Single Channel Noise Reduction System for Speech Enhancement

Mario Coutino†
Std. Num: **4410475**
M.A.Coutinomingue@student.tudelft.nl

Kris Shrishak S†
Std. Num: **4411811**
K.S.Sridaran@student.tudelft.nl

Roel Berns†
Std. Num: **4125215**
R.M.Berns@student.tudelft.nl

† Delft University of Technology

I. Introduction

In this project a single channel noise reduction (SCNR) module for a speech enhancement system is designed and tested. The proposed implementation processed the signal entirely in the frequency domain, through the Short Time Fast Fourier Transform (STFFT) which provides a way to treat the spectrum components independently.

Two power spectra density (PSD) smoothing methods for the speech signal, one in frequency domain and other in the cepstrum domain, are compared through objective measures for predicted intelligibility and speech quality. In addition, spatial filtering by means of beamforming for a multiple microphone set up is shortly described and its advantages presented.

Finally, the result of the implementation of feedback between the spatial and temporal processing modules is briefly discussed and some of its benefits are shown.

II. SCNR Module Design

The single channel noise reduction (SCNR) system was developed in Simulink platform. It consists of five main parts: Analysis/Synthesis, Noise PSD Estimation, Speech PSD Estimation, Speech PSD Smoothing and Gain. The actual design diagram block can be seen in Fig. 1. The following section of this report briefly explains each block to give some understanding of the design choices.

*A. **Analysis/Synthesis***

These are the starting and ending blocks of the system. They implement the segmentation, transformation and merging needed to process the temporal samples in the frequency domain. Basically, these blocks implement the STFT and its inverse, which means that the (Inverse) Fast Fourier Transform (I)(FFT) is performed over a segmented audio signal consisting of 512 samples $\approx 30ms$@16kHz (interval where the signal can be assumed approximately stationary and speech FFT coefficients independent [7]), with an overlap of 50% and scaled with a normalized Hamming window.

The main reason for using Fourier coefficients is due to their approximately uncorrelated nature. Under the assumptions of super-Gaussian distributed coefficients [6] and statistical independence across time and frequency bins, optimal noise suppression can be achieved.

Finally, the merging is carried out by using the *Overlap-add* method [8], which combines successive frames to construct the processed output audio signal. All these processes were implemented in Simulink through the built-in blocks provided by the Signal Processing Toolbox of Matlab.

*B. **Noise PSD Estimation**: MMSE Based*

To estimate the noise PSD we make use of the unbiased MMSE noise PSD estimator based on Speech Presence Probability (SPP). This estimator arises from the regular MMSE estimator with the difference of using a soft decision instead of a hard one.

In section 3 of [2] it is argued that the standard MMSE estimator can be seen as a VAD-based detector which results in a hard decision between noisy observation and the estimate of the spectral noise power. Furthermore it is shown that the MMSE estimator is biased when the estimated quantities are used which are not equal to the true values for noise and/or signal power.

As explained in section 4 of [2] it is possible to replace the hard decision of the VAD-based detector by a soft decision SPP with fixed priors. The advantage is that no bias compensation is necessary.

The *a posteriori* SPP is computed by:

$$\mathcal{P}(\mathcal{H}_1|y) = \left(1 + \frac{\mathcal{P}(\mathcal{H}_0)}{\mathcal{P}(\mathcal{H}_1)}(1+\xi_{\mathcal{H}_1})e^{-\frac{|y|^2}{\sigma_N^2}\frac{\xi_{\mathcal{H}_1}}{1+\xi_{\mathcal{H}_1}}}\right)$$

where we assume that it is equally likely that speech is present or not, i.e. $\mathcal{P}(\mathcal{H}_0) = \mathcal{P}(\mathcal{H}_1) = 0.5$.

As *a priori* SNR we've chosen a fixed value which has the advantage that the noise power estimator can be decoupled from other steps in the speech enhancement system. Gerkmann et al in [2] show that the optimal *a priori* SNR is 15 dB when a probability of error is given as in [11] and it is assumed that the true *a priori* SNR is uniformly distributed between $-\infty$ dB and 20 dB.

