ON THE COMBINATION OF VOICE PROMPT SUPPRESSION WITH MAXIMUM KURTOSIS BEAMFORMING

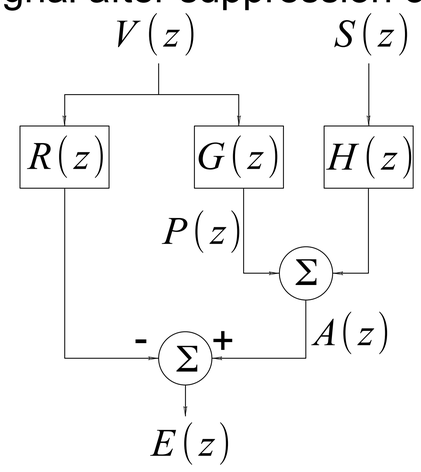
John McDonough[1], Kenichi Kumatani[2] and Bhiksha Raj[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Disney Research, Pittsburgh

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 propose a square root implementation of the information filter VPS algorithm, and 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. 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.

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.

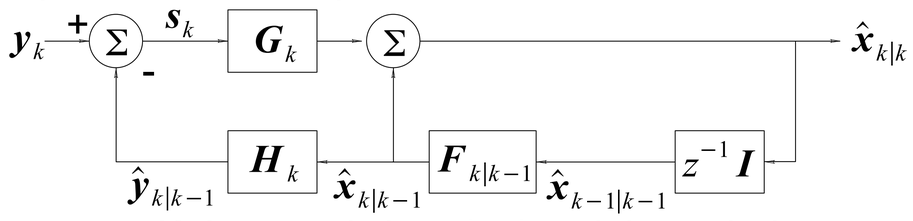


Kalman Filter

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

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

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* in information space is given by

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

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

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

Square Root Formulation

The *square root* implementation of the information filter is based on definitions

$$\boldsymbol{Z}_k = \boldsymbol{Z}_k^{1/2}\boldsymbol{Z}_k^{H/2}$$
$$\boldsymbol{z}_k \equiv \boldsymbol{Z}_k^{H/2}\boldsymbol{x}_k.$$

The temporal update involves an unitary *transformation*

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

Proving the equivalence of the direct and square root implementations involves applying the matrix factorization lemma then showing that

$$\boldsymbol{B}_{11} = \left(\boldsymbol{A}_k + \boldsymbol{U}_{k-1}^{-1}\right)^{1/2},$$
$$\boldsymbol{B}_{21} = -\boldsymbol{A}_k\left(\boldsymbol{A}_k + \boldsymbol{U}_{k-1}^{-1}\right)^{1/2},$$
$$\boldsymbol{b}_{31} = -\left(\boldsymbol{A}_k + \boldsymbol{U}_{k-1}^{-1}\right)^{-1/2}\boldsymbol{F}_{k|k-1}^{-1}\hat{\boldsymbol{d}}_{k-1|k-1}.$$

The observational update involves a second unitary transformation

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

Finally, diagonal loading can be applied to control the magnitude of the weight vector

$$\begin{bmatrix} \boldsymbol{Z}_k^{1/2} & \sigma_D\boldsymbol{e}_n \end{bmatrix}\boldsymbol{\Theta}_{diag} = \begin{bmatrix} \left(\boldsymbol{Z}'_k\right)^{1/2} & \boldsymbol{0} \end{bmatrix}$$

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 (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.

Type	L_{filt}	Pass 1 Exp.	Pass 1 Child	Pass 2 Exp.	Pass 2 Child	Pass 3 Exp.	Pass 3 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.

Conclusions

We have proposed a voice prompt suppression algorithm based on a square root implementation of the information filter. This formulation enables diagonal loading to be applied information matrix to control the magnitude weight vector. Much like in beamforming, diagonal loading provides for superior robustness. Further work is need to directly compare the proposed algorithm to convention techniques based on normalized LMS algorithms.