February 21, 2014

1 `Simulink` model

The SDW-MWF, R1-MWF and GEVD MWF where first implemented in the `MATLAB` environment as a reference for the `Simulink` model. Once the `MATLAB` model was functioning properly, this code was essentially adapted using native `simulink` blocks. While there was a previous implementation of the SDW-MWF and R1-MWF in the `simulink` model `nsb_XPC_SP15_KUL.mdl`, there were severe limitations with respect to dynamic variables (FFT frame size, initialization parameters), which, would require the end-user to find all references in the model pertaining to the variables.

All of the parameters for the `simulink` blocks can now be controlled via a mask, as shown in Figure 1[1]. The simulink block has 5 labeled inputs, 4 for the microphone signals and one for a clean speech signal reference [2] that is used for the voice activity detector, which will be described in Subsection 2.1.

The mask contains several parameters that are user adjustable :

Filter type
Filter implementation, with possible inputs (1 = SDW-MWF, 2 = R1-MWF, 3 = GEVD MWF)

Speech distortion weighting factor μ
Given in the range of $0 < \mu < \infty$, where a higher value corresponds to higher speech distortion

Forgetting factor λ
Time averaging used for the correlation matrices over approximately $\frac{1}{1-\lambda}$ blocks of L samples such that

$$T_\lambda = \frac{1}{1-\lambda}\frac{L}{f_s} \tag{1}$$

where f_s is the sampling frequency.

Window
Window used for short-time Fourier transform (STFT) analysis.

Exchange between devices
Number of signals shared between devices (0 = bilateral (each devices operates independently), 1 = front microphone is shared between devices, 2 = binaural (all signals shared between devices)).

VAD
Voice activity detector (VAD); 0 - assumes the block has access to the clean speech signal and looks at the energy of each frame, if the energy is greater than a threshold speech is present,1 - uses a log-energy VAD based on only received signal (may not function correctly at low input signal-to-noise ratios).

[1] The simulink block requires the number of samples per frame, L, as an initialization parameter.
[2] This can be removed if a VAD that does not rely on the clean speech signal is used.

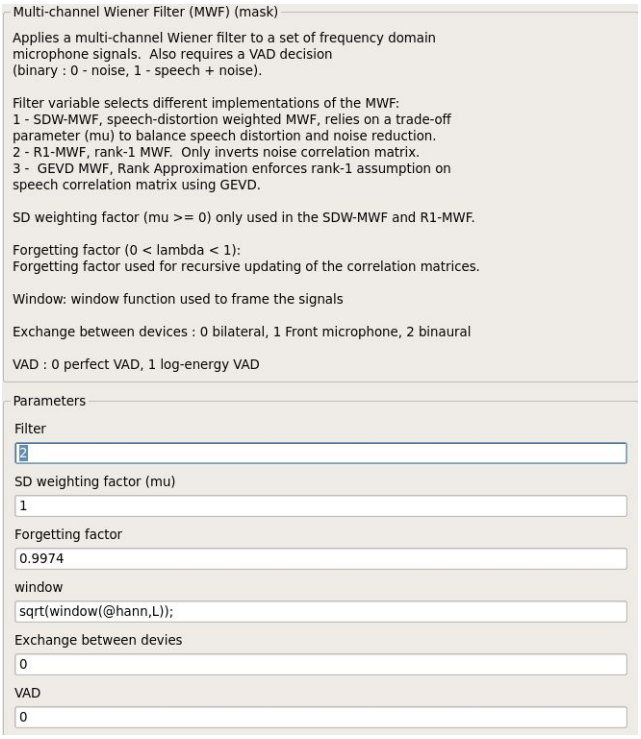

Figure 1: Simulink mask for MWF.

2 Block functionality

Figure 2 shows underneath the mask of the simulink block which will be described in the following subsections. The left and right device consist of the same process run in parallel. The signals are first multiplied by the windowing function that is given as a mask parameter and passed to a `STFT` block. The `STFT` block contains a buffer with overlap of L/2 in order to perform a weighted overlap add analysis and transforms the signals to the frequency domain.

The frequency domain signals are the concatenated together (all signals) for the left and right device. This is then followed by a `Variable Selector` block that uses the `Exchange between devices` parameter from the mask in order to only pass the desired number of signals, i.e., for bilateral filters (`Exchange between devices` = 0), only the devices microphone signals are used. This block is then followed by a `Submatrix` block to only extract the first L/2 elements of the frequency domain signal.