Defining Speech Absence Probability as:

$$\mathcal{P}(\mathcal{H}_0|y) = 1 - \mathcal{P}(\mathcal{H}_1|y)$$

the expression for the MMSE estimator results in:

$$E(|N|^2|y) = \mathcal{P}(\mathcal{H}_0|y)|y|^2 + \mathcal{P}(\mathcal{H}_1|y)\widehat{\sigma_N^2}$$

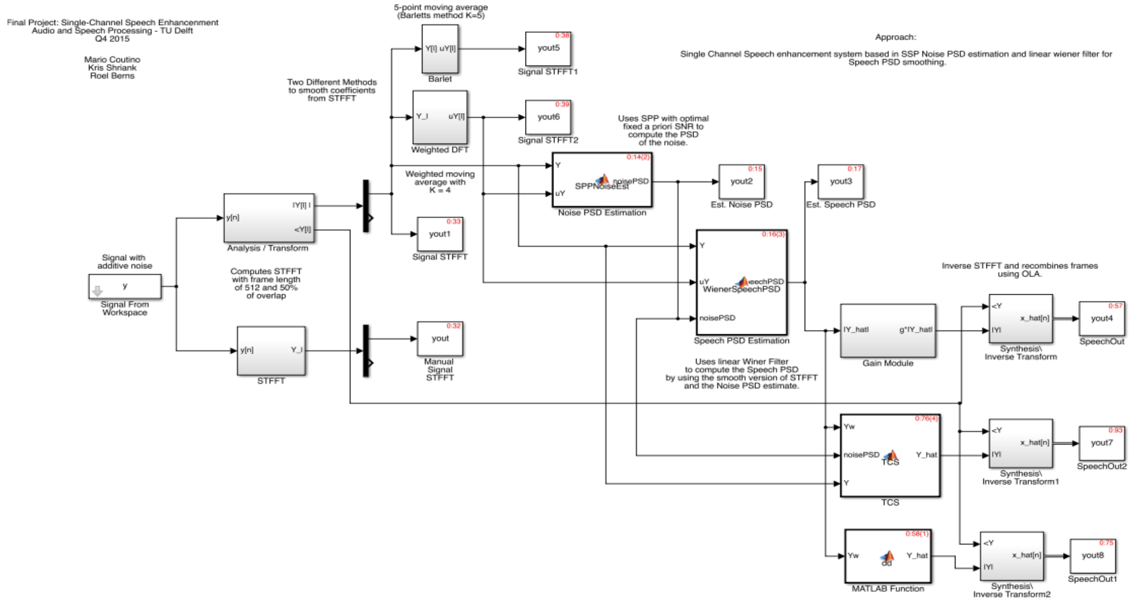


Fig. 1: Single Channel Noise Reduction Module Design

After the noise periodogram is estimated, the noise power spectral density is obtained via recursive smoothing with $\alpha_{pow} = 0.8$:

$$\widehat{\sigma_N^2}(l) = \alpha_{pow}\widehat{\sigma_N^2}(l-1) + (1-\alpha_{pow})E(|N|^2|y(l))$$

In Fig. 2 the output of the module for a particular frequency is shown. As seen in the plot, the method is able to follow the noise PSD without too much delay.

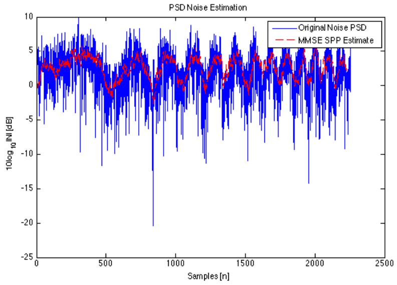


Fig. 2: Output of the MMSE-SPP Noise Estimator

A comparison between the spectograms of the original noise signal and its estimate is shown in Fig. 3. It is seen that the overall shape of the spectrogram is maintained within a small amount of distortion, which leads us to conclude that the noise is properly estimated.

