An Information Filter for Voice Prompt Suppression

John McDonough,[1] *Wei Chu,*[2] *Kenichi Kumatani,*[3] *Bhiksha Raj,*[1] *Jill Fain Lehman*[3]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Dept. of Electrical Engineering, University of California, Los Angeles, Los Angeles, CA 90024, USA
[3]Disney Research, Pittsburgh, Pittsburgh, PA 15213, USA

johnmcd@cs.cmu.edu, weichu@ee.ucla.edu, kenichi.kumatani@disneyresearch.com,
bhiksha@cs.cmu.edu, jill.lehman@disneyresearch.com

Abstract

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. In this work, we compare two techniques for such voice prompt suppression. The first is a straightforward adaptation of a conventional *Kalman filter*, which has certain advantages over the *normalized least squares algrithm* in terms of robustness and speed of convergence. The second algorithm, which is novel in this work, is also based on a Kalman filter, but differs from the first in that the update or correction step is performed in *information space* and hence allows for the use of *diagonal loading* in order to control the growth of the subband filter coefficients, and thereby add robustness to the VPS.

Index Terms: acoustic echo cancellation, speech recognition

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). Most algorithms for AEC proposed in the literature are based on the *normalized least mean squares* (NLMS) algorithm developed in the field of adaptive filtering; see [1, 2, 3, 4], for example. The first algorithm investigated here is a straightforward adaptation of a conventional Kalman filter, which has certain advantages over the NLMS algorithm in terms of robustness and speed of convergence; this algorithm is similar to that described in [5]. The second algorithm, which is novel in this work, is also based on a Kalman filter, but differs from the first in that the 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.

As the adaptive filter tends to diverge when speech from the desired speaker is present, a *double-talk detector* (DTD) is needed to halt the adaptation of filter coefficients during segments containing double-talk [6]; i.e., when both the voice prompt and desired speaker are active. Jia et al. combined local decisions of double-talk detectors on subbands to make a global decision of the presence of the near-end speaker [7]. In this paper, we also propose a *subband double-talk detection algorithm* in which the filter for a subband is only updated when the subband speaker-to-voice prompt energy ratio is sufficiently high. The proposed subband DTD is shown to be effective in increasing the rate of convergence of the subband filters and hence in improving the sound quality.

In Section 2, we review the conventional NLMS and 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. Our initial experimental results with the proposed technique are tabulated and described in Section 3. A set of *distant speech recognition* (DSR) experiments demonstrates that the information filter provides performance superior to that obtained with the conventional Kalman filter. In the final section of this work, we present our conclusions and a brief description of our plans for future research.

2. The Information Filter

In this section, we describe the components of a VPS system. We then briefly present the operational details of the standard NLMS algorithm, as well as those of both the conventional and information formulations of the Kalman filter; we discuss how all three algorithms can be used for VPS. Finally, we present a novel DTD algorithm.

2.1. Voice Prompt Suppression

Let us define the following components of our voice prompt cancellation application:

- $V(z)$ 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 cancellation is shown schematically in Figure 1. Based on the definitions above, we can express the average spectral energy as

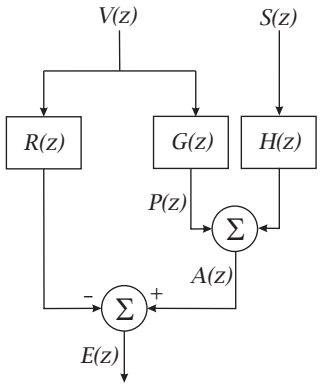


Figure 1: Block diagram for voice prompt cancellation.

$$P_E \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|E(e^{j\omega})\right|^2 d\omega, \tag{1}$$

where

$$E(z) = [R(z) - G(z)]V(z) + H(z)S(z). \tag{2}$$

Now let us partition $E(e^{j\omega})$ into discrete subband components such that

$$E_m \triangleq E(e^{j\omega_m}), \tag{3}$$

where $\omega_m \triangleq 2\pi m/M$ for all $m = 0, 1, \ldots, M-1$. Hence we can define the *expected* spectral energy as

$$P_E \approx \frac{1}{M}\sum_{m=0}^{M-1} I_m, \tag{4}$$

where $I_m \triangleq \mathcal{E}\{|E_m|^2\}$, or equivalently,

$$I_m \triangleq |R_m - G_m|^2|V_m|^2 + |H_m|^2\mathcal{E}\{|S_m|^2\}, \tag{5}$$

and we have defined discrete subband components for the quantities appearing on the right hand of (2) as in (3). Note that we have made use of $\mathcal{E}\{S(e^{j\omega})\} = 0$ in simplifying (5).

