COMPUTATIONALLY EFFICIENT SPEECH ENHANCEMENT BY SPECTRAL MINIMA TRACKING IN SUBBANDS

Gerhard Doblinger
Institut für Nachrichtentechnik und Hochfrequenztechnik
Technische Universität Wien
Gusshausstr. 25, A-1040 Vienna, Austria
E-Mail: Gerhard.Doblinger@tuwien.ac.at

Abstract— **We present an efficient algorithm for the enhancement of speech signals which are heavily corrupted by short-time stationary, acoustically or electrically added disturbances. The algorithm is based on spectral amplitude estimation using an overlap-add FFT filter bank system. Compared to other systems, the improved performance of our speech enhancement system is achieved by the combination of the best known spectral amplitude estimators of the noisy speech signal and a new efficient and reliable noise spectrum tracker. As a result, our speech enhancement system requires no speech pause detection for noise estimation and needs only 14% – 23% of the resources of a commercially available digital signal processor.**

I. Introduction

THE enhancement of noisy speech is a challenging research field with applications including suppression of environmental noise in machinery halls, mobile radio communications systems, noise suppressors for automatic speech recognition systems, and hearing aids. Depending on the specific application, the disturbing noise could be background sounds like machinery noise, traffic noise, or electrical noise added during transmission of speech via communications channels. In all cases, we will deal with the most difficult situation where only the contaminated signal is available. No additional signal (e.g. a second microphone to pick up the noise) is present which can be used to facilitate the design.

Although there are various approaches for the enhancement of noisy speech, only a few offer an acceptable performance for real-world signals and have proved to be suited for real-time implementation on today's general purpose integrated digital signal processors. Among them are methods based on spectral amplitude estimation and adaptive Wiener filtering. Due to the use of subsampled FFT filter banks, these algorithms are computationally very efficient. However, their plain application suffers from essentially two drawbacks: First, the enhancement process introduces a perceptually annoying residual noise, called *musical tones*. Second, a noticeable signal distortion of the low-energy speech portions is present, especially at high noise levels. The speech enhancement algorithm presented in this paper takes care of this, and offers a good tradeoff between residual noise suppression and speech distortion.

The key point in the design of a practicable speech enhancement system is the parameter estimation of the disturbing noise. These parameters must be extracted from the noisy speech signal. Usually, the noise measurements are performed during speech pauses employing an automatic speech pause detection. However, the detection reliability deteriorates severely in most cases for input signal-to-noise ratios below 6 dB. In addition, for the tracking of non-stationary noise the parameter update during speech pauses is not sufficient. Therefore, we will present a noise spectrum tracker which is an alternative approach to [1]. It is computationally more efficient and it can be easily adjusted to different kinds of noise disturbances by means of only two parameters.

Another important design issue of speech enhancement systems is the proper algorithm selection for spectral amplitude estimation. Our method is based on the Ephraim-Malah spectral attenuation functions [2], [3] with minor modifications regarding function parameter determination.

II. Algorithm description

The basis of our system is the well known *filter bank overlap addition method* with an FFT length of $N = 256$, a decimation factor $M = 64$, and a Hamming (or Hanning) window function $w(n)$ for input signal weighting. Due to subsampling, an FFT-frame is processed every $M = 64$ samples only. The complete system is depicted in some detail in Fig. 1.

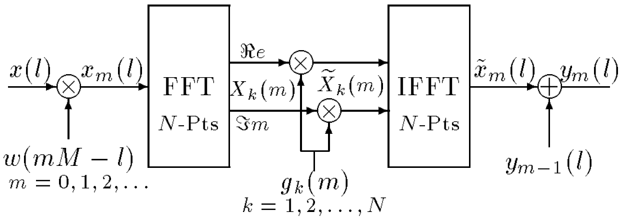


Fig. 1. Block diagram of speech enhancement algorithm (frame index m, buffer index $l = n - N + 1, \ldots, n + 1, n$, sampling index n, frequency index k), with spectral amplitude modification by $g_k(m)$.

