# Adaptive Time Segmentation for Improved Speech Enhancement

Richard C. Hendriks, Richard Heusdens, and Jesper Jensen

***Abstract*—Single-channel enhancement algorithms are widely used to overcome the degradation of noisy speech signals. Speech enhancement gain functions are typically computed from two quantities, namely, an estimate of the noise power spectrum and of the noisy speech power spectrum. The variance of these power spectral estimates degrades the quality of the enhanced signal and smoothing techniques are, therefore, often used to decrease the variance. In this paper, we present a method to determine the noisy speech power spectrum based on an adaptive time segmentation. More specifically, the proposed algorithm determines for each noisy frame which of the surrounding frames should contribute to the corresponding noisy power spectral estimate. Further, we demonstrate the potential of our adaptive segmentation in both maximum likelihood and decision direction-based speech enhancement methods by making a better estimate of the *a priori* signal-to-noise ratio (SNR) $\xi$. Objective and subjective experiments show that an adaptive time segmentation leads to significant performance improvements in comparison to the conventionally used fixed segmentations, particularly in transitional regions, where we observe local SNR improvements in the order of 5 dB.**

***Index Terms*—Adaptive time segmentation, *a priori* signal-to-noise ratio (SNR), decision directed approach, hypothesis test, speech enhancement.**

## I. Introduction

THE NEED for single-channel enhancement of speech signals degraded by noise arises frequently, e.g., in mobile communication applications. Within single-channel speech enhancement, the noise is often assumed additive, i.e., $y = x + n$, with $y$ the noisy speech signal, $x$ the clean speech signal, and $n$ the noise realization. Further, it is common to assume that the clean speech signal and the noise process are uncorrelated. Recently, enhancement methods based on short-time spectral analysis (STSA) have received significant interest, partly due to their relatively good performance and low computational complexity. These methods transform the noisy speech signal frame by frame to the spectral domain, e.g., using a discrete Fourier transform (DFT). Here, complex-valued DFT coefficients of the clean signal are estimated by applying a gain function (e.g., the Wiener gain [1] or LSA gain [2]) to the noisy DFT coefficients. Subsequently, enhanced time-domain frames are generated using the inverse DFT and the enhanced waveform is constructed by overlap-adding the enhanced frames.

Gain functions are typically computed from two quantities, namely, an estimate of the noise power spectrum $P_{nn}(k,i)$ and of the noisy speech power spectrum $P_{yy}(k,i)$, where $k$ and $i$ denote the frequency bin index and the time-frame index. The gain function can directly be expressed in terms of those two quantities, e.g., as done in the class of spectral-subtraction algorithms, e.g., [3], or indirectly using the definition of the *a priori* signal-to-noise ratio (SNR) $\xi(k,i) = P_{xx}(k,i)/P_{nn}(k,i) = (P_{yy}(k,i) - P_{nn}(k,i))/P_{nn}(k,i)$, e.g., [4], with $P_{xx}(k,i)$ the clean speech power spectrum at frequency bin $k$ and time-frame $i$. However, in both situations, it is necessary to estimate the power spectrum of the noisy speech as well as the power spectrum of the noise process. While the problem of estimating and tracking the noise power spectrum in speech presence has received significant interest recently [5], methods for accurate estimation of the noisy speech power spectrum appear to have been less explored. A classical method to estimate the noisy speech power spectrum is the periodogram, computed as $|Y(k,i)|^2$, where $Y(k,i)$ is a Fourier coefficient of noisy speech. However, the periodogram estimator suffers from a variance of $var[|Y(k,i)|^2] \propto P_{yy}^2(k,i)$ [6]. To reduce the variance of the estimated noisy speech power spectrum, smoothing methods, e.g., the Bartlett method [6], can be used. The Bartlett method computes an estimated (smoothed) power spectrum by averaging periodograms of, say $N$ consecutive frames, hereby decreasing the variance of the power spectrum estimate by a factor $N$ [6]. However, the decrease in variance comes with a side effect: the frequency resolution is decreased as well. With the Bartlett estimate each segment may consist of $N$ frames including the frame to be enhanced, as shown in Fig. 1.

In Boll's work on spectral subtraction [3], the Bartlett method was used across segments consisting of three frames located symmetrically around the frame to be enhanced. Although this leads to a significant decrease in variance, this approach has a number of disadvantages. First, the position of the segment with respect to the underlying noisy frame that needs to be enhanced is predetermined. However, if the onset of a speech sound is not aligned with the start of the segment, the onset will definitely be smeared or blurred. Second, ideally, segments should vary with speech sounds: some vowel sounds may be considered stationary up to 40–50 ms, while stop consonants may be stationary for less than 5 ms [7]. A fixed segment size, as used in Fig. 1, has

Manuscript received December 8, 2004; revised January 20, 2006. This work was supported in part by the Philips Research and the Technology Foundation STW, in part by the applied science division of NWO, and in part by the technology program of the ministry of Economics Affairs. The material in this paper was presented in part at Interspeech in 2005 and at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, in 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Li Deng.

The authors are with the Department of Mediamatics, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: r.c.hendriks@tudelft.nl; r.heusdens@tudelft.nl; j.jensen@tudelft.nl).

Digital Object Identifier 10.1109/TASL.2006.872596

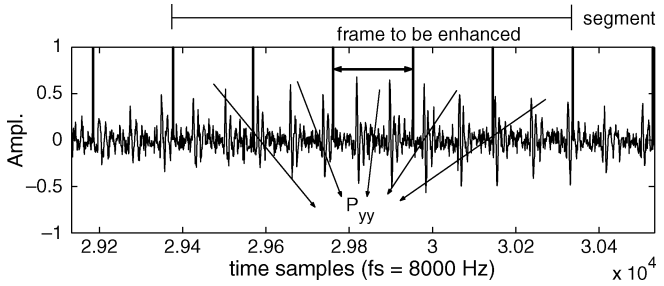



Fig. 1. Noisy speech signal with frame to be enhanced. In this example, a segment consists of five consecutive frames.

two potential drawbacks. First, in signal regions which can be considered stationary for longer time than the segment used, the variance of the spectral estimator is unnecessarily large. Second, if the stationarity of the speech sound is shorter than this fixed segment size, smoothing is applied across stationarity boundaries resulting in blurring of transients and of rapidly-varying speech components [8], leading to a degradation of the speech intelligibility.

In [8], a method was presented to overcome the above described problems using an adaptive exponential smoother. The amount of smoothing was adapted to the underlying speech process using a stationarity measure based on spectral derivatives measured over two consecutive frames.

In this paper, we propose a different approach to overcome the above mentioned problems, namely, an adaptive time segmentation. The proposed segmentation algorithm is very general. It can work as a front-end for most existing speech enhancement systems and is independent of the particular suppression rule (e.g., Wiener, LSA, etc.) that is used in the enhancement algorithm. To be more specific, the proposed method determines which noisy speech data should contribute in the estimation of the noisy speech power spectrum for a given frame, leading to better estimates of $P_{yy}(k,i)$. We then use the estimated $P_{yy}(k,i)$ within either the maximum likelihood approach [3], [4], or the decision directed (DD) approach [4] to estimate the *a priori* SNR $\xi(k,i)$, which can be used to define STSA gain functions. This leads to better estimates of $\xi(k,i)$ compared to conventional systems without adaptive segmentation. Furthermore, combining the DD approach with an adaptive segmentation will result in better approximations compared to what is currently used in practical implementations [4], [9]. Moreover, STSA gain functions based on this improved estimate of $\xi(k,i)$ lead to a decrease of residual noise and speech distortions in the enhanced signal.

The remainder of this paper is organized as follows. In Section II, we present an algorithm to determine an adaptive segmentation for speech enhancement. In Section III, we show that this adaptive segmentation leads to improved estimates of $P_{yy}(k,i)$. Further, we show how to use this estimate to improve the maximum likelihood and DD approach for speech enhancement. In Section IV, we evaluate the presented segmentation methods by means of subjective and objective experiments. In Section V, conclusions are drawn.