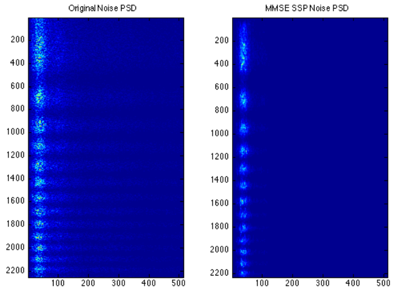


Fig. 3: Spectogram of the Original and Estimated Noise Signal

C. ***Target Estimate***: *Wiener Gain*

In order to estimate the clean speech FFT coefficients a simple approach is based on the minimization of the mean square error (MMSE). For this particular case, the estimator was constrained to be *linear*, which results in the frequency domain Wiener filter [9].
The cost function to minimize is given by

$$G_k = \underset{G_k}{\arg\min} E\{|S_k - G_k Y_k|^2\}$$

which leads to the optimal estimator given by

$$\hat{S}_k = \frac{\sigma_{s,k}^2}{\sigma_{s,s}^2 + \sigma_{n,k}^2} Y_k$$

where the subscript k indicates a specific frequency bin of the FFT.
As the gain G is always real (PSD components are always positive) it can be seen that the optimal estimator only modifies the noisy magnitude of the measurement Y, but the noisy phase is propagated without any modification.

D. ***Speech PSD Smoothing***

Decision Direct (DD)

Ephraim and Malah proposed a decision-directed (DD) estimator in [16] based on the clean-speech estimate from the previous time frame. The parameters α_{dd} and ξ_{min} control the trade-off between noise reduction and distortions of speech transients in a speech enhancement framework. The decision-directed approach allows for fast tracking of increasing levels of the speech power, and results in effective smoothing.

$$\widehat{\sigma^2_{s,k}}(l) = max(\alpha_{dd}|\widehat{S_k}(l-1)|^2 + \\ (1-\alpha_{dd})\left(|y_k(l)|^2 - \sigma^2_{N,k}(l)\right), \xi_{min}\sigma^2_{N,k}(l))$$

The problem of musical noise is observed on using this method. This is because of the high sensitivity of this approach to rising spectral amplitudes which may occur in the form of speech as well as noise. But the negative effect of musical noise much lower than observed on the usage on maximum likelihood approach.

Temporal Cepstrum Smoothing (TCS)

Breithaupt et al in [10] proposed, in contrast to DD, to perform a smoothing of the speech PSD in the cepstrum domain. They argued that as the lower cepstral coefficients represent the spectral envelop of the compressed speech, it is possible to apply a smoothing capable of maintaining the spectral envelop of the speech signal while suppressing spectral outlier due to estimation errors. By applying a quefrency dependent vector gain α, the non-speech related ceptrum coefficients can be *smoothed out* while the voice coefficients are updated *faster* with the new information.

Our modified version of the proposed algorithm in [10] is as follows

Wiener Estimate	$\hat{\sigma}^2_{s,k}$
Ceptral Domain	$\lambda_q = IFFT_q\{log(max(\hat{\sigma}^2_{s,k}, \sigma^2_{min}))\}$
Temporal Smoothing	$\bar{\lambda}_q(l) = \alpha_q(\bar{\lambda}_q(l-1) + (1-\alpha_q)\lambda_q(l)$

As our implementation follows the one from Breithaupt et al a fixed bias compensation $\mathcal{B}$ is used after we return to the frequency domain

$$\sigma^{\hat{2}}_{s,k} = \mathcal{B}exp(FFT_k\{\bar{\lambda}_q\})$$

In addition, the quefrency dependent gain α_q is also considered fixed for all the frames and it is taken as

$$\alpha[i] = \begin{cases} 0.5 & i < 4 \\ 0.7 & 4 \leq i \leq 20 \\ 0.97 & i > 20 \end{cases}$$

where i is the index $(0 < i \leq 256)$ of the array α.

