An MMSE Estimator for Speech Enhancement Under a Combined Stochastic–Deterministic Speech Model

Richard C. Hendriks, Richard Heusdens, and Jesper Jensen

***Abstract*—Although many discrete Fourier transform (DFT) domain-based speech enhancement methods rely on stochastic models to derive clean speech estimators, like the Gaussian and Laplace distribution, certain speech sounds clearly show a more deterministic character. In this paper, we study the use of a deterministic model in combination with the well-known stochastic models for speech enhancement. We derive a minimum mean-square error (MMSE) estimator under a combined stochastic–deterministic speech model with speech presence uncertainty and show that for different distributions of the DFT coefficients the combined stochastic–deterministic speech model leads to improved performance of approximately 0.8 dB segmental signal-to-noise ratio (SNR) over the use of a stochastic model alone. Evaluation with perceptual evaluation of speech quality (PESQ) shows performance improvements of approximately 0.15 on an MOS scale.**

***Index Terms*—Deterministic speech model, minimum mean-square error (MMSE), speech enhancement.**

I. Introduction

MANY digital speech communication applications, e.g., speech coding and speech recognition, are aimed at processing of noise-free speech signals. However, in practice, speech signals are often degraded by acoustical noise. To overcome degradation, noise can be removed before further processing using (single-channel) speech enhancement methods. An important class of such algorithms are the discrete Fourier transform (DFT) domain-based methods that work on a frame-by-frame basis, where relatively good quality can be obtained with relatively low complexity. Here, criteria like minimum mean square error (MMSE) [1] or maximum *a posteriori* (MAP) [2] are used to estimate the clean speech DFT coefficients. The main focus in DFT domain speech enhancement has been on the derivation of estimators relying completely on a stochastic model for the clean speech DFT coefficients. Often, speech DFT coefficients have been assumed Gaussian distributed [1], [3], although more recently estimators have been derived which assume Laplacian or Gamma distributed speech DFT coefficients [4].

Manuscript received October 6, 2005; revised March 31, 2006. This work was supported by the Philips Research and the Technology Foundation STW, the applied science division of NWO, and the technology programme of the ministry of Economics Affairs. The material in this paper was presented in part at the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hong-Goo Kang.

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

Digital Object Identifier 10.1109/TASL.2006.881666

Although most DFT domain enhancement algorithms rely on stochastic models, it can be observed that certain speech sounds have a more deterministic character. For example, it is well known that voiced speech segments may be represented well by a linear combination of sinusoidal functions with constant frequency and exponentially decaying amplitude or, as a special case, a constant amplitude [5, Ch. 4]. With this signal representation, the sequence of DFT coefficients seen across one particular frequency bin constitutes a completely deterministic time-series. In [6], a maximum likelihood-based spectral amplitude estimator was derived under a deterministic speech model. Here, the clean speech DFT coefficients are characterized by deterministic, but unknown amplitude and phase values, while the noise DFT coefficients are assumed to follow a zero-mean Gaussian probability density function (pdf). The use of this estimator leads to less suppression as compared to the case where speech DFT coefficients are assumed stochastic. Obviously, a deterministic speech model is not always appropriate. For example, for noise-like speech sounds, such as /s/, /f/, etc., the DFT coefficients should rather be represented by a stochastic model.

Assuming that speech cannot be modeled as either strictly stochastic or deterministic, we present in this paper an MMSE clean speech estimator where the speech DFT coefficients are modeled as a mixture of a deterministic and a stochastic speech model. Further, we combine this MMSE estimator with a speech presence uncertainty model, similar to [1], [6].

The remainder of this article is organized as follows. In Section II, we consider the individual deterministic and stochastic speech models and present their corresponding MMSE estimators. In Section III, we specify the deterministic model and and explain how to estimate its parameters. In Section IV, we derive the MMSE estimator under the combined stochastic–deterministic (SD) speech model. In Section V, we present experimental results, and finally in Section VI, we draw some conclusions.

II. Stochastic and Deterministic Speech Model

In this section, we introduce the stochastic and the deterministic speech model. We assume the noise process to be additive, i.e.,

$$Y(k,i) = X(k,i) + N(k,i)$$

with $Y(k,i)$, $X(k,i)$, and $N(k,i)$ the noisy speech, clean speech, and noise DFT coefficient, respectively, at frequency bin k and time frame i. Further, we assume that $X(k,i)$ and $N(k,i)$ are uncorrelated (for the stochastic model) and that the

1558-7916/$25.00 © 2006 IEEE

noise DFT coefficients have a zero-mean complex Gaussian distribution.

