Proc. Biennial Symp. Commun. (Kingston, ON), pp. 516-519, June 2002

Noise Suppression using a Perceptual Model for Wideband Speech Signals

Joachim Thiemann* and Peter Kabal

Electrical & Computer Engineering
McGill University, Montreal, Canada

Abstract

Traditional algorithms for suppressing background noise in speech signals can add annoying artefacts to the resulting denoised signal. In applications requiring better than toll quality, it is desirable that noise suppression should not add any audible artefacts. This paper describes a method that is effective for narrowband and applies these methods to wideband signals. The method presented uses a high-resolution psychoacoustic model originally developed for the evaluation of audio quality, and combines it with a method originally developed for audio signal enhancement. It is shown that while the method works well in narrowband applications, in wideband signals the quality needs to be improved.

1 Introduction

Reducing the background acoustic noise in speech signals has received a great deal of attention in literature because of the increasing popularity of mobile phones. The methods developed so far have mostly been for narrowband speech signals, and were aimed at removing as much background noise as possible while retaining or improving intelligibility.

However, traditional noise reduction algorithms tend to add unnatural-sounding artefacts. The best known of these is musical noise, although distortion of the speech signal is often evident as well. These artefacts, if strongly audible, lower the perceived quality of the signal, even though the energy of the noise relative to the energy of the speech has been lowered.

Networks transmitting wideband speech have greater perceived quality of signal due to the larger bandwidth. In these systems it is therefore desirable to have noise reduction systems that do not add annoying artefacts.

This paper describes a method that uses psychoacoustics to improve the process of noise reduction. Specifically, the effect of masking, where stronger sounds can render weaker nearby sounds inaudible, is used to focus the modification of the signal onto the audible noise. Since this reduces the total amount of signal modification necessary, fewer artifacts are generated. It has been found that this method works well for narrowband signals, and thus it has been adapted for wideband signals.

2 Description of method

This paper provides a description of the overall noise suppression algorithm. The development of this method is described in [1], which also examines similar methods and gives comparative results.

The method proposed falls into the category of basic Short-Time Spectral Amplitude (STSA) modification algorithms. This means that the processing is performed by sectioning the input signal into short frames, which are then transformed into frequency domain using a Discrete-Fourier Transform (DFT). The result of the DFT is split into the magnitude and phase components, and all further noise suppression processing is performed on the magnitude component. After processing, the time-domain signal is reconstructed from the modified magnitude and the original phase component.

The method for processing the signal and use of the psychoacoustic model is derived from a method developed by Soulodre [2] to remove camera noise from film soundtracks. Soulodre based his research on methods found in the field of audio enhancement, in particular by Tsoukalas *et al* [3].

2.1 Input Signal Analysis

Wideband speech is sampled at $f_s = 16\,000$ Hz. For noise suppression, frames should be processed every 20–40 ms, so for the proposed method, the framesize with overlap (at 50%) is set to $N_F = 1024$. Thus, the frames are 64 ms long, and the frame advance is 32 ms.

The current frame $x[n,p]$ (where p denotes the frame counter) is windowed by a window obtained from the square root of the Hann window, defined by

$$h[n] = \sqrt{\frac{1}{2}\left(1 - \cos\left(2\pi\frac{n+0.5}{N_F}\right)\right)}, \qquad (1)$$

where $n = 0, \ldots, N_F - 1$. The frame is then converted into discrete frequency domain using a scaled Discrete-

[*] now with VoiceAge in Montreal.

Fourier Transform,

$$X[k,p] = \frac{1}{N_F} \sum_{n=0}^{N_F-1} h[n]x[n,p]e^{-j2\pi nk/N_F}, \qquad (2)$$

where $k = 0, \ldots, N_F/2$, since $x[n,p]$ is assumed to be real.

2.2 Noise Spectrum Estimate

An estimate of the current noise spectrum is obtained using a Voice Activity Detector (VAD) to determine if the current frame contains energy due to speech. If the frame is considered noise only, it is used to update the current noise estimate, in addition to being processed in the normal fashion. The noise estimate is obtained by exponentially averaging the magnitude spectra of the noise only frames,

$$\hat{W}[k,p] = \lambda_W \hat{W}[k,p-1] + (1-\lambda_W)\big|X[k,p]\big|, \qquad (3)$$

where $0 \leq \lambda_W < 1$ controls the speed with which the noise estimate adapts to changes in the noise spectrum. A value of $\lambda_W = 0.97$ was used.