According to Fig. 1, we modify the spectral magnitudes only and leave the phases unchanged. The associated gain factors $g_k(m)$ can be obtained by an *minimum mean square error* (MMSE) estimation of the spectral amplitudes in the presence of additive noise, where $g_k(m)$ depends on the signal powers of speech and of noise, respectively, at the individual frequency index k [2], [3].

Denoting $X_k(m)$ as the k^{th} spectral component of frame number m of the noisy speech signal (see Fig. 1), we estimate its short-time power spectrum by

$$P_{x\,k}(m) = \alpha P_{x\,k}(m-1) + (1-\alpha)|X_k(m)|^2, \quad (1)$$

with forgetting factor α between $0.7 \ldots 0.9$ which insures an adequate tracking of the short-time stationary speech spectrum. The advantage of power spectrum estimation Eq. 1 is its computational simplicity and the fact that no measurement delay is introduced. Any additional delay in the speech enhancement algorithm would require a large extra storage space for the complex-valued spectral components. Furthermore, the resulting signal delay is normally not acceptable in telecommunications applications.

The short-time noise spectral power which is also needed for the spectral amplitude estimation of the noisy speech signal is computed by our new nonlinear estimator:

if $P_{n\,k}(m-1) < P_{x\,k}(m)$ **then**

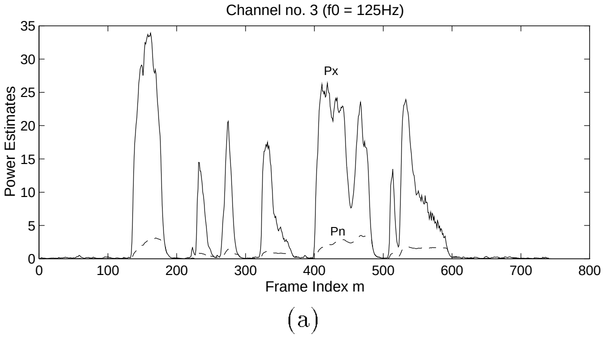
$$P_{n\,k}(m) = \gamma P_{n\,k}(m-1) + \frac{1-\gamma}{1-\beta}\left(P_{x\,k}(m) - \beta P_{x\,k}(m-1)\right) \quad (2)$$

else

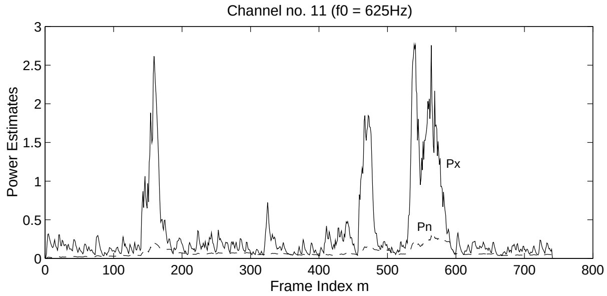
$$P_{n\,k}(m) = P_{x\,k}(m). \quad (3)$$

This noise spectrum estimation performs some type of temporal minima tracking of $P_{x\,k}(m)$ and is illustrated in Fig. 2 for subbands with center frequencies $f_0 = 125$ Hz (a), $f_0 = 625$ Hz (b), and $f_0 = 1062$ Hz (c).

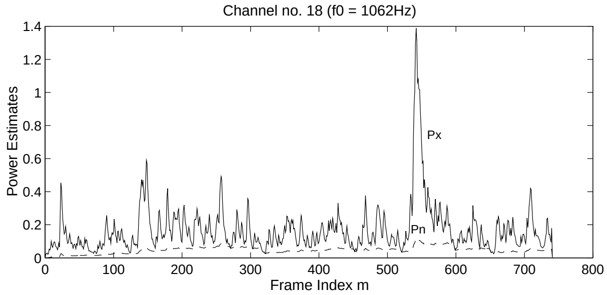
An alternative noise spectrum estimator has been presented in [1]. However, our approach is computationally more efficient. In addition, the estimator Eq. 2, 3 has a build-in "look-ahead" minimum tracking due to the β-term in Eq. 2. Typical parameter selections are $\alpha = 0.7, \beta = 0.96, \gamma = 0.998$ which yield a noise adaptation period of 0.2 to 0.4 seconds. Thus, our speech enhancement algorithm can be easily adjusted to short-time stationary disturbances by means of a few parameters.



