SOME EXPERIMENTS ON SHORT-TIME SPECTRAL ATTENUATION (STSA) ALGORITHMS AND SPEECH INTELLIGIBILITY.

[1]*Joerg Bitzer,* [2]*Uwe Simmer,* [3]*Inga Holube and* [4]*Timm Schaer*

[1,2,3,4]Institute of Hearing Technology and Audiology (IHA), University of Applied Science Oldenburg, Ofener Str. 16, 26121 Oldenburg, Germany
[joerg.bitzer, uwe.simmer, inga.holube]@fh-oldenburg.de

ABSTRACT

Short time spectral attenuation (STSA) is the most commonly used technique for single channel noise reduction. It reduces the background noise and thus increases the signal-to-noise ratio (SNR). However, only few reports are dealing with formal tests of speech intelligibility measured in percent (SIP) for denoised signals and if they do, the results indicate no improvement at all. In this contribution we will show that those dissatisfying results are not based on the fundamental principle of STSA. Introducing a-priori knowledge can increase the SIP significantly. However, all real-word algorithms have to rely on estimates of the signal components. Therefore, we tested the estimation procedure and the limiting factors. Finally, some encouraging experiments with multi-channel estimators will be presented.

1. INTRODUCTION

Short time spectral attenuation (STSA) is the most commonly used technique for noise reduction. This very broad class contains i.a. Wiener-Filtering, spectral subtraction [1] and many multi-channel post-filter algorithms [2]. In many studies and research papers the performance is measured in terms of signal-to-noise ratio (SNR) enhancement or in terms of speech degradation like Log-Area-Ratio Distance. Many papers show significant improvement of speech quality. However, an improvement of speech intelligibility has not been reported. One reason for this missing information is the great burden to estimate speech intelligibility in formal tests with statistical significance. Also, some authors suggest that no improvement at all can be achieved by using single-channel noise reduction techniques [3, 4].
The remaining question is whether this behaviour is based on the algorithm design or if it is a fundamental limitation. And if there is no real limit, what are limitation factors for real world algorithms. In order to investigate these questions, we build a very basic test setup, where the algorithms have to enhance heavily disturbed speech signals, but additionally have access to a-priori information of the desired signal in order to estimate the needed information like the power spectral density (PSD) of the undisturbed speech signal. Finally a formal speech intelligibility test is performed to measure the performance of the algorithms.

2. EXPERIMENTS

All algorithms described in this paper are based on the same baseline system for frequency domain processing. The input signals are divided into small overlapping blocks (block length = 1024 samples, overlap 50%, sampling rate = 44.1kHz). Each block is weighted with the square-root of the Hann-Window and afterwards zero-padded. A standard FFT-algorithm is used to transform the signals into the frequency domain. The complex spectrum is multiplied by the real valued transfer function of the time-varying filter and the processed spectrum is transformed back into the time domain. Finally, the standard overlap-add system with weighted overlapped blocks is applied. The weighting window is again the square-root of the Hann-Window. This procedure guarantees transparent output signals, if there is no signal processing and it prevents the signal from being disturbed by small clicks introduced by cyclic convolution, if the applied transfer function is not constrained [5].

2.1. Signal Model

The used signal model is a weighted addition of the undisturbed signal $s(k)$ and the noise signal n(k).

$$x(k) = g\, s(k) + n(k) \tag{1}$$

The weighting factor g is determined by the desired signal-to-noise ratio (SNR) and both signals are uncorrelated.
All signals are transformed individually to introduce a-priori knowledge to the different algorithms.
For the multi-channel experiments the signal model is

$$x_i(k) = g\, s(k) + n_i(k) \tag{2}$$

where $n_i(k)$ are uncorrelated noise sources. Therefore, the assumed noise field is spatially white.

