NEAR END LISTENING ENHANCEMENT: SPEECH INTELLIGIBILITY IMPROVEMENT IN NOISY ENVIRONMENTS

Bastian Sauert and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{sauert,vary}@ind.rwth-aachen.de

ABSTRACT

In contrast to common noise reduction systems, this contribution presents a digital signal processing algorithm to improve intelligibility of clean *far end* speech for the *near end* listener who is located in an environment with background noise. Since the noise reaches the ears of the *near end* listener directly and therefore can hardly be influenced, a sensible option is to manipulate the *far end* speech. The proposed algorithm raises the average speech spectrum over the average noise spectrum and takes precautions to prevent hearing damage. Informal listening tests and the Speech Intelligibility Index indicate an improved speech intelligibility.

1. INTRODUCTION

Telephone conversations over cellular phones often take place in the presence of acoustical background noise. In such a situation the speech intelligibility is reduced for both, the *far end* and the *near end* listener. Several preprocessing algorithms have been proposed to reduce the noise in the *near end* microphone signal before speech coding and transmission in order to regain speech intelligibility for the *far end* listener, e. g., [1]. However, only few solutions have been proposed to improve intelligibility of the *far end* speech for the *near end* listener.

For the problem of *near end* listening enhancement, as opposed to the problem of noise reduction, the noise signal can not be influenced because the person is located in the noisy environment and the noise reaches the ears without any possibility to intercept. Therefore a sensible option to improve intelligibility by digital signal processing is to manipulate the *far end* speech signal.

In the 1960's and 1970's some research has been done on this topic, e. g., [2]. Niederjohn et al. proposed in, e. g., [3] and [4] a high pass filtering to enhance the higher formants followed by a rapid amplitude compression to defend white noise and a power generating noise environment, respectively.

In this work a time adaptive and frequency dependent signal-to-noise ratio (SNR) recovery approach is presented combined with a limitation of the spectral amplitudes to prevent hearing damage and overload of sound equipment.

Fig. 1 shows the spectograms of a piece of speech from the TIMIT database and of destroyer engine noise from the NOISEX-92 database at an SNR of $-5\,\mathrm{dB}$. The sum of both as it would occur in the noisy environment is plotted in Fig. 1c. The speech signal is mostly covered by the noise and can hardly be identified.

2. SNR RECOVERY

A speech signal which is presented in a noisy environment is less intelligible than the same signal presented in a quiet environment due to the fact that the spectral distance between the speech signal and the noise signal is reduced.

The fundamental idea of this work is a (frequency dependent) amplification of the speech signal to reestablish the distance between the average measured speech spectrum and the average measured noise spectrum, i. e., to recover a certain signal-to-noise ratio.

2.1. Frequency Independent SNR Recovery

A first approach is to amplify the speech signal $s(k)$ in time-domain with a gain $g(k)$,

$$\hat{s}(k) = g(k) \cdot s(k)\,, \tag{1}$$

in a way that the SNR of the amplified speech signal $\hat{s}(k)$ and the environmental noise $n(k)$ is greater than or equal to a target SNR ξ:

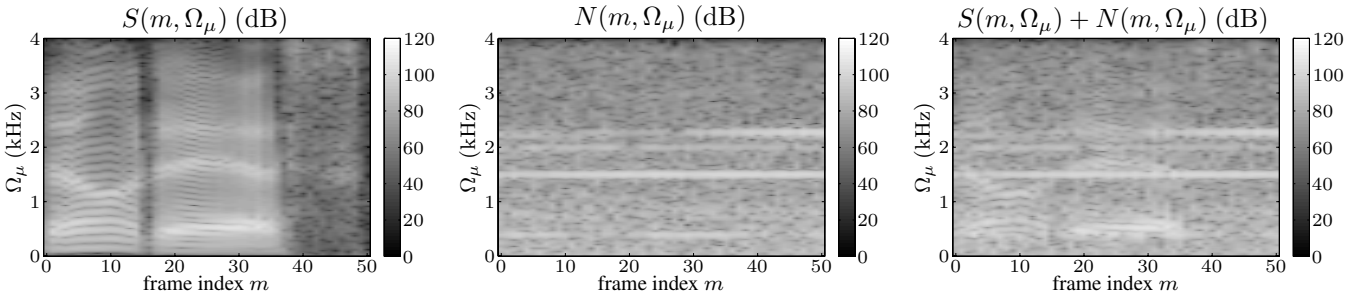
$$\frac{E\{\hat{s}^2(k)\}}{E\{n^2(k)\}} = \frac{E\{(g(k)\cdot s(k))^2\}}{E\{n^2(k)\}} \geq \xi\,. \tag{2}$$

The target SNR can, for example, be chosen to $\xi \mathrel{\hat{=}} 15\,\mathrm{dB}$. $E\{\bullet\}$ denotes the short-term expectation value of $\bullet$ and is discussed further in Section 2.3.

