SPEECH ENHANCEMENT BASED ON MASKING PROPERTIES OF THE AUDITORY SYSTEM

Nathalie Virag

Signal Processing Laboratory
Swiss Federal Institute of Technology at Lausanne
CH–1015 Lausanne, Switzerland
E–mail: virag@lts.de.epfl.ch

ABSTRACT

This paper addresses the problem of the intelligibility enhancement of speech corrupted by additive background noise in a single channel system. The proposed algorithm uses a criterion based on the human perception. It is a variation of the well–known spectral subtraction method which is attractive because of its simplicity, but introduces an unnatural and unpleasant residual noise. The proposed approach incorporates in this method considerations about noise masking of the auditory system. It succeeds in finding the best trade-off between noise reduction and speech distortion in a perceptual sense. Simulations show perceptually very satisfactory results and objective measures indicate a quality improvement. The speech processed with this new algorithm sounds more pleasant to a human listener than those obtained by the classical methods. This shows the relevance to incorporate perceptual aspects in the enhancement process.

1. INTRODUCTION

The spectral subtraction is a traditional method for removing stationary background noise in a single channel system. In addition to the basic approach of spectral magnitude subtraction [1], many variations have been developed [2]. This constitutes a family of subtractive–type algorithms which attempt to estimate the short–time spectral magnitude of speech by subtracting a noise estimation from the noisy speech, and combine it with the phase of the noisy speech.

The major drawback of these methods is the characteristics of the residual noise called musical noise. It is composed of tones at random frequencies and has an increased variance resulting in a perceptually annoying noise. Hence, there is a strong need to develop techniques for residual noise reduction. Various algorithms have been developed for that purpose : magnitude averaging [1], oversubtraction of noise and introduction of a spectral floor [3], soft–decision noise suppression filtering [4], optimal MMSE estimation of the short–time spectral amplitude [5], introduction of morphological-based spectral constraints in the enhancement algorithm [6].

The residual noise is annoying because of its unnatural characteristic and is even sometimes as disturbing as the original additive noise. One of the objectives of the algorithm proposed in this paper is to render this residual noise perceptually white. Furthermore, subtractive–type algorithms are in general effective in reducing the noise but not in improving intelligibility. It is important for a system aiming at improving intelligibility to take into account properties of the human auditory system. Some methods have been developed in this direction, by modeling several aspects of the enhancement function present in the auditory system [7, 8, 9].

In this paper, it is proposed to incorporate a human hearing model which is already widely used in perceptual wideband audio coding [10]. This model is based on the masking phenomenon. It is related to the notion of critical band analysis, which is a central notion since the auditory perception is based on a similar analysis in the inner ear. The noise masking properties are modeled by calculating a *noise masking threshold.* A listener tolerates additive noise as long as it remains below the masking threshold.

In audio coding this threshold helps to find the optimal quantization parameters which produce an inaudible distortion. In the enhancement process it allows to determine the audible noise components, based on the knowledge that below the masking threshold the noise is masked.

2. THE ENHANCEMENT METHOD

Consider a single channel system corrupted by additive background stationary noise. The noisy speech can be

0-7803-2431-5/95 $4.00

expressed as following :

$$y(n) = s(n) + n(n) \quad (1)$$

where $s(n)$ is the original speech signal and $n(n)$ the additive noise. The processing is done on a frame–by–frame basis in the frequency domain. It is mainly composed of two phases : calculation of the noise masking threshold according to the method described in [11], and noise subtraction as detailed in Section 2.2. The speech signal processed by the enhancement algorithm is denoted $\hat{s}(n)$. Frames with an overlap of 50% are used and Hamming windowing is applied. For each frame, the optimal parameters of subtraction are computed whithin each frequency band, depending on the masking threshold. The corresponding block diagram is described in Figure 1.