2.2. Estimation

In this contribution only algorithms using short-time spectral attenuation based on power spectral densities (PSD) are considered. The PSD estimation procedure is based on the recursive Welch-periodogram, which is the standard technique for speech signal processing.

$$\Phi_{XX}(n,\ell) = \alpha_{(.)}\Phi_{XX}(n,\ell-1) + (1-\alpha_{(.)})|X(n,\ell)|^2, \quad (3)$$

where $X(n,k)$ is the spectrum of the signal x(k) with block-index ℓ and frequency index n. $\alpha_{(.)}$ is a smoothing constant. Equation 3 is used to estimate the PSD of the speech signal $\Phi_{SS}(n,\ell)$ and of the noise signal $\Phi_{NN}(n,\ell)$. The smoothing constants α_S and α_N are set individually.

2.3. Tested Algorithms

We tested the following basic algorithms.

- Wiener-Filter:

$$H(n,\ell) = \frac{\Phi_{SS}(n,\ell)}{\Phi_{SS}(n,\ell) + \Phi_{NN}(n,\ell)} \quad (4)$$

- Spectral subtraction without any modifications:

$$H(n,\ell) = \frac{\Phi_{XX}(n,\ell) - \Phi_{NN}(n,\ell)}{\Phi_{XX}(n,\ell)} \quad (5)$$

- Ephraim and Malah (EM) logSTSA noise suppressor [6]

- For comparison purposes the delay and sum beamformer (D&S), which is the optimal Minimum Variance Distortionless Response (MVDR)-Beamformer for the given uncorrelated noise field[7].

- A multi-channel post-filter given by [2]

$$H_{PF}(n,\ell) = \frac{\Phi_{Y_bY_b}(n,\ell)}{\Phi_{X_1X_1}(n,\ell)} \quad (6)$$

  where $\Phi_{X_1X_1}$ is the PSD of one input channel and $\Phi_{Y_bY_b}$ is the output of the beamformer. This Post-Filter (PF) is applied to a single input channel and not to the beamformer output, in oder to be comparable to the other algorithms.

2.4. Speech Intelligibility Measurement

The measurement of SIP can be done by asking listeners how many and which words they understand in a nonsense sentence. The Oldenburger speech sentence test is based on this procedure [8]. It ist based on a closed set of 50 words connected to sentences, which may have no meaning. The order of the words is always a name, a verb, a number, an adjective and a noun. The noise is designed

Figure 1: Graphical User Interface for a speech sentence test with a closed set of words.

to have exactly the same long-term spectral density as the speech signal. It is generated by mixing hundreds of the speech signals with different starting points. Normally, these tests are designed to find the SNR at which 50% of the words are correctly identified (Speech Reception Threshold (SRT)). We are using a slightly different version of the test in order to measure the percentages of correct words, and we are not using any operator. See figure 1 for the Graphical User Interface (GUI). The words are in German, since our five test persons are German students. All of them have normal hearing capabilities, checked with a tonal audiometry test.
The mixed and enhanced signals are provided via headphones (closed system , Sennheiser HDA 200). The headphone is calibrated to 65 dB SPL (Sound Pressure Level) for the noise. The speech signal gets softer for negative SNRs, starting at -3dB.

2.5. Tested Situations

The following settings for the algorithms have been tested:

1. Test: For the very basic test we used full a-priori knowledge ($\Phi_{SS}(n,\ell)$ is known) and the Wiener-Filter. The smoothing constants are set to zero. Therefore, all PSDs are estimated by simple periodograms.

2. Test: In this test the basic spectral subtraction algorithm is tested with all smoothing constants set to zero

3. Test: Since the Wiener-Filter depends heavily on a-priori knowledge we decided to perform the next test with spectral subtraction only. We varied the smoothing constants α_N from zero to close to one, whereas α_X was set to zero. The noise estimate $\Phi_{NN}(n,\ell)$ is based on the pure noise signal $n(k)$ and is not estimated from the disturbed speech signal $x(k)$. The SNR was set to -12dB, which means

that unprocessed signals are not intelligible (see figure 3).

4. Test: For comparison purposes we included a well tuned Ephraim and Malah algorithm [6]. For noise estimation the pure noise signal with a high smoothing factor ($\alpha_N = 0.96$) was used. α_X was set to 0.4.

