Improving Speech Intelligibility in Noisy Environments by Near End Listening Enhancement

Bastian Sauert and Peter Vary

Institute of Communication Systems and Data Processing (ive), 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 improve speech quality 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. Additionally, the available information about the environmental noise at the ears is imprecise and in most telephony situations only one ear is covered by a loudspeaker of the handset, which

both make active noise control techniques ineffective. 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 against several noise environments.

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.

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. One obvious solution would be to raise the speech beyond the masking threshold

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

caused by the noise. A simpler solution is 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.

The speech signal s(k) and the noise signal n(k) are divided into half-overlapping blocks of 20 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 $G(m, \Omega_{\mu})$:

$$\hat{S}(m,\Omega_{\mu}) = G(m,\Omega_{\mu}) \cdot S(m,\Omega_{\mu}).$$
(1)

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

The gain should be chosen in a way that 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 environmental noise signal $\Phi_{NN}(m, \Omega_{\mu})$ is greater or equal to a target SNR ξ :

$$\frac{\Phi_{\hat{S}\hat{S}}(m,\Omega_{\mu})}{\Phi_{NN}(m,\Omega_{\mu})} \ge \xi.$$
(2)

Since $G(m, \Omega_{\mu})$ is deterministic, (2) can be written as

$$\frac{\Phi_{\hat{S}\hat{S}}(m,\Omega_{\mu})}{\Phi_{NN}(m,\Omega_{\mu})} = \frac{G^2(m,\Omega_{\mu}) \cdot \Phi_{SS}(m,\Omega_{\mu})}{\Phi_{NN}(m,\Omega_{\mu})} \ge \xi,$$
(3)

which leads to the first constraint on the gain:

$$G(m, \Omega_{\mu}) \ge \sqrt{\xi \cdot \frac{\Phi_{NN}(m, \Omega_{\mu})}{\Phi_{SS}(m, \Omega_{\mu})}}.$$
 (4)

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

$$G(m, \Omega_{\mu}) \ge 1. \tag{5}$$

ITG-Fachtagung Sprachkommunikation 2006

which comes into effect if $\frac{\Phi_{SS}(m,\Omega_{\mu})}{\Phi_{NN}(m,\Omega_{\mu})} \geq \xi$. Choosing the smallest allowed gain, this leads to the smallest allowed gain

$$G(m, \Omega_{\mu}) = \max\left\{\sqrt{\xi \cdot \frac{\Phi_{NN}(m, \Omega_{\mu})}{\Phi_{SS}(m, \Omega_{\mu})}}, 1\right\}.$$
 (6)

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., $G_{\text{max}} = 30 \text{ dB}$. It finally follows that

$$G(m, \Omega_{\mu}) = \begin{cases} 1 & \text{if } \sqrt{\xi \frac{\Phi_{NN}(m, \Omega_{\mu})}{\Phi_{SS}(m, \Omega_{\mu})}} < 1 \\ G_{\max} & \text{if } \sqrt{\xi \frac{\Phi_{NN}(m, \Omega_{\mu})}{\Phi_{SS}(m, \Omega_{\mu})}} > G_{\max} \\ \sqrt{\xi \frac{\Phi_{NN}(m, \Omega_{\mu})}{\Phi_{SS}(m, \Omega_{\mu})}} & \text{otherwise.} \end{cases}$$
(7)

2.1 Short-Term PSD and Expectation Value

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

$$\Phi_{SS}(m, \Omega_{\mu}) = \alpha_S \cdot \Phi_{SS}(m - 1, \Omega_{\mu}) + (1 - \alpha_S) \cdot |S(m, \Omega_{\mu})|^2,$$
(8)

where $\alpha_S \in [0,1]$ is the time constant of the recursive average. $\Phi_{NN}(m, \Omega_{\mu})$ is defined analogously to (8) 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. At first the two extreme cases $\alpha_S = \alpha_N = 0$ and $\alpha_S = \alpha_N = 1$ will be discussed:

With $\alpha_S = \alpha_N = 0$ it follows from (8) that

$$\Phi_{SS}(m,\Omega_{\mu}) = |S(m,\Omega_{\mu})|^2 \text{ and }$$
(9)

$$\Phi_{NN}(m,\Omega_{\mu}) = |N(m,\Omega_{\mu})|^2.$$
(10)

Assuming that $\frac{\Phi_{NN}(m,\Omega_{\mu})}{\Phi_{SS}(m,\Omega_{\mu})}$ is such that neither of the two limitations in (7) takes effect, it follows from (1) and (7) that

$$\hat{S}(m,\Omega_{\mu}) = \sqrt{\xi} \cdot |N(m,\Omega_{\mu})| \cdot \frac{S(m,\Omega_{\mu})}{|S(m,\Omega_{\mu})|} .$$
(11)

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.

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:

$$\hat{S}'(m,\Omega_{\mu}) = \begin{cases} \hat{S}_{\max}(\Omega_{\mu}) \frac{\hat{S}(m,\Omega_{\mu})}{|\hat{S}(m,\Omega_{\mu})|} \\ \text{if } \frac{|\hat{S}(m,\Omega_{\mu})|}{\hat{S}_{\max}(\Omega_{\mu})} > 1 \\ \hat{S}(m,\Omega_{\mu}) & \text{otherwise. (12)} \end{cases}$$

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 Hz and 4 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].

4. Results

As explained in Section 2.1, the spectral shape correction 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 $\hat{S}_{\max}(\Omega_{\mu}) = 120 \,\mathrm{dB}$ in relation to the least significant bit of a 16 bit representation turned out to be appropriate at our sound reproduction system. A target SNR of $\xi = 15 \,\mathrm{dB}$ 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 $G_{\text{max}} = 30 \, \text{dB}$ 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. 1).

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

ITG-Fachtagung Sprachkommunikation 2006

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. 1 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 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 ms length, averaged in decibeldomain, 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].

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.





0

10

20

30

-10

Fig. 1. Speech Intelligibility Index before and after processing with frequency dependent SNR recovery algorithm.

References

- [1] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error shorttime 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.