Time-Frequency Filtering of Speech Signals in Hands-Free Telephone Systems

Srdjan Stanković

Time-varying filtering of noisy speech signals is a very attractive challenge, with the main question: how does the most appropriate time-varying filter scheme look like? Speech signals are of highly nonstationary and multicomponent nature. If we deal with filtering of noisy speech signals, as they occur in hands-free telephone systems, then the desired scheme would provide a signal-to-noise ratio (SNR) greater than approximately 12dB. At the same time it should be suited for real-time implementation, with time delay less than 39ms for mobile telephony, and 2ms for circuit-switched telephony. The second requirement can cause additional difficulties and restrictions on finding an appropriate time-varying filter procedure.

The most commonly used approach in the filtering of speech signals is the so called quasi-stationary approach, where it is assumed that the signal is stationary in the time interval T, with T between 20 ms and 40ms being often used [1], [14]. In this interval of time, classical speech enhancement schemes such as those given in Table I are used [6], [16].

The noise is reduced by applying frequency-dependent suppression factors according to the various filtering rules given in Table I. From the aspect of time-varying filtering we can say that this technique is quasi time-varying filtering. Thus, we have a sliding window of duration T along the signal where the filtering is performed after every T or after every $T/2$ (the second case is used in an overlap-add scheme in order to avoid block effects).

From the point of view of time-frequency analysis, having in mind the high nonstationarity of speech signals, we can conclude that the quasi-stationary approach of filtering is approximate in nature, and that it will more or less satisfy subjective perception requirements. If we want to achieve more accurate and more objective filtering of speech signals, time-varying filtering needs to be applied. Since a unique definition of time-frequency spectra does not exist, several approaches to time-varying filtering have been proposed. We will use the one based on the Wigner distribution. It uses the Wigner spectrum, where the statistically independent cross-terms in Wigner distribution are averaged out. However, in order to calculate the Wigner spectrum it is necessary to have many different realizations of the same random process at a given instant. Obviously, in the case of real time applications, the processing has to be based on a single noisy speech realization. It is the reason for using an approximation, in the sense that the Wigner spectrum is replaced by a cross-terms free (reduced) time-frequency distribution. According to the additional criterion of realization simplicity, special attention will be devoted to the filtering based on the spectrogram and distributions whose realization is directly related to the spectrogram. The use of other reduced interference time-frequency distributions in the filtering, instead of the spectrogram, is straightforward.

A. Time-Variant Filtering of Speech Signals

By analogy with the filtering of stationary signals, nonstationary time-varying filtering of a noisy signal can be defined by [7], [11], [15]:

$$(Hx)(t) = \int_{-\infty}^{\infty} h(t+\frac{\tau}{2}, t-\frac{\tau}{2})x(t+\tau)d\tau. \quad (1)$$

Time-Frequency Signal Analysis and Processing, ed. B. Boashash, Elsevier, 2003.

The signal

$$x(t) = s(t) + \epsilon(t)$$

is a noisy one with the desired signal $s(t)$ and the noise $\epsilon(t)$. Impulse response of the time-varying filter is $h(t+\frac{\tau}{2}, t-\frac{\tau}{2})$. The optimal transfer function

$$L_H(t,f) = \int_{-\infty}^{\infty} h(t+\frac{\tau}{2}, t-\frac{\tau}{2}) x(t+\tau) e^{-j2\pi f\tau} d\tau$$

is defined by the relation [4], [5], [8], [10]:

$$\overline{W}_{sx}(t,f) = L_H(t,f)\overline{W}_{xx}(t,f), \qquad (2)$$

where

$$\overline{W}_{xx}(t,f) = E\{W_{xx}(t,f)\}$$
$$= \int_{-\infty}^{\infty} E\{x(t+\frac{\tau}{2}) x^*(t-\frac{\tau}{2})\} e^{-j2\pi f\tau} d\tau \quad (3)$$

is the mean value of the Wigner distribution $W_{xx}(t,f)$ of signal $x(t)$ (i.e. the Wigner spectrum of signal $x(t)$ [3]). We can conclude that (2) is of the same form as the Wiener filter for the stationary case.