(a)



(b)



(c)

Fig. 2. Examples of spectral power estimation using Eq. 1 – 3 (solid lines for P_x, dashed lines for P_n) for a noisy speech signal with SNR = 0 dB, plotted at different subbands with center frequencies $f_0 = 125$ Hz (a), $f_0 = 625$ Hz (b), and $f_0 = 1062$ Hz (c).

In order to achieve the noise reduction of the FFT-based speech enhancement system, the spectral amplitudes $|X_k(m)|$ are modified by real-valued gain factors $g_k(m)$ (see Fig. 1). Normally, spectral subtraction or Wiener filtering methods are applied to compute these gain factors using the estimated spectral powers $P_{x\,k}(m)$ and $P_{n\,k}(m)$. However, these basic algorithms suffer from a very pronounced residual noise (see [4] for a study on the *musical tones phenomenon*).

A better residual noise behavior can be achieved by more advanced techniques based on simplified speech models [5], [2], [3]. We have implemented and compared these methods in our speech enhancement system. According to our listening tests, the Ephraim-Malah gain factors are superior to those of McAulay-Malpass, since they introduce less speech signal distortion. However, the original gain factor algorithms have to be refined to further reduce the residual noise.

We will focus here only on one of the Ephraim-Malah gain factors which are derived in [3] by minimizing a logarithmic spectral amplitude error criterion. With these gain factors, we have obtained the best musical tone reduction in a variety of applications. Using this minimization procedure, the following gain factors (attenuation curves) can be found:[1]

$$g_k = f(R_{post_k}, r_k) = r_k \exp\left(\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-x}}{x}\,dx\right), \quad (4)$$

[1] The frame index m is omitted for clearness.

with

$$r_k = \frac{R_{prio_k}}{1+R_{prio_k}}, \tag{5}$$

and

$$v_k = (1+R_{post_k})\, r_k. \tag{6}$$

The function parameters $R_{prio} = \frac{P_s}{P_n}$ and $R_{post} = \frac{P_x}{P_n} - 1$ are called *a prior SNR* and *a posteriori SNR*, respectively [2], [5]. The role of these parameters is somewhat confusing, especially when the gain factors of McAulay-Malpass are compared with those of Ephraim-Malah. Some illumination on this fact can be found in [6]. In our speech enhancement system the gain factors of Eq. 4 are precomputed and stored in tables. A plot of the attenuation curves is shown in Fig. 3.

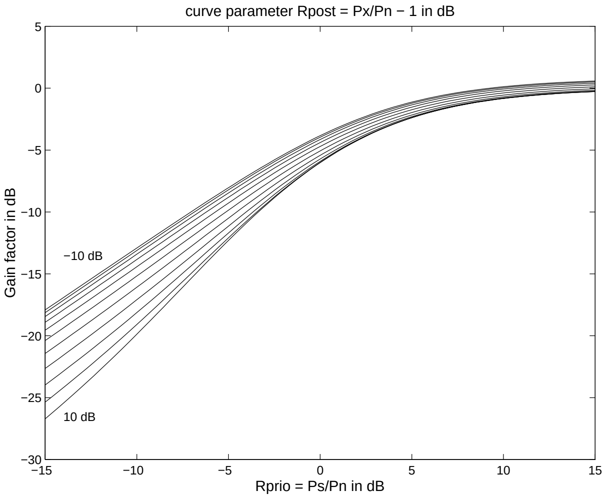


Fig. 3. Example of attenuation curves according to Eq. 4 with curve parameter R_{post} in 2 dB steps.

The problem involved with the gain factor computation given in Eq. 4 – 6 is the estimation of the parameters R_{prio} and R_{post}. On one hand, this estimation must take into account the short-time stationarity of speech signals, and on the other hand it must allow for sufficient smoothing to reduce the residual noise. Based on the power estimates given in Eq. 1 – 3 and on our results of extensive listening tests, we adopted the heuristic approach proposed in [2].

The detailed algorithm which combines the spectral amplitude minima tracking and the improved gain factor computation is summarized in Table I.