E. Gain Module

Most human beings are capable of hearing sounds in the frequency range 20Hz-20KHz. Human beings are most sensitive to sounds in the range 3kHz-5KHz. In terms of the loudness, sounds from 0dB to 130 dB sound pressure level can be heard by humans. But in the case of a hearing impaired person, the hearing profile changes. Though it differs from person to person, typically hearing loss is characterized by increased threshold of hearing and almost unchanged threshold of pain. This means that the dynamic range of hearing is reduced.
As the goal of speech enhancement system is to improve the intelligibility and the pleasantness of sound, its design for hearing impaired cannot use a constant gain function. Instead a frequency dependent gain function is used in order to compensate for the reduced dynamic range. Common rationales set by standardization authorities are used to design the frequency dependent gain function [15].
These rationales may include gender and language among others. In order to keep the sounds below the threshold of pain, a compressor based on standard compression tables is also used.

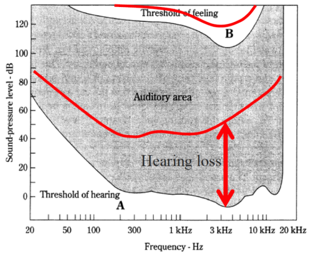


Fig. 4: The auditory area of the human ear [15]

III. Evaluation of Speech Enhancement System

Although listening tests are the ultimate way to validate an audio system, the high cost and the length of the trials are the main motivation to look for automatic ways to validate.

For the evaluation of our speech enhancement system three objective measures are used. Two of them are related to the distortion and audio quality perceived in the processed signal and one with the intelligibility of the output speech. In this section these measures will be briefly described and the result of the system under these metrics will be discussed. The keen reader is referred to [5], [12] and [13] for an in depth treatment of the matter.

	SEG-SNR	STOI	PESQ
Noisy Signal	-6.9277	0.5066	1.1592
Wiener	-3.8139	0.5165	1.4689
DD	-3.9775	0.5223	1.4645
TCS	-1.5905	0.5165	1.4689

TABLE I: Results of different objective quality and intelligibility measures

Segmental SNR (SEG-SNR)

This measure makes a comparison of the waveforms in time domain, which can lead to a situation in which the result does not represent properly the SNR of the individual frames. This measure is defined as

$$\text{SEG-SNR} = \frac{1}{L}\sum_{l\in\mathcal{L}} 10\log_{10}\left(\frac{\|s_l\|^2}{\|s_l - \hat{s}_l\|^2}\right)[dB]$$

The first column of Table I shows that TCS has the best performance under this measure even though it is under 0 dB. As mentioned before, this measure does not correlate very well with actual quality, leading to think that probably the low SEG-SNR given by the other two methods are by sign inversion of the time-domain signal or similar effects, which are imperceptible to us but affect the measure result.

Perceptual Evaluation of Speech Quality (PESQ)

PESQ is a measure to evaluate distortions in the speech signal. It is reported in [13] that it correlates well with the perceived quality of the speech. In this project the PESQ based on the ITU and implemented by C. Loizou [14] is used.
From the results in Table I it is seen that Wiener Filter, DD and TCS give similar results, slightly higher values than the PESQ of the original noisy signal.

Short Time Objective Intelligibility (STOI)

As actual listening tests are expensive, STOI measure provides a way to predict (objectively) the intelligibility for the processed speech signal.
STOI filters signal similar as the cochlea, then the silent regions are removed and finally the temporal envelops are correlated with the ones from the original speech signal. It is expected a monotonic relation of the measure with the average intelligibility [5].

In Table I the STOI for the three methods are slightly higher than the one from the noisy signal. From the processed signals, DD has the highest predicted value ($\approx 52\%$), just above the original STOI of the noisy signal. The other two outputs have the same STOI increase range $\approx +1\%$, which makes us doubt if intelligibility is increased after the noise reduction process.