By deriving an MMSE estimator under an SD speech model, we exploit the idea that certain speech DFT coefficients can be better modeled with a deterministic model while others can be better modeled by a stochastic model. In the following derivations, we use the complex zero-mean Gaussian distribution as stochastic representation for the clean speech DFT coefficients. However, we note that this work is general and can also be extended to other distributions like the Laplace and Gamma distribution as shown in the experimental results in Section V.

A. Probability Density Function of Noisy DFT Coefficients

In this section, we consider the probability density functions of the noisy DFT coefficients under both the stochastic and the deterministic model, respectively.

- Stochastic Model:
  Under the stochastic speech model and using the assumption that clean speech DFT coefficients have a complex zero-mean Gaussian distribution, the noisy speech DFT coefficients have the following zero-mean complex Gaussian distribution:

$$p_{Y|S}\left(y(k,i)|s\right) = \frac{1}{\pi\sigma_Y^2(k,i)} \exp\left\{-\frac{|y(k,i)|^2}{\sigma_Y^2(k,i)}\right\} \quad (1)$$

  where S indicates that speech is produced with a stochastic model, and where $\sigma_Y^2(k,i)$ is the variance of the noisy DFT coefficient $Y(k,i)$ which equals the sum of the noise variance and the clean speech variance, that is $\sigma_Y^2(k,i) = \sigma_X^2(k,i) + \sigma_N^2(k,i)$.
- Deterministic Model:
  Under the deterministic speech model, we assume that $Y(k,i)$ can be written as the sum of a deterministic variable (due to $X(k,i)$) and a stochastic variable (due to $N(k,i)$). Using the assumed (zero-mean) Gaussian distribution of the noise DFT coefficients, this leads to a nonzero mean Gaussian distribution for the noisy DFT coefficients

$$p_{Y|D}\left(y(k,i)|d\right) = \frac{1}{\pi\sigma_N^2(k,i)} \times \exp\left\{-\frac{|y(k,i) - E\left[Y(k,i)\right]|^2}{\sigma_N^2(k,i)}\right\} \quad (2)$$

  with $E[Y(k,i)] = x(k,i)$, and where D indicates that speech is produced with a deterministic model. Apart from having a nonzero mean, we note that the variance of $Y(k,i)$ under the deterministic model may be significantly smaller than that of $Y(k,i)$ under a stochastic model.

B. MMSE Estimators

In order to derive an MMSE estimator for the clean speech DFT coefficients under an SD speech model, we first consider the individual MMSE estimators for stochastic and deterministic representations.

- Stochastic Model:
  Under the stochastic Gaussian speech model it is well known that the Wiener filter is the MMSE estimator, that is

$$\hat{x}(k,i) = E\left[X(k,i)|y(k,i)\right] = \frac{\xi(k,i)}{1+\xi(k,i)} y(k,i) \quad (3)$$

  with $\xi(k,i) = E[X(k,i)^2]/E[N(k,i)^2] = \sigma_X^2(k,i)/\sigma_N^2(k,i)$ the *a priori* SNR.
- Deterministic Model:
  Under the deterministic speech model, the clean speech DFT coefficients are assumed to be deterministic, but unknown. This means that $p_X(x(k,i)) = \delta(x(k,i) - x'(k,i))$ with $x'(k,i)$ the value of the deterministic clean speech DFT coefficient itself and where $\delta(\cdot)$ is a delta function. The MMSE estimator then is

$$\hat{x}(k,i) = E\left[X(k,i)|y(k,i)\right] = x'(k,i) \quad (4)$$

  where we observe that $x'(k,i) = E[Y(k,i)]$.
  Notice that both estimators in (3) and (4) are MMSE and expressed in terms of expected values. Since in practice these expected values are unknown, estimation is necessary. For estimation of $\xi(k,i)$, the decision-directed approach or maximum likelihood approach is often used [1]. Estimation of $E[Y(k,i)]$ will be considered in the next section.