TABLE I

Algorithm for computing gain factors $g_k(m)$.

```
for each frame $m$ do
    compute $P_{x_k}(m)$, $P_{n_k}(m)$ using Eq. 1 – 3
    $P_{post_k}(m) = \max(0, P_{x_k}(m) - \delta_k P_{n_k}(m))$
    $P_{prio_k}(m) = (1-\eta) P_{post_k}(m) +$
                   $+ \eta g_k^2(m-1) P_{x_k}(m-1)$
    $R_{post_k}(m) = \frac{P_{post_k}(m)}{\delta_k P_{n_k}(m) + \varepsilon}$
    $r_k(m) = \frac{R_{prio_k}(m)}{1 + R_{prio_k}(m)}$
           $= \frac{P_{prio_k}(m)}{P_{prio_k}(m) + \delta_k P_{n_k}(m) + \varepsilon}$
    $g_k(m) = f(R_{post_k}(m), r_k(m)).$
```

The estimation of the *a posteriori SNR* which requires knowledge of the speech spectral power is performed by subtracting a noise power estimate from the signal plus noise power (first equation in Table I). A one-way rectification operation is applied to insure that this estimate is always greater or equal to zero. The noise bias factor δ_k has been included for further musical tone suppression. This approach to obtain a speech spectral power estimate is common to spectral subtraction algorithms. In addition, a small constant ε in the respective equations avoids division by zero during silence periods.

As argued in [2], the determination of the *a priori SNR* should be chosen as a weighted sum of the current *a posteriori SNR* and an SNR computed with a speech power estimate of the previous frame. Since the weighting factor η is selected close to one, the latter term dominates in the equation for $P_{prio_k}(m)$ (Table I). The role of η on the musical tone suppression performance is investigated in some detail in [6].

III. Experimental results

In order to visualize the functioning of our speech enhancement algorithm, a typical spectral power estimation $P_{x_k}(m)$ and the associated gain factors $g_k(m)$ are presented in Fig. 4 (frame index m plotted on horizontal axis, frequency corresponding to index k shown on vertical axis). A relatively high input SNR of 15 dB has been chosen only to clearly illustrate the system behavior with a monochrome picture. Our speech enhancement algorithm can handle input SNRs down to -8 dB (white gaussian noise disturbance).

The upper image shown in Fig. 4 displays all the information which is used to compute the gain factors. Light gray areas correspond to the speech components. The dark gray textured background associates from additive white noise. The lower picture shows the gain factors $g_k(m)$, where the dark areas are the suppression regions ($g_k(m) \approx 0$) and the light gray bordered parts are the pass band regions ($g_k(m) \approx 1$) in the time-frequency plane. This picture clearly indicates that only speech portions are passed by the system and the noise is suppressed. Therefore, the system operation can be interpreted as a time-variant filtering process matched to the short-time spectrum of the speech signal. At the beginning of the time axis, the adaptation phase of the noise suppression is visible. Light spots in the dark region are fluctuations

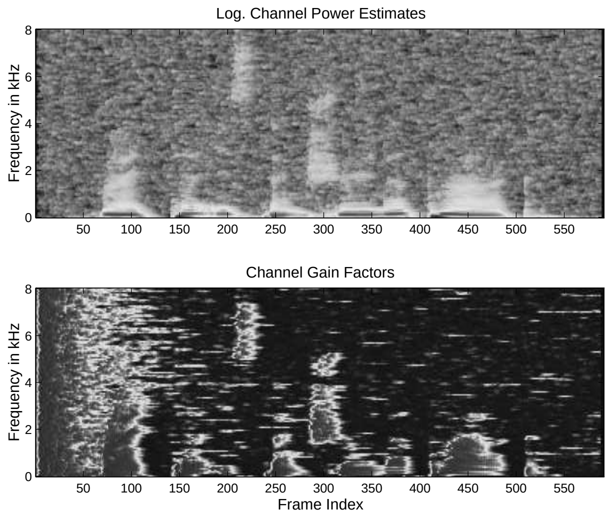


Fig. 4. Example of short-time power spectrum $P_{xk}(m)$ **(above)**, and gain factors $g_k(m)$ **(below)** (16 kHz sampling frequency, input SNR = 15 dB (additive white noise), frame index m, frequency index k).

which cause very low residual noise.