IV. Multiple Microphones: Spatial Filtering

The possibility of spatially filtering of noisy sources gives a great advantage in terms of Signal to Interference-Noise Ratio (SINR) as now the beam can be focused towards the desired source and null can be oriented towards the interference. Unfortunately, this increase in SINR does not come without a cost. In order to be able to perform spatial filtering, more than one microphone is needed. In addition, higher computational cost is piled up as most of this methods rely on recursive estimation of the inverse signal/noise covariance matrix.

A. *Microphone Array Manifold*

In this project a two microphone system is assumed for spatial filtering. The elements of the array were selected to be cardioid microphones spaced at 5mm from each other. The joint response of the array is given by Fig. 5 for the range of audible frequencies.

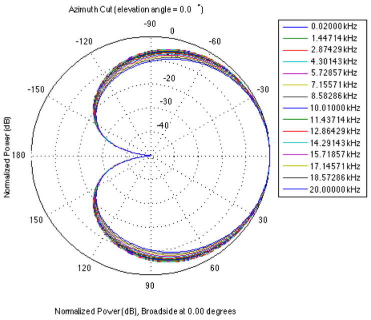


Fig. 5: Microphone array response in azimuthal plane

As expected, the shape is still a cardiod, with a null at π radians. However, for the tested frequencies, the response is not uniform anymore. This effect can be seen more clearly if instead of cardioid elements omnidirectional microphones are used as in Fig. 6.

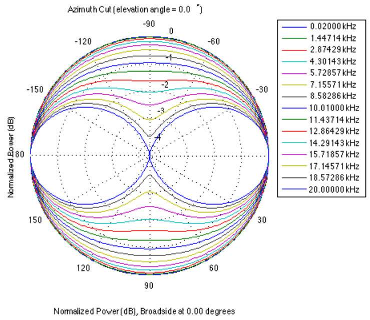


Fig. 6: Microphone array azimuthal response with omnidirectional elements.

This implies that the spatial filter has to be designed taking into account frequency dependencies. This will lead to an implementation of a set of narrow band beam formers centered at the frequency bins of the FFT.

B. *Beam Forming*

As the speech signal has a certain bandwidth, a straight forward beam former cannot be applied to it. In order to deal with this problem, a set of beam formers are designed for each frequency bin of the FFT of the noisy signal.
The Zero Forcing (ZF) and Minimum Variance Distortionless Response (MVDR) beam formers are implemented to compare their performance under the assumption of one speech source and one noise source at known angles (or its angular separation) plus extra noise (i.e thermal noise or/and ambient noise).

The beam formers W are given by the following expressions (under complete knowledge of the array manifold $\mathcal{A}$)

Zero Forcing

$$\mathrm{W}_k = (\mathrm{A}^{\dagger}_{\Phi,k})^H$$

where $A_{\Phi,k}$ is the array response for direction $\Phi = [\theta_s, \theta_n]$ at frequency bin k and $A^{\dagger}$ represents the Moore-Penrose pseudoinverse of A.

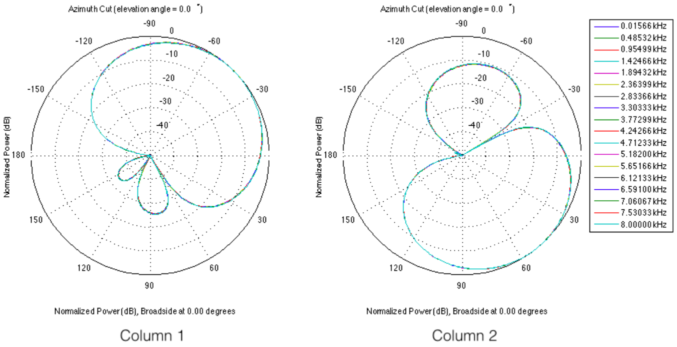
MVDR

$$\mathrm{w}_k = \mathrm{R}_{YY,k}^{-1}\mathbf{a}_{\theta_s}(\mathbf{a}_{\theta_s}^H \mathrm{R}_{YY,k}^{-1}\mathbf{a}_{\theta_s})^{-1}$$

where $\mathrm{R}_{YY,k}$ is the covariance matrix of the FFT coefficients at frequency bin k.