Since $g(k)$ is deterministic and not a random variable, (2) can be written as

$$\frac{E\{(g(k)\cdot s(k))^2\}}{E\{n^2(k)\}} = \frac{g^2(k)\cdot E\{s^2(k)\}}{E\{n^2(k)\}} \geq \xi\,, \tag{3}$$

This work was funded by Siemens AG, Munich, Germany.



(a) Clean unprocessed speech from speaker DSC0 saying "Don't ask".

(b) Destroyer engine noise.

(c) Noisy unprocessed speech from (a) with noise from (b).

Fig. 1. Spectograms of speech and of noise without further processing.

which leads to the first constraint on the gain $g(k)$:

$$g(k) \geq \sqrt{\xi \cdot \frac{E\{n^2(k)\}}{E\{s^2(k)\}}}\,. \tag{4}$$

Furthermore the algorithm should not modify the speech signal $s(k)$ if there is no or only negligible background noise. This is accomplished by introducing a second constraint

$$g(k) \geq 1\,, \tag{5}$$

which comes into effect if $\frac{E\{s^2(k)\}}{E\{n^2(k)\}} \geq \xi$. Choosing the smallest allowed gain, the combination of (4) and (5) leads to

$$g(k) = \max\left\{\sqrt{\xi \cdot \frac{E\{n^2(k)\}}{E\{s^2(k)\}}}, 1\right\}. \tag{6}$$

The frequency independent SNR recovery approach gives good results for white noise. But the disturbed frequency components may still be hidden under narrow band noise whereas frequency components with low or no noise are over-amplified. In order to overcome this problem a frequency dependent solution is proposed.

2.2. Frequency Dependent SNR Recovery

The speech signal $s(k)$ and the noise signal $n(k)$ are divided into half-overlapping blocks of $20\,\mathrm{ms}$ length, which are denoted with the frame index m. Each frame is multiplied with a Hann window and transformed to the frequency-domain representations $S(m,\Omega_\mu)$ and $N(m,\Omega_\mu)$ using the discrete Fourier transform (DFT), where Ω_μ is the discrete DFT frequency and μ is the frequency index. Afterwards the speech signal is amplified with a gain $G(m,\Omega_\mu)$:

$$\hat{S}(m,\Omega_\mu) = G(m,\Omega_\mu) \cdot S(m,\Omega_\mu)\,. \tag{7}$$

Finally the amplified speech coefficients $\hat{S}(m,\Omega_\mu)$ are transformed back to time-domain using the inverse DFT and reassembled with the overlap-add technique.

Analogously to (2) the ratio of the short-term power spectral density (PSD) of the amplified speech $\Phi_{\hat{S}\hat{S}}(m,\Omega_\mu)$ and the short-term PSD of the noise signal $\Phi_{NN}(m,\Omega_\mu)$ should be greater or equal to a target SNR ξ:

$$\frac{\Phi_{\hat{S}\hat{S}}(m,\Omega_\mu)}{\Phi_{NN}(m,\Omega_\mu)} \geq \xi\,. \tag{8}$$

Since $G(m,\Omega_\mu)$ is deterministic, (8) can be written as

$$\frac{\Phi_{\hat{S}\hat{S}}(m,\Omega_\mu)}{\Phi_{NN}(m,\Omega_\mu)} = \frac{G^2(m,\Omega_\mu)\cdot\Phi_{SS}(m,\Omega_\mu)}{\Phi_{NN}(m,\Omega_\mu)} \geq \xi\,, \tag{9}$$

which leads to the first constraint on the gain $G(m,\Omega_\mu)$:

$$G(m,\Omega_\mu) \geq \sqrt{\xi \cdot \frac{\Phi_{NN}(m,\Omega_\mu)}{\Phi_{SS}(m,\Omega_\mu)}}\,. \tag{10}$$