III. Estimation of Deterministic Speech Model Parameters

So far, we considered the use of a deterministic speech model; however, we did not specify the exact model itself. If the clean speech signal can be represented by a sum of P (exponentially damped) sinusoids with constant frequency, that is

$$x_t(m) = \sum_{p=1}^{P} a_p e^{j\phi_p} e^{(-d_p + j\nu_p)m}$$

where $x_t(\cdot)$ indicates a time domain sample, m is the time sample index, a_p the amplitude, ϕ_p the phase, d_p the exponential decay factor, and ν_p the frequency of component p, then the DFT coefficients at each frequency bin k can be described by a sum of P complex exponentials seen across time. However, under the assumption of sufficiently long frame sizes, there will be no more than one dominant exponential, say component p, per frequency bin. Let us, therefore, assume that our deterministic model for a clean speech DFT coefficient $x(k,n)$ is a single complex exponential, that is

$$x(k,n) = \sum_{m=0}^{K-1} a_p e^{j\phi_p} e^{(-d_p + j\nu_p)(m-nM)} w(m) e^{-j\omega_k m} \quad (5)$$

$$= e^{-(-d_p + j\nu_p)nM} x(k,0) \quad (6)$$

with $w(m)$, $m = 0, \ldots, K-1$ the analysis window (of length K) used to define the signal frame, M ($M \leq K$) the frame shift and $\omega_k = (2\pi/L)k$, where L is the DFT size ($L \geq K$), and n the index of a certain frame. We can write (6) in the form

$x(k,n) = z^n x(k,0)$, with $z = e^{(d_p - j\nu_p)M}$. If the noise is wide sense stationary for $n = i - n_1, \ldots, i + n_2$ and if M is sufficiently large with respect to the correlation time of the noise, then the observed noise sequence $N(k,n)$ for $n = i - n_1 \ldots i + n_2$ is white. Estimation of d_p and ν_p is then known as a standard harmonic retrieval problem [7], and estimation of d_p and ν_p can be done from the noisy DFT coefficients using the ESPRIT algorithm [8].

When $n = i - n_1, \ldots, i + n_2$ is the time span across which we assume the deterministic model to be valid, then in practice we can approximate (4) using the relation in (6) as

$$\hat{x}(k,i) = E[Y(k,i)] \tag{7}$$

$$\approx \frac{1}{n_2 + n_1 + 1} \sum_{n=i-n_1}^{i+n_2} y(k,n) e^{(-d_p + j\nu_p)(i-n)M} \tag{8}$$

where each term is corrected for the decay in amplitude and phase shift. The values for n_1 and n_2 should be chosen such that the deterministic model is valid. This could be done by using fixed values such that the deterministic model is valid over the interval $n = i - n_1, \ldots, i - n_2$ or adaptively, e.g., by using an adaptive segmentation as in [9]. Note that for $d_p = 0$ we have a special case of the aforementioned presented model

$$\hat{x}(k,i) = E[Y(k,i)] \tag{9}$$

$$\approx \frac{1}{n_1 + n_2 + 1} \sum_{n=i-n_1}^{i+n_2} y(k,n) e^{j\nu_p(i-n)M}. \tag{10}$$

We see that the estimators in (8) and (10) modify magnitude as well as phase of the noisy DFT coefficient $y(k,i)$. Further, notice that when $n_1 = n_2 = 0$, the estimate of $x(k,i)$ is $\hat{x}(k,i) = y(k,i)$.

A. Simulation Examples

To illustrate the idea of using a deterministic speech model, we conducted two simulation experiments. As a first experiment, we generate a synthetic clean speech signal consisting of five (deterministic) sinusoidal components and a (stochastic) autoregressive process. Then we generate a noisy signal by adding white Gaussian noise at an SNR of 10 dB to the clean synthetic signal. We now compute DFT coefficients seen across time and plot in Fig. 1 in the complex plane the values of $y(k,n)$ with $n = i - n_1 \ldots i + n_2$ originating from a frequency bin containing only the stochastic components (cloud centered around the origin) and values of $y(k,n)$ originating from a frequency bin containing only one of the deterministic components (cloud with an offset from the origin). The variance of the latter is smaller than the variance of the first cloud and is only due to the noise variance in (2). Notice, that for the cloud containing noisy deterministic components, it is sufficient to compute the mean of the cloud to estimate the clean deterministic signal component.