2.1 Voice activity detector (VAD)

The voice activity detector has two modes of operation : assuming that the clean speech signal is available and where only the received signal is available. For the case when the clean speech signal is available, the energy of the frame, E_{frame}, is calculated and compared to a threshold. If the frame energy is greater than this threshold, then the frame contains speech, and a 1 is sent to the other blocks in the model, otherwise a 0 is sent to the blocks in the model.

In the case when the clean speech signal is not available a log-energy VAD is implemented in order to differentiate between speech and noise frames [2]. This is accomplished by adaptively estimating a noise log-energy threshold over previous frames and comparing to the log-energy of the current frame. It is noted

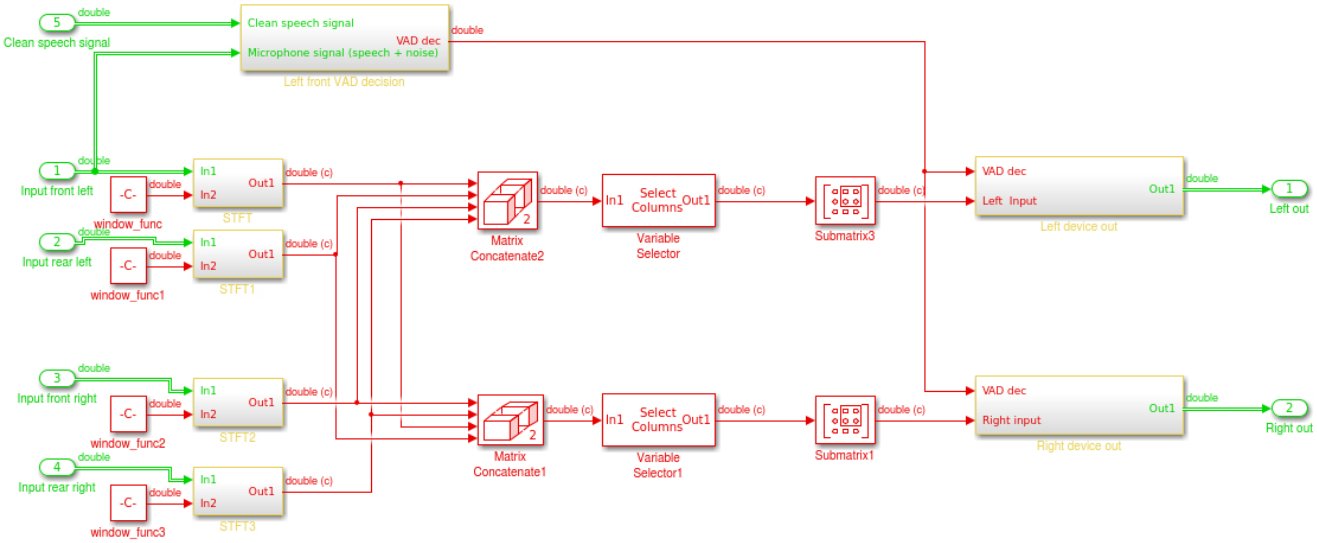

Figure 2: Under the mask of the simulink block.

that in environments with low input signal-to-noise ratios or that are highly non-stationary the log-energy VAD will often fail due to not being able to correctly estimate the log-energy of the noise and speech signals.

2.2 Correlation matrices

The correlation matrices are updated using the current signals passed through the `STFT` block and the VAD decision passed from the VAD block. The correlation matrices use the `Forgetting factor`, λ, parameter from the mask to update the correlation matrices which, for example, when speech+noise is present is updated as

$$\mathbf{R_{yy}} = \lambda \mathbf{R_{yy}} + (1 - \lambda)\mathbf{y}\mathbf{y}^H \tag{2}$$

where $\mathbf{y}$ is the current frequency domain representation of the signal that contains speech+noise and H is the conjugate transpose. When only noise is present the noise only correlation matrix $\mathbf{R_{nn}}$ is updated in a similar fashion.

At the beginning of the simulation, both of the correlation matrices are null matrices until the first samples are received. They will only be properly initialized and contain the signal statistics once the VAD has been triggered by both a noise and speech+noise frame, i.e., the VAD has been on and off.