The main problem with ZF is the fact that the angles of the sources should be known in order to create the matrix $A_{\Phi,k}$. In contrast, MVDR only requires knowledge of the direction of the speech source. However, MVDR needs to compute a covariance matrix which has to be estimated as it is not known a priori. In addition, if the process is not stationary (as in most of the cases) techniques to follow the changes in the covariance matrix should be implemented.

Both beam formers are implemented as described before considering knowledge of the positions of both source and noise, its output is then propagated towards the SCNR module for further processing. Particularly for MVDR, the covariance matrix is estimated using a sliding window large enough to obtain an invertible matrix.



Column 1

Column 2

Fig. 7: ZF Beamforming response for each of the columns of **W**

In Fig 7 the array response after the application of the ZF beam former is shown. In this example each of the columns stirs a null to one of the directions $\theta = [120^o, -30^0]$ respectively. It should be noticed that the spreading in the response present in Fig. 5 is not seen anymore.

C. *Results*

In order to evaluate the performance of the addition of a spatial filtering stage to our project, the same instrumental distortion measures are used. The result of this test are shown in Tables II, III and IV.

SEG-SNR

BeamFormer	Wiener	DD	TCS
ZF	3.9955	3.0445	-0.2494
MVDR	-1.4698	-1.8084	-1.1826

TABLE II: SEG-SNR for the different BeamFormers

STOI

BeamFormer	Wiener	DD	TCS
ZF	0.8280	0.8186	0.8280
MVDR	0.6659	0.6646	0.6659

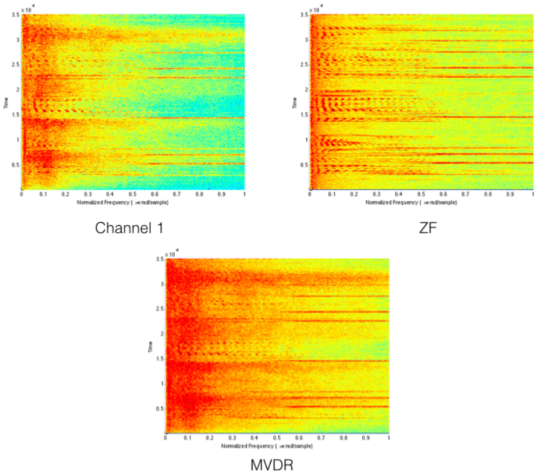
TABLE III: STOI for different BeamFormers

PESQ

BeamFormer	Wiener	DD	TCS
ZF	2.4400	2.4213	2.4400
MVDR	1.7657	1.7538	1.7657

TABLE IV: PESQ for different BeamFormers

From the tables it is seen that the addition of the beamformer improves SEG-SNR and PESQ measurements, compare with the results presented in Table I. When STOI is measured after the beamformer is applied, a considerable increase in intelligibility is forecast by the metric, specially in the case of ZF. In order to obtain proper results from any of these measures, the maximum correlation lag has to be obtained, otherwise wrong measures are delivered.



Channel 1

ZF

MVDR

Fig. 8: Comparison of the Spectogram from the Output of the two used Beamformers and the Channel 1 of the system

In addition to these measures, for the sake of visual understanding, in Fig. 8 and 9 the spectograms and time-domain output of the beamformers are shown. From Fig. 8 it can be seen that the spectogram becomes *blurrier* but at the same time the structure of the speech is reinforced and better defined. In Fig. 9 the beamformed output shows a huge reduction in noise and almost perfectly match the speech signal shape (microphone noise and contributions from other directions are present).

