Variable Span Filtering for Speech Enhancement

ICASSP, 20–25 March 2016

Jesper Rindom Jensen[*,1], Jacob Benesty[1,2], and Mads Græsbøll Christensen[1]

[*]jrj@create.aau.dk

[1]Audio Analysis Lab,
AD:MT
Aalborg University
Denmark

[2]INRS-EMT
University of Quebec
Canada

Partly funded by the Danish Council for Independent Research, grant ID: DFF – 1337-00084, and the Villum Foundation.

AALBORG UNIVERSITY
DENMARK

Agenda

Introduction

- Noise reduction/enhancement is essential in many multichannel applications, such as hearing-aids.
- Have been tackled using many methods, e.g., linear filtering, spectral subtractive, and subspace methods.
- We propose a new class of methods that combines the linear filtering and subspace approaches.
- This is achieved by designing filters using a joint diagonalization.
- These variable span filters give explicit control over noise reduction versus signal distortion.

Signal Model

Model:

M sensors captures a convolved source signal and noise:

$$y_m(t) = g_m(t) * s(t) + v_m(t) = x_m(t) + v_m(t), \tag{1}$$

$m = 1, \ldots, M$, where

- g_m: m'th acoustic impulse response,
- s: desired source,
- v_m: noise captured by sensor m,

Problem Formulation

Goal

Extract x_1 from y_m, $m = 1, \ldots, M$ with little/no distortion and residual noise using optimal filters.

A few assumptions to facilitate the task:

1. x_m and v_m uncorrelated, zero mean, stationary, real and broadband,
2. sensor signals are aligned wrt. the source direction.

Frequency Domain Model

Using the STFT, we get:

$$Y_m(k,n) = X_m(k,n) + V_m(k,n), \tag{2}$$

$m = 1, \ldots, M$, where

k & n: frequency and time indices,

Y, X, V: STFTs of y, x and v at k'th frequency.

In vector format:

$$\begin{aligned} \mathbf{y}(k,n) &= \begin{bmatrix} Y_1(k,n) & Y_2(k,n) & \cdots & Y_M(k,n) \end{bmatrix}^T \quad (3) \\ &= \mathbf{x}(k,n) + \mathbf{v}(k,n). \quad (4) \end{aligned}$$

Observation

Common assumption:

$$X_m(k,n) = G_m(k)S(k,n), m = 1, 2, \ldots, M, \tag{5}$$

where G_m and S are STFTs of g_m and s.

Observation:

- only valid with inf. long windows $N_{\text{win}} \to \infty$ or periodic sources,
- window length limited in practice,
- leads to $\text{rank}(\mathbf{\Phi}_\mathbf{x}(k,n)) = P > 1$,
- not accounted for in conventional methods.

Interframe Correlation

Successive time frames (N) taken into account:

$$\begin{aligned} \underline{\mathbf{y}}(k,n) &= \begin{bmatrix} \mathbf{y}^T(k,n) & \mathbf{y}^T(k,n-1) & \cdots & \mathbf{y}^T(k,n-N+1) \end{bmatrix}^T && (6) \\ &= \underline{\mathbf{x}}(k,n) + \underline{\mathbf{v}}(k,n). && (7) \end{aligned}$$

Correlation matrix of $\underline{\mathbf{y}}(k,n)$:

$$\begin{aligned} \mathbf{\Phi}_{\underline{\mathbf{y}}}(k,n) &= \mathrm{E}\left[\underline{\mathbf{y}}(k,n)\underline{\mathbf{y}}^H(k,n)\right] && (8) \\ &= \mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n) + \mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n), && (9) \end{aligned}$$

where

$\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)$: correlation matrix of $\underline{\mathbf{x}}$ of rank $P < MN$,

$\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)$: correlation matrix of $\underline{\mathbf{v}}$ of rank MN.

Joint Diagonalization

Joint diagonalization of correlation matrices:

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\mathbf{B}(k,n) = \mathbf{\Lambda}(k,n), \tag{10}$$

$$\mathbf{B}^H(k,n)\mathbf{\Phi}_{\underline{\mathbf{v}}}(k,n)\mathbf{B}(k,n) = \mathbf{I}_{MN}, \tag{11}$$

where

B: full rank, $MN \times MN$ matrix,

Λ: diagonal matrix with P real, positive entries (sorted),

$\mathbf{I}_{MN}$: identity matrix with dimensions $MN \times MN$.