2.3 Filter type

The filter type is selected from the mask parameter `Filter type`. The parameter is then used with a switch statement block to calculate the desired filter as shown in Figure 4. One benefit of using the switch system is that only the selected filter type is calculated, i.e., all three filters are not calculated simultaneously which would slow down the simulink model. Also in order to achieve proper output dimensions of the subsystem a `Merge` block is used, however, the output still only contains the desired filter output and is not a combination of the three.

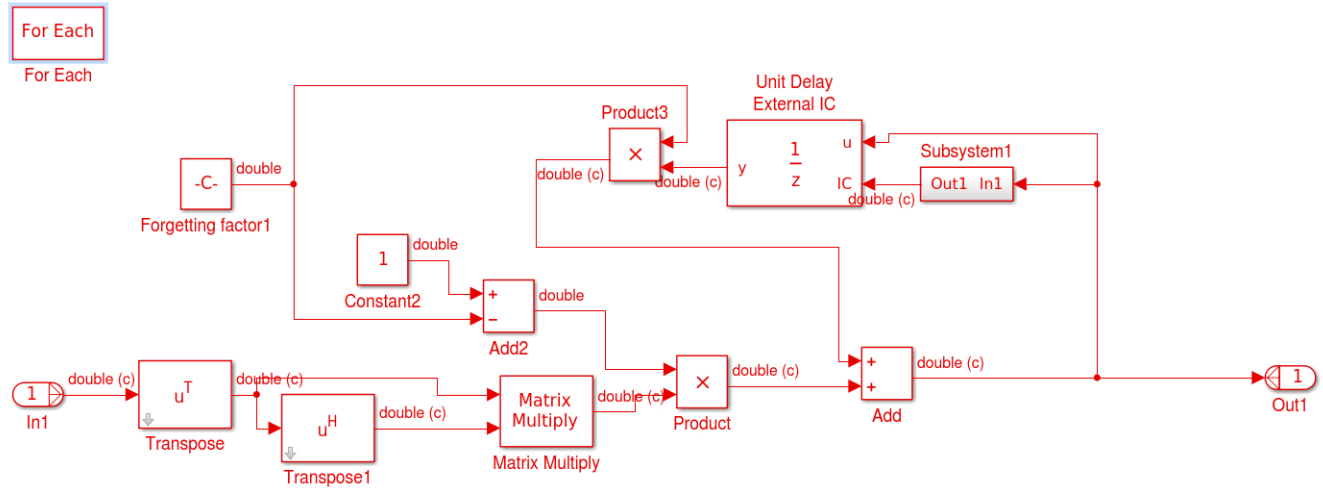

Figure 3: Correlation matrix initialization and updating.