If the signal and noise are not correlated, we have:

$$L_H(t,f) = \frac{\overline{W}_{ss}(t,f)}{\overline{W}_{ss}(t,f) + \overline{W}_{\epsilon\epsilon}(t,f)}. \qquad (4)$$

Consider now relation (4). Obviously, the mean value $E\{W_{ss}(t,f)\} = \overline{W}_{ss}(t,f)$ will eliminate uncorrelated cross-terms in the Wigner distribution, since

$$E\{s_i(t+\frac{\tau}{2}) s_j^*(t-\frac{\tau}{2})\} = 0 \quad for \quad i \neq j,$$

as long as components $s_i(t)$ and $s_j(t)$ are not correlated [3]. However, if we have to perform filtering on the base of a single realization, the Wigner distribution should be instead of the Wigner spectrum in (4). For filtering of multicomponent signals equation (4) is useless because emphatic cross-terms will appear. The problem of cross-terms will be partially overcome if we modify the definition (4) so that we apply some of the cross-terms reduced distributions $\rho(t,f)$ instead of the Wigner distribution. In this case we have:

$$L_H(t,f) = \frac{\rho_{ss}(t,f)}{\rho_{ss}(t,f) + \rho_{\epsilon\epsilon}(t,f)}. \qquad (5)$$

It is clear that definition (5) is an approximation of (4) with $\rho(t,f)$ approximating the Wigner spectrum.

In order to obtain a more efficient filter, for numerical implementation, the previous definitions can be slightly modified by using their pseudo form:

$$(Hx)(t) = \int_{-\infty}^{\infty} h(t+\frac{\tau}{2}, t-\frac{\tau}{2}) w(\tau) x(t+\tau) d\tau. \qquad (6)$$

Here, a lag window $w(\tau)$ is introduced. It can be shown that, for frequency modulated signals, $w(\tau)$ does not influence the output signal $(Hx)(t)$ if $w(0) = 1$ [11]. By using Parseval's theorem, (6) can be written in the form:

$$(Hx)(t) = \int_{-\infty}^{\infty} L_H(t,f) F_x(t,f) df \qquad (7)$$

where

$$F_x(t,f) = \int_{-\infty}^{\infty} x(t+\tau) w(\tau) e^{-j2\pi f\tau} d\tau$$

is the short-time Fourier transform of the signal $x(t)$.

The choice of $\rho(t,f)$, in (5) will play a crucial role in the time-varying filter scheme. Obviously, for an efficient time-varying filtering, it is desired that the chosen $\rho(t,f)$ satisfies three main conditions:

1) satisfactory noise reduction,
2) appropriateness for real-time realization,
3) its auto-terms are close to those in the Wigner spectrum.

The simplest and most commonly used $\rho(t,f)$, for which the real time application is very well studied, is the spectrogram. It is the square module of the short-time Fourier transform:

$$S_x(t,f) = |F_x(t,f)|^2$$
$$= \left| \int_{-\infty}^{\infty} x(t+\tau) w(\tau) e^{-j2\pi f\tau} d\tau \right|^2$$

The main problem of using the short-time Fourier transform (and the spectrogram) is in determination of the window width $w(t)$. A narrow window produces better time resolution, while a wider window gives better frequency resolution. The window should be chosen by a compromise of these two opposite requirements.

TABLE I

FILTER TRANSFER FUNCTION FOR DIFFERENT ALGORITHMS, WHERE $S_{xx}^2(\omega)$ AND $S_{\epsilon\epsilon}^2(\omega)$ ARE THE POWER SPECTRA OF THE NOISY SIGNAL AND NOISE RESPECTIVELY, AND λ IS AN OVERESTIMATION FACTOR.

Algorithm	*Wiener*	*Spectral subtraction*	*Maximum likelihood*	*Magnitude subtraction*
Filter tran. fun.	$1-\frac{S_{\epsilon\epsilon}^2(\omega)}{S_{xx}^2(\omega)}$	$\sqrt{1-\frac{\lambda S_{\epsilon\epsilon}^2(\omega)}{S_{xx}^2(\omega)}}$	$\frac{1}{2}[1+\sqrt{1-\frac{S_{\epsilon\epsilon}^2(\omega)}{S_{xx}^2(\omega)}}]$	$1-\sqrt{\frac{S_{\epsilon\epsilon}^2(\omega)}{S_{xx}^2(\omega)}}$

Having in mind that a speech signal is approximately stationary within the interval T between 20ms and 40ms, for a sampling rate of $f_s = 8$kHz, we conclude that we can use a lag window width of $N = 256$ samples, corresponding to $T = 32$ms. In order to achieve a more accurate calculation of integral (7), zero padding up to 1024 samples will be used.