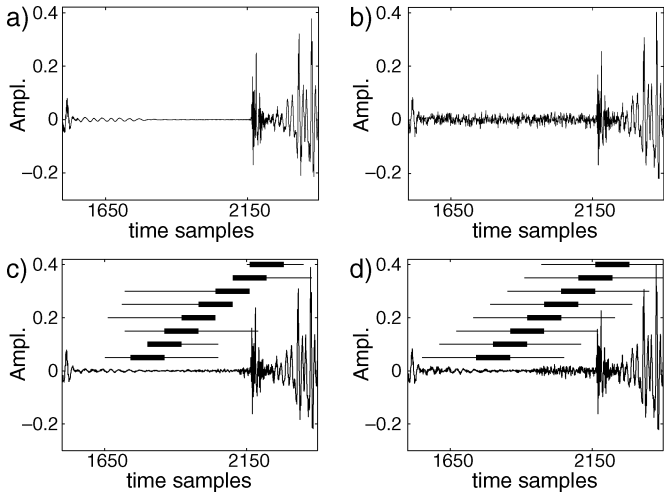



Fig. 2. (a) Clean speech signal. (b) Noisy speech signal with SNR of 10 dB. (c) Enhanced noisy speech using an adaptive segmentation. (d) Enhanced noisy speech using a fixed segmentation. (sample frequency is 8 kHz).

## II. Adaptive Time Segmentation

To illustrate the impact of an adaptive segmentation within a speech enhancement context, we show in Fig. 2 an example, where we compare time domain waveforms of a noisy speech signal enhanced using an adaptive segmentation and a fixed segmentation, respectively. Fig. 2(a) shows the clean speech signal that contains a stop consonant. Fig. 2(b) shows the noisy speech signal (white Gaussian noise) at an SNR of 10 dB. Fig. 2(c) shows the enhanced signal using a Wiener filter where the noisy speech power spectrum was estimated using a Bartlett estimate with an adaptive segmentation. For ease of illustration, the adaptive segmentation was here found under an ideal situation (i.e., using the clean speech signal). The adaptive segmentation that is used is shown in Fig. 2(c), where the thick lines mark the location of the signal frames, and the thin lines the adaptive segments that are used to estimate the noisy speech power spectrum for each frame. Fig. 2(d) shows the enhanced speech signal using a fixed segmentation. Comparing Fig. 2(c) and Fig. 2(d), it is clear that with the fixed segmentation in Fig. 2(d), the enhancement leads to a pre-echo present in front of the transient. With the adaptive segmentation in Fig. 2(c), no pre-echoes are present, because segments are adapted to the speech signal.

Our goal in this section is to develop an adaptive segmentation algorithm that finds for each frame a corresponding segment containing noisy speech samples and which can be assumed roughly stationary. We see that the segments found in this way depend on both speech and noise statistics. For a completely stationary noise source, the resulting segments will be decided by the stationary regions of the speech signal only, while if the noise is nonstationary as well, the resulting segments will generally become shorter in order to ensure stationarity of both speech and noise within each segment. To find an adaptive segmentation based on the noisy speech signal, we propose here a segmentation algorithm based on a probabilistic framework. The segments, used for estimation of the noisy power spectrum for a given frame, are formed based on the outcome of a hypothesis test. We test the hypotheses whether two consecutive sequences

of time-samples should be merged to form one segment or not, by regarding the sequences as outcomes of random processes and search for sequences that are stationary to a certain degree. In particular, we will use a test statistic based on a necessary condition for stationarity, namely that zero-lag correlation coefficients of the random process

$$R[0] = E\left\{|y|^2\right\}$$

where $R[0]$ is the correlation coefficient with lag 0 and $E$ the expectation operator, must remain invariant over time. Let $s_1$ and $s_2$ be two neighboring wide sense stationary segments, both consisting of independent frames with frame numbers $i \in \{n, \ldots, n+n_0-1\}$ and $j \in \{n+n_0, \ldots, n+N-1\}$, respectively, and let $\hat{R}_1^i[0]$ and $\hat{R}_2^j[0]$ denote estimates of $R[0]$ for each such frame. We can view $\hat{R}_1^i[0]$ and $\hat{R}_2^j[0]$ as realizations of random variables $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively. The two hypotheses then are as follows:

$$\begin{aligned} H_0 &: \mathcal{R}_1 \text{ and } \mathcal{R}_2 \text{ have the same distribution} \\ &\quad ([s_1, s_2] \text{ is considered stationary}) \\ H_1 &: \mathcal{R}_1 \text{ and } \mathcal{R}_2 \text{ do not have the same distribution} \\ &\quad ([s_1, s_2] \text{ can be considered not stationary}). \end{aligned} \tag{1}$$

Let $\hat{R}_1[0] \in \mathbb{R}^{n_0}$ and $\hat{R}_2[0] \in \mathbb{R}^{N-n_0}$ be vectors containing $n_0$ (iid) realizations of $\mathcal{R}_1$ and $N-n_0$ (iid) realizations of $\mathcal{R}_2$, respectively, and let $\hat{R}_{12}[0] = [\hat{R}_1[0]^T, \hat{R}_2[0]^T]^T \in \mathbb{R}^N$. The decision between the two hypotheses is made using the likelihood ratio test (LRT) [10]

$$\text{Reject } H_0 \text{ if } \quad \frac{p\left(\hat{R}_{12}[0]|H_1\right)}{p\left(\hat{R}_{12}[0]|H_0\right)} > \lambda \tag{2}$$

with $\lambda$ a decision threshold and $p(\hat{R}_{12}[0]|H_0)$ and $p(\hat{R}_{12}[0]|H_1)$ the likelihood of the sequence $\hat{R}_{12}[0]$ under hypothesis $H_0$ and $H_1$, respectively. In order to apply (2), pdfs, $p(\hat{R}_{12}[0]|H_1)$ and $p(\hat{R}_{12}[0]|H_0)$ must be determined. From the assumption of independent frames it then follows that $p(\hat{R}_{12}[0]) = p(\hat{R}_1[0])p(\hat{R}_2[0])$, $p(\hat{R}_1[0]) = \prod_{i=n}^{n+n_0-1} p(\hat{R}_1^i[0])$ and $p(\hat{R}_2[0]) = \prod_{j=n+n_0}^{n+N-1} p(\hat{R}_2^j[0])$.

We will argue in Section II-A that under certain assumptions the probability density function (pdf) $p(\hat{R}^i[0])$ and $p(\hat{R}^j[0])$ are Gaussian distributed and use the standard procedure of the Generalized LRT [10] and substitute unknown pdf parameters with their maximum likelihood estimates. In Section II-B, we will comment on the relation between the threshold $\lambda$ and the significance level $\alpha$ under the Neyman–Pearson theorem. In Section II-C, we present the algorithm that is used in combination with the LRT to find for each frame a corresponding segment.

### A. Distribution of $\hat{R}[0]$

To determine the distribution type of $\hat{R}[0]$, we assume that the noisy speech frames $y(i)$ are sufficiently long and that their statistics can be described by a multidimensional Gaussian distribution, i.e., $y(i) \sim N(0, R_y)$, with $R_y$ the noisy speech covariance matrix. Under this assumption, the covariance matrix $R_y = E[yy^H]$, with $H$ denoting hermitian transposition, is asymptotically equivalent to a circulant matrix [11]. The covariance matrix of $Y = \mathcal{F}y$, with $\mathcal{F}$ the DFT transform, can be written as $E[YY^H] = \mathcal{F}R_y\mathcal{F}^H$. Because the DFT is known to diagonalise a circulant matrix [11], $\mathcal{F}R_y\mathcal{F}^H$ is asymptotically diagonal, $Y$ is an uncorrelated multidimensional Gaussian sequence, and consequently also an independent sequence. The estimate of $R^i[0]$

$$\hat{R}^i[0] = \frac{1}{K}\sum_{k=1}^{K} |Y(k,i)|^2$$

where $K$ is the frame length and where $Y(k,i)$ are the noisy speech DFT coefficients, is therefore a sum of independent random variables. Using the central limit theorem, it follows that $\hat{R}^i[0]$ approaches a Gaussian distribution. In Fig. 3(a), a measured distribution of $\hat{R}^i[0]$, based on a synthetic speech