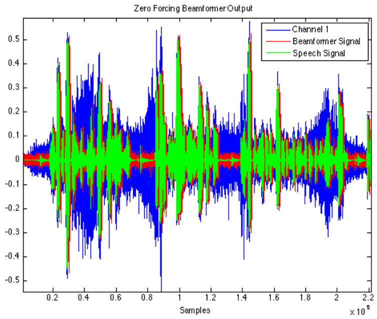


Fig. 9: Time Domain output for the Zero Forcing Beamforming

V. Adaptative Beamforming: Feedback Loop

One of the main problems while doing beamforming, particularly using MVDR, is the estimation of the covariance matrix. As discussed before, the covariance matrix (or its inverse) has to be estimated by means of a sliding window and a smoothing temporal filter. Due to the fact that speech and noise are not stationary in general, the estimation process of this matrix will be degraded.
In order to overcome this issue, the idea of adding feedback from the SCNR module towards the MVDR beam forming is proposed. As enough effort is already made to estimate the residual noise present in the single channel signal, the authors proposed to use knowledge of both estimated speech and noise PSD in order to improve the beamforming, updating online the covariance matrix with this information.

Lets consider the following signal model in the frequency domain for two sources S and N and two microphones:

$$X_{1,k} = S_k + N_k$$
$$X_{2,k} = \alpha_k(\theta_s)S_k + \alpha(\theta_n)N_k$$

where $\alpha_k(\theta) := exp(-jk\tau(\theta)$ is the related phase shift for the frequency k.

Now, the data model can be compactly written as

$$\mathbf{X}_k = \mathbf{A}_{k,\Phi}\mathbf{I}_k$$

where $\mathbf{I}_k$ is a vector containing the sources at frequency k and $\mathbf{A}_{k,\Phi}$ contains the array response for the frequency k and directions $\Phi = [\theta_s \ \theta_n]$.

The covariance matrix for the measurements $\mathbf{X}_k$ is then given by

$$\mathbf{R}_{XX,k} = \begin{bmatrix} \sigma^2_{s,k} + \sigma^2_{n,k} & \alpha_k(\theta_s)^*\sigma^2_{s,k} + \alpha_k(\theta_n)^*\sigma^2_{n,k} \\ \alpha_k(\theta_s)\sigma^2_{s,k} + \alpha_k(\theta_n)\sigma^2_{n,k} & \sigma^2_{s,k} + \sigma^2_{n,k} \end{bmatrix}$$

and by assuming independence between speech and noise FFT coefficients

$$\mathbf{R}_{XX,k} = \mathbf{R}_{SS,k} + \mathbf{R}_{NN,k}$$

with

$$\mathbf{R}_{NN,k} = \begin{bmatrix} \sigma^2_{n,k} & \alpha_k(\theta_n)^*\sigma^2_{n,k} \\ \alpha_k(\theta_n)\sigma^2_{n,k} & \sigma^2_{n,k} \end{bmatrix}$$

Now, it is possible to exploit the information retrieved from the SCNR module in order to estimate and update the noise covariance matrix online, to then use it for the MVDR beamformer.

Recalling from the output of the MVDR beam former

$$Z_k = S_k + w_k^H[1, \ \alpha_k(\theta_n)]^T N_k$$
$$Z_k = S_k + n_k$$

where n_k is the residual noise present after the beam former is applied, which is nothing more than a scaled version of N_k.

Under the assumption of additive noise, the SCNR module computes an estimate $|\hat{n}_k|^2$ (through the SPP approach described before), which implies that

$$|\hat{N}_k|^2 = \frac{|\hat{n}_k|^2}{|w_k^{(0)} + w_k^{(1)}\alpha_k(\theta_n)|^2}$$

Finally, the parameter $\sigma^2_{n,k}$ can be taken to be our compensated estimate $|\hat{N}_k|^2$ by using the angle at which the null is of the array is going to be stirred. However, if this angle is not known an estimate for $\alpha_k(\theta_n)$ should be made. For this matter a proposed estimate, if no more information is available, could be the mean angular response of the array at the given frequency k.

The objective measures of the *feedbacked* MVDR beamformer are shown in Table V.