Clearly (4–5) can be minimized by minimizing each I_m individually with respect to the FIR subband coefficients R_m. Taking partial derivatives on both sides of (5) yields

$$\frac{\partial I_m}{\partial R_m^*} = (R_m - G_m)|V_m|^2. \tag{6}$$

In order to use (6) in a *normalized least mean square* (NLMS) update formulae [8, §7], we need to estimate G_m, which can be achieved as follows. Assuming that S_m is a zero-mean stochastic process, we have $\mathcal{E}\{A_m\} = G_m V_m$. Hence, we can make the instantaneous estimate

$$\hat{G}_m = \frac{A_m}{V_m} = G_m + \frac{S_m}{V_m}H_m. \tag{7}$$

From (7) it is apparent that the approximation will be better for segments wherein only the voice prompt is active, or for subbands where the voice prompt dominates the desired speech. Substituting $\hat{G}_m = A_m/V_m$ for G_m in (6) yields the update rule

$$R'_m = R_m - \eta_m \frac{\partial I_m}{\partial R_m^*} = R_m + \eta_m(V_m^* A_m - |V_m|^2 R_m), \tag{8}$$

where η_m is a step size determined with a NLMS strategy.

2.2. Kalman Filter Formulation

The *state model* of the Kalman filter can be expressed as

$$\mathbf{x}_k = \mathbf{F}_{k|k-1}\,\mathbf{x}_{k-1} + \mathbf{u}_{k-1}, \tag{9}$$

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k, \tag{10}$$

where $\mathbf{F}_{k|k-1}$ and $\mathbf{H}_k$ are the known *transition* and *observation* matrices. The noise terms $\mathbf{u}_k$ and $\mathbf{v}_k$ in (9–10) are by assumption zero mean, white Gaussian random vector processes with covariance matrices $\mathbf{U}_k \triangleq \mathcal{E}\{\mathbf{u}_k\mathbf{u}_k^T\}$ and $\mathbf{V}_k = \mathcal{E}\{\mathbf{v}_k\mathbf{v}_k^T\}$, respectively. Moreover, by assumption $\mathbf{u}_k$ and $\mathbf{v}_k$ are statistically independent.

Once more 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

$$\mathbf{s}_k \triangleq \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1} \tag{11}$$

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.

We begin by stating how the predicted observation may be calculated based on the current state estimate, according to,

$$\hat{\mathbf{y}}_{k|k-1} = \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}. \tag{12}$$

In light of (11) and (12), we may write

$$\mathbf{s}_k = \mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k|k-1}. \tag{13}$$

Substituting (10) into (13), we find

$$\mathbf{s}_k = \mathbf{H}_k\boldsymbol{\epsilon}_{k|k-1} + \mathbf{v}_k, \tag{14}$$

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 $\mathbf{v}_k$ [8, §10.1]. Using (14) and exploiting the statistical independence of $\mathbf{u}_k$ and $\mathbf{v}_k$, the covariance matrix of the innovations sequence can be expressed as

$$\mathbf{S}_k \triangleq \mathcal{E}\left\{\mathbf{s}_k\mathbf{s}_k^T\right\} = \mathbf{H}_k\mathbf{K}_{k|k-1}\mathbf{H}_k^T + \mathbf{V}_k, \tag{15}$$

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

$$\mathbf{K}_{k|k-1} \triangleq \mathcal{E}\left\{\boldsymbol{\epsilon}_{k|k-1}\boldsymbol{\epsilon}_{k|k-1}^T\right\}. \tag{16}$$

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

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

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_{k|k-1}\hat{\mathbf{x}}_{k-1|k-1}. \tag{17}$$

  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

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{G}_k\mathbf{s}_k, \tag{18}$$

  where the *Kalman gain* is defined as

$$\mathbf{G}_k \triangleq \mathcal{E}\{\mathbf{x}_k\mathbf{s}_k^T\}\mathbf{S}_k^{-1}, \tag{19}$$

for $\mathbf{x}_k$, $\mathbf{s}_k$, and $\mathbf{S}_k$ given by (9), (13), and (15), respectively. Note that (18) 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 (19), whereupon the state estimate can be updated according to (18). From (17) and (18), we deduce that the KF has the predictor-corrector

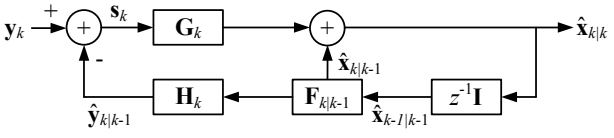


Figure 2: Predictor-corrector structure of the Kalman filter.

structure shown in Figure 2.