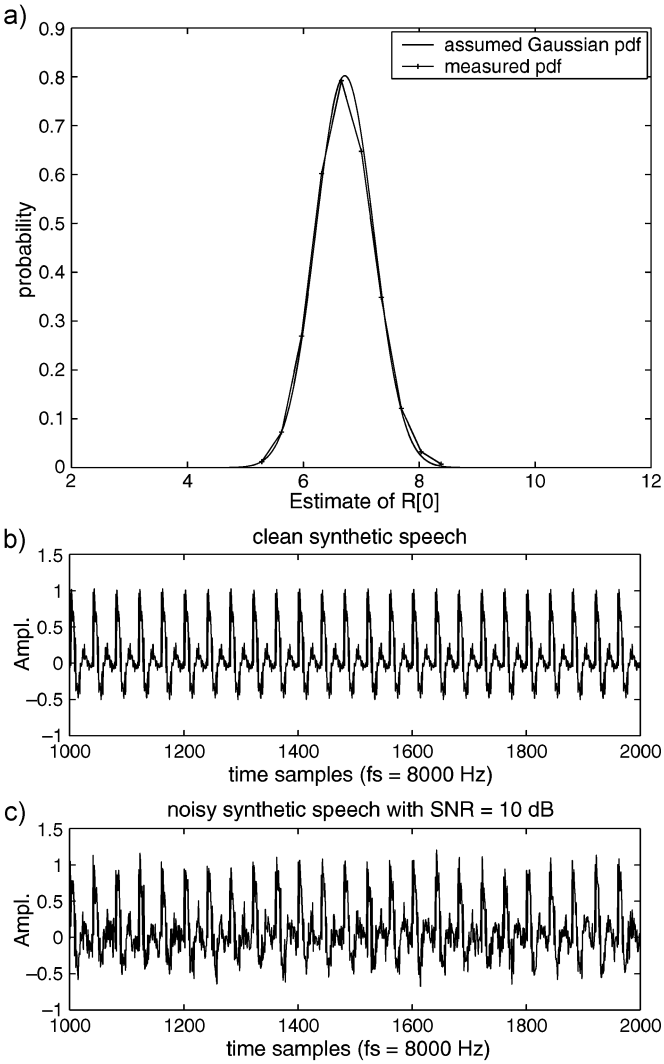



Fig. 3. (a) Measured distribution of $R[0]$ based on synthetic speech. (b) Synthetic clean speech signal. (c) Synthetic noisy speech signal.

signal shown in Fig. 3(b), is compared with a Gaussian distribution whose mean and variance are computed using $\hat{R}^i[0]$ values from the synthetic noisy speech data. It is shown that the measured histogram approximates the Gaussian distribution quite closely. The synthetic speech signal was created by filtering an impulse train through a time-invariant LPC-synthesis filter whose coefficients were extracted from a speech signal. The pdf was measured by windowing the noisy speech data followed by computation of $\hat{R}^i[0]$ per window. The reason to use a synthetic speech signal is to be able to create a long stationary sequence of speech with enough data to draw the histogram.

### *B. Relation Between $\lambda$ and $P_{fa}$ Under Neyman–Pearson Theorem*

Using the argumentation from Section II-A, it follows that $\hat{R}^i[0]$ can be assumed Gaussian distributed. This makes it possible to compute the likelihood ratio (2). However, it is not possible yet to decide whether $H_0$ should be accepted or rejected, since the threshold $\lambda$ is still unknown. Under the Neyman–Pearson criterion, the threshold is related to the false alarm probability $P(H_1|H_0)$, also referred to as the significance level $\alpha$, as follows:

$$\alpha = \int_{\lambda}^{\infty} p\left(\hat{R}_{12}[0]|H_0\right) d\hat{R}_{12}[0]. \tag{3}$$

This means that the threshold is dependent on the pdf of $\hat{R}_{12}[0]$ under $H_0$. However, in this particular situation, it is impossible to solve (3) for $\lambda$ for a given $\alpha$, because $p(\hat{R}_{12}[0]|H_0)$ is not completely specified. A common way to overcome this problem is to rewrite the likelihood ratio in the form $f(\hat{R}_{12}[0]) > \lambda'$,, where $f$ is a function of data dependent terms only, and $\lambda'$ is a threshold that is generally different from $\lambda$ in (3). Then, the relation between the false alarm probability $\alpha$ and the threshold $\lambda'$ can be made by derivation of the distribution of $f(\hat{R}_{12}[0])$.

With the Gaussian distribution of $\hat{R}^i[0]$ as derived in Section II-A, we can write the $H_0$ and $H_1$ hypotheses from (1) as follows:

$$\begin{aligned} &H_0 : \sigma_1 = \sigma_2,\ \mu_1 = \mu_2 \\ &H_1 : \text{otherwise} \end{aligned} \tag{4}$$

with $\sigma_1$ and $\sigma_2$ the standard deviations of $\hat{R}^i[0]$ in sequence $s_1$ and sequence $s_2$, respectively, and $\mu_1$ and $\mu_2$ the mean of $\hat{R}^i[0]$ in sequence $s_1$ and sequence $s_2$, respectively. Because $\hat{R}^i[0]$ is Gaussian distributed under both $H_0$ and $H_1$, it is possible to rewrite

$$\frac{p\left(\hat{R}_{12}[0]|H_1\right)}{p\left(\hat{R}_{12}[0]|H_0\right)} > \lambda$$

such that the left-hand side contains only data dependent terms

$$\frac{\hat{\sigma}_{12}^N}{\hat{\sigma}_1^{n_0}\hat{\sigma}_2^{N-n_0}} > \lambda. \tag{5}$$

Here, $\hat{\sigma}_{12}^2 = (1/N)\sum_{i=0}^{N-1}\left(R[0]^i - \hat{\mu}_{12}\right)^2$ is the maximum likelihood estimate of the variance of $R[0]^i$ over the concatenated sequence $s_1$ and $s_2$ with mean $\hat{\mu}_{12} = (1/N)\sum_{i=0}^{N-1} R[0]^i$, $\hat{\sigma}_1^2 = (1/n_0)\sum_{i=0}^{n_0-1}\left(R[0]^i - \hat{\mu}_1\right)^2$ the maximum likelihood estimate of the variance of $R[0]^i$ over sequence $s_1$ with mean $\hat{\mu}_1 = (1/n_0)\sum_{i=0}^{n_0-1} R[0]^i$ and $\hat{\sigma}_2^2 = (1/N-n_0)\sum_{i=n_0}^{N-1}\left(R[0]^i - \hat{\mu}_2\right)^2$ the maximum likelihood estimate of the variance of $R[0]^i$ over sequence $s_2$ with mean $\hat{\mu}_2 = (1/N-n_0)\sum_{i=n_0}^{N-1} R[0]^i$. However, we cannot express the distribution of the left-hand side of (5) in terms of known pdfs and as a consequence, we cannot express analytically the link between the value of $\lambda$ and the resulting false alarm probability.

However, by a slight modification of the hypotheses in (1) into a less strict form, we are able to relate $\lambda$ to a significance level $\alpha$. With the modified hypotheses, we test for a change in variance of $R[0]^i$ over sequence $s_1$ and $s_2$, while the means $\mu_1$ and $\mu_2$ for, respectively, $s_1$ and $s_2$ under both $H_0$ and $H_1$ are unspecified and possibly different. The two hypotheses in this situation are written as

$$\begin{aligned} &H_0 : \sigma_1 = \sigma_2 \\ &H_1 : \sigma_1 \neq \sigma_2. \end{aligned} \tag{6}$$

This leads to the same expression as in (5), but the maximum likelihood estimate of $\hat{\sigma}_{12}^2$ must now be computed as a pooled variance

$$\hat{\sigma}_{12}^2 = \frac{1}{N}\left(\sum_{i=0}^{n_0-1}\left(R[0]^i - \hat{\mu}_1\right)^2 + \sum_{i=n_0}^{N-1}\left(R[0]^i - \hat{\mu}_2\right)^2\right).$$

With this pooled variance, it is possible to write (5) in terms of $\hat{\sigma}_1$, $\hat{\sigma}_2$, $n_0$, and $N$ only [12]

$$\begin{aligned} \frac{\hat{\sigma}_{12}^N}{\hat{\sigma}_1^{n_0}\hat{\sigma}_2^{N-n_0}} &= \frac{\left(\frac{n_0}{N}\right)^{\frac{N}{2}}\left(\hat{\sigma}_1^2 + \frac{N-n_0}{n_0}\hat{\sigma}_2^2\right)^{\frac{N}{2}}}{\hat{\sigma}_1^{n_0}\hat{\sigma}_2^{N-n_0}} \\ &= \frac{\left(\frac{n_0}{N}\right)^{\frac{N}{2}}\left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} + \frac{N-n_0}{n_0}\right)^{\frac{N}{2}}}{\left(\frac{\hat{\sigma}_1}{\hat{\sigma}_2}\right)^{n_0}} \\ &> \lambda. \end{aligned} \tag{7}$$

It can be shown that (7) is a concave function [12]. Therefore it is sufficient to consider $\hat{\sigma}_1^2/\hat{\sigma}_2^2$ only and reject hypothesis $H_0$ if $\hat{\sigma}_1^2/\hat{\sigma}_2^2 < \lambda_1$ or $\hat{\sigma}_1^2/\hat{\sigma}_2^2 > \lambda_2$, with $0 < \lambda_1 < \lambda_2 < \infty$. The ratio $\hat{\sigma}_1^2/\hat{\sigma}_2^2$ has an $F$-distribution with $n_0, N-n_0$ degrees of freedom [13]. Therefore, it is now possible to relate the threshold to a significance level $\alpha$. This means that hypothesis $H_0$ is rejected when $\hat{\sigma}_1^2/\hat{\sigma}_2^2 < F_{n_1,n_2,\alpha/2}$ or $\hat{\sigma}_1^2/\hat{\sigma}_2^2 > F_{n_1,n_2,1-\alpha/2}$.

