A Speech Enhancement Method for Improved Intelligibility in the Presence of an Ambient Noise

Margaret Lech and Glen Stanley Hawksworth

School of Electrical and Computer Systems Engineering,
RMIT University, Melbourne, Australia
margaret.lech@rmit.edu.au

Abstract

A speech enhancement method for the improvement of speech intelligibility in the presence of a background noise was developed and tested. The enhancement method is based on psychoacoustic criteria. The Auditory Masking Threshold (AMT) of noise was calculated and used to extract information about the spectral components of speech, which don't mask the noise and thus, are not audible. Only these inaudible components of speech were enhanced. Since the method modifies the spectral components of speech in an adaptive way, and only within a small number of selected frequency bands, the distortion of the original acoustic features of speech is minimal. Numerical simulations of the enhancement process were performed and tested using the subjective Modified Rhyme Test (MRT). The results of the MRT were analyzed statistically with respect to the subjects' scores and with respect to the word sets scores. A statistically significant improvement of speech intelligibility was observed.

1. Introduction

Listening to speech in noisy environments requires an increased level of concentration from the listener and reduces the intelligibility of the speech. The obstruction of speech by ambient noise is often encountered by mobile phone and hearing aid users in noisy environments, such as in cars, railway stations, airports, and motorcycle riders and plane pilots listening to speech through speakers mounted inside their helmets. Over the past three decades, many techniques have been developed for speech enhancement and ambient noise cancellation (Tsaukolas et al 1997), (Virag 1995, 1999). Most of these techniques concentrate on improvement of speech quality in terms of the signal–to-noise ratio (SNR) parameter. The SNR parameter provides an objective measure of speech quality, but it has a low correlation with the subjective perception of speech. Additionally, many of the SNR improvement techniques introduce annoying distortions in the enhanced signal, which are very hard to remove. This paper presents a perceptual approach based upon the calculation of the Auditory Masking Threshold (AMT), which minimizes the artifacts and produces high quality speech. The proposed speech enhancement technique assumes that the noise in not present in the transmitted speech. It is assumed that the noise mixes with the speech signal at the receiver end of the transmission channel. Thus, the enhancement problem, addressed here, is to change the transmitted speech characteristics in such a way that an addition of an ambient noise does not reduce the speech intelligibility below an acceptable level.

2. Speech enhancement method

The noisy speech samples $x_n[k]$ produced in the presence of an additive noise n[k] are given as

$$x_n[k] = x[k] + n[k] \tag{1}$$

The presented speech enhancement method aims to modify the spectral components of speech to produce an enhanced speech signal $\hat{x}[k]$ such that the noisy enhanced speech $\hat{x}_n[k]$ given as,

$$\hat{x}_n[k] = \hat{x}[k] + n[k] \tag{2}$$

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

has an improved intelligibility compared to the noisy speech $x_n[k]$. Figure 1 contains the flowchart of the proposed speech enhancement scheme.

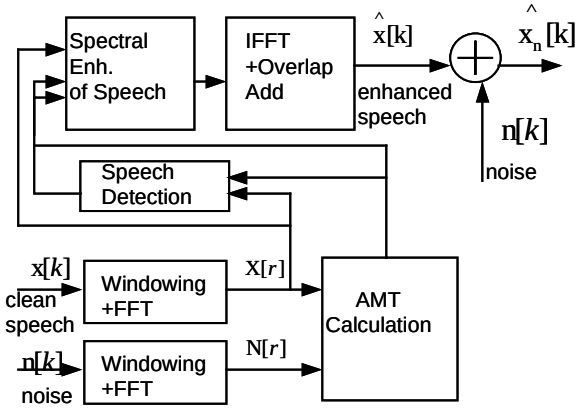

Figure 1: *Block diagram of the speech enhancement method; X[r] represent spectral samples of speech and N[r] represent spectral samples of noise.*

The speech processing was performed in the frequency domain on a frame-to-frame basis. For each frame, the Auditory Masking Threshold (AMT) was calculated and used to detect the speech activity, localize the enhancement bands and calculate the enhancement parameters.

2.1. Calculation of the short-time power spectra

The short time spectra N[r] and X[r] (r=0,1,...,255) of the noise and speech signal, respectively, were calculated using the FFT algorithm and the Hamming window with 50% overlap between frames. The power spectra $P_N[r]$ and $P_X[r]$ were then calculated for each frame using

$$P_N[r] = \mathrm{Re}(N[r])^2 + \mathrm{Im}(N[r])^2 \quad (3)$$

$$P_X[r] = \mathrm{Re}(X[r])^2 + \mathrm{Im}(X[r])^2 \quad (4)$$

and transformed to the logarithmic scale using

$$P_{NdB}[r] = 10\log_{10} P_N[r] \quad (5)$$

$$P_{XdB}[r] = 10\log_{10} P_X[r] \quad (6)$$

The sampling frequency of 16kHz was used for both the noise and the speech signal.