In Fig. 2, we present a second simulation example where the potential of distinguishing between a stochastic and a deterministic model on a natural speech signal is demonstrated. In Fig. 2(a) and (b), an original clean speech time domain signal and its spectrogram are shown, respectively. The signal was degraded by white noise at an SNR of 10 dB and enhanced using

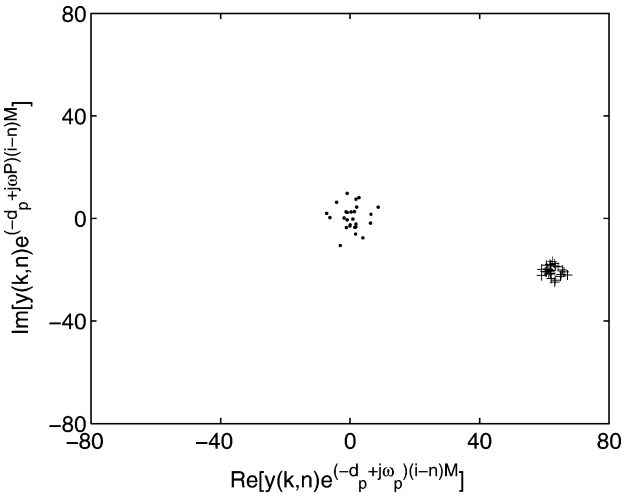


Fig. 1. $y(k,n)e^{(-d_p + j\omega_p)(i-n)M}$ at a frequency k containing a deterministic signal component $(+)$ and at a frequency k containing a stochastic signal component $(\cdot)$, respectively.

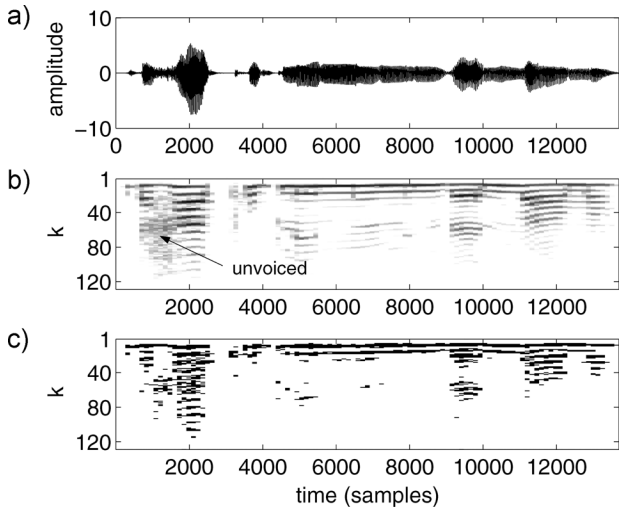


Fig. 2. (a) Clean speech signal. (b) Clean speech spectrogram. (c) Black: deterministic model is optimal in terms of local SNR. White: stochastic model is optimal in terms of local SNR.

two different enhancement systems, one using the stochastic model and one using the deterministic model. We compute for each time-frequency point for each method the resulting SNR and evaluate which of the two models lead to the highest SNR. This is shown in Fig. 2(c); a preference for the deterministic model is expressed as a black dot and a preference for the stochastic model as a white dot. As expected, the deterministic model performs better at the spectral lines that are visible in the spectrogram (voiced regions), while in the unvoiced speech regions, the stochastic model is preferred. For this experiment, we averaged the result over 100 different noise realizations and used the maximum likelihood (ML) approach [1] to estimate the *a priori* SNR ξ, where the number of frames that is used in the ML approach is set at $L = 1$. We use the ML approach instead of the often-used decision-directed [1] (DD) approach to overcome a dependency on past frames, as will be the case with the DD approach. Such dependencies can lead to wrong, biased, estimates of the suppression gain under the stochastic speech

model when speech sound changes take place and as a result can lead to too much suppression or even complete removal of low energy speech components that only last for a short time.

IV. MMSE Estimation Under Combined Stochastic-Deterministic Speech Model

To find an MMSE estimator of the clean speech DFT coefficients under a combined SD speech model, we present in this section two different setups: a completely general model where a soft decision is made based on estimated probabilities between the stochastic and deterministic model and where speech presence uncertainty is taken into account, abbreviated with SOFT–SD–U. Second, a special case of the first model where instead of a soft decision, a hard decision between the stochastic and deterministic model is made without speech presence uncertainty, abbreviated with HARD–SD. To do so, we introduce the set $\alpha = \{A, D, S\}$. Here $\alpha = A$, $\alpha = D$, and $\alpha = S$ indicate speech absence, that speech was generated with a deterministic model, and that speech was generated with a stochastic model, respectively. Although all derivations in this section are per frequency bin k and frame index i, we leave out these indices for notational convenience. This means that $P_{D|Y}(d|y(k,i))$ is written as $P_{D|Y}(d|y)$.