Although this procedure relates the significance level with the threshold, it tests for a weaker necessary condition for stationarity, because the modified hypothesis in (6) is weaker than the necessary condition to reach stationarity. In the following, we will, therefore, use the original hypothesis test as defined in (4)

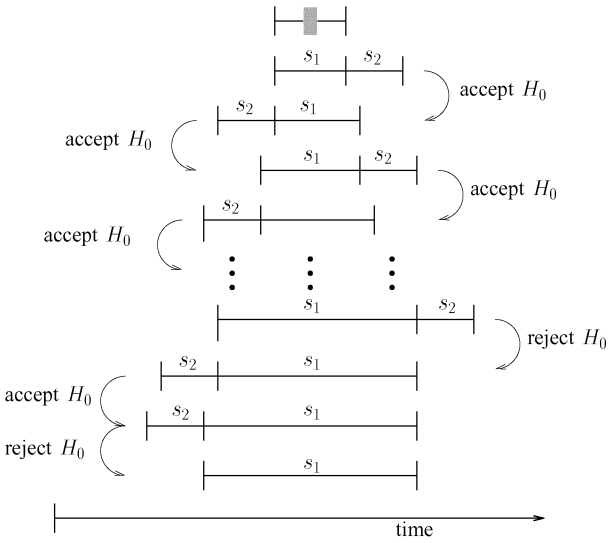



Fig. 4. Segmentation algorithm based on hypothesis.

without directly relating the threshold to the significance level; we will, however, show the results obtained with the weaker hypothesis test (6) as well, presenting our results in Section IV.

### C. Segmentation Procedure

Knowing the distribution of $\hat{R}[0]^i$, we are now in a position to compute the likelihood ratio given sequences $s_1$ and $s_2$

$$\frac{p\left(\hat{R}_{12}[0]|H_1\right)}{p\left(\hat{R}_{12}[0]|H_0\right)} = \frac{p\left(\hat{R}_1[0]|H_1\right)p\left(\hat{R}_2[0]|H_1\right)}{p\left(\hat{R}_{12}[0]|H_0\right)}$$

where the three probability distributions $p(\hat{R}_1[0]|H_1)$, $p(\hat{R}_2[0]|H_1)$, and $p(\hat{R}_{12}[0]|H_0)$ are all three Gaussian pdfs with possibly a different mean and variance. To find for a given frame a corresponding segment, we should in principle perform an exhaustive search over all possible segments. To avoid this computationally demanding full-search approach, we propose instead a computationally simpler algorithm which simulation experiments have shown to lead to the same performance as the full search algorithm. Fig. 4 describes this simplified algorithm. Start with a minimum segment $s_1$, which is assumed to be stationary and contains the frame under consideration (shaded area in Fig. 4). Then, extend this minimum segment with one frame at a time in an iterative process. Whether the segment should be extended with a neighboring frame is decided using the hypothesis test over sequence $s_1$ and a neighboring sequence $s_2$. We continue this process until on both sides of $s_1$, $H_0$ is rejected. The final sequence $s_1$ shown at the bottom of Fig. 4 is considered as the stationary segment that can be used in a Bartlett estimate of the noisy speech power spectrum.

This segmentation algorithm can be generalized by dividing the frequency range into subbands and determining a segmentation for each band independently. However, in this case, less information is present per band to estimate maximum likelihood parameters of the Gaussian pdf. This, in turn, means that the variance of these estimates will be larger than in the full-band case. We expect that increasing the number of bands may be beneficial for a small number of bands, but for a larger number of bands, the advantage of having many bands may be overshadowed by the increased variance of the parameters of the Gaussian pdf estimate in each band. Note that in a setup where the segmentation is determined per frequency band, the assumption of independent time-samples becomes less valid. Nevertheless, we will proceed with this setup, because, as we will see, it leads to improvements.

Fig. 5 shows a block scheme of the proposed segmentation algorithm in combination with an enhancement algorithm. First, a noisy speech signal frame $y$ is divided into $L$ frequency bands with an $L$-channel filterbank. Then, for each frequency band, the energy or $R[0]$-coefficient is computed and a segmentation is determined. This segmentation is then used to estimate the noisy power spectrum $P_{yy}$ in that band. Then, the resulting spectra of the subbands are combined to form the full band estimate $\hat{P}_{yy}$, which then is used together with an estimate $\hat{P}_{nn}$ to enhance the noisy signal resulting in the clean speech estimate $\hat{x}$.

In Figs. 6 and 7, we show the result of the above described hypothesis-based segmentation algorithm applied to speech signals degraded by white noise at an SNR of 15 and 5 dB, respectively. In the figures, the original clean speech signal is shown together with the resulting segmentation. The thick lines mark the frames in which the signal is divided for enhancement. The thin lines represent for each frame the corresponding segment that is found by the hypothesis based algorithm. In those examples, we used a full-band version of the previously described algorithm. In Fig. 6, the speech signal under consideration consists of four parts: an initial silence part, a transient, some ringing after the transient, and a voiced part. We see that frames in the silence and voiced part have long segments associated which cover, respectively, the whole silence and voiced part. Frames in the transient part have rather short segments, which prevents smearing of the transient. Further, the onset of the voiced part is resolved, preventing it from smearing into the ringing of the transient. In Fig. 7, the speech signal under consideration consists also of four parts. A voiced speech sound, a silence region, another voiced sound, and again a silence region. In Fig. 7, we see that the frames in the two voiced regions have corresponding segments that completely cover the whole voiced region, and frames in the two silence regions have segments that also cover the complete silence region.

Notice that the segments are found using future information. The more future information can be used, the more secure the estimates of $P_{yy}$ will become. The use of future information implies a certain latency. However, the latency in the presented segmentation algorithm is adjustable. In Section IV-B, we demonstrate that also with no latency or a limited latency, the use of an adaptive time segmentation leads to performance improvements.

## III. *A Priori* SNR Estimation Using Adaptive Segmentation

The Bartlett method reduces the variance of the estimate of $P_{yy}$ with a factor $N$ by averaging $N$ periodograms. In principle, the Bartlett estimate assumes no overlap and rectangular-windowed frames. However, other methods are developed that do allow overlap and other windows than the rectangular window,

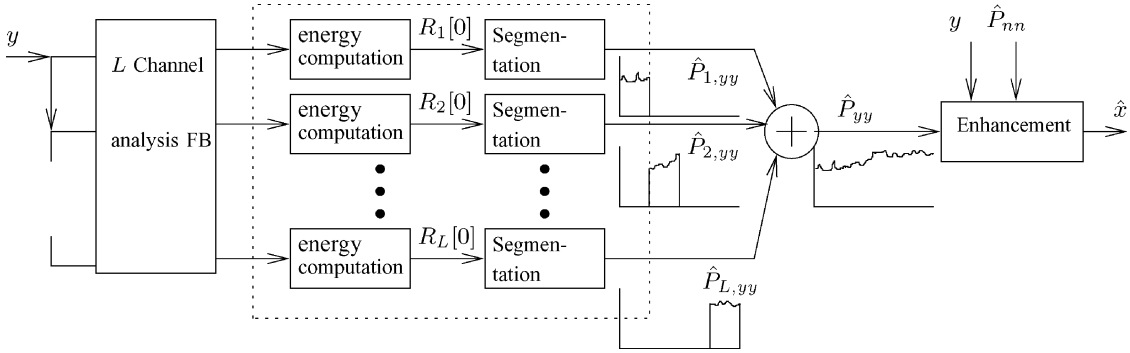



Fig. 5. Block diagram of adaptive segmentation speech enhancement system.

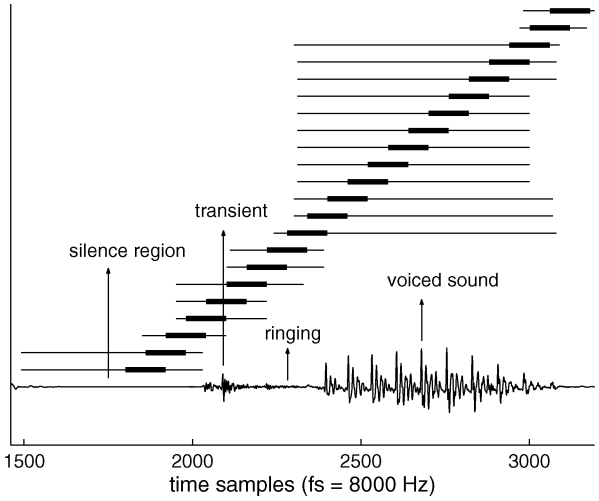



Fig. 6. Example segmentation. Thick horizontal lines: duration of frames. Thin horizontal lines: corresponding segments. Input SNR = 15 dB.