2.2. Calculation of the ERB power spectrum

The speech and noise power spectra for each frame were partitioned into the critical bands of the Equivalent Rectangular Bandwidth (ERB) scale (Moore et al 1996). The total signal energy within each of the critical bands was calculated using

$$B_{NdBi} = 10\log_{10}\left(\sum_{r=b_{li}}^{r=b_{hi}} P_N[r]\right) \quad (7)$$

$$B_{XdBi} = 10\log_{10}\left(\sum_{r=b_{li}}^{r=b_{hi}} P_X[r]\right) \quad (8)$$

where i=1,2,...,N_{ERB}, b_{li} is the lower boundary of the critical band i, b_{hi} is the upper boundary of the critical band i, and B_i is the total energy in the critical band i. The value of N_{ERB} represents the number of ERBs falling into the analyzed frequency range from 0 to 8kHz. The values B_i (i=1,2,...,N_{ERB}) are called the critical-band spectrum or the ERB power spectrum.

2.3. Calculation of the Auditory Masking Threshold

The short-time spectrum of the noise signal was used to calculate the Auditory Masking Threshold (AMT). The AMT of the noise is a power threshold given as a frequency domain function describing the ability of the reference signal to mask the noise. Since the speech signal was interpreted as the reference signal masking the noise, the frequency components of speech that fall above the AMT of the noise mask the noise, and thus are audible. The frequency components of speech that fall below the AMT of the noise don't mask the noise and are not audible. The AMT levels were calculated using the Johnston's method (Johnston 1988). The values of AMT were given as constant levels AMT_i (i=1,2,...N_{ERB}) corresponding to the more recent ERB bands (Moore et al 1996), whereas the Johnston's procedure calculates the AMT levels across the Bark spectral bands (Zwicker et al 1990). The blue stepping line in Figure 2 shows an example of the AMT for one frame of the speech signal; the green dashed line represents the power spectrum of speech in dB.

2.4. Speech detection

A noisy speech signal with SNR value above or equal to 5dB is intelligible (Gold et al 2000). Based upon this observation, a simple method of detecting the presence of speech was derived. It is clear that for speech to be intelligible, it would need to be audible. It was therefore assumed, that for a noisy speech with a SNR of 5dB, at least part of the clean speech power spectrum

Accepted after full paper review

must fall above the noise AMT level. Furthermore, if the clean speech component of the noisy speech at a SNR of 0dB was amplified to four times its power, the resulting SNR would be between 5dB and 6dB, and the resulting clean speech power spectrum should appear at least partially above the AMT of noise. Therefore, any frame of a noisy signal at 0dB SNR with the clean speech component amplified four times and not containing audible speech can be assumed to represent a non-speech interval. An amplification factor of four was used when the SNR of the original speech was 0dB. For different SNR values, an appropriate multiplication factor increasing the SNR value to at least 5dB would have to be determined. It was observed that during the unvoiced speech intervals, the highest power levels were occurring within the 3000 Hz to 6500 Hz range. During the voiced speech intervals, the power peaks moved towards the range 600-5000 Hz. Therefore, to speed up the process of speech detection, the monitoring range of frequencies was limited to the range from 600 Hz to 4800 Hz. This region of the speech spectrum plays an important role in speech intelligibility (Moore 1989).

2.5. Speech enhancement

The audible and non-audible components of the speech spectrum were determined by the comparison of the auditory masking threshold levels AMT_i (i=1,2,...N_{ERB}) of the noise with the power spectral levels P_{XdBi} of the clean speech, within a limited frequency band ranging from 600 Hz to 4800 Hz (see Figure 2). The spectral components of speech with the power spectrum P_{XdBi} exceeding the auditory masking threshold AMT_i, (i=1,2,...N_{ERB}) were assumed to be audible; spectral components of speech with the power spectrum falling below the AMT levels were considered to be non-audible. The audible components of the speech spectrum were kept intact, whereas the non-audible spectral components of speech were amplified. Such adaptive frame-by-frame processing helped to minimize the distortion of the original acoustic structure of speech, while maintaining a high level of speech intelligibility. Additionally, the speech amplitude level could be kept within a range preferred by the listener. To aid in the enhancement decisions, the speech power peaks within the 600 Hz to 4800 Hz region were identified. The spectral envelope of the power peaks PeakLine[i] for $(N_{600} \le i \le N_{4800})$, was calculated by determining the maximum power values within each of the ERB bands, as shown in Eq. 9.

