OPTIMIZING SPEECH ENHANCEMENT BY EXPLOITING MASKING PROPERTIES OF THE HUMAN EAR

A. AKBARI AZIRANI - R. LE BOUQUIN JEANNÈS - G. FAUCON

Laboratoire de Traitement du Signal et de l'Image - Université de Rennes 1
Bât. 22 - Campus de Beaulieu - 35042 RENNES CEDEX - FRANCE

ABSTRACT

The problem of speech enhancement and mainly noise reduction in speech remains a key-point of hand-free telecommunications. A great number of techniques have been already put forward and for a few years an auditory model has been investigated in noise reduction. In this paper a new approach for enhancing a speech signal degraded by uncorrelated stationary additive noise is developed. In this approach the simultaneous masking effect of the human ear is exploited. Two states "noise masked/noise unmasked" are derived from a noise masking threshold computed using a rough estimate of the speech signal. Then a speech signal estimator is proposed as a weighted sum of the individual estimators in each state. The gain in the signal to noise ratio (SNR) and a distortion measure indicate some improvement in real noise conditions. Subjectively this improvement is noticeable only at high input SNRs.

1. INTRODUCTION

One of the drawbacks of the speech enhancement methods in noisy environments is the distortion of the useful signal. The compromise between the signal distortion and the level of residual noise is a well known problem in speech enhancement [1]. In fact these two quantities cannot be minimized simultaneously. When one tries to remove noise from speech, some components of the useful signal may be cancelled or attenuated due to the strictness of the noise reduction filter.

Recent developments in modeling the masking effects of the human ear seem to have attained the required accuracy to be used in different speech processing applications. They have been used successfully in speech coding and in the evaluation of speech quality [2].

In our noise reduction approach some noise components are inaudible and we have no interest to suppress them for fear of introducing distortion or musical noise. So, an auditory model is considered.

2. BASIC PRINCIPLE

The general idea consists in considering that the clean speech signal can be in different "states", in the present instance masking or not masking the disturbing noise. In consequence, the speech estimator should be composed of individual estimators each corresponding to the previous states. This estimator is obtained by a weighted sum of the individual estimators, where the weights are given by the probabilities of the two states.

The idea of using different estimators for different classes or states of the signal was first introduced by Drucker [3] who considered five categories of speech signals, comprising fricatives, stops, vowels, glides and nasals. It was also exploited by McAulay and Malpass [4] who distinguished the two states "silence" and "non-silence". We carry on this idea to take into account the fact that noise can be inaudible.

Using the auditory masking effects for development of noise reduction techniques has interested some researchers in the past years. Mourjopoulos et al. [5] calculated the audible noise using the masking threshold to improve spectral subtraction. Tsoukalas et al. [6] used the frequency masking properties to extract some information on the audible noise before performing a non-linear spectrum filtering. In our case the auditory masking threshold allows us to compute the posterior probabilities of the two states "noise masked/noise unmasked" given the noisy observation.

3. AUDITORY MASKING MODEL

3.1. Presentation

The probability of masking noise by speech must be computed from an auditory masking model. Among the three kinds of known masking effects (simultaneous, forward and backward masking) we consider only the simultaneous masking. It is a frequency domain phenomenon where a low level signal (masked) can be made inaudible by a simultaneously occuring stronger signal (masker). First of all we compute the masking threshold in order to derive the amount of noise that can remain inaudibly in the enhanced speech signal. On the other hand this threshold is an indicator of the probability of masking noise by speech. In our application only the noisy speech signal is accessible, so the clean speech power spectrum must be estimated before computing the masking threshold. We use a spectral subtraction to obtain a rough estimate of the speech signal power spectrum. No noise overestimation is introduced to avoid any supplementary distortion of the estimated speech spectrum. Among the different approaches for the computation of the masking threshold we retain the method introduced in [7] in which the author calculates

0-7803-2431-5/95 $4.00 © 1995 IEEE

the masking threshold using following steps. First of all he computes the speech signal energy in each critical band. Secondly a spreading function is applied to the critical band spectrum to take into account the effects of masking across critical bands. Then an offset is added to distinguish the case where noise masks a tone from the inverse case. Finally he accounts for absolute threshold of ear and relates the spread masking threshold to the critical band masking threshold.