A. SOFT–SD–U Estimator

To find the MMSE optimal estimator SOFT–SD–U, we compute the conditional expectation $E[X|y]$. That is

$$\begin{aligned}\hat{x} &= E[X|y] \\ &= \int_x x p_{X|Y}(x|y)dx \\ &= \int_x x \sum_\alpha p_{X|Y,\alpha}(x|y,\alpha)p_{\alpha|Y}(\alpha|y)dx \\ &= \int_x x \left\{p_{X|Y,D}(x|y,d)P_{D|Y}(d|y)\right. \\ &\qquad \left. + p_{X|Y,S}(x|y,s)P_{s|y}(S|Y)\right\} dx \qquad (11) \\ &= E[X|y,d]P_{D|Y}(d|y) + E[X|Y,S]P_{S|Y}(s|y) \qquad (12)\end{aligned}$$

where in (11) we used the fact that $x = 0$ when $\alpha = a$. The conditional probabilities $P_{D|Y}(d|y)$ and $P_{S|Y}(s|y)$ can be computed using Bayes rule as

$$\begin{aligned}&P_{D|Y}(d|y) \\ &= \frac{p_{Y|D}(y|d)P_D(d)}{p_{Y|D}(y|d)P_D(d) + p_{Y|S}(y|s)P_S(s) + p_{Y|A}(y|a)P_A(a)} \qquad (13) \\ &= \frac{\Lambda_D}{\Lambda_D + \Lambda_S + 1} \qquad (14) \\ &P(S|Y) \\ &= \frac{p_{Y|S}(y|s)P_S(s)}{p_{Y|D}(y|d)P_D(d) + p_{Y|S}(y|s)P_S(s) + p_{Y|A}(y|a)P_A(a)} \qquad (15) \\ &= \frac{\Lambda_S}{\Lambda_D + \Lambda_S + 1} \qquad (16)\end{aligned}$$

with

$$\Lambda_D = \frac{p_{Y|D}(y|d)P_D(d)}{p_{Y|A}(y|a)P_A(a)}$$

and

$$\Lambda_S = \frac{p_{Y|S}(y|s)P_S(s)}{p_{Y|A}(y|a)P_A(a)}$$

respectively. Here $P_A(a)$, $P_D(d)$, and $P_S(s)$ denote the prior probabilities, that in a frequency bin speech is absent, that a frequency bin contains speech and is deterministic and that a frequency bin contains speech and is stochastic, respectively. The values chosen for those *a priori* probabilities will be discussed in Section V. Further, $p_{Y|A}(y|a)$ is given by

$$p_{Y|A}(y|a) = \frac{1}{\pi\sigma_N^2}\exp\left\{-\frac{|y|^2}{\sigma_N^2}\right\}$$

and $p_{Y|S}(y|s)$ and $p_{Y|D}(y|d)$ are given by (1) and (2), respectively. Computation of (2) can be done by substitution of (10) in (2). Notice that Λ_S can efficiently be written in terms of the *a priori* and *a posteriori* SNR $\xi(k,i)$ and $\gamma(k,i)$, respectively, as presented in [1]. For an outline of the SOFT–SD–U algorithm see the Appendix.

B. HARD-SD Estimator

With the HARD-SD estimator we assume that speech is always present, that is $p_A(a) = 0$. The estimator HARD-SD follows from (12) by setting $P_{D|Y}(d|y)$ either equal to 1 (deterministic speech model), or to 0 (stochastic speech model). This means that

$$\hat{X} = \begin{cases} E[X|y,d], & \text{if deterministic speech} \\ E[X|y,s], & \text{if stochastic speech.} \end{cases} \qquad (17)$$

The decision between the deterministic and stochastic speech model is made by the following hypothesis test:

$$\begin{aligned}H_0: &\quad E[Y(k,i)] = 0 \\ H_1: &\quad E[Y(k,i)] = x(k,i) \text{ and } VAR[Y(k,i)] = \sigma_n^2(k,i).\end{aligned}$$

Under the H_0 hypothesis, the stochastic model is chosen ($P_{D|Y}(d|y) = 0$), and under the H_1 hypothesis, the deterministic model is chosen ($P_{D|Y}(d|y) = 1$). We decide between H_0 and H_1 using the Bayes criterion [10], that is

$$T = \frac{p_{Y|D}(y|d)}{p_{Y|S}(y|s)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \qquad (18)$$

where the threshold $\lambda = (P_S(s)/P_D(d))$, and $P_S(s)$ and $P_D(d)$ are the likelihood of the stochastic and deterministic model to occur, respectively. For an outline of the HARD-SD algorithm see the Appendix.