$$PeakLine[i] = \max_{b_{li} \le r \le b_{hi}} (P_{XdB}[r]) \qquad (9)$$

The index i in Eq. 9 represents the ERB bands that contain frequencies within the 600 Hz to 4800 Hz interval. The red continuous line in the example shown in Figure 2 interpolates the values of the PeakLine function. Since the speech intelligibility depends mostly on the audibility of the power peaks corresponding to the formant frequencies, the goal of the proposed enhancement scheme was to ensure that most of the PeakLine is above the AMT and thus, the formants represented by the PeakLine are audible.

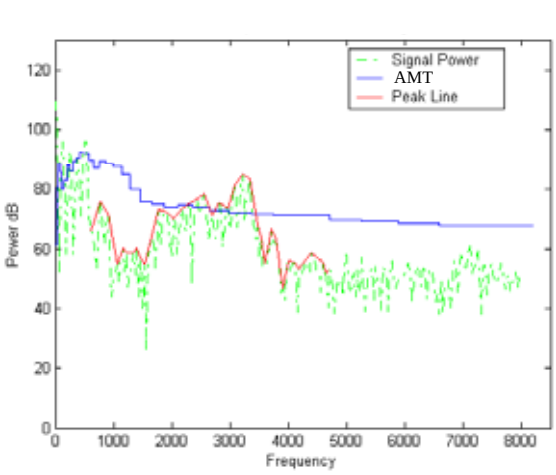

Figure 2: *An example of the speech power spectrum (green dashed line) and the speech AMT levels (blue step-line) for one frame of the speech signal. Peak values are outlined (red continuous line) from 600-4800Hz.*

The individual amplification factors for the spectral components of speech within the 600 Hz to 4800 Hz range, were calculated using

$$\text{AmpFactor}[i] = \begin{cases} \dfrac{AMT_i}{PeakLine[i]} & \text{if} \quad \text{PeakLine}[i] < \text{AMT}_i \\ 1 & \text{if} \quad \text{PeakLine}[i] \ge \text{AMT}_i \end{cases} \qquad (10)$$

The individual amplification factors AmpFactor[i] $(N_{600} \le i \le N_{4800})$ were then used to determine a common amplification factor for all components of the PeakLine[i] using

$$\text{EnhFactor} = \begin{cases} (\text{AmpFactor})_{\max} & \text{if} \quad (\text{AmpFactor})_{\max} \le \text{Limit} \\ (\text{AmpFactor})_{\text{L}} & \text{if} \quad (\text{AmpFactor})_{\max} > \text{Limit} \end{cases} \qquad (11)$$

Accepted after full paper review

where

$$(\text{AmpFactor})_{\max} = \max_{N_{600} \le i \le N_{4800}} (\text{AmpFactor}[i]) \quad (12)$$

$$(\text{AmpFactor})_{L} = \max_{\substack{N_{600} \le i \le N_{4800} \\ \text{AmpFactor}[i] \le \text{Limit}}} (\text{AmpFactor}[i]) \quad (13)$$

To lower the computational load, the threshold used to limit the amplification value was the same as the threshold used in the process of speech detection. By limiting the amplification in this way, under maximum amplification conditions, the SNR would rise to about 5dB-6dB, ensuring speech audibility (Gold et al 2000).

3. Numerical simulations

Numerical simulations of the enhancement process were used to generate the test sequences of speech samples. Two sets of test signals were generated. The first set represented the noisy non-enhanced speech signal. The second set represented the noisy enhanced speech signal. The noisy non-enhanced speech was produced by adding motorcycle noise to the clean non-enhanced speech in a proportion yielding a given value of the Signal-to-Noise Ration (SNR). Such mixing produced a sequence of weighted noise samples. These weighted noise samples were saved and added to the enhanced speech signal to generate the noisy enhanced speech samples. This way the same noise was added to the original non-enhanced speech and to the same speech after enhancement. The SNR value of 0dB was chosen for testing purposes as a level with moderately low intelligibility but sufficient to demonstrate the benefits of the proposed enhancement scheme. An alternative, non-adaptive approach in which the speech signal is amplified uniformly across all frequency bands, lifting the average speech level above the AMT of noise, has also been tested against the proposed adaptive enhancement procedure. This approach does not take into consideration subjective aspects of auditory perception, and has been used only to generate control signals for the test procedures.