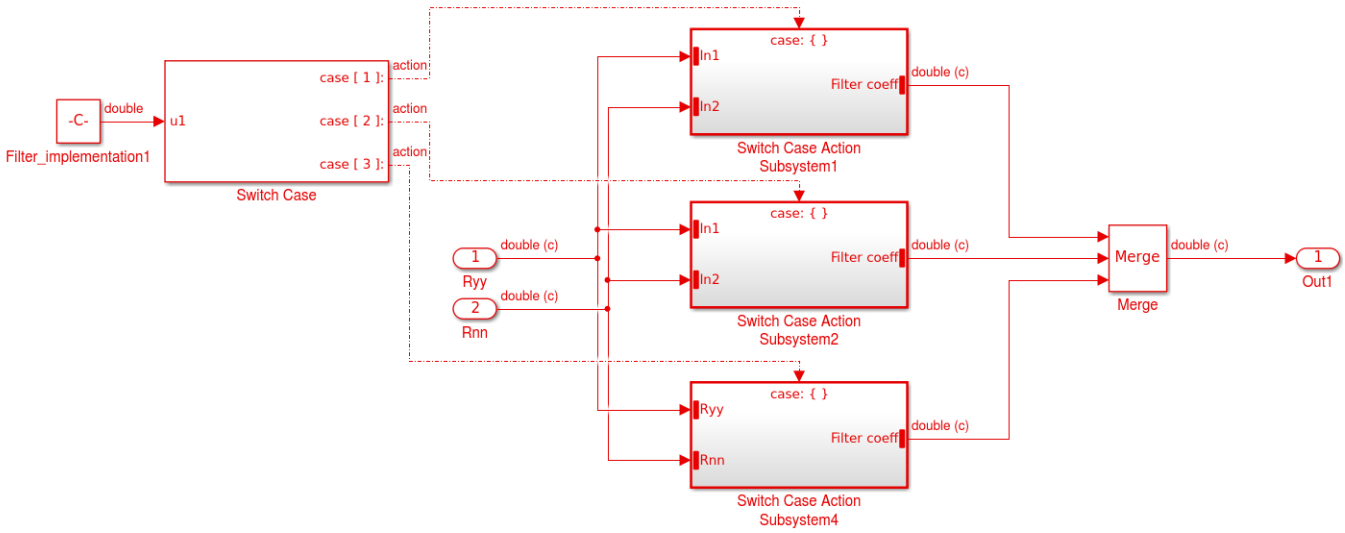

Figure 4: Selection of filter.

2.4 GEVD

Surprisingly, `simulink` does not come equipped with a native way to perform eigenvalue decomposition. Obviously, this was the major hurdle for implementation of the GEVD algorithm in the `simulink` model. In order to overcome this shortcoming, a user defined `fcn` was used that incorporated `MATLAB`'s existing GEVD algorithm. The block takes as inputs the current $\mathbf{R_{yy}}$ and $\mathbf{R_{nn}}$ matrices and outputs the rank-1 approximation of the speech correlation matrix $\mathbf{R_{xx}}$. The code for the `simulink` block is given below

```
function Rxx = fcn(Ryy,Rnn)
%#codegen

freq_bins = size(Ryy,3);

Rxx = zeros(size(Ryy,1),size(Ryy,1),freq_bins)+1i.*zeros(size(Ryy,1),size(Ryy,1),freq_bins);

for freq=1:freq_bins
    [V,D] = eig(Ryy(:,:,freq),Rnn(:,:,freq));
    Q = inv(V');
    [¬,idx] = sort(real(diag(D)));
    D_Rnn = diag(diag(real(Q\Rnn(:,:,freq)/Q')));
    D_Ryy = diag(diag(real(Q\Ryy(:,:,freq)/Q')));
    D_Rxx = diag(D_Ryy - D_Rnn);
    Rxx(:,:,freq) = Q(:,idx(end))*Q(:,idx(end))'*D_Rxx(idx(end));
end
```

2.5 Signal synthesis

Once the filter has been calculated it is applied in the frequency domain to the current input signal, which again utilizes a `For Each` block for each frequency. After the signal has been filtered it passes through an inverse-STFT block which utilizes an over-lap add procedure to compute the framed signal in the time-domain. This process is shown Figure 5 and is contained in the `Filter synthesis` subsystem of the model. The resultant signal is then passed as the output of the block.

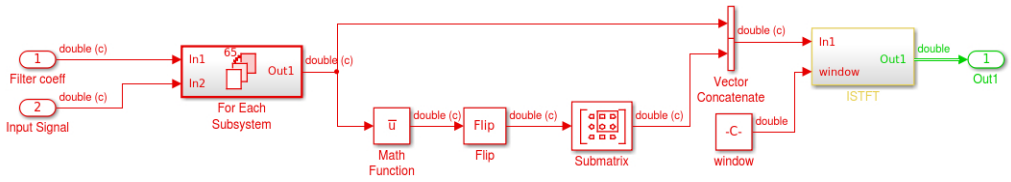

Figure 5: WOLA signal synthesis.

3 Objective Speech Quality Measures

In order to test the functionality of the proposed `simulink` model several objective measures were used to compare the filtering algorithms, specifically the R1-MWF and the GEVD MWF.

3.1 Speech intelligibility weighted SNR (SIW-SNR)

The speech intelligibility weighted signal-to-noise ratio (SIW-SNR) is a SNR measurement which has been shown to correlate well with perceptual speech intelligibility improvements. The signals are first filtered though $\frac{1}{3}$ octave band bandpass filters around center frequencies f_i^c and the SNR improvement is calculated per band as

$$\Delta\text{SIW-SNR} = \sum_i I_i \text{SNR}_{\text{OUT}} - \text{SNR}_{\text{IN}} \tag{3}$$

where I_i expresses the band importance. The center frequencies and weights are defined in [1].

