough, K. Kumatani, B. Raj, and J. F. Lehman, "A mutual information criterion for voice activity detection," in *Proc. Interspeech*, submitted for publication, 2011.