Proc. IEEE Workshop Speech Coding (Tsukuba, Japan), pp. 172-174, Oct. 2002

LOW DISTORTION ACOUSTIC NOISE SUPPRESSION USING A PERCEPTUAL MODEL FOR SPEECH SIGNALS

Joachim Thiemann and Peter Kabal*

Department of Electrical & Computer Engineering
McGill University, Montreal, Canada

ABSTRACT

Algorithms for the suppression of acoustic noise in speech signals are generally Short-Time Spectral Amplitude (STSA) methods such as Spectral Subtraction. These methods have been effective at reducing or removing the background noise, but have a tendency (at low SNR) to add annoying artefacts, such as musical noise, and distortion of the speech signal. By employing an auditory model, psychoacoustic effects such as simultaneous masking can be used to apply spectral modification in a more effective manner, reducing the amount of overall modification necessary. In this way, the artefacts introduced by the processing are reduced. This paper proposes a method to significantly improve the reduction in the background acoustic noise in narrowband and wideband speech signals, even at low SNR. Here we show that the use of a subtraction strategy and psychoacoustic model originally intended for audio signals yields an output signal with little or no audible distortion.

1. INTRODUCTION

With the increasing popularity of mobile phones, a need for acoustic noise suppression algorithms for speech signals has arisen since these phones are often used in environments where large amounts of background noise are present. This noise lowers the perceived quality of the signal, and is tiring to listen to for a prolonged period. For example, a typical situation is a hands-free set in a vehicle, where one would have engine noise, road noise, wind noise, etc. This problem has already received much attention in the literature, with algorithms typically having the goal of removing the background noise while retaining speech intelligibility. However, these methods tend to corrupt the speech signal by introducing artefacts that can sound unnatural. This paper proposes a method which focuses instead on maximally reducing noise without affecting the perceived quality of the speech signal. By using techniques from audio enhancement and a sophisticated perceptual model, a large degree of noise reduction is possible.

Commonly, noise suppression algorithms are based on Short-Time Spectral Analysis (STSA). Using an estimate of the background noise spectrum and of the current noisy speech, an estimate of the clean speech is obtained. (The estimate of the background noise spectrum is updated during speech pauses.) However, without further processing, this estimate can exhibit musical noise and speech signal distortion.

Certain properties of the human auditory system can be exploited to improve the quality or effectiveness of noise reduction algorithms. One such property is the effect of masking, whereby stronger sounds can render weaker nearby sounds inaudible. Based on this, parts of the background noise that are inaudible due to the presence of the speech signal itself do not need to be processed. Since this reduces the overall amount of signal modification, there is a corresponding reduction in introduced artefacts.

In audio processing, particularly in the restoration of archive material, it is desirable that the resulting signal exhibits no artefacts from the noise reduction process. Thus in this field methods have been developed that focus not on the complete removal of the noise, but on retaining the perceived quality of the signal.

2. NOISE SUPPRESSION USING SHORT-TIME SPECTRAL ANALYSIS METHODS

STSA methods segment the input into frames short enough that the speech signal can be assumed to be stationary within a frame. This frame is then transformed into the frequency domain using a DFT or similar transform, with appropriate windowing. Typically, frames are about 20 ms long, and a windowed overlap-add method is used to avoid discontinuities at frame boundaries.

Spectral Subtraction is a special case of STSA noise suppression, where the following assumptions are made. It is assumed that the noise and the speech signal are uncorrelated, and thus the power spectrum of the noisy signal is the sum of the power spectra of the signal and the noise. It is also assumed that the noise is relatively stationary, such that a periodogram obtained in previous frames is a good estimate of the current noise spectrum. Finally, it is assumed that the human hearing is insensitive to small phase distortions.

Using these assumptions, Spectral Subtraction generates an estimate of the clean speech spectrum $\hat{S}(f)$ from the noisy speech spectrum $X(f)$ and an estimate of the noise spectrum $\hat{W}(f)$ using

$$|\hat{S}(f)|^2 = \max(|X(f)|^2 - |\hat{W}(f)|^2, 0). \qquad (1)$$

Generalizing (1) and accounting for the reconstruction of the phase from the noisy signal, we get

$$\hat{S}(f) = \left(\max(|X(f)|^a - k|\hat{W}(f)|^a, 0)\right)^{\frac{1}{a}} e^{j\phi_x(f)}, \qquad (2)$$