3.2 Speech intelligibility weighted signal distortion (SIW-SD)

The speech intelligibility weighted signal distortion (SIW-SD) measures the average spectral distortion between the filtered signal and the clean speech signal received at the microphone. The signals are again first filtered though $\frac{1}{3}$ octave band bandpass filters around center frequencies f_i^c, where the spectral distortion SD_i per band i is then calculated as

$$\Delta\text{SIW-SD} = \sum_i I_i \mathrm{SD}_i \tag{4}$$

where SD_i as

$$\mathrm{SD}_i = \frac{1}{(2^{1/6} - 2^{1/6}) f_i^c} \int_{2^{-1/6} f_i^c}^{2^{1/6} f_i^c} |10\log_{10} G^s(f)| df \tag{5}$$

and $G^s(f)$ is given by

$$G^s(f) = \frac{P_s}{P_{z_s}} \tag{6}$$

where P_s and P_{z_s} are the power of the speech component of the input and power of filtered speech component respectively.

3.3 Perceptual Evaluation of Speech Quality (PESQ)

The PESQ is an objective voice quality measurement standard that cover a scale of 1 (bad) to 5 (excellent). The one main drawback of the PESQ measurement is that it can only be calculated with signals that have a sampling frequency of $f_s = 16kHz$. Therefore, all signals were resampled from the given sampling rate of$f_s = 15659Hz$.

4 Simulations

Broadband measurements were calculated using the entire length of the signal, i.e., the algorithm was run on the full audio signals and then the corresponding filter at the end of the simulation was then used to filter the clean speech signal and the noise signal separately. It was assumed that a perfect VAD was used for all measurements.

Scenario : 2-SWN-rear

	Device + link	SIW-SNR$_{\mathrm{IN}}$	SIW-SNR$_{\mathrm{OUT}}$	ΔSIW-SNR	SIW-SD	PESQ
Input	Left reference	-4.4	-	-	-	1.4
R1-MWF	Left - bilateral		7.3	11.7	3.4	1.9
	Left - front		7.3	11.7	2.9	1.9
	Left - binaural		9.3	13.7	2.7	2.0
GEVD-MWF	Left - bilateral		7.2	11.6	3.2	1.9
	Left - front		7.4	11.8	3	1.9
	Left - binaural		9.2	13.6	2.5	2.0
Input	Right reference	-4.4	-	-	-	1.5
R1-MWF	Right - bilateral		4.4	8.8	4.9	1.9
	Right - front		6.0	10.4	3.8	2.0
	Right - binaural		9.4	13.8	2.7	2.1
GEVD-MWF	Right - bilateral		4.3	8.7	4.6	1.9
	Right - front		5.8	10.2	3.9	2.0
	Right - binaural		9.1	13.5	2.6	2.1

For the remaining scenarios only the left device results are given.
Scenario : 4-TB-rear

	Device + link	SIW-SNR$_{IN}$	SIW-SNR$_{OUT}$	ΔSIW-SNR	SIW-SD	PESQ
Input	Left reference	-3.2	-	-	-	1.5
R1-MWF	Left - bilateral		0.0	3.2	4.1	1.7
	Left - front		-0.4	2.8	1.1	1.8
	Left - binaural		2.9	6.1	0.9	2.1
GEVD-MWF	Left - bilateral		4.4	7.6	6.1	1.8
	Left - front		6.1	9.3	4.9	2.9
	Left - binaural		7.2	10.4	3.6	2.1

Scenario : 4-TB-full

	Device + link	SIW-SNR$_{IN}$	SIW-SNR$_{OUT}$	ΔSIW-SNR	SIW-SD	PESQ
Input	Left reference	-3.4	-	-	-	1.5
R1-MWF	Left - bilateral		-2.9	0.5	4.6	2.1
	Left - front		-6.0	-2.6	1.1	1.9
	Left - binaural		-0.9	2.5	2.9	1.6
GEVD-MWF	Left - bilateral		0.7	4.1	11.9	1.8
	Left - front		1.9	5.3	10.9	1.8
	Left - binaural		2.1	5.5	7.3	1.8

Scenario : 20-TB-full

	Device + link	SIW-SNR$_{IN}$	SIW-SNR$_{OUT}$	ΔSIW-SNR	SIW-SD	PESQ
Input	Left reference	-3.5	-	-	-	1.5
R1-MWF	Left - bilateral		0.4	3.9	7.7	1.6
	Left - front		1.8	5.3	4.4	1.6
	Left - binaural		2.8	6.3	3.0	1.7
GEVD-MWF	Left - bilateral		1.7	5.2	9.2	1.7
	Left - front		3.1	6.5	7.0	1.8
	Left - binaural		4.4	7.9	5.7	1.9

We see that for the first scenario the R1-MWF and GEVD MWF have similar filtering properties which is due to the stationarity of the noise signals. In the 3 other simulated environments with competing babble, the R1-MWF performs significantly worse than the GEVD MWF.