A rather hard test for the tracking behavior of a speech enhancement system is a disturbance with synthetic helicopter noise. Such a signal can be generated by sinusoidal amplitude modulation of white gaussian noise, e.g. with a modulation index 0.5 and a modulation frequency of 7 Hz. An example of this test situation is shown in Fig. 5. The striped pattern of the noise spectrum is clearly visible in the upper image. The gain factor image shows that most of the disturbances are suppressed. However, some speech signal degradation may be inferred from the lower picture in Fig. 5. Nevertheless, listening tests indicate that this distortion is acceptable and that the quality of the noisy speech is significantly improved.

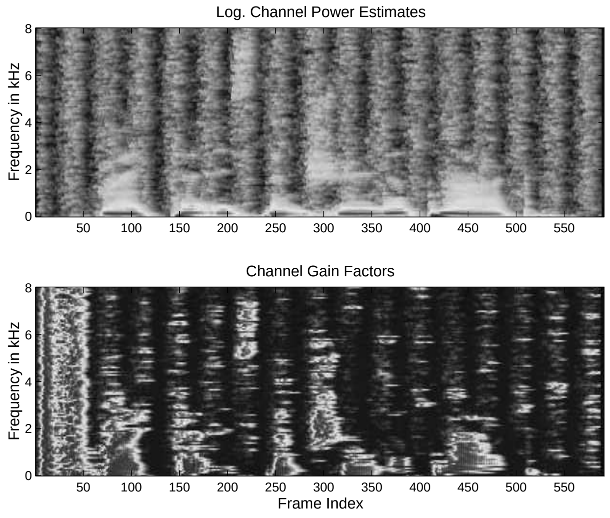


Fig. 5. Short-time power spectrum $P_{xk}(m)$ **(above)**, and gain factors $g_k(m)$ **(below)** (speech signal disturbed by synthetic helicopter noise with (overall) input SNR = 15 dB).

In order to optimize and extensively test our speech enhancement system in real-time operation, we have implemented it as an assembly language program on a single integrated floating-point digital signal processor (Analog Devices' ADSP-21020). The selection of a floating-point processor is mainly motivated by the ease of programming. Since there is no large dynamic range for data representation, a fixed-point arithmetic would also be sufficient. Be means of the real-time implementation, we have been able to optimize the parameters of our speech enhancement system for a wide variety of noise disturbances. To mention a few, we can significantly reduce synthetic $1/f$-noise, helicopter noise, and real-world noise like cockpit noise in cars. Due to the highly efficient algorithm, the processor utilization is only 14% – 23%, depending on the selected gain factor algorithm. Thus, there is a large reserve of computer power left for future improvements and extensions of our speech enhancement system.

IV. Conclusions

In this paper, we have presented a highly efficient speech enhancement algorithm with a novel noise estimation scheme, combined with an advanced spectral amplitude estimation taken from literature. The noise estimation performs a spectral minima tracking in subbands, and thus allows for suppression of short-time stationary noise disturbances. Furthermore, no speech pause detection is required.

According to informal listening test with various speech material, the speech enhancement algorithm offers a performance superior to conventional spectral subtraction methods and to Wiener optimum filters.

References

[1] R. Martin, “Spectral subtraction based on minimum statistics”, *in Proc. Seventh European Signal Processing Conference*, pp. 1182–1185, Sept. 1994.

[2] Y. Ephraim, D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator”, *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim, D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator”, *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, pp. 443–445, April 1985.

[4] P. Vary, “Noise suppression by spectral magnitude estimation – mechanism and theoretical limits”, *Signal Processing*, vol. 8, pp. 387–400, 1985.

[5] R. J. McAulay, M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter”, *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 137–145, April 1980.

[6] O. Cappé, “Elimination of the musical noise phenomenon with Ephraim and Malah noise suppressor”, *IEEE Trans. Speech, Audio Processing*, vol. 2, pp. 345–349, April 1994.