If the VAD detects speech in the current frame, the noise estimate is not modified, so $\hat{W}[k,p] = \hat{W}[k,p-1]$.

2.3 Perceptual Based Spectral Subtraction

An initial estimate of the clean speech is obtained by subtracting the noise estimate magnitude spectrum from the magnitude spectrum of the current unprocessed speech frame. Similar to traditional spectral subtraction noise suppression,

$$\hat{S}_1[k,p] = \max(|X[k,p]| - \hat{W}[k,p], 0). \qquad (4)$$

At this stage, the perceptual model is applied to both the clean speech estimate $\hat{S}_1$ and the noise estimate $\hat{W}$, separately. The model used in this case is the basic model of ITU-R BS.1387 [4], described below. It outputs an excitation pattern at a resolution of 0.25 Bark, that is, about four “bins” per critical band. This excitation pattern relates closely to the masking threshold, the frequency dependent threshold below which additional sounds are inaudible.

The Bark domain gain is obtained using Soulodre’s formula,

$$G_{\text{Bark}}[b,p] = \frac{\text{PE}(\hat{S}_1[k,p])}{\text{PE}(|X[k,p]|)}, \qquad (5)$$

where $\text{PE}(\cdot)$ denotes applying the perceptual model, and $b = 0, \ldots, B_{f_{\max}}$ is the index of sub-critical bands in perceptual domain (see below). This formula will attenuate the signal at the frequencies where the noise is very audible (low excitation from the clean speech, but high excitation from the noisy speech).

Figure 1 shows the overall gain calculation structure. The perceptual domain gain $G_{\text{Bark}}[b,p]$ is mapped into linear domain by the following procedure. Linear gain bins $G_1[k,p]$ that are fully within a perceptual gain bin $G_{\text{Bark}}[b,p]$ assume the gain of that bin. Linear bins that fall on the boundary of two perceptual bins are assigned a gain that is linearly interpolated from the gains of the perceptual bins.

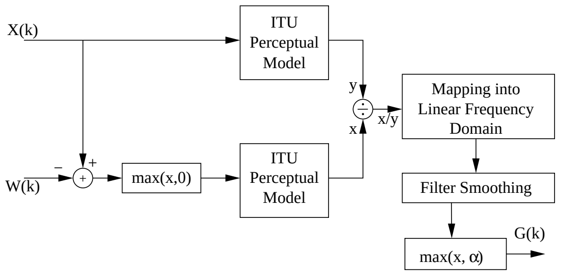

Fig. 1 Gain calculation using Soulodres method

2.4 Overview of the Perceptual Model

As mentioned above, the perceptual model is derived from the basic model used by the ITU-R recommendation BS.1387. Since a full description of all the steps required to obtain the excitation pattern is outside the scope of this paper, only an outline of the algorithm is given here, with some key points explained in more detail.

First, since some aspects of sound perception are dependent on absolute sound pressure, the input to the perceptual model is scaled to the assumption that a full-range input sine wave is at 92 dB_{SPL}. Also, a frequency dependent weighing function is applied to model the frequency response of the outer and middle ear. The weighing function appears to be based on work done by Terhard [5].

The resulting scaled spectrum is now converted into perceptual domain. In the case of the basic model of BS.1387, the linear spectrum is mapped into a discrete Bark domain, where each bin represents 0.25 Bark. The center frequency of each bin is given by

$$f_c = 650 \sinh(z/7), \qquad (6)$$

where f_c is in kHz and z is in Bark. Note that where b is used, it is meant to represent a bin in the discrete Bark domain, where the difference between bins is $\Delta z = 0.25$Bark.

To model the internal noise of the auditory system, a frequency-dependent offset is added to the perceptual spectrum. The resulting *pitch patterns* are denoted $E[b,p]$.

At this stage, the critical step of *spreading* is performed. Here, it is calculated how sounds close to each other in time and frequency affect each other. The resulting *Excitation Pattern* will be independent of sounds that are inaudible. The excitation pattern is calculated in two steps: frequency-domain spreading and temporal smearing.