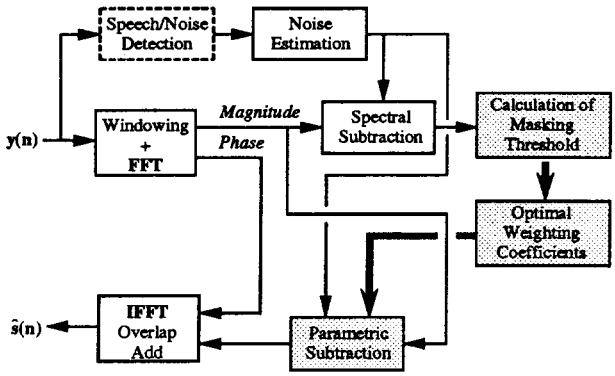

Figure 1: *Block diagram of the enhancement method.*

2.1. Noise masking threshold

The masking model used in perceptual audio coding allows to compute a threshold, which is then used to control the quantization process of the spectral components. This model takes into account only *simultaneous masking* (masking in the frequeny domain) : a weak signal is made inaudible by a stronger signal occuring simultaneously. This model is simple enough to avoid a great computational load. It shows good performances in speech coding, even though it does not take into account temporal masking (backward and forward masking).

The calculation of the masking threshold is described in [11]. It is composed of the following steps, as illustrated in Figure 2 :

1. Critical band analysis : the energies of each critical band are added up. Frequencies within the same critical band are equally perceived by the human ear.
2. Masking across critical bands : the application of a spreading function allows to take into account the masking between signals in different critical bands. A convolution is performed with the spreading function, which operates on a Bark scale.
3. Calculation of the noise masking threshold : subtraction of a threshold offset.
4. Renormalization and comparison with the absolute hearing threshold.

The calculation of the threshold is based on a rough estimate of the short–time spectral magnitude. It is a good way of incorporating properties of the auditory model in the enhancement process. The threshold is used to control the spectral subtraction process. The suppression of noise is not necessary at the frequencies where the noise components are masked by speech.

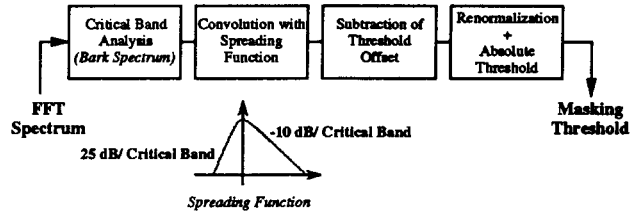

Figure 2: *Calculation of the masking threshold.*

2.2. Noise subtraction rule

The enhanced speech is formed from the estimation of the short–time spectral magnitude obtained with spectral subtraction according to the following equation :

$$\hat{s}(n) = \text{IFFT}\left[|\hat{S}(\omega)| \cdot e^{j\theta_y(\omega)}\right] \quad (2)$$

where $\theta_y(\omega)$ represents the phase of the noisy speech signal. In this paper the generalized spectral subtraction [2, 3] is used. Hence, the enhanced spectral magnitude is given by :

$$|\hat{S}(\omega)| = [|Y(\omega)|^\gamma - \alpha\,|N(\omega)|^\gamma]^{1/\gamma} \quad (3)$$

where $|Y(\omega)|$ represents the spectral magnitude of the noisy speech and $|N(\omega)|$ represents the estimation of noise, updated during speech pauses. This formulation includes previously developed algorithms such as spectral magnitude subtraction, power subtraction or Wiener filtering.

This parametric formulation is chosen because of its flexibility in varying the noise suppression characteristics and because it allows an improvement of the

performances. Spectral components have to be positive and negative components are set to the masking threshold.

Single channel enhancement systems are limited by a tradeoff between noise reduction and speech distortion. This tradeoff determines the choice of the parameters in Equation (3). The originality of this algorithm resides in the fact that the noise masking threshold is used to control the enhancement process adaptively, hence helping to find the best tradeoff.