3.2. Implementation

We use a 256 points FFT analysis for computing the estimated signal power spectrum. Then we consider the frequency scale in BARK (z) that is related to the frequency scale in Hertz with the following nonlinear relation as in [8]:

$$z = 13 \times \mathrm{Arctan}(0.76 \times f_{kHz}) + 3.5 \times \mathrm{Arctan}\left(\frac{f_{kHz}}{7.5}\right)^2.$$

For a 0 - 4 kHz speech signal we have 18 critical bands. The spectral lines are distributed nonuniformly in the BARK domain and there are more spectral lines in the last critical band than in the first one. In fact we have a high resolution in computing the masking threshold in the last critical bands. We consider a simple triangular spreading function with slopes of +25 dB and -10 dB. The slope of this function is steeper towards lower frequencies i.e. the higher frequencies are more easily masked. The BARK domain signal power spectrum is convolved with the above spreading function and the following steps are as in [7].

Fig. 1 plots the spectrum of a frame of speech signal and its noise masking threshold. The threshold value is not constant for all the frequency components in a critical band. We have a resolution of 1/3 BARK in the first critical band (corresponding to 3 spectral lines in this band) and 1/18 BARK in the last critical band.

Fig. 2 shows the noise masking threshold computed from the original clean speech spectrum (same as in Fig. 1) and the one calculated from the estimated speech spectrum for an input signal to noise ratio (SNR) equal to 10 dB. There is some error in the estimated masking threshold due to the estimation of the speech spectrum. So, we try to prevent the masking curve from being above the curve deduced from the clean speech in subtracting a constant value from the estimated masking threshold. Given this new threshold and the noisy speech spectrum the audible noise spectrum is computed as in [6]. Let $s(k)$ and $n(k)$ denote the speech and noise processes respectively. The observed signal $x(k)$ may be written:

$$x(k) = s(k) + n(k).$$

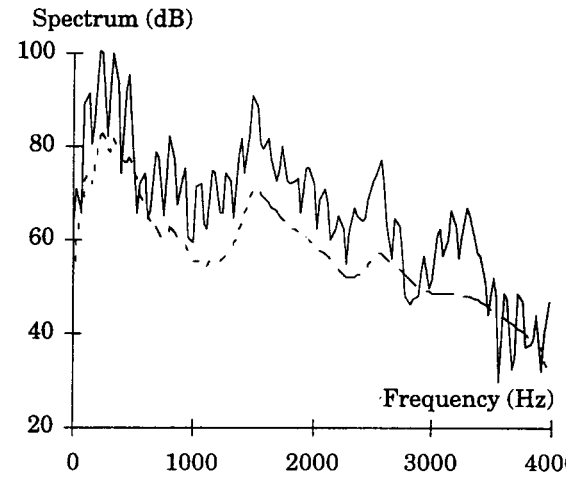

Fig. 1. Short time power spectrum of the clean speech signal (——) and noise masking threshold (– – –).

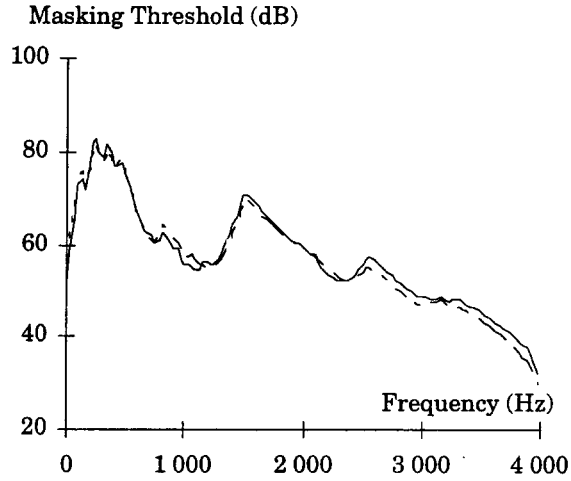

Fig. 2. Noise masking threshold: computed from clean speech (——), computed from estimated speech (– – –).

The noisy signal in the frequency domain is:

$$X(f)\exp(j\,\varphi_X(f)) = S(f)\exp(j\,\varphi_S(f)) + N(f)\exp(j\,\varphi_N(f))$$

where $X(f)$, $S(f)$ and $N(f)$ denote the amplitudes of short time Fourier transforms of the signals $x(k), s(k), n(k)$ respectively and $\varphi_X(f), \varphi_S(f), \varphi_N(f)$ determine their phases. Let $\tilde{S}(f)$ and $T(f)$ denote the initial estimate of $S(f)$ and the noise masking threshold respectively. The audible noise spectrum can be computed as:

$$\begin{cases} |X(f)|^2 - |\tilde{S}(f)|^2 & \text{if } |X(f)|^2 > T(f) \text{ and } |\tilde{S}(f)|^2 > T(f) \\ |X(f)|^2 - T(f) & \text{if } |X(f)|^2 > T(f) \text{ and } |\tilde{S}(f)|^2 < T(f) \\ 0 & \text{otherwise.} \end{cases}$$

Now the probability of noise masked state can be determined as a function of the audible noise. A simple approach consists in considering a probability of one for the noise masked state when the audible noise is null and a probability of zero otherwise. The results given in section 5 are obtained using this simplified approach.