In Fig. 3, the hypothesis test to distinguish between a stochastic (Gaussian) and deterministic speech model in (18) is demonstrated using the same speech signal as used in Fig. 2, degraded by white noise at an SNR of 10 dB. The top figure

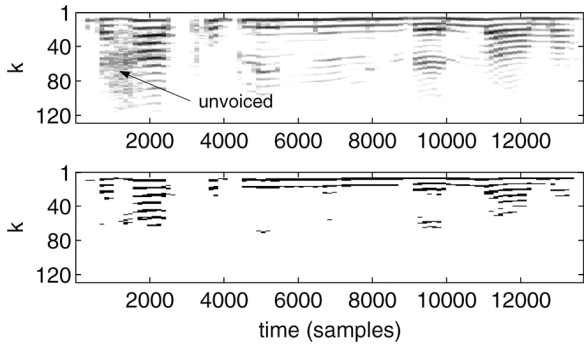


Fig. 3. Top: clean speech signal spectrogram. Bottom: Outcome of hypothesis test, Black: speech component is classified as deterministic, White: speech component is classified as stochastic.

shows the clean speech spectrogram. The bottom figure shows in the time-frequency plane the outcome of the hard decision of (18), where a black dot means that the speech component is classified as deterministic and a white dot that it is classified as stochastic. The hypothesis test appears to perform as expected: DFT coefficients representing harmonics are classified as deterministic, while, e.g., the DFT coefficients in the indicated region are classified as stochastic.

V. Experimental Results and Discussion

In this section, we compare the proposed SOFT–SD–U and HARD-SD enhancement methods with a traditional enhancement method which relies on a stochastic speech model alone with and without speech presence uncertainty, respectively. For evaluation, we use an extended version of the perceptual evaluation of speech quality (PESQ) measure [11] and segmental SNR defined as [12]

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{\mathcal{L}} \sum_{i=0}^{\mathcal{L}-1} \mathcal{T} \left\{ 10 \log_{10} \frac{\|\mathbf{x}_t(i)\|^2}{\|\mathbf{x}_t(i) - \hat{\mathbf{x}}_t(i)\|^2} \right\}$$

where $\mathbf{x}_t(i)$ and $\hat{\mathbf{x}}_t(i)$ denote frame i of the clean speech signal and the enhanced speech signal, respectively, $\mathcal{L}$ is the number of frames within the speech signal in question, and $\mathcal{T}(x) = \min\{\max(x, -10), 35\}$, which confines the SNR at each frame to a perceptually meaningful range between -10 and 35 dB. All objective results presented below are averaged over 24 different speech signals originating from the TIMIT database.

In all experiments, we use speech fragments sampled at 8 kHz and frame sizes of 256 samples taken with 50% overlap. To have good time resolution in the estimation of (4), the DFT samples $y(k,n)$, $n = i - n_1 \ldots i + n_2$, are computed from frames with an overlap of 84%. This overlap was chosen based on a tradeoff, where on one hand, a small overlap is desirable to better satisfy the assumption made in Section III, i.e., frame shift M is sufficiently large with respect to the correlation time of the noise. On the other hand, a large overlap is necessary when using multiple samples in (8), i.e., n_1, $n_2 > 0$, because approximation of (4) by (8) is only valid over relatively short time intervals. In all experiments, noise statistics are measured during silence regions preceding speech activity.

Initial experiments have shown that in terms of $\mathrm{SNR}_{\mathrm{seg}}$, the difference between the use of (8) and (10) for estimating $x(k,i)$ is negligible. Therefore, we use in all our experiments (10). Furthermore, $n_1 = n_2 = 2$ is chosen based on initial experiments.

TABLE I
Probabilities Used in Experiments

Probability	value
$P_D(d)$	0.041
$P_S(s)$	0.22
$P_A(a)$	0.74

Eqs. (13) and (15) require knowledge of the prior probabilities $P_A(a)$, $P_D(d)$, and $P_S(s)$. To compute these probabilities, we assume that for English speech on average speech is voiced in 78% of the time [13], that the fundamental frequency of speech is between $f_0 = 50$ and $f_0 = 500$ Hz [12] and that for most voiced speech sounds, speech energy is dominantly present up to approximately $f_c = 2000$ Hz. We then can compute the prior probabilities as

$$P_D(d) = 0.78 * \frac{f_c}{f_0} / \frac{K}{2} \tag{19}$$

$$P_S(s) = 0.22 \tag{20}$$

$$P_A(a) = 1 - P_D(d) - P_S(s), \tag{21}$$

where K is the window size. For a sample frequency $f_s = 8000$ Hz, window size $K = 256$ samples, and a typical fundamental frequency of $f_0 = 300$ Hz, this leads to the values as listed in Table I, which are the ones used in the experiments.