The second constraint guarantees that the speech signal is not attenuated in a noise-free environment and results in

$$G(m,\Omega_\mu) \geq 1\,. \tag{11}$$

This leads to the smallest allowed gain

$$G(m,\Omega_\mu) = \max\left\{\sqrt{\xi \cdot \frac{\Phi_{NN}(m,\Omega_\mu)}{\Phi_{SS}(m,\Omega_\mu)}}, 1\right\}. \tag{12}$$

The speech signal is weighted according to the spectral characteristics of the noise signal and thereby accounts for non-white noise environments. However, the solution over-amplifies low speech signal components since it tries to raise anything over the noise by the same amount ξ independent of the original signal strength. This effect can be reduced by limiting the gain $G(m,\Omega_\mu)$ to a maximum gain $G_{\max}$ with, e. g., $G_{\max} \mathrel{\hat{=}} 30\,\mathrm{dB}$. It finally follows that

$$G(m,\Omega_\mu) = \min\left\{\max\left\{\sqrt{\xi\frac{\Phi_{NN}(m,\Omega_\mu)}{\Phi_{SS}(m,\Omega_\mu)}}, 1\right\}, G_{\max}\right\}. \tag{13}$$

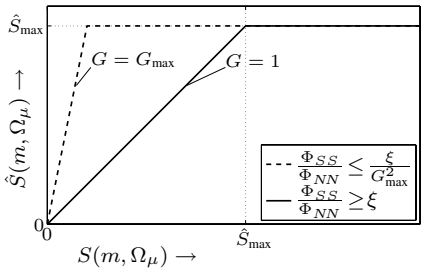
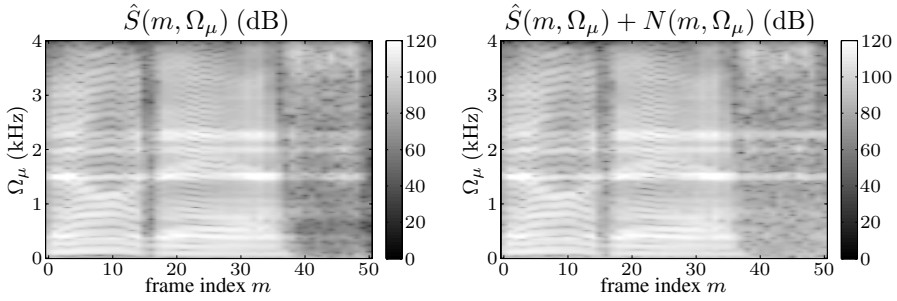


Fig. 2. Diagram of the effective gain $\hat{S}(m,\Omega_\mu)/S(m,\Omega_\mu)$ of the frequency dependent SNR recovery algorithm.



(a) Clean processed speech from speaker DSC0 saying "Don't ask".

(b) Noisy processed speech from (a) with noise from Fig. 1b.

Fig. 3. Spectograms after processing with frequency dependent SNR recovery algorithm (see also Fig. 1).

2.3. Short-Term PSD and Expectation Value

The short-term PSD $\Phi_{SS}(m,\Omega_\mu)$, which was introduced in Section 2.2, is computed as the recursive average of the periodogram $|S(m,\Omega_\mu)|^2$:

$$\begin{aligned}\Phi_{SS}(m,\Omega_\mu) = \alpha_S \cdot \Phi_{SS}(m-1,\Omega_\mu) \\ + (1-\alpha_S) \cdot |S(m,\Omega_\mu)|^2 ,\end{aligned} \tag{14}$$

where $\alpha_S \in [0,1]$ is the time constant of the recursive average. $\Phi_{NN}(m,\Omega_\mu)$ is defined analogously to (14) with the time constant α_N.

The choice of the time constants α_S and α_N is crucial for the performance of the algorithm. In the following the two cases $\alpha_S = \alpha_N = 0$ and $\alpha_S = \alpha_N = 1$ will be discussed:

With $\alpha_S = \alpha_N = 0$ it follows from (14) that

$$\Phi_{SS}(m,\Omega_\mu) = |S(m,\Omega_\mu)|^2 \text{ and} \tag{15}$$

$$\Phi_{NN}(m,\Omega_\mu) = |N(m,\Omega_\mu)|^2 . \tag{16}$$

Assuming that $\frac{\Phi_{NN}(m,\Omega_\mu)}{\Phi_{SS}(m,\Omega_\mu)}$ is such that neither of the two limitations in (13) takes effect, it follows from (7) and (13) that

$$\hat{S}(m,\Omega_\mu) = \sqrt{\xi} \cdot |N(m,\Omega_\mu)| \cdot \frac{S(m,\Omega_\mu)}{|S(m,\Omega_\mu)|} . \tag{17}$$

The amplified speech preserves only the phase from the original speech signal, but gets the amplitude of the noise.