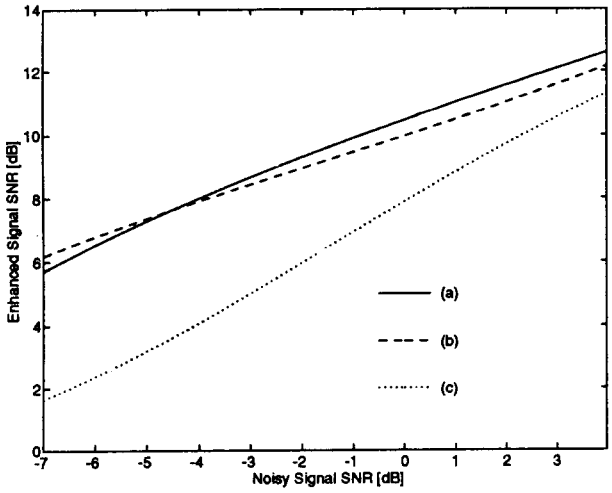

Figure 3: *Performance comparison in terms of input–output SNR : (a) Proposed algorithm. (b) Power spectral subtraction with residual noise reduction. (c) Spectral magnitude subtraction.*

3. RESULTS

Experiments have been made with speech corrupted by additive white noise. The performances of the proposed algorithm are compared to two other subractive–type algorithms :

1. Basic spectral magnitude subtraction [1].
2. Power spectral subtraction including residual noise reduction [3].

Figure 3 shows an objective comparison. If we compare the performances of the proposed algorithm with those of the algorithm proposing residual noise reduction, it can be observed that the background noise is well reduced. It is shown that in terms of SNR, the improvement is comparable. However, subjective listening tests show that the proposed algorithm produces a perceptually more pleasant output.

This can be explained by the observation of the speech spectrograms in Figures 4 to 7. These spectrograms show the distribution of noise in the time–frequency plane. Figure 4 shows the spectrogram of the original test signal. Figure 5 illustrates the noisy speech signal. Figure 6 shows the speech enhanced by power subtraction and residual noise reduction, where the musical structure of the residual noise can be observed. On the other hand, Figure 7 shows the result obtained with the proposed algorithm. It is clearly visible that the residual noise is reduced and that it has lost its structure, leading to a perceptually less annoying result.

These results show that if the end–user is human, the goal should not only be the maximization of the output SNR, since it does not measure the perceptual quality of the enhanced speech signal.