Estimation of ν_p in (10) is done using the ESPRIT algorithm as mentioned in Section III. Under very low SNRs, estimation of ν_p can lead to insecure estimates and, consequently, insecure values for (13) and (15). This in turn leads to a perceptually annoying switching between the deterministic and stochastic model. To overcome this, we discard the deterministic model when $\hat{\xi}(k,i) < -7$ dB and use a stochastic model alone instead. To estimate the *a priori* SNR $\xi(k,i)$ under the stochastic speech model, the decision-directed approach [1] is used with a smoothing factor $\alpha = 0.98$ with $\hat{\xi}(k,i) = \min(\hat{\xi}(k,i), -15\ \mathrm{dB})$.

To demonstrate that the proposed method is general and can also work with other distributions under the stochastic speech model, we present experimental results for the Gaussian and Laplace distribution. The reference methods used in the experiments are named: Stoch–Gauss, which is when speech DFT coefficients are always assumed to be Gaussian distributed and speech is always assumed to be present. When speech presence uncertainty is taken into account, this is referred to as Stoch–Gauss–U. Similarly, when speech DFT coefficients are always assumed to be Laplace distributed and speech is always assumed to be present, this is referred to as Stoch–Lap. When speech presence uncertainty is taken into account, this is referred to as Stoch–Lap–U.

A. Objective Results Under Gaussian Stochastic Model

In this section, we present objective results for the proposed algorithms, where we model the clean speech DFT coefficients under the stochastic speech model with a Gaussian distribution.

In Fig. 4, we compare the performance of the proposed algorithms with the reference methods in terms of improvement in

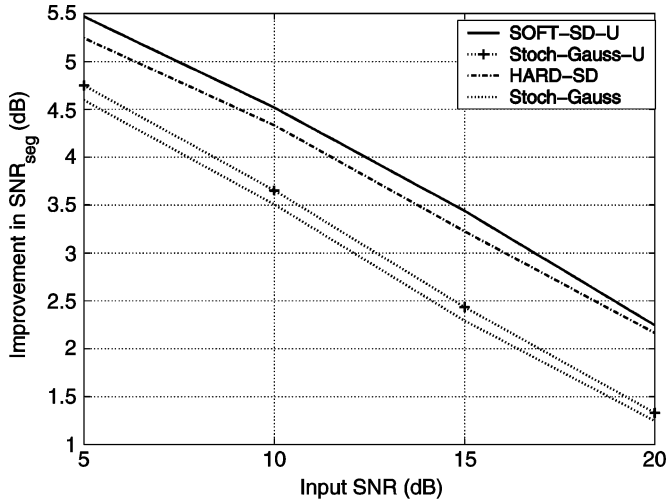


Fig. 4. Performance comparison for Gaussian stochastic model versus combined Gaussian stochastic/deterministic model for speech signals degraded by white noise in terms of input SNR versus improved $\mathrm{SNR}_{\mathrm{seg}}$.

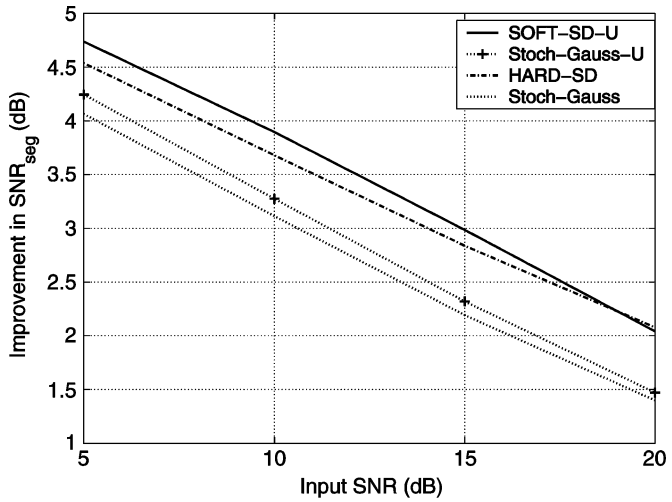


Fig. 5. Performance comparison for Gaussian stochastic model versus combined Gaussian stochastic/deterministic model for speech signals degraded by F16-fighter cockpit noise in terms of input SNR versus improved $\mathrm{SNR}_{\mathrm{seg}}$.