The Kalman gain (19) can be efficiently calculated according to

$$\mathbf{G}_k = \mathbf{K}_{k|k-1}\mathbf{H}_k^T\mathbf{S}_k^{-1}, \tag{20}$$

where the covariance matrix $\mathbf{S}_k$ of the innovations sequence is defined in (15). The *Riccati equation* then specifies how $\mathbf{K}_{k|k-1}$ can be sequentially updated, namely as,

$$\mathbf{K}_{k|k-1} = \mathbf{F}_{k|k-1}\,\mathbf{K}_{k-1}\,\mathbf{F}_{k|k-1}^T + \mathbf{U}_{k-1}. \tag{21}$$

The matrix $\mathbf{K}_k$ in (21) is, in turn, obtained through the recursion,

$$\mathbf{K}_k = (\mathbf{I} - \mathbf{G}_k\mathbf{H}_k)\mathbf{K}_{k|k-1}. \tag{22}$$

This matrix $\mathbf{K}_k$ can be interpreted as the covariance matrix of the *filtered state estimation error* [8, §10], such that, $\mathbf{K}_k \triangleq \{\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^T\}$, 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

$$\hat{\sigma}_{u,m}^2(k) = (1-\lambda)\hat{\sigma}_{u,m}^2(k-1) - \lambda\,|E_m(k)|^2, \tag{23}$$

where $0 < \lambda < 1$ is a *forgetting factor* that controls how quickly past observations are discounted. Note that the update in (23) 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 $\mathbf{F}_{k|k-1}$ can be assumed to be the identity matrix. Finally, the covariance matrix $\mathbf{U}_{k-1}$ of the process noise appearing in (21) can be treated as a system parameter to be tuned for optimal performance.

2.3. Information Filter Formulation

In the hybrid implementation of the information filter we have adopted for our initial studies, the prediction step is performed as in (17) and (21). At this point, however, we calculate the *Fisher information matrix* and *information vector* according to

$$\mathbf{Z}_{k|k-1} \triangleq \mathbf{K}_{k|k-1}^{-1}, \tag{24}$$

$$\hat{\mathbf{z}}_{k|k-1} \triangleq \mathbf{Z}_{k|k-1}\hat{\mathbf{x}}_{k|k-1}, \tag{25}$$

respectively. The update then occurs in information space, as given by,

$$\mathbf{Z}_k = \mathbf{Z}_{k|k-1} + \mathbf{H}_k^H\,\mathbf{V}_k^{-1}\,\mathbf{H}_k, \tag{26}$$

$$\hat{\mathbf{z}}_{k|k} = \hat{\mathbf{z}}_{k|k-1} + \mathbf{H}_k^H\,\mathbf{V}_k^{-1}\,\mathbf{y}_k. \tag{27}$$

Comparing the forms of the prediction and update, it is clear that in the covariance formulation of the Kalman filter, the prediction is trivial while the update is fairly involved, primarily due to the Riccati equation and the necessity of calculating the Kalman gain. In the information formulation, however, the update is fairly trivial. The prediction step, however is not straightforward, which is why we have chosen to implement the latter in the state space as indicated in (17) and (21), and the update in information space as specified by (26–27). Both direct form [9, §6.2] and square-root [10, §6.8] implementations of the information filter have appeared in the literature, however.

Once $\mathbf{Z}_k$ has been calculated from (26) 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

$$\mathbf{K}_k = \left(\mathbf{Z}'_{k|k}\right)^{-1}, \tag{28}$$

$$\hat{\mathbf{x}}_{k|k} = \mathbf{K}_k\,\hat{\mathbf{z}}_{k|k}. \tag{29}$$

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 (28–29).

2.4. Double Talk Detection

At frame k, the estimated subband *a priori speaker-to-voice prompt energy ratio* for the mth subband is calculated as

$$\hat{\xi}_k[m] = \frac{\|E_k[m]\|^2}{\|A_k[m] - E_k[m]\|^2}$$

where $E_k[m]$ denotes the estimated energy of the speaker's voice, and $A_k[m] - E_k[m]$ represents the energy of the estimated voice prompt. Let $\tilde{\hat{\xi}}_k[m]$ denote a "smoothed" version of $\hat{\xi}_k[m]$ defined as

$$\tilde{\hat{\xi}}_k[m] = (1-\eta)\tilde{\hat{\xi}}_{k-1}[m] + \eta\hat{\xi}_k[m]$$

where η denotes the smoothing factor. We used a value of $\eta = 0.05$ for our experiments. If $\tilde{\hat{\xi}}_k[m] > \xi^{th}$, then the adaptation of the filter coefficients is halted for the mth subband and kth frame.