5. Test: The D&S was implemented in the time domain and it is no STSA algorithms. We used four channels. In this case, the D&S will reduce the noise level by 6 dB by averaging the uncorrelated noise without introducing any artefacts.

6. Test: Finally, the Post-Filter (PF) was estimated without any a-priori knowledge with time constants set to zero.

3. RESULTS AND CONCLUSIONS

The following results and conclusions can be given for the different tests:

1. The results for the unprocessed speech are given in figure 3. You can see the results (diamonds) for five input SNRs and the psycho-metric function [8] which was computed to fit for the five measurements. One important result is, that speech intelligibility is a very steep function, which starts at -15dB and ends at -5dB for normal hearing persons. All algorithms have to deal with negative input SNRs.

2. The results for the Wiener-Filter are 100% speech intelligibility for all tested SNRs down to -24dB. This result can be explained by the fact that Φ_{SS} is known a-priori. From -9dB down the mixed signal is more or less pure noise. However, the Wiener-filter based on the true speech signal shapes this noise into the corresponding spectrum of the speech and attenuates the whole signal. The result is some artificial whisper speech, without any tonal components. This kind of speech is completely intelligible for the average person. A very similar signal can easily be constructed by randomizing the phase of a speech signal. This is known as whisperization as a special effect [9]. If the SNR gets very low, parts of the signal spectrum are getting below the hearing threshold in quiet, and eventually the SIP will decrease below 100%.

3. The results for the spectral subtraction algorithm are very similar as for the Wiener-Filter, because the estimate of the speech signal is very close to the one based on the speech signal directly. The speech intelligibility rate is 100% at all SNRs. However, the signals are not that pleasant and some musical tones are introduced. This behaviour can be explained by the mismatch between Φ_{NN} and the noise contained in Φ_{XX}, because of constructive and destructive interference between signal and noise in the time-domain mixing process.

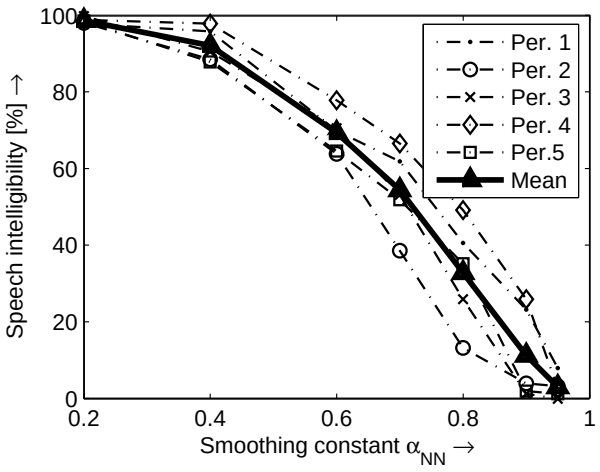

Figure 2: Speech intelligibility in percent for varying α_N (spectral subtraction, SNR = -12dB, $\alpha_X = 0$)

4. Figure 2 shows the SIP vs. smoothing constant α_N for five individuals and the resulting average. It can be seen clearly that increasing α_N decreases the SIP. This can be explained by the increasing discrepancy between the noise estimate Φ_{NN} based on averages and the noise contained in Φ_{XX} which is still a simple periodogram estimate. In order to verify this conclusion, we changed both α_X and α_N equally and found that the SIP is very close to 100% up to $\alpha_{N,X} = 0.95$. The resulting signal only sounded more reverberated.

5. For the EM no significant improvement in terms of speech intelligibility can be reached (see figure 3). This complies with the results reported in [Wit01].

6. The results for the D&S are shown in figure 3 on the left side. The SNR shift compared to the unprocessed signal is given by the 6dB SNR-enhancement of the D&S with no additional artefacts in the output signal.

7. The post-filter applied to a single input channel gives comparable performance to the D&S (see figure 3 crosses), which is interesting since it has introduced typical artefacts of STSA-algorithms. These results indicate that STSA algorithms based on multi-channel systems are promising candidates for future extensions in order to overcome the problem of noise estimation.