4. Subjective test of speech intelligibility

The Modified Rhyme Test (MRT) (ANSI S3.2-1989) was used to provide a subjective measure of speech intelligibility. Three test sequences based on the same set of 50 different words were produced. The test sequences represented: the noisy non-enhanced speech (control series), the noisy amplified speech (amplified control series) and the noisy enhanced speech (enhanced series). The three test sequences were used to generate a randomly ordered list of 150 English words read by a native English male speaker. A subjective test of speech intelligibility was then conducted. The test group of 17 volunteers included 11 male and 6 female, native English speaking subjects. The experiment was conducted in a sound shielded room. The words were presented to the subjects, one by one, through headphones. The subjects had to identify the presented words from the lists of four alternatives.

5. Statistical analysis of results

The results of the MRT were analysed statistically with respect to the subjects' scores and with respect to the word set scores.

Table 1: *Results of the paired t-tests for the subject scores.*

alpha =0.05	Amplified control series versus control series	Enhanced series versus control series	Enhanced. series versus amplified control series
t Stat	0.453683041	5.938230913	4.697133548
P(T<=t) one-tail	0.326029996	1.45668E-07	1.08358E-05
P(T<=t) two-tail	0.652059991	2.91335E-07	2.16715E-05

The statistical analysis of the subjects' scores based on the paired t-test (see Table 1), indicates that the proposed speech enhancement scheme yields a statistically significant improvement of the speech intelligibility compared to both the control series and the amplified control series. In other words, the proposed speech enhancement technique yields a significant improvement of speech intelligibility, not only compared to the nosy signal, but also compared to the simple approach of lifting the amplitude gain at a uniform rate across all frequencies.

The t-test analysis for the word sets scores (see Table 2) indicates that there is a statistically significant improvement of the median value for the word scores due to the enhancement process.

Table 2: *Results of the paired t-tests for the word set scores.*

alpha =0.05	Amplified control series versus control series	Enhanced series versus control series	Enhanced. series versus amplified control series
t Stat	0.646233913	11.81192326	12.18537296
P(T<=t) one-tail	0.263644622	1.29286E-09	8.23587E-10
P(T<=t) two-tail	0.527289243	2.58571E-09	1.64717E-09

6. Discussion and conclusions

A new speech enhancement method for the improvement of speech intelligibility in the presence of an ambient noise was developed and tested. The originality of the presented approach in comparison to other techniques described in the literature can be summarized as follows: firstly, the proposed enhancement method represents an approach based on perceptual (subjective) criteria rather than objective measures of speech quality; secondly, the improvement of speech intelligibility is achieved through modifications applied to the clean speech signal, rather than to the noisy signal; and thirdly, the enhancement decision-making process is based on the comparison of the speech power spectrum with the AMT calculated for the noise signal, rather than for the speech signal. In addition, the enhancement method requires very little initial information about the nature of the noise. Since the method modifies the spectral components of speech in an adaptive way, and only within a small number of selected frequency bands, the distortion of the original acoustic features of the speech signal is minimal. Numerical simulations of the enhancement process were performed and tested using the subjective MRT procedure. The test results were analysed statistically showing a significant improvement of speech intelligibility.

7. References

Tsaukolas, D. E. and Mourjopoulos,. J. N. (1997). Speech Enhancement Based on Audible Noise Supression, *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 6, pp. 497-514.

Virag, Nathalie (1999) Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System, *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137.

Virag, Nathalie (1995) Speech Enhancement Based on Masking Properties of the Auditory System, *International Conference on Acoustics, Speech, and Signal Processing, ICASSP 95*, vol. 1, pp. 796–799.

Johnston, James D. (1988) Transform Coding of Audio Signals Using Perceptual Noise Criteria, *IEEE Journal of Selected Areas of Communications*, vol. 6, no. 2, pp. 314-323.

Moore, B. C. J. and Glasberg, (1996) B. R. A Revision of Zwicker's Loudness Model, *ACTA Acustica*, vol. 82, pp. 335-345.

Zwicker, Z.and Fastl, H. (1990) *Psychoacoustics, Facts and Models*, Berlin: Springer Verlag.

Gold, B. and Morgan, N. (2000) *Speech and Audio Signal Processing*, John Willey & Sons, Inc.

Moore, B. C. J. (1989) *Introduction to the Psychology of Hearing*, 3rd edition, London, Academic Press, 1989.

ANSI S3.2-1989, The American National Standards Institute's approved procedure, "Method for Measuring the Intelligibility of Speech Over Communication Systems".

Accepted after full paper review