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. Further details of the sensor configuration used to capture the far-field data are given in [11].

Our basic DSR system was trained on three corpora of children's speech:

1. the CMU Kids' Corpus, which contains 9.1 hours of speech from 76 speakers;
2. the Center for Speech and Language Understanding (CSLU) Kids' Corpus, which contains 4.9 hours of speech from 174 speakers.
3. A set of *Copycat* data collected at the Carnegie Mellon Childrens' School in June, 2010.

HMM training was conducted along the lines suggested in [12]. 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 [13, §8.1.3]. Details of the front end used for feature extraction in our system are given in [11].

Our experiments involved two passes of speech recognition:

1. Recognize with the unadapted conventionally-trained model;
2. Estimate *vocal tract length normalization* (VTLN) [13, §9.1.1], *maximum likelihood linear regression* (MLLR) [13, §9.2.1] and *constrained maximum likelihood linear regression* (CMLLR) [13, §9.1.2] parameters, then recognize once more with the adapted conventially-trained model;

Unsupervised speaker adaptation prior to the second was performed on the word lattices written during the first 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 child subjects. A total of 356 utterances and 1,305 words were spoken by the experimenter.

The results of the DSR experiments using the standard Kalman filter for VPS are given in Table 1. From these results, it is clear that increasing the length of the subband filter beyond one degrades performance. This is due to the uncontrolled growth of the filter coefficients which leads to poor robustness.

A comparable set of results obtained with the information filter with a diagonal loading level of 10^{-4} are presented in Table 2. From these results, it is clear that using a filter of length four yields the lowest WERs. This is due to the fact that the diagonal loading prevents the filter coefficients from growing uncontrollably, thereby improving robustness. Moreover, increasing the length of the subband FIR filters enables the VPS to model the long impulse responses of typical acoustic environments. The effect of VPS in the time and spectral domains is illustrated in Figure 3: Figure 3 a) shows the clean voice prompt; Figure 3 b) shows the speech of the desired speaker corrupted with the voice prompt, and Figure 3 c) shows the corrupted signal after VPS. It is clear that VPS has removed much of the distorting voice prompt.

Table 1: *Word error rates (WERs) for several subband filter lengths using the standard Kalman filter for VPS.*

	%WER	
Filter Length	Experimenter	Child
1	52.3	74.3
4	54.0	75.2
16	68.7	77.5

Table 2: *Word error rates (WERs) for several subband filter lengths using the information filter for VPS with a diagonal loading level of* 10^{-4}.

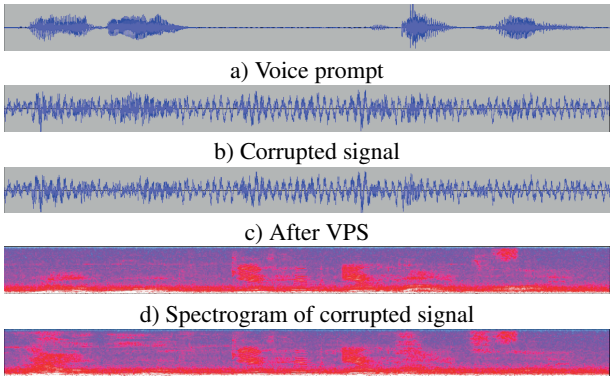
	%WER	
Filter Length	Experimenter	Child
1	54.2	79.0
4	50.7	71.2
8	51.6	71.8
16	55.0	73.6

As reported in McDonough, et al. [14] Figure 4, drastic reductions in error rate can be achieved through the combination of VPS with *maximum kurtosis* (MK) beamforming; indeed, the cascade of square-root implementation of the information filter after maximum MK beamforming reduced WER to 16.1% and 40.0% for the experimenter and child subjects, respectively. The effect of the beamforming was equally evident in the time and spectral domains, as shown in Figure 4. The portions of this figure are analogous to those in Figure 3. In addition to doubling the numerical accuracy of the direct form implementation of the information filter, the square-root implementation eliminates the possibility of *explosive divergence* to which the direct form is prone [8, §11].