$\mathbf{\Lambda}$ and $\mathbf{B}$ are eigenvalue and -vector matrices of $\mathbf{\Phi}_{\underline{\mathbf{v}}}^{-1}\mathbf{\Phi}_{\underline{\mathbf{x}}}$, i.e.,

$$\mathbf{\Phi}_{\underline{\mathbf{v}}}^{-1}(k,n)\mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\mathbf{B}(k,n) = \mathbf{B}(k,n)\mathbf{\Lambda}(k,n). \tag{12}$$

Filtering

Desired signal X_1 estimated from $\underline{\mathbf{y}}$ through filtering:

$$Z(k,n) = \underline{\mathbf{h}}^H(k,n)\underline{\mathbf{y}}(k,n), \tag{13}$$

where

$$\underline{\mathbf{h}}(k,n) = \begin{bmatrix} \mathbf{h}^T(k,n) & \mathbf{h}^T(k,n-1) & \cdots & \mathbf{h}^T(k,n-N+1) \end{bmatrix}^T. \tag{14}$$

With $\mathbf{B}$ as basis, the filter is

$$\underline{\mathbf{h}}(k,n) = \mathbf{B}(k,n)\underline{\mathbf{a}}(k,n), \tag{15}$$

where $\underline{\mathbf{a}}$ is a filter in $\mathbf{B}$, and

$$\underline{\mathbf{a}}(k,n) = \begin{bmatrix} A_1(k,n) & \cdots & A_{MN}(k,n) \end{bmatrix}^T. \tag{16}$$

Filtering

Using $\underline{\mathbf{a}}$, the signal estimate is

$$
\begin{aligned}
Z(k,n) &= \underline{\mathbf{a}}^H(k,n)\mathbf{B}^H(k,n)\underline{\mathbf{x}}(k,n) + \underline{\mathbf{a}}^H(k,n)\mathbf{B}^H\underline{\mathbf{v}}(k,n) && (17)\\
&= X_{\text{fd}}(k,n) + V_{\text{rn}}(k,n). && (18)
\end{aligned}
$$

Variance of Z:

$$
\begin{aligned}
\phi_Z(k,n) &= \underline{\mathbf{a}}^H(k,n)\mathbf{\Lambda}(k,n)\underline{\mathbf{a}}(k,n) + \underline{\mathbf{a}}^H(k,n)\underline{\mathbf{a}}(k,n) && (19)\\
&= \phi_{X_{1,\text{fd}}}(k,n) + \phi_{V_{1,\text{rn}}}(k,n) && (20)
\end{aligned}
$$

Performance

Output SNR:

$$\mathrm{oSNR}[\underline{\mathbf{h}}(k,n)] = \frac{\phi_{X_{1,\mathrm{fd}}}(k,n)}{\phi_{V_{1,\mathrm{rn}}}(k,n)} = \frac{\underline{\mathbf{a}}^H(k,n)\mathbf{\Lambda}(k,n)\underline{\mathbf{a}}(k,n)}{\underline{\mathbf{a}}^H(k,n)\underline{\mathbf{a}}(k,n)} \tag{21}$$

Signal reduction factor:

$$\xi_{\mathrm{sr}}[\underline{\mathbf{h}}(k,n)] = \frac{\phi_{X_1}(k,n)}{\phi_{X_{1,\mathrm{fd}}}(k,n)} = \frac{\phi_{X_1}(k,n)}{\underline{\mathbf{a}}^H(k,n)\mathbf{\Lambda}(k,n)\underline{\mathbf{a}}(k,n)} \tag{22}$$

Mean Squared Error

Error given by $\mathcal{E}(k,n) = Z(k,n) - X_1(k,n)$, leads to MSE

$$J\left[\mathbf{a}'(k,n)\right] = E\left[\left|\mathcal{E}(k,n)\right|^2\right] = \underbrace{J_{\text{ds}}\left[\mathbf{a}'(k,n)\right]}_{\text{distortion MSE}} + \underbrace{J_{\text{rs}}\left[\mathbf{a}'(k,n)\right]}_{\text{residual noise MSE}},$$

where

$$J_{\text{ds}}\left[\mathbf{a}'(k,n)\right] = E\left[\left|X_1(k,n) - \mathbf{a}'^H(k,n)\mathbf{B}'^H(k,n)\underline{\mathbf{x}}(k,n)\right|^2\right] \tag{23}$$

$$J_{\text{rs}}\left[\mathbf{a}'(k,n)\right] = E\left[\left|\mathbf{a}'^H(k,n)\mathbf{B}'^H(k,n)\underline{\mathbf{v}}(k,n)\right|^2\right]. \tag{24}$$

Tradeoff Filter

A general filter is obtained by solving:

$$\min_{\mathbf{a}'} J_{\mathrm{ds}}[\mathbf{a}'(k,n)] \quad \text{s.t.} \quad J_{\mathrm{rs}}[\mathbf{a}'(k,n)] = \beta \phi_{V_1}(k,n), \tag{25}$$

where $0 \leq \beta \leq 1$ controls the level of noise reduction.

Tradeoff filter design:

$$\underline{\mathbf{h}}_{\mathrm{T},\mu}(k,n) = \sum_{p=1}^{P} \frac{\underline{\mathbf{b}}_p(k,n)\underline{\mathbf{b}}_p^H(k,n)}{\mu + \lambda_p(k,n)} \mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}, \tag{26}$$

with μ a Lagrange multiplier adjusting noise level.