The frequency domain (actually in perceptual domain) spreading is calculated by

$$E_s[b,p] = \frac{1}{B_{\mathrm{SP}}[b]}\left(\sum_{l=0}^{N_b-1}(E[l,p]S(b,l,E[l,p]))^{0.4}\right)^{\frac{1}{0.4}}, \tag{7}$$

where N_b denotes the number of perceptual domain bins. $B_{\mathrm{SP}}[b]$ is a normalization factor given by

$$B_{\mathrm{SP}}[b] = \left(\sum_{l=0}^{N_b-1} S(b,l,1)^{0.4}\right)^{\frac{1}{0.4}}, \tag{8}$$

and $S(\cdot,\cdot,\cdot)$ denotes the actual spreading function, given by

$$S(b,l,E) = \frac{1}{A(l,E)}10^{S_{\mathrm{dB}}(b,l,E)/10}, \tag{9}$$

where $A(l,E)$ is a normalization term to give a unit area to each center frequency l. The slopes of the spreading functions are expressed in dB, as

$$S_{\mathrm{dB}}(b,l,E) = \begin{cases} 27(b-l)\Delta z, & b \leq l, \\ \left[-24 - \frac{230}{f_c[l]} + 2\log_{10}(E)\right](b-l)\Delta z, & b > l, \end{cases} \tag{10}$$

where $\Delta z = 0.25$.

The second step is to calculate the frequency dependent temporal smearing, giving the final Excitation Pattern $\tilde{E}_S[b,p]$. The temporal smearing is calculated by

$$\begin{aligned} E_f[b,p] &= \alpha[b]E_f[b,p-1] + (1-\alpha[b])E_S[b,p] \\ \tilde{E}_S[b,p] &= \max(E_f[b,p], E_S[b,p]). \end{aligned} \tag{11}$$

The parameter $\alpha[b]$ controls the time constant for the averaging for the decaying energies, to model the effect of backwards masking. It is calculated by

$$\alpha[b] = \exp\left(-\frac{N_F/2}{f_s\tau[b]}\right), \tag{12}$$

where

$$\tau[b] = \tau_{\min} + \frac{100}{f_c[b]}(\tau_{100} - \tau_{\min}), \tag{13}$$

where $\tau_{100} = 0.030$ s and $\tau_{\min} = 0.008$ s.

2.5 Output Signal Synthesis

The gains $G_1[k,p]$ are smoothed by an exponential average similar to (3),

$$G_2[k,p] = \lambda_F G_2[k,p-1] + (1-\lambda_F)G_1[k,p], \tag{14}$$

to reduce any musical noise that might still appear. A low value of $\lambda_f = 0.01$ was used successfully. The resulting smoothed gains are constrained to a minimum to add a natural-sounding "noise floor",

$$G[k,p] = \max(G_2[k,p], \alpha), \tag{15}$$

where $\alpha = 0.1$. The gains $G[k,p]$ are applied to the speech spectrum to get the estimated speech spectrum by

$$\hat{S}[k,p] = G[k,p]X[k,p], \tag{16}$$

from which the time-domain signal is reconstructed by computing the Inverse Discrete-Fourier Transform. The result is windowed to avoid discontinuities at frame boundaries, giving

$$\hat{s}[n,p] = h[n]\sum_{k=-N_F/2}^{N_F/2}\hat{S}[|k|,p]e^{j2\pi nk/N_F}. \tag{17}$$

Finally, frame overlap is handled by the overlap-add method, that is, the last half of the previous frame is added to the first half of the current frame.

3 Results

Informal listening tests have suggested that with narrowband speech this method works well even at low Signal-to-Noise Ratio (SNR). Compared with more traditional noise suppression algorithms, there are fewer audible artefacts in the resulting noise suppressed speech, for the same degree of noise suppression. Here it is tested if the same applies for wideband speech signals.

To compare, two speech files (one male speaker and one female speaker) were mixed with two types of noise, at two levels of SNR (0 dB and 20 dB)[1]. The first noise was recorded in a car driving at 120 km/h, and is strongly lowpass (at 5000 Hz, 60 dB below maximum at 125 Hz). The second noise is "room noise" consisting mainly of fan noise from a desktop computer, and is more white (at 5000 Hz, 25 dB below maximum at 150 Hz).