4. SPEECH ENHANCEMENT METHOD

A well known amplitude estimator of the signal $\hat{S}(f)$ that minimizes the mean-squared spectral error is the conditional mean:

$$\hat{S}(f) = E\left[S(f) \mid X(f)\right]$$

where the expectation operator is used to indicate averaging over the ensemble of noise sample functions, speech envelopes and phases and the ensemble of speech events. The signal estimate is obtained using the phase of the noisy observation and inverse Fourier transform. We consider a binary hypothesis model where H_0 is the hypothesis that signal masks noise and H_1 is the opposite hypothesis. By developing the expectation operator for these two speech events we obtain:

$$\hat{S}(f) = E\left[S(f) \mid X(f), H_0\right] \times P\left[H_0 \mid X(f)\right] + E\left[S(f) \mid X(f), H_1\right] \times P\left[H_1 \mid X(f)\right]$$

where $P\left[H_k \mid X(f)\right]$ is the probability that the speech signal is in the state H_k given that the observed noisy signal amplitude in the frequency domain is $X(f)$. Since, under hypothesis H_0, speech signal masks noise, the best estimator $\hat{S}(f)$ that introduces neither speech signal distortion nor audible noise is directly the observation $X(f)$. Hence the estimator is given by:

$$\hat{S}(f) = P \times X(f) + (1-P) \times E\left[S(f) \mid X(f), H_1\right]$$

where

$$P = P\left[H_0 \mid X(f)\right] = 1 - P\left[H_1 \mid X(f)\right]$$

is the probability of noise masked state that was determined from our auditory masking model in the previous section.

Under hypothesis H_1 a short time spectral amplitude estimator that minimizes the mean-squared error is used. This one is a simplified version of the estimator proposed by Ephraïm and Malah [9] that is shown to work much better than the classical spectral subtraction estimator [10]. This estimator takes the uncertainty of signal presence in the noisy observation $X(f)$ into account:

$$E\left[S(f) \mid X(f), H_1\right] = G_1(f) \times G_2(f) \times X(f).$$

$G_1(f)$ is the gain function of the MMSE estimator and $G_2(f)$ represents the gain function due to the uncertainty of speech presence:

$$G_1(f) = \frac{R_{pr}(f)}{1 + R_{pr}(f)}$$

$$G_2(f) = \frac{\Lambda(f)}{1 + \Lambda(f)}.$$

$R_{pr}(f)$ can be interpreted as the a priori SNR and $\Lambda(f)$ is the generalized likelihood ratio that is calculated as:

$$\Lambda(f) = \frac{1-q}{q} \times \frac{1}{1 + R_{pr}(f)} \times \exp\left[\frac{R_{pr}(f)}{1 + R_{pr}(f)} R_{po}(f)\right].$$

$R_{po}(f)$ is the a posteriori SNR and q is the probability of signal absence in $X(f)$. The estimation of $R_{pr}(f)$ and $R_{po}(f)$ and the performance of the above estimator are detailed by Ephraïm [9] and Cappé [10]. When the spectrum of the estimated signal is below the masking threshold, it is set to this last value in order to decrease the signal distortion.

5. RESULTS

We have tested the proposed estimator using objective measures and informal listening tests and compared it with two other estimators. The first one is the amplitude spectral subtraction with a noise overestimation factor equal to 1.5. The second one is the simplified version of the estimator by Ephraïm presented in section 4 which doesn't take the masking properties of the human ear into account. The disturbing noise is recorded in a car moving at 130 km/h and added to a speech signal recorded in a stopped car to obtain the noisy signal. Four input SNRs (from 5 to 20 dB) have been considered to simulate different real noise conditions. In terms of objective measures, we evaluate the segmental gain in the SNR and a segmental distortion that is computed from the useful signal. To this end, each of the three estimators is expressed as a noisy speech filtering. Then this filtering is applied to the clean speech signal $s(k)$ to obtain the filtered signal $s'(k)$ and we compute the distortion D by:

$$D = \frac{1}{M} \sum_{m=0}^{M-1} \frac{\sum_{k=0}^{K-1} \left[s(k+Km) - s'(k+Km)\right]^2}{\sum_{k=0}^{K-1} \left[s(k+Km)\right]^2}$$

where K is the number of signal samples in a frame and M is the number of frames.