Consider now the spectrogram-based filtering of noisy speech signals, recorded in a car cruising along the highway. Estimations of the spectrogram of noise are performed in only one time instant during a speech pause. This assumption is made in order to have the worst filtering situation as in a real case. Since the noisy signal contains significant noise components in the low frequency range (below 98Hz) where, in our application, no speech components exist, the signal is prefiltered by using a high-pass filter with cut-off frequency 98 Hz. In this realization we will apply the time-varying Wiener filter definition (5), with the spectrogram instead of $\rho(t,f)$, and the time-varying version of the spectral subtraction definition:

$$L_{HW}(t,f) = 1 - \frac{S_\epsilon(t,f)}{S_x(t,f)}. \tag{8}$$

$$L_{HSS}(t,f) = \sqrt{1-\lambda\frac{S_\epsilon(t,f)}{S_x(t,f)}} \tag{9}$$

In (9) λ is an overestimation factor applied in order to give some correction of the errors caused by the assumption that the noise is stationary in the interval between two pause estimations. The value $\lambda = 4$ is used. Modifications of the equations (8),(9) are used after introducing a spectral floor[14]:

$$L_{HW}(t,f) = \max\{L_{HW}(t,f), \beta\} \tag{10}$$

and

$$L_{HSS}(t,f) = \max\{L_{HSS}(t,f), \beta\}. \tag{11}$$

In our examples the spectral floors are set to $\beta = 0.12$ and $\beta = 0.08$ (in (10), and (11) respectively).

Note that by increasing λ, better noise reducing is obtained, but the distortion of signal becomes significant. By increasing β more noise remains in signal, but speech distortion is audible. Thus, these two factors are chosen by compromise.

The time-frequency representations of a clean signal and a noisy signal are shown in Fig.1(a) and (b), respectively. Time-frequency representation of denoised signal, filtered by using the time-varying Wiener filtering, and the time-varying spectral substraction filtering, based on the spectrogram, are shown in Fig.2(a) and (b). It is obvious that the noise suppression is better when the time-varying spectral substraction filter definition is used, because overestimation factor λ provides better estimation of the spectrogram of noise.

Now, there is the question whether it is possible to use some other time-frequency distributions, in order to further improve the filtering results. The answer is yes. Namely, we can use reduced interference distributions which belong to the general Cohen class of distributions [2]:

$$\rho_{xx}(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\tau,\nu)x(u+\frac{\tau}{2})$$

$$\times x^*(u-\frac{\tau}{2})e^{-j2\pi\nu t}e^{-j2\pi f\tau}e^{j2\pi\nu u}dud\theta d\tau \tag{12}$$

where the kernel $g(\tau,\nu)$ specifies the distribution. The most commonly used distributions

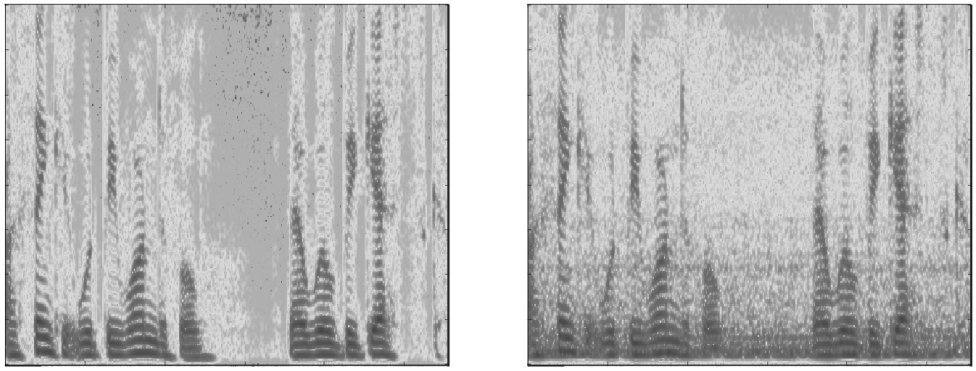

Fig. 1. Spectrogram of: a) the clean speech signal, b) the noisy signal filtered by a high-pass filter.

are: Choi-Williams distribution, Zao-Atlas-Marks distribution, Born-Jordan distribution, Zhang-Sato distribution, S-method...