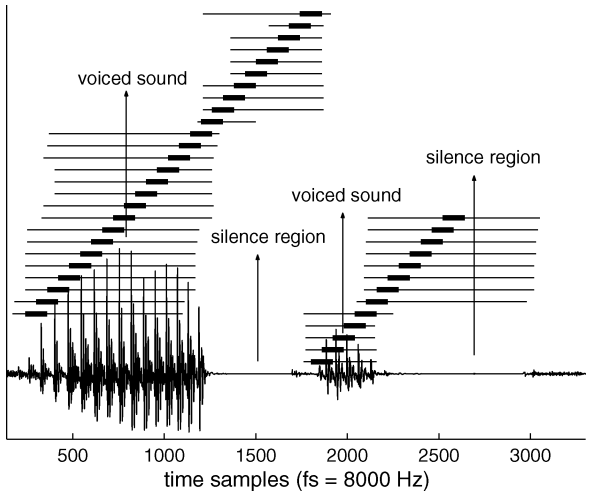



Fig. 7. Example segmentation. Thick horizontal lines: duration of frames. Thin horizontal lines: corresponding segments. Input SNR = 5 dB.

e.g., the Welch and Blackman–Tukey approach [6]. Note that a side effect of increasing the overlap is that the decrease in variance will become smaller than a factor $N$. To compute the Bartlett estimate in practice, a segment of $M$ samples is divided in frames of length $K$, and the periodograms of these $N = M/K$ frames are then averaged. Note also that the decrease in variance comes with a side effect, the frequency resolution of a periodogram based on a single frame is smaller than that of a periodogram based on the entire segment.

In conventional systems, $P_{yy}$ is estimated using either a Bartlett estimate with fixed segment length and fixed start and end positions of the segment [3], [4] or using a periodogram estimator [4]. The power spectral estimate can be improved by combining the Bartlett estimate with an adaptive segmentation, that is

$$\hat{P}^a_{yy}(k,i) = \frac{1}{N}\sum_{i=n_1}^{n_2} |Y(k,i)|^2 \tag{8}$$

with $n_1$ and $n_2$ the frame numbers that denote the start and end points of the segment found with the adaptive segmentation algorithm and $N = n_2 - n_1 + 1$ the number of frames in the segment. Estimation of $P_{yy}(k,i)$ using (8) leads to a reduced variance while preserving the transitional regions in a speech fragment, because it is adapted to the underlying speech signal.

In Sections III-A and III-B, we will present an improved maximum likelihood approach and an improved DD approach to estimate $\xi$ using this improved power spectral estimate $\hat{P}^a_{yy}$.

### *A. A Priori SNR Estimation Based on Improved Maximum Likelihood Approach*

The *a priori* SNR $\xi(k,i)$ for a frequency bin $k$ and time frame $i$ can be defined as

$$\xi(k,i) = \frac{P_{yy}(k,i) - P_{nn}(k,i)}{P_{nn}(k,i)}. \tag{9}$$

Under the assumption that speech DFT coefficients follow a Gaussian distribution, (9) is in [4] approximated by the maximum likelihood estimate of $\xi$

$$\hat{\xi}(k,i) = \frac{\frac{1}{N}\sum_{n=0}^{N-1} |Y(k,i-n)|^2 - \hat{P}_{nn}(k,i)}{\hat{P}_{nn}(k,i)}$$

where $\hat{P}_{nn}(k,i)$ is an estimate of the noise power spectrum, and $(1/N)\sum_{n=0}^{N-1} |Y(k,i-n)|^2$ is a Bartlett estimate of $P_{yy}(k,i)$ with fixed segment boundaries and fixed number of terms $N$. This makes it impossible to adapt appropriately to changes in the speech signal. Instead, the improved Bartlett estimate of (8)

can be used. This provides an estimate of $P_{yy}(k,i)$, where the start and end positions of the segment may vary for different frames. Consequently, the gain function will be better able to adapt to changes in the noisy speech signal, leading to a more efficient use of the data and less smearing of transitional areas in the speech signal. Insertion of (8) into (9) then leads to

$$\hat{\xi}(k,i) = \frac{\hat{P}^a_{yy}(k,i) - \hat{P}_{nn}(k,i)}{\hat{P}_{nn}(k,i)}.$$

As an example, the Wiener filter can be combined with the adaptive Bartlett estimate. This leads then to

$$G(k,i) = \frac{\hat{\xi}(k,i)}{1+\hat{\xi}(k,i)} = \frac{\hat{P}^a_{yy}(k,i) - \hat{P}_{nn}(k,i)}{\hat{P}^a_{yy}(k,i)}. \quad (10)$$

In order to obtain an upper bound of the achievable enhancement performance when $P_{yy}(k,i)$ is estimated using the Bartlett approach combined with the adaptive segmentation, we consider now an idealized situation, where optimal segments are found using knowledge of the clean signal. Clearly, in a practical situation, such an approach is not possible. The ideal segmentation is found by

$$\min_{s \epsilon S} E\left[D\left(x, \hat{x}(s)\right)\right] \quad (11)$$

where $s$ is a segmentation from the set $S$ of all allowed segmentations, $x$ is the clean speech signal, $\hat{x}$ is the estimated clean speech signal, $D$ is a distance measure between the clean speech signal $x$ and the estimated clean speech signal, and $E$ is the statistical expectation operator. The expectation operator is used to eliminate the influence of the noise realization on the distortion measure. We assume that distortions across frames are additive and independent. We can then write (11) as

$$\sum_i \min_{s_i \epsilon S_i} E\left[D\left(x(i), \hat{x}(i, s_i)\right)\right] \quad (12)$$

where $i$ is the frame index, $x(i)$ is frame $i$ of the clean speech signal, $\hat{x}(i)$ is the estimated clean speech for frame $i$, and $s_i$ is a certain segment from the set of all allowed segments for frame $i$. The purpose of (12) is to find for each frame a corresponding segment such that $D$ is minimized. The distortion measure we minimize here is the $l_2$ difference between the clean speech and the estimated clean speech frames; $D(i) = \|x(i) - \hat{x}(i)\|^2$, where $x(i) \in \mathbb{R}^K$ and $\hat{x}(i) \in \mathbb{R}^K$ and $K$ is the frame size. As described in Section II-C for the hypothesis based segmentation, the ideal segmentation can also be generalized by dividing the frequency range in subbands. In contrast to the hypothesis based segmentation, increasing the number of bands for the ideal case will result in a better segmentations always.

### *B. A Priori SNR Estimation Based on Improved Decision Directed Approach*

The decision directed approach [4] is another way of estimating $\xi$, which is well known and often used, because it results in less unnatural residual noise than the maximum likelihood based scheme.

Originally in [4], the DD approach was defined as a linear combination between two equally valid definitions of the *a priori* SNR

$$\xi(k,i) = \frac{E\left[|X(k,i)|^2\right]}{P_{nn}(k,i)}$$

and

$$\xi(k,i) = E\left[\gamma(k,i) - 1\right]$$

where $|X(k,i)|$ is the clean speech amplitude of frame $i$ and frequency bin $k$, and $\gamma(k,i) = |Y(k,i)|^2/P_{nn}(k,i)$ is the *a posteriori* SNR which is based on a periodogram estimate. With a smoothing factor $0 \le \alpha \le 1$, the linear combination results in

$$\xi(k,i) = E\left[\alpha \frac{|X(k,i)|^2}{P_{nn}(k,i)} + (1-\alpha)\left[\gamma(k,i) - 1\right]\right] \quad (13)$$

which is a mathematically correct expression without approximations, but which is hard to implement in practice. Therefore, in [4], at first the expectation operator was simply neglected. Leaving out this expectation operator results in an estimate $\hat{\xi}(k,i)$ with large variance, because the periodogram $|Y(k,i)|^2$ in $\gamma(k,i) = |Y(k,i)|^2/P_{nn}(k,i)$ has a variance as large as $var(|Y(k,i)|^2) \propto P^2_{yy}(k,i)$. To reduce the influence of this large variance, the smoothing factor $\alpha$ is chosen close to one. Second, since the estimate of the amplitude $|X(k,i)|$ of frame $i$ is not available, the estimate of the amplitude at frame $i-1$ was used. This results in a delay in the estimate of $\xi(k,i)$, which especially in transitional speech regions, is of large influence. Finally, a third approximation was necessary to overcome a side effect of the first approximation. Specifically, after neglecting the expectation operator $\gamma(k,i) - 1$, and thus $\xi(k,i)$ may become negative. Realizing that this is unreasonable for an SNR estimate, a max operator was introduced. Altogether, this led to [4]

$$\hat{\xi}(k,i) = \alpha \frac{\left|\hat{X}(k,i-1)\right|^2}{P_{nn}(k,i-1)} + (1-\alpha)\max\left[\gamma(k,i) - 1, 0\right] \quad (14)$$

which is the DD approach as it is used in practice.