*Now with VoiceAge Corp. in Montreal

where $\phi_x(f)$ is the phase information of the current noisy speech frame, and k is parameter to allow for *oversubtraction* to account for the variance of the noise spectrum. Generally, $1 < k < 2$. The parameter a may be set to 1 for magnitude subtraction or 2 for power subtraction. The above can be interpreted as a gain function, and viewed as a zero-phase filter. By rewriting (2) as a filter to X, we can additionally include a noise floor parameter α, such that

$$\hat{S}(f) = X(f)H(f), \tag{3}$$

where

$$H(f) = \Big(\max\big(1 - k\frac{|\hat{W}(f)|^a}{|X(f)|^a}, \alpha\big)\Big)^{\frac{1}{a}}. \tag{4}$$

The output of (2) exhibits the artefacts commonly found in spectral subtraction noise suppression. The severity of artefacts depends on the choice of k, a, α, and the method used to obtain the noise estimate $\hat{W}$. The difference of the noise estimate $\hat{W}$ from the actual noise component in the short time spectrum of X leads to musical noise and spectral magnitude distortion. The contribution of the phase of the noise leads to phase modulation.

There are other gain formulas for spectral subtraction algorithms. Most of these also attenuate the signal more strongly at frequencies with low SNR.

3. USING PSYCHOACOUSTIC MODELS IN STSA ALGORITHMS

There are various ways in which psychoacoustic models are used in noise suppression algorithms. Generally, the psychoacoustic model is used to modify the gain function as shown in Fig. 1. In some algorithms, such as the noise suppressor for the EVRC standard, the SNR-based gain is calculated for each of a set of distinct frequency groups that correspond roughly to critical bands. Overall noise level is also taken into account when calculating the attenuation.

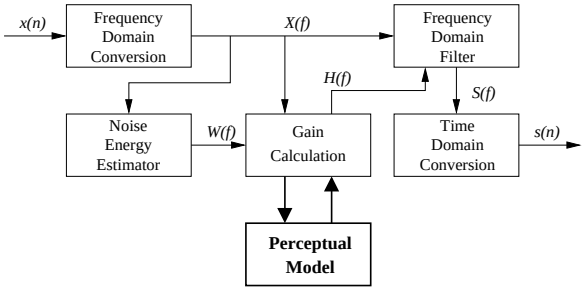

Figure 1: STSA algorithms using perceptual model

A more direct use of a psychoacoustic model is to calculate the masking threshold of the resulting signal, and then either directly remove elements which are determined to be musical noise, or dynamically change the parameters of the subtraction gain function [1].

While the above methods tend to solve the problem of reducing or even removing the musical noise, distortion of the clean signal is not addressed. For low-distortion noise suppression, it is better to look at methods developed for noise reduction of audio signals. The method presented here is based on an algorithm developed for the suppression of broad-band noise in audio signals [2]. This algorithm calculates the psychoacoustic representation (denoted here $\mathrm{PE}(\cdot)$) of the noise estimate and of the noisy signal, and then obtains a per-critical band attenuation by

$$H(b) = 1 - \frac{\mathrm{PE}(|\hat{W}|)}{\mathrm{PE}(|X|)}, \tag{5}$$

where b is the the critical band index. The similarity to (3) is apparent. However, in [3] Soulodre noted that the masking pattern is dependent on the absolute level of the *clean* signal and modified (5) to

$$H(b) = \frac{\mathrm{PE}(|X| - |\hat{W}|)}{\mathrm{PE}(|X|)}. \tag{6}$$

3.1. Masking Models

The psychoacoustic model greatly influences the quality of the noise suppression algorithms. Psychoacoustic models were first proposed in conjunction with compression algorithms to mask quantization noise. Since then these models have undergone significant refinement. Most significantly, it has been shown that it is insufficient to calculate levels at the resolution of a single critical band. While some masking models are calculated in the linear frequency domain provided by the DFT, it is still advantageous (from a computational complexity perspective) to transform the frequency values into the Bark domain, a frequency scaling that is based on critical band widths. The key point is that the width of a critical band increases with its center frequency. In Bark domain, critical bands have equal width. The specific mapping operation of frequency to bark varies with the model used.

For the algorithm presented here, the basic model of ITU-R recommendation BS.1387 (PEAQ) [4] is used. This model calculates an excitation pattern and masking threshold at the resolution of 0.25 Bark. At this resolution, it is not necessary to differentiate between noise-like maskers and tone-like maskers, thereby avoiding this problematic aspect of single-bark resolution models.