When we use the reduced interference distributions, it is important to know that, in the case of a noisy signal, the distance between two auto-terms during voiced segments of speech is approximately equal to the value of the fundamental frequency in the case of a signal without noise. In the noisy case, we also have harmonically shaped components of noise, which can occur between the auto-terms of speech, causing additional cross-terms and errors in filtering [12].

A very simple and flexible implementation can be obtained by using the S-method (SM) [9], [Article 6.2], whose realization is straightforwardly based on the short-time Fourier transform. The result of this fact is that time-varying filtering based on the SM is a simple extension of the spectrogram-based filtering. Additionally, the SM of multicomponent signals:

$$x(t) = \sum_{i=1}^{N} x_i(t), \tag{13}$$

can assume the form $SM_{xx}(t,f) \cong \sum_{i=1}^{N} W_{x_i x_i}(t,f)$, being a desired approximation of the Wigner distribution auto-terms.

The SM is defined in the form:

$$SM_{xx}(t,f) = 2\int_{-\infty}^{\infty} P(\theta) F_x(t, f+\theta) F_x^*(t, f-\theta) d\theta. \tag{14}$$

where $P(\theta)$ is a rectangular window in the frequency domain.

Discretization of the SM (14), taking a rectangular window for $P(l)$, produces:

$$SM_{xx}(n,k) = \sum_{l=-L}^{L} F_x(n, k+l) F_x^*(n, k-l)$$

$$= |F_x(n,k)|^2 + 2Re\{\sum_{l=1}^{L} F_x(n, k+l) F_x^*(n, k-l)\}. \tag{15}$$

From the previous equation we see that the SM realization is based on the spectrogram. Thus, filtering based on the SM will be a straightforward extension of the previously considered filter schemes:

$$L_{HW}(t,f) = \max\left\{1 - \frac{SM_\epsilon(t,f)}{SM_x(t,f)}, \beta\right\} \tag{16}$$

and

$$L_{HSS}(t,f) = \max\left\{\sqrt{1 - \lambda\frac{SM_\epsilon(t,f)}{SM_x(t,f)}}, \beta\right\}. \tag{17}$$

In our experiments we have used the SM with $L = 3$, and spectral floors $\beta = 0.12$ and $\beta = 0.08$, respectively [13].