It is possible to derive a DD approach that is closer to the definition of (13) and with less approximations than (14), using the improved Bartlett estimate $\hat{P}^a_{yy}(k,i)$. Replacing the periodogram $|Y(k,i)|^2$ in $\gamma(k,i)$ in (14) with $\hat{P}^a_{yy}(k,i)$, we can write

$$\hat{\xi}(k,i) = \alpha \frac{\left|\hat{X}(k,i-1)\right|^2}{P_{nn}(k,i-1)} + (1-\alpha)\max\left[\frac{\hat{P}^a_{yy}(k,i)}{P_{nn}(k,i)} - 1, 0\right] \quad (15)$$

where it is assumed that the power spectrum of the noise $P_{nn}$ is given. An advantage of (15) is that because the variance of the second term is decreased, it is possible to decrease $\alpha$, which means less influence of the first term (15) and as a result, less tracking delay and less speech distortions.

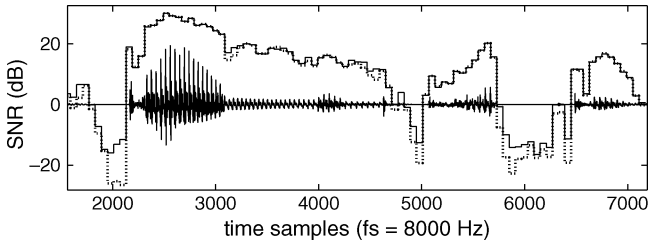



Fig. 8. SNR per frame after enhancement of noisy speech with 15-dB input SNR using the fixed segmentation (dotted) and the hypothesis based segmentation with four subbands (solid).

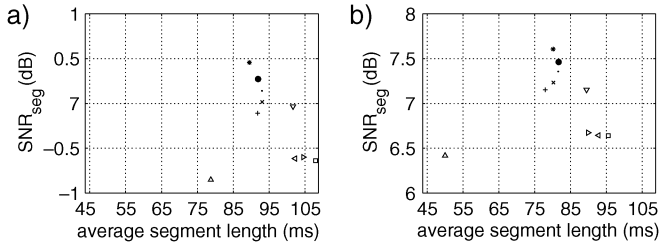



Fig. 9. Comparison between fixed segmentation ($\triangle$); hypothesis-based segmentation with four subbands ($\triangledown$), ideal segmentation with 1 ($+$), 2 (x), 4 ($\cdot$), 8 ($\bullet$), and 16 ($*$) subbands; and segmentation based on modified hypothesis for fixed $\alpha$ with 1 ($\triangleleft$), 2 ($\triangleright$), and 4 ($\square$) subbands. a) Input SNR of 5 dB. ($SNR_{\text{seg}} = -8.47$ dB) b) Input SNR of 15 dB ($\text{SNR}_{\text{seg}} = 1.53$ dB).

## IV. Objective and Subjective Simulation Experiments

We evaluate the presented segmentation algorithms by means of objective and subjective simulation experiments. In all experiments, we use speech fragments sampled at 8 kHz and frame sizes of 120 samples taken with 50% overlap. Noise statistics are measured during silence regions preceding speech activity and are assumed to be stationary. For objective evaluation, we use SNR per frame, defined as $\text{SNR}(i) = 10\log_{10}(\|x(i)\|^2/\|x_i - \hat{x}(i)\|^2)$ and segmental SNR defined as $\text{SNR}_{\text{seg}} = (1/N)\sum_{i=0}^{N-1}\text{SNR}(i)$ [7], where $x(i)$ and $\hat{x}(i)$ denote frame $i$ of the clean and the enhanced speech signal, respectively.

### *A. Objective and Subjective Results for Maximum Likelihood-Based Speech Enhancement*

In this section, we evaluate the objective and subjective quality improvement of the presented segmentation algorithms within a maximum likelihood-based speech enhancement scheme. We use the proposed segmentation algorithm as well as a fixed segmentation approach as front-ends for maximum likelihood Wiener filter-based enhancement algorithms with a gain function as in (10). In all experiments presented in this section, $\lambda$ was chosen offline and kept fixed at $\lambda = 10^{6.9}$ for all speech sentences, independent of input SNR. Both the threshold for the hypothesis-based segmentation and the segment length for the fixed segmentation were chosen such that they lead to optimal segmental SNR after enhancement averaged over a representative speech database.

*1) Objective Results:* In Fig. 8, the impact of our hypothesis-based adaptive segmentation algorithm is demonstrated on a female speech signal and compared with a fixed segmentation. We show the clean speech signal together with the SNR per frame after enhancement of a noisy speech signal for both the fixed segmentation and the hypothesis based algorithm with four subbands. Here, clean speech was degraded by white noise at an SNR of 15 dB. Especially at locations where the speech signal changes abruptly, the proposed scheme improves performance. The local improvements of 10 dB around the onsets and endings of speech sounds are due to less smearing of the speech sound.

As a second objective evaluation, we compared fixed segmentation, ideal adaptive segmentation as described in Section III-A, and the hypothesis-based segmentation algorithm in terms of segmental SNR. Fig. 9(a) and (b) shows the results averaged over six different speakers, three male and three female, for input SNRs of 5 and 15 dB, respectively. Both figures show the segmental SNR versus the average segment length for the fixed segmentation, the hypothesis-based segmentation with four equal-width subbands and the ideal segmentation with 1, 2, 4, 8, and 16 equal-width subbands. For the hypothesis-based segmentation, we used four subbands, because that led to an optimal segmental SNR. Comparing Fig. 9(a) and (b), it can be seen that for lower input SNR, all methods have longer segment lengths. The proposed scheme with four subbands leads to an increase of the segmental SNR of 0.82 and 0.73 dB for an input SNR of 5 and 15 dB, respectively. Further, it can be seen that the proposed algorithm with four subbands approximately reaches the segmental SNR of the full-band version of the ideal segmentation. From Fig. 9, it is clear that the segmental SNR for the ideal segmentation increases with the number of subbands that is used.

We also experimented with the modified hypothesis test of (6) to find segmentations under a fixed significance level $\alpha$ instead of under a fixed threshold $\lambda$. Those results are also shown in Fig. 9. It can be seen that the performance in terms of segmental SNR is much lower in comparison to the segmentation based on the hypothesis of (1). The reason for this is that the modified hypothesis of (6) does not fulfill the minimum requirements of stationarity, necessary to find a proper segmentation.

While in the experiments reported so far, we used the same value of the threshold $\lambda$ for each frame and for each frequency band, we might expect a performance gain if we allow different $\lambda$-values for different frames or frequency bands dependent on the SNR. However, by experiments with 30 different speech signals, it was observed that the optimal $\lambda$ is fairly insensitive to the SNR.

*2) Subjective Results:* For subjective evaluation, a listening test was performed with nine participants, the authors not included. To this listening test we will refer as OAB test. Here, O is the original signal and A and B are two enhanced signals. We implemented a Wiener filter based on the maximum likelihood approach combined with a fixed segmentation and a Wiener filter based on the maximum likelihood approach combined with the hypothesis-based segmentation algorithm with four subbands. Six speech signals were used, three male and three female speakers, all degraded by white noise at an SNR of 5 and 15 dB. We presented the listeners first the original signal followed by two versions enhanced with a fixed or hypothesis based segmentation. Each series was repeated four times with

TABLE I
WILCOXON TEST RESULTS TO DETERMINE THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE METHODS USED IN THE LISTENING EXPERIMENT

| noise source | input snr | P-value | significant |
|---|---|---|---|
| white noise | 5 dB | $5.8 * 10^{-7}$ | yes |
| | 15 dB | $7.3 * 10^{-6}$ | yes |

the enhanced versions played in random order. For all speech sentences and all SNRs, the maximum likelihood approach with the hypothesis based segmentation was preferred above the maximum likelihood approach with a fixed segmentation. The average relative preference for the maximum likelihood approach with the hypothesis-based segmentation over a fixed segmentation was 77.3% and 74.1% for 5- and 15-dB input SNR, respectively. A statistical significance test (Wilcoxon test [14]) is used determine whether the difference between the two methods was significant. In Table I, the P-values of this Wilcoxon test are given. The P-value is the significance level at which the $H_0$ hypothesis of the Wilcoxon test (this is the hypothesis in which the two tested methods have equal quality) would be rejected. Here, we compare the P-values with a significance level of $0.5 * 10^{-3}$ and can conclude that for both SNRs the difference is statistically significant.