Below are the real-time outputs of the R1-MWF and GEVD MWF compared to that of the input microphone signal (speech+noise). The first five seconds are removed due to the initialization of the filters. Again we can see in the first scenario that contained 2 uncorrelated sources of speech shaped noise, the performance of the R1-MWF and GEVD MWF are comparable. In the other three scenarios, which contain competing babble noise, the R1-MWF often becomes unstable, whereas the GEVD MWF remains stable throughout the estimation process.

References

[1] Acoustical Society of America. ANSI S3.5-1997 American National Standard Methods for calculation of the speech intelligibility index. June 1997.

[2] Stefaan Van Gerven and Fei Xie. A comparative study of speech detection methods. In *EUROSPEECH*, 1997.

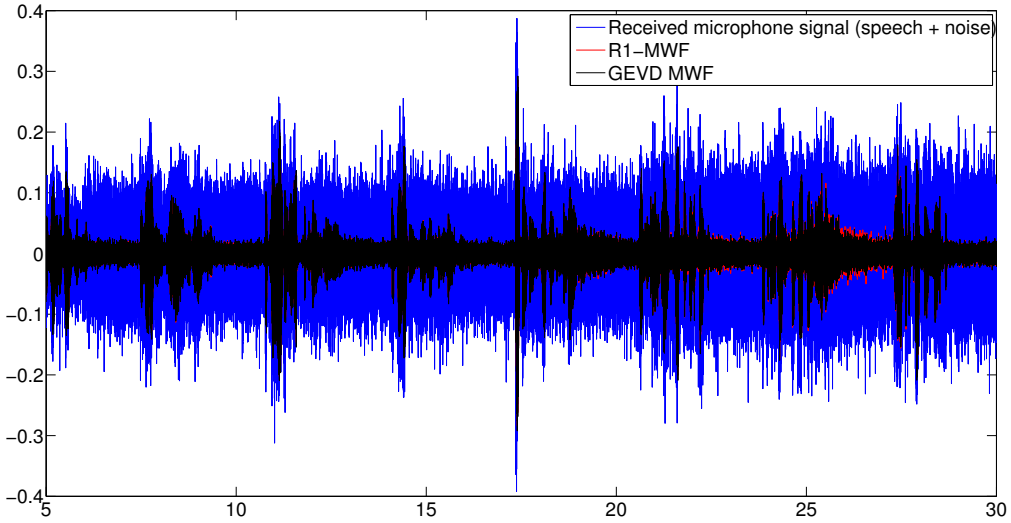

Figure 6: Two sources of uncorrelated speech shaped noise (rear hemi-field) (2-SWN-rear).

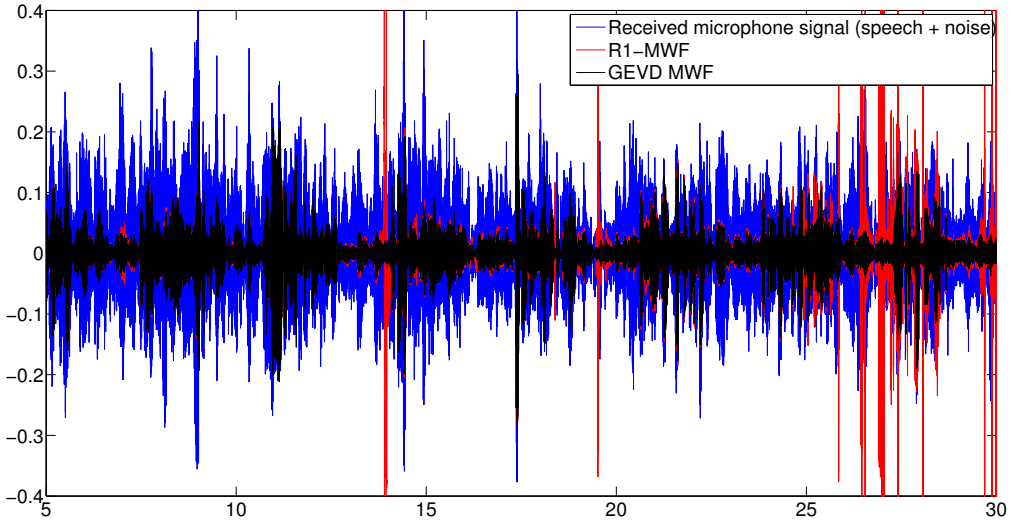

Figure 7: Four sources of competing speech (rear hemi-field) (4-TB-rear).