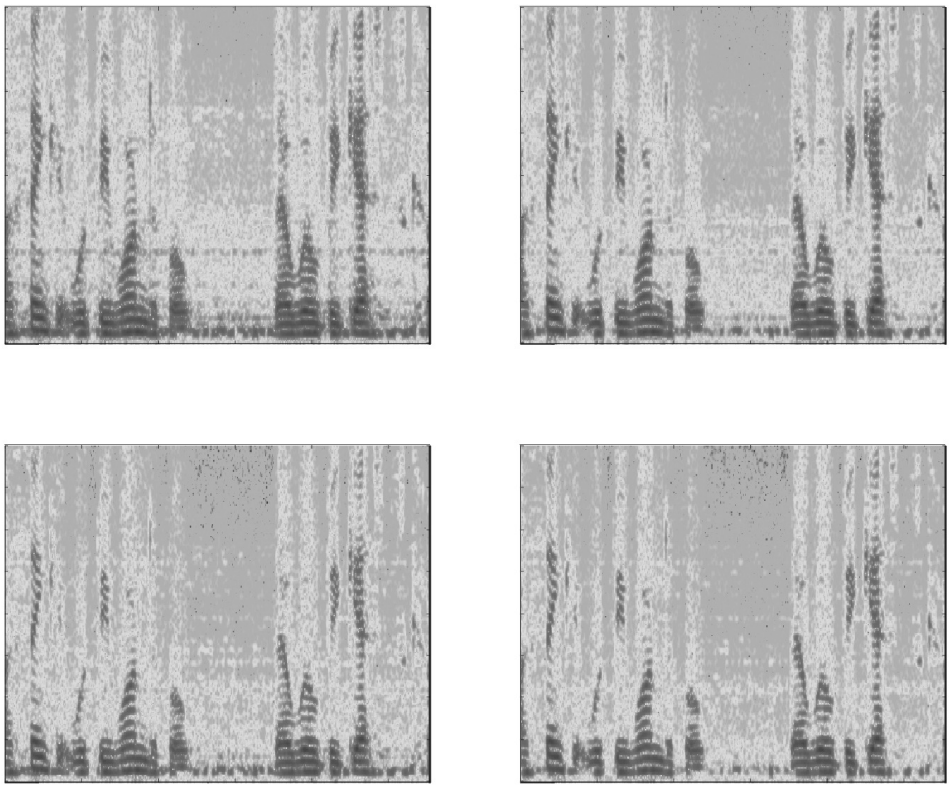

Fig. 2. Denoised signal obtined by filtering based on: a) time-varying Wiener filter using the spectrogram, b) time-varying spectral subtraction using the spectrogram, c) time-varying Wiener filter using the SM, d) time-varying spectral substraction using the SM.

The denoised signals by using the time-varying Wiener filtering and the time-varying spectral substraction filtering, based on the SM, are shown in Fig.2(c) and (d). By comparing the results with the ones produced by using the spectrogram based filtering, the improvements are obvious. It is important to note that the SM has a form very suitable for simple hardware realization. This property is attractive for on-line applications.

B. Summary and Conclusion

Time-varying filtering of speech signals disturbed by car noise is presented. On the base of the time-varying Wiener filter form, the time-varying spectral subtraction form of filtering is introduced. The filtering is performed on the base of the spectrogram and the S-method. The proposed filter schemes are efficient and suitable for hardware realization.

References

[1] C. Breining, P. Dreiseitel, E. Haensler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt and J. Tilp, "Acoustic echo control: An application of very-high-order adaptive filters," *IEEE Signal Processing Magazine*, vol.16, pp.42-69, July 1999.

[2] L. Cohen, *Time-frequency analysis*, Prentice-Hall, 1995.

[3] P. Flandrin and W. Martin, "The Wigner-Ville spectrum of nonstationary processes," in *Wigner distribution: Theory and applications in signal processing*, eds. W. Mecklenbrauker, F. Hlawatsch, Elsevier, 1997.

[4] H. Kirchauer, F. Hlawatsch and W. Kozek, "Time-frequency formulation and design of non-stationary Wiener filters," *IEEE Int. Conf. on ASSP*, pp.1549-1552, 1995.

[5] W. Kozek, "Time-frequency signal processing based on the Wigner-Weyl framework," *Signal Processing*, vol.29, No.1, pp.77-92, Oct.1992.

[6] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noise speech," *Proc. of IEEE*, vol.67, pp.1586-1604, Dec.1979.

[7] G. Matz, F. Hlawatsch and W. Kozek, "Generalized evolutionary spectral analysis and the Weyl spectrum of nonstationary random processes," *IEEE Transactions on Signal Processing*, vol.45, pp.1520-1534, Jun 1997.

[8] A. Papoulis, *Signal Analysis*, McGrow-Hill Book Company, 1977.

[9] LJ. Stanković, "A method for time frequency analysis," *IEEE Transactions on Signal Processing*, vol.42, pp.225-229, Jan.1994.

[10] LJ. Stanković, "On the time-frequency analysis based filtering," *Annales des Telecommunications*, vol.55, No.5-6, pp.216-225, May/June 2000.

[11] LJ. Stanković, S. Stanković and I. Djurović, "Space/spatial frequency based filtering," *IEEE Transactions on Signal Processing,* vol.48, pp.2343-2352, Aug.2000.

[12] S. Stanković, "About time-variant filtering of speech signals with time-frequency distributions for hands-free telephone systems," *Signal Processing*, vol.80, No.9, pp.1777-1785, 2000.

[13] S. Stanković and J. Tilp, "Time-varying filtering of speech signals using linear prediction," *Electronics Letters*, vol.36, No.8, pp.763-764, April 2000.

[14] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction,* Wiley and Teubner, 1996.

[15] H.L. Van Tres, *Detection, Estimation and Modulation Theory* , New York: Wiley, 1968.

[16] J. Yang, "Frequency domain noise suppression approaches in mobile telephone systems," *IEEE Int. Conf. on ASSP*, vol.2, pp.363-366, 1993.