If α_S and α_N are equal or very close to one, $\Phi_{SS}(m,\Omega_\mu)$ and $\Phi_{NN}(m,\Omega_\mu)$ basically stays constant independent of the actual signals. Thereby the system could not react on changing speech and noise situations.

Setting the time constants to $\alpha_S = 0.996$ and $\alpha_N = 0.96$ turned out to be a reasonable compromise. They bring the speech on average into line with the noise characteristics and at the same time guarantee adaption to changing situations.

A similar discussion can be done for the short-term expectation values $E\{s^2(k)\}$ and $E\{n^2(k)\}$, which are also computed as the recursive average of $s^2(k)$ and $n^2(k)$, respectively.

3. LIMITING OUTPUT

In most cases there are at least two reasons to limit the output power. Firstly, most sound production systems, i. e., amplifier and speaker, can only transfer limited output power without being damaged. Secondly, only limited output power may be presented to the human ear without risking pain and hearing loss.

In order to restrict the output the amplitude of all DFT coefficients is limited to a maximum DFT amplitude $\hat{S}_{\max}(\Omega_\mu)$ while preserving the phase:

$$\hat{S}'(m,\Omega_\mu) = \begin{cases} \hat{S}_{\max}(\Omega_\mu) \frac{\hat{S}(m,\Omega_\mu)}{|\hat{S}(m,\Omega_\mu)|} \\ \qquad\qquad \text{if } |\hat{S}(m,\Omega_\mu)| > \hat{S}_{\max}(\Omega_\mu) . \\ \hat{S}(m,\Omega_\mu) \qquad \text{otherwise} \end{cases} \tag{18}$$

Note that in general the threshold of pain of the human ear is not constant over frequency. However, it is sufficiently flat in the range of interest between $100\,\mathrm{Hz}$ and $4\,\mathrm{kHz}$ to use a frequency independent maximum DFT amplitude.

If the amplitudes of the DFT coefficients are restricted, mainly the periodic structure of the speech spectrum at the first formant is flattened. This results in a rougher and less voiced speech but should not effect speech intelligibility too much [5].

The resulting effective gain $\frac{\hat{S}(m,\Omega_\mu)}{S(m,\Omega_\mu)}$ of the frequency dependent SNR recovery algorithm is sketched in Fig. 2.

4. RESULTS

Fig. 3 depicts the same situation as Fig. 1 after processing with the frequency dependent SNR recovery approach. The speech spectrum can clearly be identified in Fig. 3b as opposed to Fig. 1c and is not covered in general.

The spectral shape of the speech follows in average the shape of the noise. This effect of the optimization criterion was audible in the listening tests but not annoying.

4.1. Parameter Choice

The maximum DFT amplitude $\hat{S}_{\max}(\Omega_\mu)$ can, for example, be experimentally determined so that no unpleasantly loud signal peaks occur even during very loud signal segments. During informal listening tests a maximum DFT amplitude of $\hat{S}_{\max}(\Omega_\mu) \mathrel{\hat{=}} 120\,\mathrm{dB}$ in relation to the step size of a 16 bit quantizer turned out to be appropriate at our sound reproduction system. A target SNR of $\xi \mathrel{\hat{=}} 15\,\mathrm{dB}$ resulted in a highly increased intelligibility depending on the input SNR. Besides, the algorithm was not restricted due to $\hat{S}_{\max}(\Omega_\mu)$ in most common speech and noise situations. Over-amplification of low speech components was prevented with a maximum gain of $G_{\max} \mathrel{\hat{=}} 30\,\mathrm{dB}$ without reducing the overall effect. These settings were used for all processings in this section.