### *B. Objective and Subjective Results for Decision Directed Approach Based Speech Enhancement*

In this section we evaluate the objective and subjective quality improvement of DD approach based speech enhancement combined with the proposed segmentation algorithm over the standard DD approach. In all experiments, both methods are combined with a Wiener filter. For the adaptive segmentation based DD approach we used a smoothing factor $\alpha = 0.94$. This choice is based on experiments presented in Fig. 11. For the standard DD approach we used $\alpha = 0.97$ as proposed in [4]. In all experiments presented in this section $\lambda$ was chosen off line and kept fixed at $\lambda = 10^{7.5}$ for all speech sentences.

*1) Objective Results:* Fig. 10 shows a performance comparison in terms of SNR per frame between the standard DD approach and DD approach improved with an adaptive segmentation. The improved approach is implemented using (15). The noisy speech signal was degraded by white noise with a SNR of 15 dB. It can be seen that the improved DD approach generally leads to better performance: in onset regions, during sustained speech sounds, and during silence intervals.

As a second objective evaluation, we compared the standard DD approach with the improved approach in terms of segmental SNR for different values of the smoothing factor $\alpha$ [see (14) and (15)]. The results are averaged over six different speakers, three male and three female with three different SNR levels, 5, 10, and 15 dB. The results are shown in Fig. 11. It is shown that combining the DD approach with an adaptive segmentation leads to an improved segmental SNR in the order of 0.7 dB. Further, it can be seen that the improved DD approach has its optimum at a lower $\alpha$ than the standard DD approach, which means less tracking delay in the estimation of $\xi$.

In Fig. 12, we demonstrate the influence of the allowed latency in the segmentation algorithm on the performance after enhancement. The signals used in this experiment were degraded by white noise at an input SNR of 10 dB. It is shown

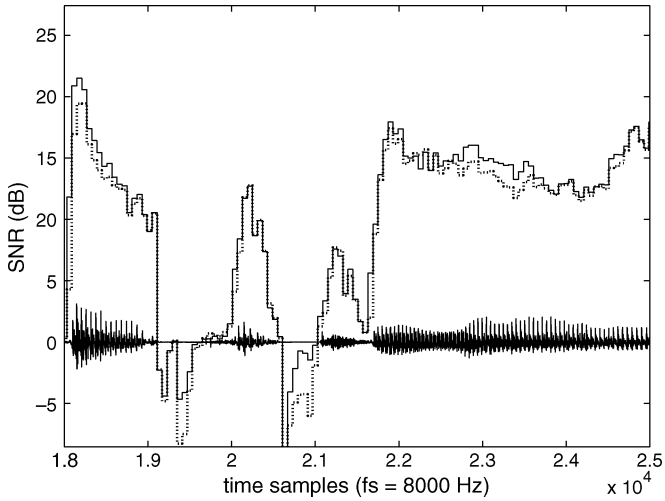



Fig. 10. SNR per frame after enhancement of noisy speech with 15-dB input SNR after using standard DD approach (dotted line) and the with an adaptive segmentation improved DD approach (solid line).

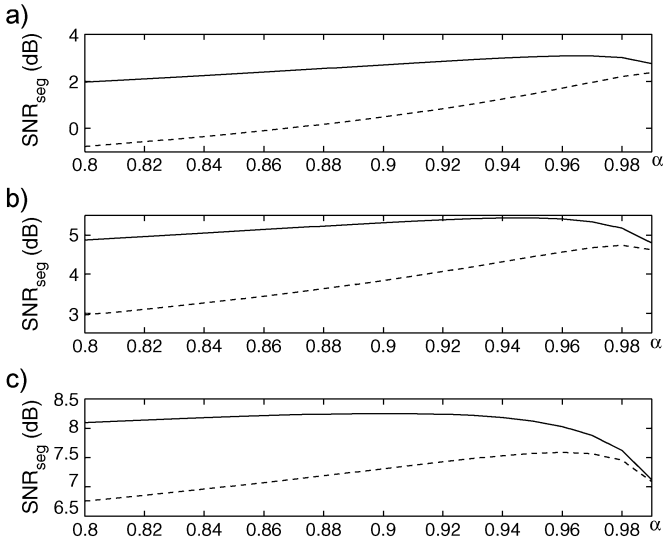



Fig. 11. Comparison between the DD approach (dashed line) and the DD approach combined with an adaptive segmentation (solid line) in terms of segmental SNR as a function of the smoothing factor $\alpha$. a) Input SNR of 5 dB ($\mathrm{SNR_{seg}} = -8.23$ dB). b) Input SNR of 10 dB ($\mathrm{SNR_{seg}} = -3.23$ dB). c) Input SNR of 15 dB ($\mathrm{SNR_{seg}} = 1.77$ dB).

that the segmentation algorithm with a latency of 30 ms has almost a similar performance as with an infinite latency. Even with a latency of 0 ms, the segmentation algorithm still leads to improvement compared to the DD approach without any adaptive segmentation.

*2) Residual Noise Analysis:* An important aspect of the quality of a speech enhancement algorithm is the nature of the residual noise, because many enhancement methods suffer from a disturbing and unnatural sounding character of the residual noise. Often, the power spectrum of the residual noise consists of frequency components where the noise energy occurs on and off at almost random frequencies and is, therefore, called musical noise. Fig. 13 shows a comparison of the energy of the residual noise between the standard DD approach and the improved DD approach using an excerpt of a female speech signal. The speech signal was degraded by white noise with an SNR of 5 dB. In Fig. 13(a), we show the noisy speech signal. We

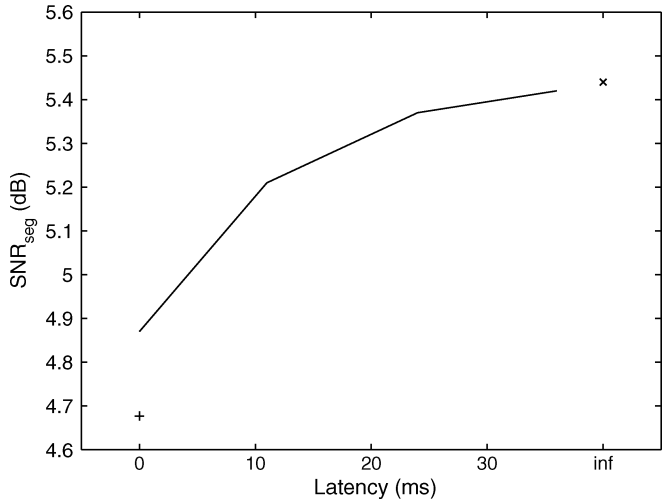



Fig. 12. Performance in terms of segmental SNR versus latency for the DD approach (+), DD approach combined with an adaptive segmentation (solid line), and DD approach combined with an adaptive segmentation and infinite latency (×).