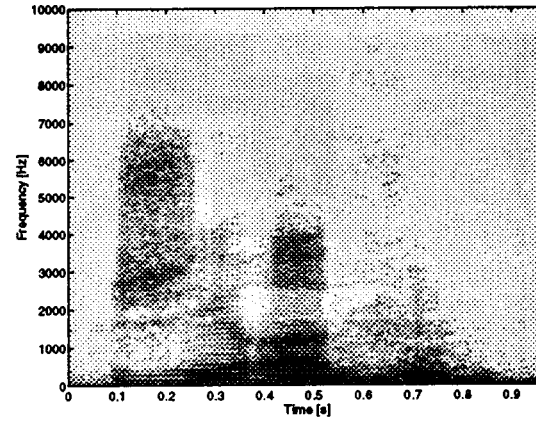

Figure 4: *Original speech signal sampled at 20 kHz.*

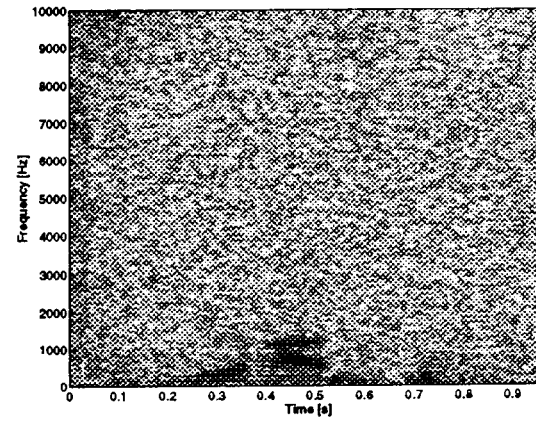

Figure 5: *Noisy speech signal with a SNR=-2dB.*

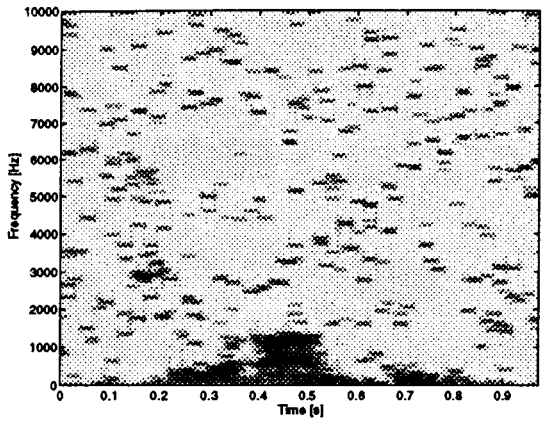

Figure 6: *Speech resulting from the enhancement with power spectral subtraction and residual noise reduction.*

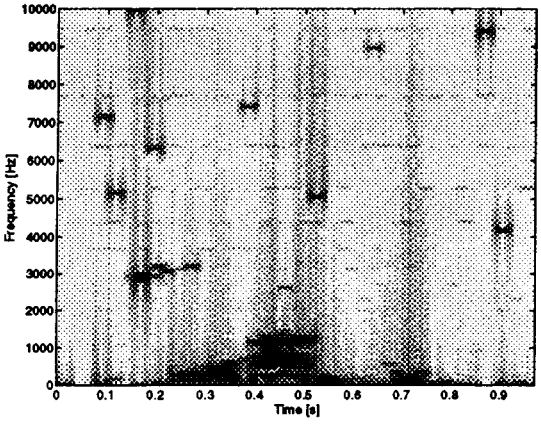

Figure 7: *Speech resulting from the enhancement with the proposed algorithm.*

4. CONCLUSION

Subtractive–type enhancement methods introduce a perceptually annoying residual noise. In this paper, a simple but efficient way to take into account properties of the auditory system in the enhancement process is proposed. Instead of using a mathematical criterion which is not well correlated with perception, a new algorithm introducing a criterion based on auditory masking is developed. Results show that the background noise is reduced and that the residual noise is less structured than with the classical methods, while the distortion of speech remains acceptable. Hence, we can conclude that the introduction of considerations based on perceptual properties in the enhancement process is very promising.

5. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. on Acoust., Speech, Signal Proc.*, vol. ASSP-27, pp. 113–120, April 1979.

[2] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, vol. 67, pp. 1586–1604, December 1979.

[3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise", *in Proc. IEEE ICASSP*, pp. 208–211, Washington, DC, April 1979.

[4] R.J. McAulay and M. L. Malpass, "Speech enhancement using a soft decision noise suppression filter", *IEEE Trans. on Acoust., Speech, Signal Proc.*, vol. ASSP-28, pp. 137–145, April 1980.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. on Acoust., Speech, Signal Proc.*, vol. ASSP-32, pp. 1109–1121, December 1984.

[6] J.H.L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect", *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 598–614, October 1994.

[7] Y.M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence", *IEEE Trans. on Signal Processing*, vol. 39, pp. 1943–1954, September 1991.

[8] S. Nandkumar and J.H.L. Hansen, "Speech enhancement based on a new set of auditory constrained parameters", *in Proc. IEEE ICASSP*, vol. I, pp. 1–4, Adelaide (Australia), April 1994.

[9] D. Tsoukalas, M. Paraskevas, and J. Mourjopoulos, "Speech enhancement using psychoacoustic criteria", *in Proc. IEEE ICASSP*, pp. 359–362, Minneapolis, April 1993.

[10] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception", *Proc. IEEE*, vol. 81, pp. 1385–1422, October 1993.

[11] J.D. Johnston, "Transform coding of audio signal using perceptual noise criteria", *IEEE J. on Select. Areas Commun.*, vol. 6, pp. 314–323, February 1988.