Fig. 3 shows that the gain of the proposed estimator is greater than (or equal to) the gain obtained by the two others. The gain improvement in the higher input SNRs is more noticeable because of the more accurate masking

threshold. The signal distortion caused by the proposed method is lower than the one brought by the Ephraïm's estimator or the spectral subtraction (see Fig. 4).

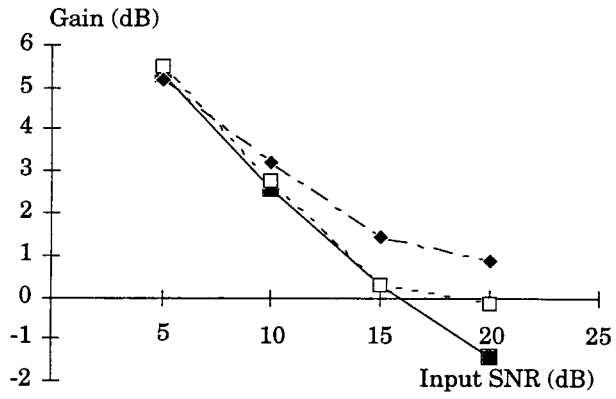

Fig. 3. Gain in the SNR: Sp. subt. (——) Ephraïm (– – –) Proposed estimator (– - – -).

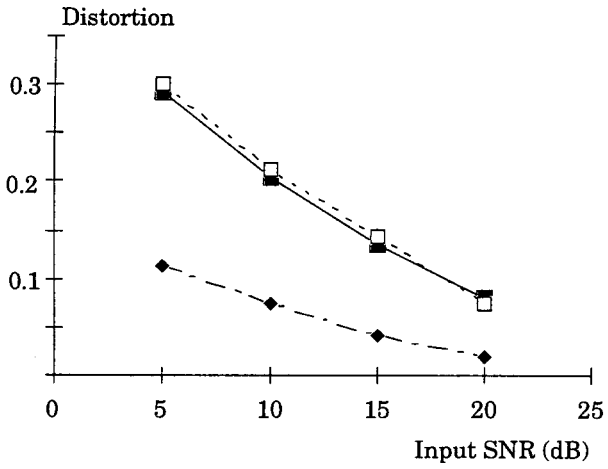

Fig. 4. Distortion: Sp. subt. (——) Ephraïm (– – –) Proposed estimator (– - – -).

As it could be judged by informal listening tests, the spectral subtraction suffers from the well known residual "musical noise" and the enhanced speech seems to be distorted. The residual noise of Ephraïm's estimator doesn't sound musical and is less annoying but the speech signal appears distorted too. The proposed estimator has a colorless residual noise compared with Ephraïm's estimator; the enhanced speech seems less distorted at high input SNRs.

6. CONCLUSION

In conclusion we obtain some improvement considering two states "noise masked/noise unmasked" for the speech signal and optimal estimators for each state. This is a general approach that can be used for optimizing different speech estimators to give a less distorted enhanced speech. It would be interesting to see the behavior of the proposed estimator when forward and backward masking are considered. Finally the computation of the probability of noise masked state as an experimental function of audible noise must be investigated.

REFERENCES

[1] Y. Ephraïm, H. L. Van Trees "A Signal Subspace Approach for Speech Enhancement", ICASSP, pp. II.355-II.358, 1993.

[2] K. Brandenburg, G. Stall, "The ISO/MPEG - Audio Codec : a Generic Standard for Coding of High Quality Digital Audio", AES, 92nd convention, Vienna, March 1992.

[3] H. Drucker, "Speech Processing in a High Ambient Noise Environment", IEEE Trans. Audio Electroacoust., vol. AU-16, n°2, pp. 165-168, June 1968.

[4] R. J. Mc Aulay, M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. on ASSP, vol. ASSP-28, n°2, pp. 137-145, April 1980.

[5] J. Mourjopoulos et al., "Noisy Audio Signal Enhancement Using Subjective Spectra", AES, 92nd Convention, Vienna, March 1992.

[6] D. Tsoukalas et al., "Speech Enhancement Using Psychoacoustic Criteria", IEEE ICASSP, pp. II.359-II.362, 1993.

[7] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", IEEE Journal on Selected Areas in Communications, vol. 6, n°2, pp. 314-323, February 1988.

[8] C. Colomès, "Étude d'un Modèle d'Audition et d'une Mesure Objective de la Qualité d'un Signal Sonore dans le Contexte du Codage à Réduction de Débit", Thèse de l'Université de Rennes 1, 1994.

[9] Y. Ephraïm and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. on ASSP, vol. ASSP-32, n°6, pp. 1109-1121, December 1984.

[10] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraïm and Malah Noise Suppressor", IEEE Trans. on Speech and Audio Processing, vol. 2, n°2, pp. 345-349, April 1994.