An Information Filter for Voice Prompt Suppression

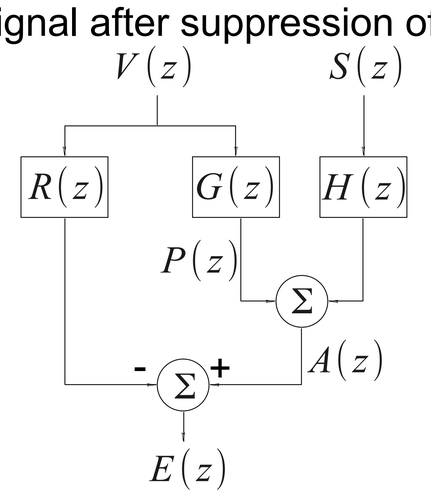
John McDonough[1], Wei Chu[2], Kenichi Kumatani[2], Bhiksha Raj[1], Jill Fain Lehman[2]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Disney Research, Pittsburgh

Abstract

In this work, we propose a *voice prompt suppression* (VPS) algorithm based on an information 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. We also investigate the effectiveness of cascading VPS after maximum kurtosis beamforming. In distant speech recognition experiments, we demonstrate that our system can improve recognition accuracy.

Voice Prompt Suppression

- *V(z)* denotes the transform of the known voice prompt;
- *S(z)* denotes the transform of the unknown desired speech;
- *R(z)* denotes the FIR filter simulating the *room impulse response* (RIR);
- *G(z)* is the transform of the RIR for the voice prompt *V(z)*;
- *H(z)* is the transform of the actual, unknown RIR for the speech *S(z)*;
- *A(z)* is the combined signal at single channel of the microphone array;
- *E(z)* is the residual signal after suppression of the voice prompt.

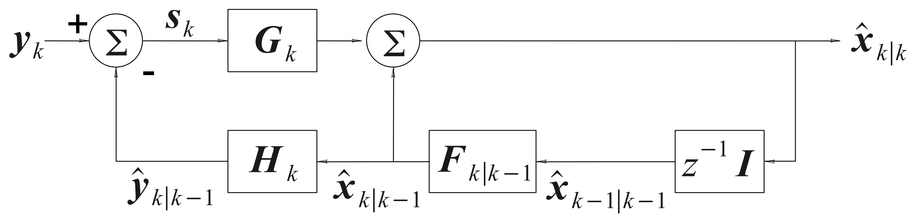


Kalman Filter

The Kalman filter is governed by a *state* and an *observation equation*

$$\boldsymbol{x}_k = \boldsymbol{F}_{k|k-1}\boldsymbol{x}_{k-1} + \boldsymbol{u}_k,$$
$$\boldsymbol{y}_k = \boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{v}_k,$$

The state update involves a *prediction* and a *correction*:



The all important *Kalman gain* is calculated through the recursion

$$\boldsymbol{S}_k = \boldsymbol{H}_k\boldsymbol{K}_{k|k-1}\boldsymbol{H}_k^H + \boldsymbol{V}_k$$
$$\boldsymbol{G}_k = \boldsymbol{K}_{k|k-1}\boldsymbol{H}_k^H\boldsymbol{S}_k^{-1}$$
$$\boldsymbol{K}_{k|k-1} = \boldsymbol{F}_{k|k-1}\boldsymbol{K}_{k-1}\boldsymbol{F}_{k|k-1}^H + \boldsymbol{U}_{k-1}$$
$$\boldsymbol{K}_k = \left(\boldsymbol{I} - \boldsymbol{G}_k\boldsymbol{H}_k\right)\boldsymbol{K}_{k|k-1}$$

Hybrid Information Filter

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

$$\boldsymbol{Z}_k \equiv \boldsymbol{K}_k^{-1},$$
$$\hat{\boldsymbol{d}}_{k|k-1} \equiv \boldsymbol{Z}_{k|k-1}\hat{\boldsymbol{x}}_{k|k-1}.$$