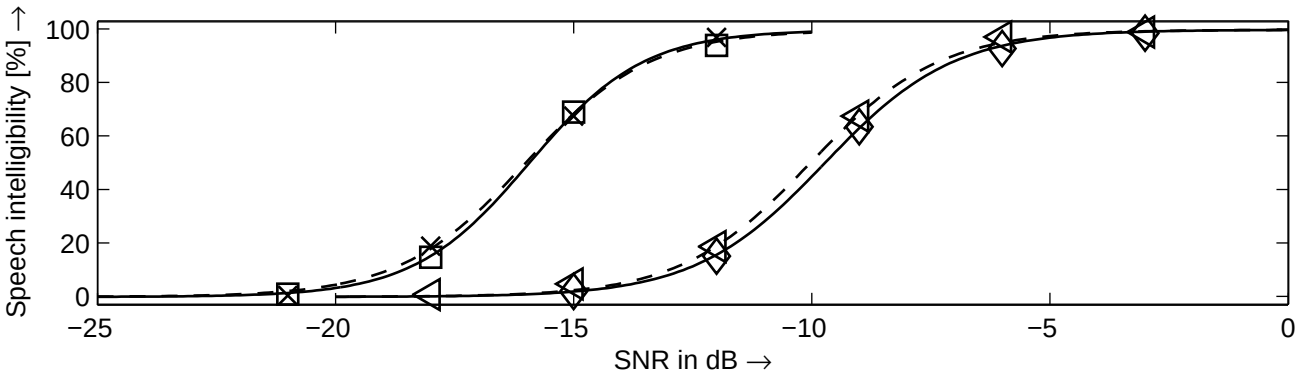

Figure 3: Speech intelligibility in percent for different algorithms ($\diamond$ = Unprocessed, $\triangleleft$= Ephraim and Malah $\square$ =, Delay and Sum, x = PF(1chn))

Some meaningful examples will be provided on the homepage of the Institute of Hearing-technology and Audiology (IHA).

4. FINAL CONCLUSIONS

In this paper we have shown that STSA algorithms are able to improve speech intelligibility (SI) in principal. However, the underlying estimate of the PSDs is the critical aspect. Every deviation from the exact noise estimate to the noise estimated in the disturbed signal will decrease the performance significantly. This deviation can be introduced by smoothing in the time or frequency domain or other estimation errors. Therefore, known single-channel algorithms will fail to improve SI in real world scenarios, where the noise has to be estimated in speech pauses or by using clever smoothing techniques. In contrast, it is possible to improve SI, if multi-channel based algorithms are used to estimate the filter transfer functions. However, very first results suggest that the combination of the beamformer output with the post-filter will not improve the SI further. We have no explanation for this behaviour so far.

5. ACKNOWLEDGEMENT

Listen to nonsense sentences near SRT is very stressful, therefore, we thank all students for their willingness to take part in this laborious test procedure.

6. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Washington DC, Apr. 1979, pp. 208–211.

[2] K. U. Simmer, J. Bitzer (Meyer), and C. Marro, "Post-filtering Techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 3, pp. 39–57. Springer, Berlin, Heidelberg, New York, May 2001.

[3] M. Marzinzik, *Digital Hearing Aids and their use for the Hearing Impaired*, Ph.D. thesis, University of Oldenburg, 2000.

[4] T. Wittkop, *Two-channel noise reduction algorithm motivated by models of binaural interaction*, Ph.D. thesis, University of Oldenburg, 2001.

[5] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 28, no. 1, pp. 99–102, 1980.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error log spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[7] J. Bitzer (Meyer) and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds., chapter 2, pp. 19–37. Springer, Berlin, Heidelberg, New York, May 2001.

[8] K. Wagener, *Factors Influencing Sentence Intelligibility in Noise*, Ph.D. thesis, University of Oldenburg, 2003.

[9] U. Zoelzer, *DAFX - Digital Audio Effects*, John Wiley and Son, 2002.