	SEG-SNR	STOI	PESQ
Wiener	-0.3723	0.6222	1.9630
DD	-0.7083	0.6176	1.9491
TCS	-0.5475	0.6222	1.9630

TABLE V: Evaluation measures results from Adaptative MVDR

Comparing these values with the ones from the tables in the previous section, an increase in the SEG-SNR and PESQ while a reduction in predicted intelligibility due to lower STOI can be observed. In the subjective evaluation it was found that even though the result presents a clearer speech, new artifacts appear which are considered *bothersome*. A similar situation as with musical noise, now the output of the system gives a speech signal polluted with a kind of *electrical* noise, chirp-like sounds. This problem makes the listening process annoying for some of the listeners, hence reducing its pleasantness.

VI. Conclusion

In this report we discussed the design of a single channel noise reduction system consisting of different processing blocks (e.g. gain, noise estimation etc.). Three methods of speech PSD estimation were discussed and a comparison was made. From the evaluation we've seen that the TCS method has best performance. From listening tests done by the authors we conclude that the output of the model with TCS is most pleasant. The last part of the report consists of the evaluation of the beamformer with an array of two microphones. In the evaluation we've seen that the zero-forcing beamformer has best performance. We tried to improve the results of the MVDR beamformer by making it adaptive. An improvement of PESQ score was observed but eventually the pleasantness (and seg-SNR + STOI score) degraded.

References

[1] R. Hendriks, T. Gerkmann and J. Jensen *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement"*, Synthesis Lectures on Speech and Audio Processing, 2013

[2] T. Gerkmann and R. Hendriks *"Unbiased MMSE-based noise power estimation with low complexity and low tracking delay"*, IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 4, pp. 13831393, May 2012.

[3] T.Gerkmann, M.Krawczyk, and R.Martin *Speech presence probability estimation based on temporal cepstrum smoothing* IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Dallas, TX, USA, Mar. 2010, pp. 42544257

[4] T.Gerkmann and R.Martin *On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling* IEEE Trans. Signal Process., vol. 57, no. 11, pp. 41654174, Nov. 2009

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen *An algorithm for intelligibility prediction of time-frequency weighted noisy speech* IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 7, pp. 21252136, Sep. 2011.

[6] S. Gazor and W. Zhang *Speech enhancement employing Laplacian-Gaussian mixture* IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 896904, Sep. 2005.

[7] I. Cohen *Relaxed statistical model for speech enhancement and a priori SNR estimation* IEEE Trans. Speech Audio Process., vol. 13, no. 5, pp. 870881, Sep. 2005.

[8] Oppenheim, Alan V.; Schafer, Ronald W. *Digital signal processing* Englewood Cliffs, N.J.: Prentice-Hall, 1975

[9] J.S.LimandA.V.Oppenheim *Enhancement and band width compression of noisy speech* Proc. of the IEEE, vol. 67, no. 12, pp. 15861604, Dec. 1979.

[10] C.Breithaupt,T.Gerkmann,and R.Martin *A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing* IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Las Vegas, NV, USA, Apr. 2008, pp. 48974900.

[11] H. L. Van Trees *Detection, Estimation, and Modulation Theory: Part I.* New York, NY, USA: John Wiley & Sons, 1968.

[12] P.C.Loizou *Speech quality assessment* in Multimedia Analysis,Processing and Communications, Lin et al., Eds. Springer Verlag, 2011, vol. 346, pp. 623654

[13] ITU-T *Perceptual evaluation of speech quality (PESQ)* ITU-T Recommendation P.862, 2001

[14] Loizou, P. *Speech Enhancement: Theory and Practice* CRC Press, Boca Raton: FL. 2007

[15] Alton Everest and Ken C. Pohlmann, *Master Handbook of Acoustics* McGraw Hill, 5th edition, 2009

[16] Y. Ephraim and D. Malah, *Speech enhancement using a minimum mean-square error shorttime spectral amplitude estimator*, IEEE Trans. Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109-1121, Dec. 1984.