4. LOW DISTORTION NOISE SUPPRESSION

The implementation of the proposed low-distortion noise suppression (LDNS) method is a Short-Time Spectral analysis, modification, and synthesis block. Some of the parameters were dictated by the PEAQ model, specifically the frame size, overlap and analysis/synthesis windows. The basic model of PEAQ is designed to use frames of about 20 ms, with a 50% overlap, using an FFT sized such that the width of a frequency bin is less than the smallest quarter-Bark bin (ca. 25 Hz). Each frame is windowed by a Hanning window.

Figure 2 expands the "Gain Calculation" block of Fig. 1, showing the implementation of (6). The block at the output is a piecewise linear mapping to constrain the per-bin attenuation. The upper limit is 1 to avoid overload distortion, and the lower limit is set to $\alpha = 0.2$ to provide a noise floor similar to Eq. (4) above. This block can be further

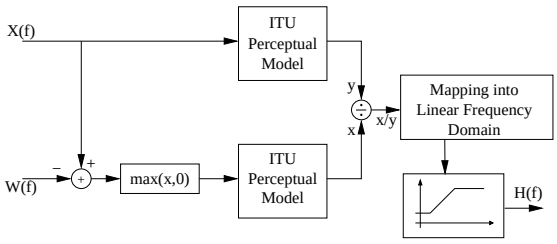

Figure 2: Gain calculation using Soulodres method

generalized to "tune" the noise suppressor performance by using some alternate (monotonically increasing) function. For example, it is possible to further reduce the residual noise at the expense of adding some signal distortion.

5. RESULTS

Tests of the proposed method were performed by mixing speech files (using one male speaker and one female speaker) with two types of noise, at two levels of SNR (3 dB and 12 dB)[1]. The first noise was recorded in a car driving at 120 km/h, and is strongly lowpass (at 3000 Hz, 45 dB below maximum at 125 Hz). The other noise is "room noise" consisting mainly of fan noise from a desktop computer, and is more white (at 3000 Hz, 25 dB below maximum at 150 Hz). It also has a noticeable tonal component.

The proposed method is compared to the EVRC noise suppressor[6], since it is a widely-used standardized noise suppression algorithm, and also uses some perceptual properties. Thus, it provides a suitable baseline reference. The samples[2] were presented to 10 listeners in an A/B comparison test. The listeners would indicate whether they preferred file "A", "B" or if no preference exists for either sample.

	Room Noise		Car Noise	
Subtraction Type	12 dB	3 dB	12 dB	3 dB
LDNS	23	20	13	9
EVRC	10	10	16	13
no preference	6	10	11	18

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

It is interesting to note that the results from the room noise shows a stong bias towards preference of LDNS, while the results with the car noise show a statistically small difference ($p > 0.5$ using the sign test). While the sample size is small, it can be presumed that LDNS can perform at least as good as EVRC, and can in some cases even outperform the EVRC noise suppressor.

To illustrate, Fig. 3 shows a sample frame of noisy speech with the current speech estimate and the resulting LDNS attenuation $H(f)$. The masking provided by the speech harmonics below 500 Hz reduces the attenuation at nearby frequencies, while the smoothing effect becomes more pronounced at higher frequencies, since the width of the critical bands increase.

[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 [5]

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

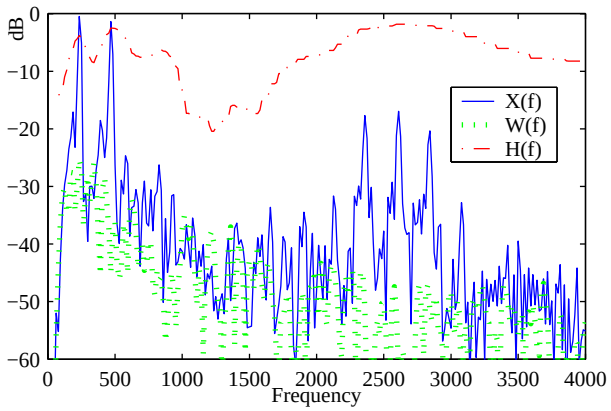

Figure 3: Noisy signal, Noise estimate, and resulting attenuation for voiced frame at 12 dB SNR

6. CONCLUSION

The use of an auditory model in noise suppression algorithms can lead to an improvement of the perceived quality of the resulting signal. This paper presents a method that uses the basic model of the PEAQ algorithm to provide noise suppression with low audible distortion, even at low SNR. It is found that this method performs well when compared to an established standard algorithm.

7. REFERENCES

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[2] 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.

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

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

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

[6] "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," Jan. 1996, TR-45.5, PN-3292 AD3 (published as IS-127-3).