General Tradeoff Filter

Even more general filter obtained by using an arbitrary number of eigenvalues instead

$$\underline{\mathbf{h}}_{\mu,Q}(k,n) = \sum_{q=1}^{Q} \frac{\underline{\mathbf{b}}_q(k,n)\underline{\mathbf{b}}_q^H(k,n)}{\mu + \lambda_q(k,n)} \mathbf{\Phi}_{\underline{\mathbf{x}}}(k,n)\underline{\mathbf{i}}. \tag{27}$$

Observations:

- $\underline{\mathbf{h}}_{0,1}(k,n) = \underline{\mathbf{h}}_{\mathrm{max}}(k,n),$ (max SNR filter)
- $\underline{\mathbf{h}}_{1,P}(k,n) = \underline{\mathbf{h}}_{\mathrm{W}}(k,n),$ (general Wiener filter)
- $\underline{\mathbf{h}}_{0,P}(k,n) = \underline{\mathbf{h}}_{\mathrm{MVDR}}(k,n),$ (distortionless filter)
- $\underline{\mathbf{h}}_{0,Q}(k,n) = \underline{\mathbf{h}}_{\mathrm{MD}}(k,n),$ (minimum distortion filter)
- $\underline{\mathbf{h}}_{\mu,P}(k,n) = \underline{\mathbf{h}}_{\mathrm{T},\mu}(k,n).$

Experimental Results

Estimation of statistics

The statistics were estimated directly from the speech and noise signals, respectively.

It was conducted recursively using the following general equation for approximating the correlation matrix of a vector $\underline{\mathbf{a}}(k, n)$:

$$\widehat{\mathbf{\Phi}}_{\underline{\mathbf{a}}}(k, n) = (1 - \xi)\widehat{\mathbf{\Phi}}_{\underline{\mathbf{a}}}(k, n - 1) + \xi\underline{\mathbf{a}}(k, n)\underline{\mathbf{a}}(k, n)^H, \qquad (28)$$

where ξ is a forgetting factor.

Forgetting factors for the signal and noise statistics estimation were chosen as $\xi_{\mathrm{s}} = 0.05$ and $\xi_{\mathrm{n}} = 0.05$, respectively.

Simulation Setup

Parameters:

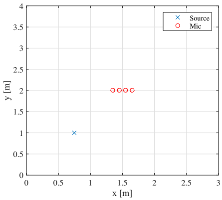
- sensor distance: 5 cm
- sound speed: 343 m/s
- reverb time: 0.2 s
- RIR length: 2,048
- mic type: omnidirectional

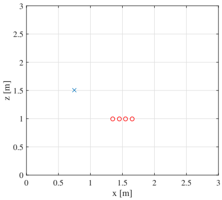
Signals:

Desired: speech (2 female and 2 male),

Noise: diffusive (babble) + sensor noise (white).

Room layout:



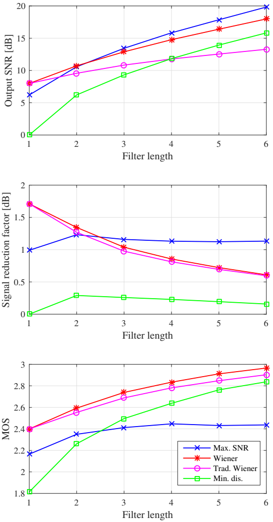


Results

Evaluation vs. filter length

Parameters:

- # of sensors: 3
- SDNR: 0 dB
- SSNR: 30 dB
- window length: 40
- FFT length: 64
- Assumed rank: 3

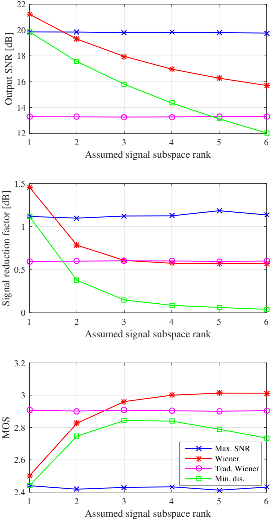


Results

Evaluation vs. filter rank

Parameters:

- # of sensors: 3
- SDNR: 0 dB
- SSNR: 30 dB
- window length: 40
- FFT length: 64
- Filter length: 6



Examples

Parameters:
# of sensors: 3, SDNR: 0 dB, SSNR: 30 dB, window length: 40, FFT length: 64, filter length: 4.

Clean	Noisy	Max. SNR	Min. Dis.	Wiener	Trad. Wiener

Conclusions

- We considered the topic of multichannel speech enhancement.
- Proposed a new class of variable span filters (STFT domain).
- Unifies the filtering and subspace approaches.
- Provides explicit control over noise reduction versus signal distortion.
- Encompasses many well-known filter designs.
- Can outperform the traditional filtering counterparts according to experimental results (oSNR, PESQ, distortion).