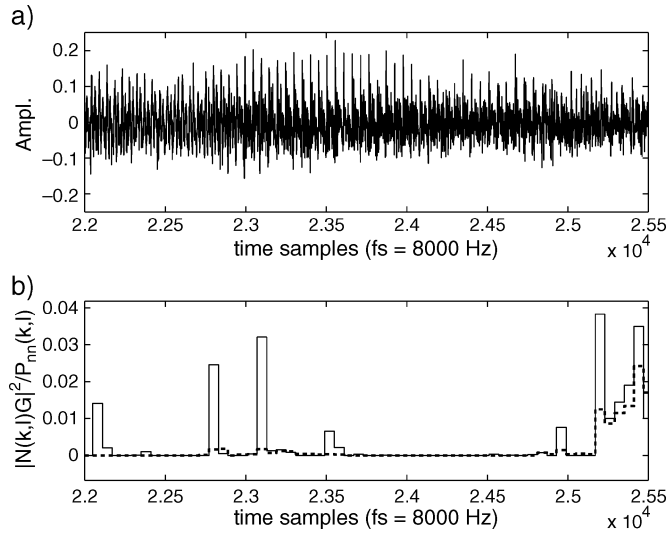



Fig. 13 a) Noisy speech signal at 5-dB SNR. b) Comparison between standard DD approach (solid line) and with the adaptive segmentation improved DD approach (dashed line) of the energy of the residual noise $|r(k,i)|^2$.

compare in Fig. 13(b) the energy of the residual noise $|r(k,i)|^2$ for a certain frequency bin $k$ over several consecutive frames for both the standard (solid line) and the improved (dashed line) DD approach. The energy of the residual noise was computed by first applying a decomposition of the total residual signal in a residual noise and a speech distortion component [15], that is

$$\begin{aligned} X(k,i) - \hat{X}(k,i) &= X(k,i) - Y(k,i)G(k,i) \\ &= X(k,i) - \{X(k,i) + N(k,i)\}\, G(k,i) \\ &= X(k,i)\{1 - G(k,i)\} - N(k,i)G(k,i) \\ &= d(k,i) + r(k,i) \end{aligned}$$

where $X(k,i)$, $Y(k,i)$, and $N(k,i)$ are Fourier coefficients, $G(k,i)$ is the value of the gain function, $d(k,i)$ speech distortion, and $r(k,i)$ the residual noise. From Fig. 13(b), it is clear that the energy of the residual noise has a smoother character when using the improved DD approach. With the standard DD approach, the energy of the residual noise shows jumps and irregularities. Informal listening tests confirmed that the DD approach with the adaptive segmentation results in less residual noise. Moreover, informal listening tests confirmed that the character is less musical, because of the decreased variance of the estimate of $E[\gamma - 1]$. Consequently, the DD approach improved with an adaptive segmentation has less disturbing sounding residual noise then without adaptive segmentation.

TABLE II
WILCOXON TEST RESULTS TO DETERMINE THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN THE METHODS USED IN THE LISTENING EXPERIMENT

| noise source | input snr | P-value | significant |
|---|---|---|---|
| white noise | 5 dB | $1.8 * 10^{-5}$ | yes |
| | 15 dB | $3.53 * 10^{-3}$ | yes |
| car noise | 5 dB | 0.33 | no |
| | 15 dB | 0.55 | no |
| F16 noise | 5 dB | $1.7 * 10^{-4}$ | yes |
| | 15 dB | $1.5 * 10^{-4}$ | yes |

*3) Subjective Results:* For subjective evaluation, an OAB listening test similar to the test in Section IV-A2 was performed with nine participants, the authors not included. We implemented a Wiener filter where the *a priori* SNR was determined with the standard DD approach from (14) and a Wiener filter where the *a priori* SNR was determined with the improved DD approach from (15). In this listening test, we used three different types of additive noise at two different SNRs, white noise, car noise, and F16-cockpit noise at SNRs of 5 and 15 dB. For each noise type and noise power level, we presented the listeners two female sentences and two male sentences. The listeners were presented first the original signal followed by the two different enhanced signals. Each series was repeated three times with the enhanced versions played in random order. For speech signals corrupted with white noise at an SNR at 5 and 15 dB, the relative preference of the improved DD approach over the standard DD approach was 80.6% and 70.4%, respectively. For speech signals corrupted with F16-cockpit noise at an SNR of 5 and 15 dB, the improved DD approach was preferred above the standard DD approach with 77.8% and 75%, respectively. A statistical Wilcoxon significance test revealed that the difference between the two methods is indeed significant at a significance level of $0.5 * 10^{-3}$. The P-values of this test are tabulated in Table II. For speech signals corrupted with car noise, the outcome of the listening test was close to 50%. In this case, the statistical significance test (Wilcoxon test) was applied and revealed that the difference between the two methods indeed is insignificant, although objective tests done by the authors showed improvement in terms of SNR. This result can be explained by the fact that the energy of car noise is concentrated mainly in a small frequency band where in general the majority of speech energy is present. This means that most of the residual noise that is left after using the standard DD approach will be masked by the speech energy. As a result, the perceptual difference between the standard DD approach and the improved DD approach becomes smaller.

## V. CONCLUSION

We presented an adaptive time segmentation for speech enhancement to improve the estimation of the noisy speech power spectrum. The segmentation algorithm only needs

knowledge of the noisy speech signal to determine for each frame which segment of data should be used to estimate the noisy speech power spectrum. The segments are formed based on the outcome of a sequence of hypothesis tests. We used this adaptive estimate of the noisy speech power spectrum to improve the maximum likelihood and decision directed-based speech enhancement methods. Moreover, by combining the decision directed approach with an adaptive segmentation based Bartlett estimate instead of the periodogram estimate, we showed that less approximations are needed. Objective experiments showed that usage of the adaptive time segmentation to improve the maximum likelihood and decision directed-based speech enhancement methods leads to a better quality in terms of SNR. Transitional regions gain in terms of SNR because of a better adaptivity of the gain function to the speech signal. Also, in stationary regions the SNR is improved because of a more efficient use of the data. Simulation experiments showed that the improved decision directed approach results in less residual noise, but also with a less musical character. Furthermore, subjective listening tests with speech signals degraded with various noise sources and noise levels showed that in terms of perceptual quality for both the maximum likelihood and the decision directed approach the adaptive segmentation algorithm is preferred over the usage of a fixed segmentation. For car noise, the perceptual difference was negligible, because most of the residual noise energy is then masked by the speech energy.

## REFERENCES


[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[6] J. G. Proakis, C. M. Rader, F. Ling, C. L. Nikias, M. Moonen, and I. K. Proudler, *Algorithms for Statistical Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2002.

[7] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Piscataway, NJ: IEEE Press, 2000.

[8] T. F. Quatieri and R. B. Dunn, "Speech enhancement based on auditory spectral change," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, May 2002, vol. 1, pp. 257–260.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[10] H. L. van Trees, *Detection, Estimation and Modulation Theory*. New York: Wiley, 1968, vol. 1.

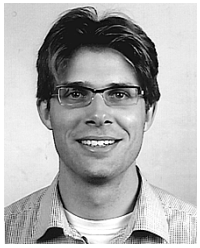
[11] R.-M. Gray, Toeplitz and Circulant Matrices: A Review Stanford Univ., Stanford, CA, 2002, Tech. Rep..

[12] N. Mukhopadhyay, *Probability and Statistical Inference*. New York: Marcel Dekker, 2000.

[13] S. K. Kay, *Fundamentals of Statistical sSgnal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1998, vol. 2.

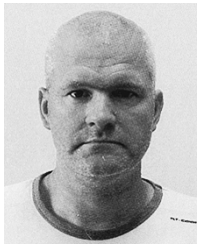
[14] D. J. Sheskin, *Parametric and Nonparametric Statistical Procedures*, 3rd ed. London, U.K.: Chapman & Hall/CRC, 2004.

[15] Y. Ephraim and H. L. van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

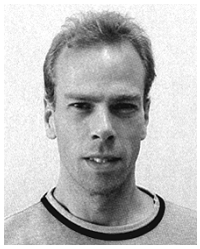
**Richard C. Hendriks** received the B.Sc. and M.Sc. degrees in electrical engineering from Delft University of Technology, Delft, The Netherlands, in 2001 and 2003, respectively. In September 2003, he joined the Department of Mediamatics, Delft University of Technology, where he has been working toward his Ph.D. degree.

His main interests are digital speech and audio processing, including acoustical noise reduction and speech enhancement.



**Richard Heusdens** received the M.Sc. and Ph.D. degrees from Delft University of Technology, Delft, The Netherlands, in 1992 and 1997, respectively.

Since 2002, he has been an Associate Professor in the Department of Mediamatics, Delft University of Technology. In the spring of 1992, he joined the Digital Signal Processing Group at the Philips Research Laboratories, Eindhoven, The Netherlands. He has worked on various topics in the field of signal processing, such as image/video compression and VLSI architectures for image processing algorithms. In 1997, he joined the Circuits and Systems Group, Delft University of Technology, where he was a Postdoctoral Researcher. In 2000, he moved to the Information and Communication Theory (ICT) Group, where he became an Assistant Professor responsible for the audio and speech processing activities within the ICT group. He is involved in research projects that cover subjects such as audio and speech coding, speech enhancement, and digital watermarking of audio.



**Jesper Jensen** received the M.Sc and Ph.D degrees in electrical engineering from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively.

From 1996 to 2001, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Researcher, Ph.D. student, and Assistant Research Professor. In 1999, he was a Visiting Researcher at the Center for Spoken Language Research, University of Colorado, Boulder. Currently, he is an Assistant Professor at Delft University of Technology, Delft, The Netherlands. His main research interests are digital speech and audio signal processing, including coding, synthesis, and enhancement.