4. Conclusions

In this work, we compared two techniques for voice prompt suppression. The first was a straightforward adaptation of a conventional *Kalman filter*, which has certain advantages over the *normalized least squares algrithm* in terms of robustness and speed of convergence; this algorithm is similar to that described in [5]. The second algorithm, which was first proposed in this work, is also based on a Kalman filter, but differs from the first in that the update or correction step is performed in *information space* and hence allows for the use of *diagonal loading* in order to control the growth of the subband filter coefficients, and thereby add robustness to the VPS. This added robustness enabled the effective use of longer subband filters. Distant recognition experiments showed that the information filter reduced word error

	Word Error Rate (%WER)			
Subject	Instructor	Children	Instructor	Children
Filter Length	Standard Kalman Filter		Information Filter	
1	54.3	74.3	54.2	79.0
4	54.0	75.2	50.7	71.2
8			51.6	71.8
16	68.7	77.5	55.7	73.6



a) Voice prompt

b) Corrupted signal

c) After VPS

d) Spectrogram of corrupted signal

e) Spectrogram after VPS

Figure 3: Waveforms from a single array channel.

rates from 52.3% and 74.3% for an adult experimenter and child subjects to 50.7% and 71.2%, respectively. As reported in [14], further reductions in WER to 16.1% and 40.0% were achieved by cascading VPS after maximum kurtosis beamforming. In future, we plan to investigate the relative effectiveness of cascading maximum kurtosis after VPS, and vice versa.

5. References

[1] W. Kellermann, "Acoustic echo cancellation for beamforming microphone arrays," in *Microphone Arrays*, M. Branstein and D. Ward, Eds., pp. 281–306. Springer, Heidelberg, 2001.

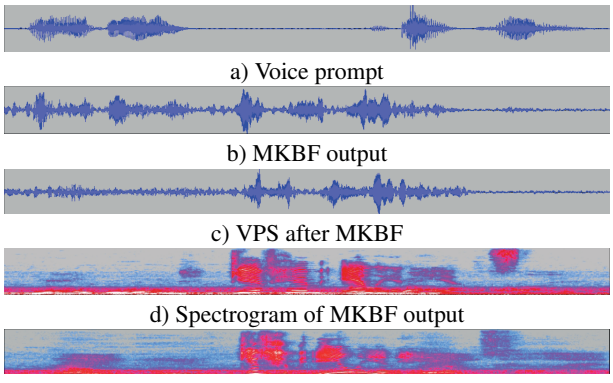
[2] Constantin Paleologu, Silviu Ciochina, and Jacob Benesty, "Double-talk robust VSS-NLMS algorithm for under-modeling acoustic echo cancellation," in *ICASSP*, 2008, pp. 245–248.

[3] Irina Dornean, Marina Topa, Botond Sandor Kirei, and Marius Neag, "Sub-band adaptive filtering for acoustic echo canellation," in *Proc. ICASSP*, 2009, vol. 5, pp. 810–813.

[4] Zhao Yue, Li Nian Qiang, and Zhong Tian Yu, "A subband acoustic echo cancellation coupled with double talk detector," in *Proc. ICSPS*, 2010, pp. 232–236.

[5] Gerald Enzner and Peter Vary, "Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones," *Elsevier Signal Processing*, vol. 86, no. 6, pp. 1140–1156, 2006.

[6] T. Gander, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Trans. on Communication*, vol. 44, no. 11, pp. 1421–1427, 1996.



a) Voice prompt

b) MKBF output

c) VPS after MKBF

d) Spectrogram of MKBF output

d) Spectrogram of VPS after MKBF

Figure 4: Waveforms processed with maximum kurtosis beamforming.

[7] J. Tao, J. Ying, L. Jian, and Y. Hu, "Subband doubletalk detector for acoustic echo cancellation systems," in *proc. ICASSP*, 2003, vol. 5, pp. 604–607.

[8] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, New York, fourth edition, 2002.

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

[10] Mohinder S. Grewal and Angus P. Andrews, *Kalman Filtering: Theory and Practice*, Prentice Hall, Upper Saddle River, NJ, 1993.

[11] John McDonough, Kenichi Kumatani, Bhiksha Raj, and Jill Fain Lehman, "A mutual information criterion for voice activity detection," in *Proc. Interspeech*, submitted for publication, 2011.

[12] Steve Young, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland, *The HTK Book*, Entropic Software, Cambridge, 1999.

[13] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Wiley, London, 2009.

[14] John McDonough, Kenichi Kumatani, and Bhiksha Raj, "On the combination of voice prompt suppression with maximum kurtosis beamforming," in *Proc. of WASPAA*, New Paltz, NY, USA, October 2011.