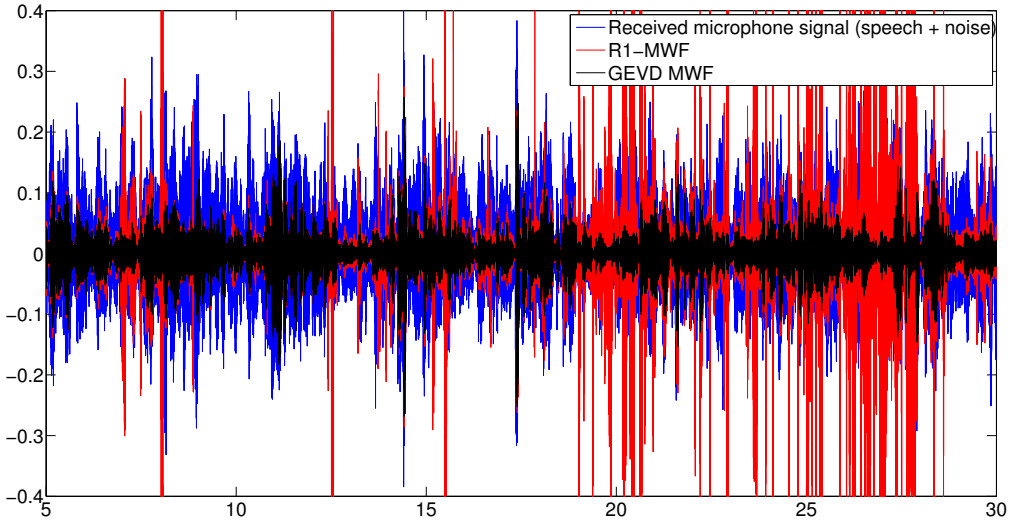

Figure 8: Four sources of competing speech (full circle - restricted.) (4-TB-full).

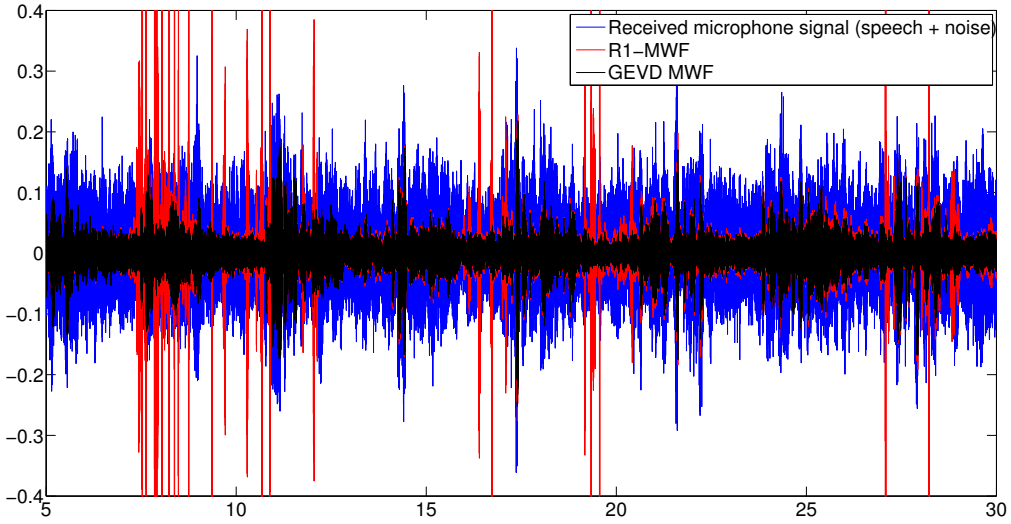

Figure 9: Five uncorrelated segments from each of four sources of competing speech (full circle) (4-TB-full).