The proposed method is compared to the common spectral subtraction method, based on Boll [7]. The implementation is similar to the perceptual method described above, except the perceptual part described in Section 2.3 is replaced with

$$G_1[k,p] = \max\left(1 - k_B\left(\frac{(\hat{W}[k,p])^a}{|X[k,p]|^a}\right), 0\right)^{\frac{1}{a}}, \tag{18}$$

where $k_B = 1.5$ is the oversubtraction factor. a is a parameter to choose energy or power domain spectral subtraction (by setting a to 1 or 2, respectively). In these tests, $a = 1$ was used. This gain formula is a

[1] For the purposes of calculating the level at which noise is added to the speech, the speech level was calculated according to ITU-T recommendation P.56 [6]

generalization of Eq. (4). In addition, to reduce the musical noise, λ_F was increased to 0.2.

The samples[2] were presented to 6 listeners in a A/B comparison test. The listeners would indicate whether they preferred file "A", "B" or if they were equally preferred. While the sample is too small to draw solid conclusions, it is nevertheless informative, based on the feedback from the listeners.

	Room Noise		Car Noise	
Subtraction Type	20 dB	0 dB	20 dB	0 dB
Perceptual	12	11	6	13
Boll	5	11	6	11
no preference	7	12	12	0

Table 1 Preferences of subtraction methods versus type and level of background noise

Unlike the results in [1], no strong preference can be shown either for or against the perceptual noise reduction method. Table 1 show the results of the tests, where the responses for the male and female speech samples have been summed.

However, some observations can be made from the tests. Some listeners will quickly focus on one single type of artefacts, and ignore any others. Two listeners found musical noise annoying and preferred the perceptual method almost all the time. One listener considered speech distortion annoying and mostly preferred the non-perceptual method. The samples processed by the perceptual method are almost completely free of audible musical noise, but exhibit more speech distortion than the nonperceptual method.

There are several possible explanations for the lack of improvement by the perceptual processing for wideband speech in contrast to similar experiments with narrowband speech. The most likely explanation is that distortions at higher frequencies are more audible, or rather, are considered more annoying.

The choice of the parameters λ_F, α, and k_B also poses a problem in spectral subtraction, since these parameters can be tuned depending on the signal, noise and personal preference. One advantage of perceptual processing is that the effect of changing these parameters was found to be much smaller than for the nonperceptual method.

3.1 Future Work

From the above results, it is clear that the perceptual noise reduction method needs to be improved in order to be useful in wideband applications, in contrast to Boll's method, which is much lower complexity. The focus should be on reducing the audible distortion of the speech signal, which may be achieved by nonlinear scaling of the gain function $G_1[k,p]$, or improving the initial clean speech estimate $\hat{S}_1[k,p]$. Another possibility is to add a preemphasis and deemphasis stage that would cause strong noise suppression at low frequencies, and less at high frequencies.

4 Conclusion

Use of a perceptual model in noise subtraction has been shown to improve the resulting signal in both narrowband signals and audio-quality signals. The perceptual model calculates the exitation pattern in the perceptual domain, of the original noisy signal and an initial estimate of the clean speech. The ratio of these exitation patterns is used to obtain a frequency dependent attenuation which is applied to the noisy speech signal.

In this paper, this method was used with wideband speech signals, at relatively high and low SNR. However, it was found that while musical noise is rendered inaudible, the denoised speech exhibits significant distortion. This distortion is highly annoying to some listeners, and further work to reduce the distortion is necessary.

References

[1] J. Thiemann, "Acoustic noise suppression for speech signals using auditory masking effects," M.S. thesis, McGill University, Montréal, Canada, May 2001.

[2] G. Soulodre, *Adaptive Methods for Removing Camera Noise from Film Soundtracks*, Ph.D. thesis, McGill University, Montréal, Canada, 1998.

[3] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Perceptual filters for audio signal enhancement," *J. Audio Eng. Soc.*, vol. 45, no. 1/2, pp. 22–35, Jan/Feb 1997.

[4] International Telecommunications Union, "Method for objective measurements of perceived audio quality," 1998, Recommendation ITU-R BS.1387.

[5] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Am.*, vol. 71, no. 3, Mar. 1982.

[6] International Telecommunications Union, "Objective measurement of active speech level," 1993, Recommendation ITU-T P.56.

[7] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-27, no. 2, Apr. 1979.

[2] The sample files can be found on-line at `http://www.tsp.ece.mcgill.ca/Kabal/papers`.