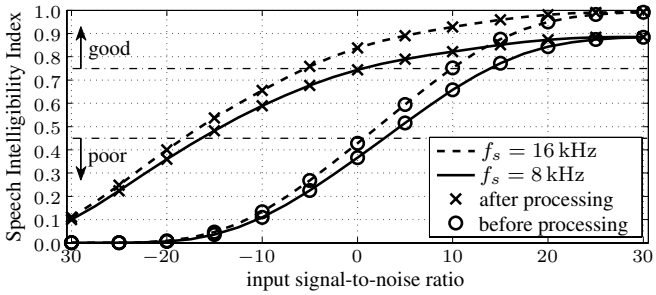
4.2. Speech Intelligibility Index

The performance of the proposed algorithm was evaluated in terms of the Speech Intelligibility Index (SII) as defined in [6]. The SII is supposed to be correlated with the intelligibility of speech under a variety of adverse listening conditions. It is computed by adding the speech-to-noise ratio in each contributing frequency band weighted according to its contribution to speech intelligibility. According to [6], good communication systems have an SII of 0.75 or above, while poor communication systems have an SII below 0.45 (see Fig. 4).

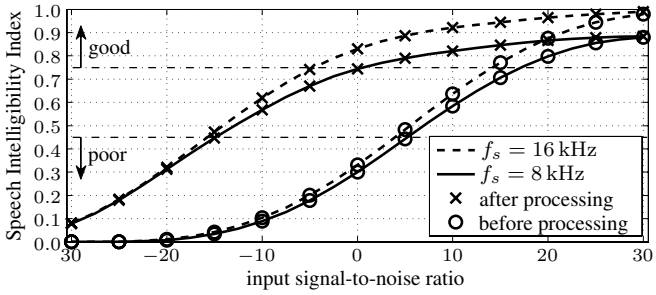
However, the SII does not account for any spectral fine structure or temporal envelope. Therefore it does not consider speech parameters like roughness or voiceness. This makes it more suitable for rating intelligibility of standard speech in different noise situations than for optimizing parameters of speech manipulation techniques like the proposed algorithm. Nevertheless, it is a good measure of speech intelligibility for a set of parameters that have been proven to sound well in informal listening tests in Section 4.1.

For this evaluation the SII was calculated for every speech file of the TIMIT database, in total 5.4 hours, together with destroyer engine noise and white noise from the NOISEX-92 database. The mean of the SIIs before and after processing with the frequency dependent SNR recovery algorithm is depicted in Fig. 4 for several overall signal-to-noise ratios at the sample rates $f_s = 8\,\mathrm{kHz}$ and $f_s = 16\,\mathrm{kHz}$. The proposed algorithm increases the SII by up to 0.5 or, conversely, keeps the same SII at a $15\,\mathrm{dB}$ to $20\,\mathrm{dB}$ lower input SNR.

Prior to any processing all speech files of the TIMIT database were amplified by $10\,\mathrm{dB}$ for adjustment to our sound reproduction system. Afterwards the SII was calculated with the critical band procedure. In order to calculate the speech and noise spectrum level of each sound file, the spectrum level is calculated for frames of $20\,\mathrm{ms}$ length, averaged in decibel-domain, and normalized to match the overall level. Thereby an average speech spectrum level of the TIMIT database was achieved which is comparable to the standard speech spectrum level for normal vocal effort specified in [6].



(a) Disturbed with destroyer engine noise.



(b) Disturbed with white noise.

Fig. 4. Speech Intelligibility Index before and after processing w. frequency dependent SNR recovery algorithm.

5. CONCLUSIONS

In this contribution we presented an efficient algorithm to enhance the speech intelligibility of clean speech presented in noisy environments. The algorithm raises the average speech spectrum over the average noise spectrum and thereby regains speech intelligibility. Precautions have been taken to avoid hearing damage by output signals which are too loud. However, if the listener is in a noise-free environment the algorithm keeps the speech signal as it is.

6. REFERENCES

[1] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] Jae S. Lim, *Speech Enhancement*, Prentice-Hall Signal Processing Series, 1983.

[3] Russell J. Niederjohn and James H. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," in *Proc. of ICASSP*, Aug. 1976, vol. 24, pp. 277–282.

[4] Russell J. Niederjohn and James H. Grotelueschen, "Speech intelligibility enhancement in a power generating noise environment," in *Proc. of ICASSP*, Aug. 1978, vol. 26, pp. 378–380.

[5] Ian B. Thomas, "The influence of first and second formants on the intelligibility of clipped speech," *Journal of the Audio Engineering Society*, vol. 16, no. 2, pp. 182–185, Apr. 1968.

[6] American National Standard, "Methods for the Calculation of the Speech Intelligibility Index," *ANSI S3.5-1997*, 1997.