The *temporal update* or *prediction* is performed in *state space* as

$$\hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}_{k|k-1}\hat{\boldsymbol{x}}_{k-1|k-1}$$
$$\boldsymbol{K}_{k|k-1} = \boldsymbol{F}_{k|k-1}\boldsymbol{K}_{k-1}\boldsymbol{F}_{k|k-1}^H + \boldsymbol{U}_{k-1}$$

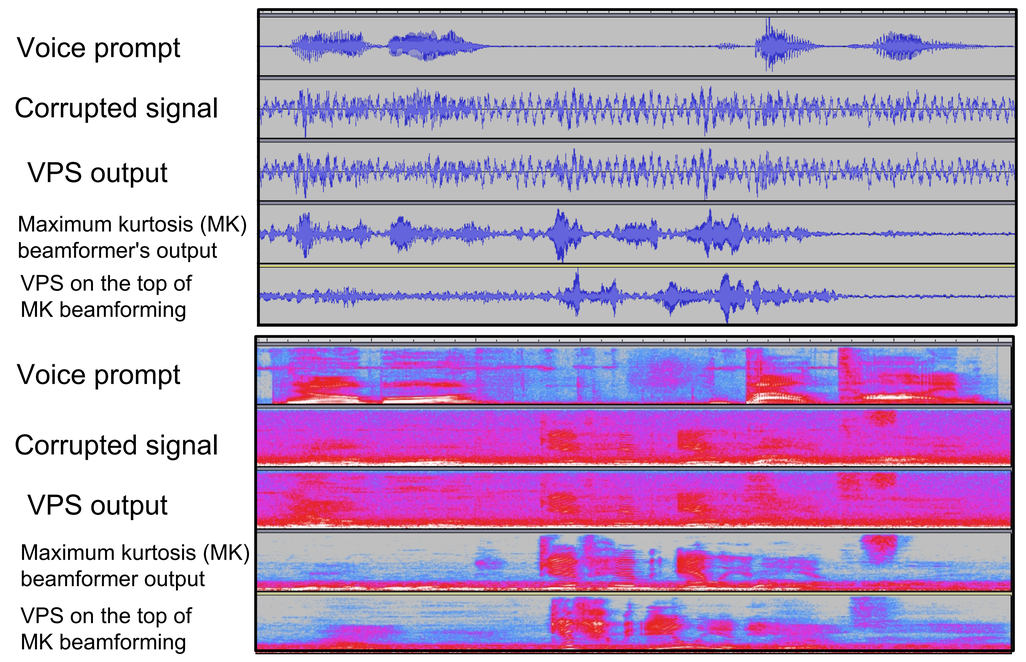
The *observational update* or *correction* can be expressed as

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

Diagonal loading can then be applied according to

$$\boldsymbol{K}_k = \left(\boldsymbol{Z}_k + \sigma_D^2\boldsymbol{I}\right)^{-1},$$
$$\hat{\boldsymbol{x}}_{k|k} = \boldsymbol{K}_k\hat{\boldsymbol{z}}_{k|k}.$$

Processed Samples



These results indicate

- The information filter can suppress the voice prompt,
- The speech signal can be enhanced with maximum kurtosis (MK) beamforming,
- VPS on the beamformed signal provides better speech enhancement and suppression performance.

Experiments and Results

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. 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.

	Word Error Rate (WER)			
Subject	Instructor	Children	Instructor	Children
Filter Length \ Type	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 %

Table. Word error rates (WERs) for several subband filter lengths using the standard Kalman and information filters.

Information filtering on top of MK beamforming can further reduce WER to 16.9 % and 41.7% for the instructor and children, respectively. The lowest WERs of 16.1 % and 40.0% were obtained with a square-root implementation of the information filter applied after MK beamforming.

Conclusions

We have proposed a voice prompt suppression algorithm based on an information filter. This formulation enables diagonal loading to be applied to the information matrix to control the magnitude of the subband filter coefficients. Much like in beamforming, diagonal loading provides for superior robustness. Further work is needed to directly compare the proposed algorithm to conventional techniques based on normalized LMS algorithms.