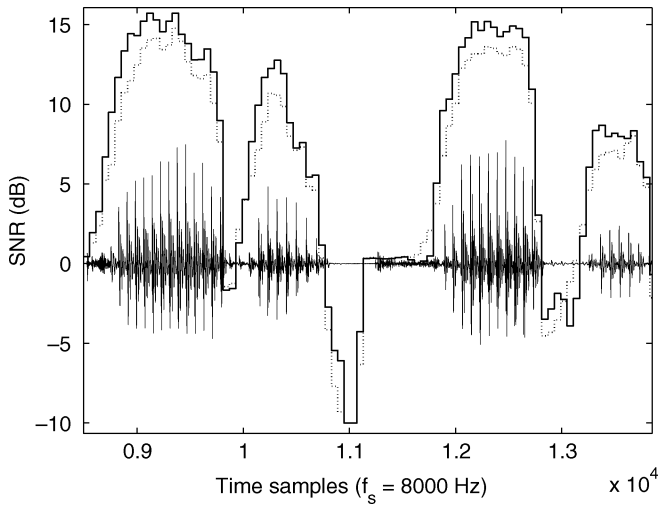


Fig. 6. Performance in terms of SNR per frame, SOFT–SD–U (solid) versus Stoch–Gauss–U (dotted).

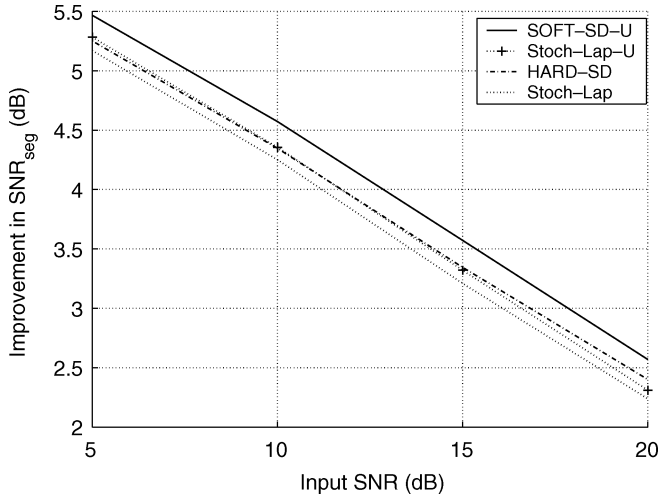


Fig. 7. Performance comparison for Laplacian stochastic model versus combined Laplacian stochastic/deterministic model for speech signals degraded by white noise in terms of input SNR versus improved $\mathrm{SNR}_{\mathrm{seg}}$.

$\mathrm{SNR}_{\mathrm{seg}}$ when speech signals are degraded by white noise at an SNR in the range from 5 to 20 dB.

Over the whole range of input SNRs, the proposed methods improve the performance compared to the use of a stochastic model alone. In terms of $\mathrm{SNR}_{\mathrm{seg}}$, the performance improvement of HARD-SD over Stoch-Gauss is approximately 0.82 dB. Incorporating the soft decision model between speech absence, the deterministic speech model and the stochastic speech model, i.e., SOFT–SD–U over HARD–SD leads to an additional 0.2 dB improvement. The improvement of SOFT–SD–U over Stoch–Gauss–U is approximately 0.87 dB.

In Fig. 5, objective results are shown for signals degraded by F16-fighter cockpit noise, where similar performance is shown as for the white noise case.

In Fig. 6, a performance comparison between SOFT–SD–U and Stoch–Gauss–U is shown in terms of SNR per frame over time together with the clean speech signal. The clean speech signal was degraded by white Gaussian noise at an SNR of 10 dB. It is clear that using the SD speech model leads to an increase in performance mainly in voiced signal regions with improvements in local SNR up to 2.5 dB.

B. Objective Results Under Laplace Stochastic Model

In this section, we present objective results for the proposed algorithms, for the case that clean speech DFT coefficients are modeled as Laplace distributed random variables under the stochastic model.

In Fig. 7, we compare the performance of the proposed algorithms with the reference methods in terms of improvement in $\mathrm{SNR}_{\mathrm{seg}}$, for speech signals degraded by white noise in the range from 5 to 20 dB. Similarly, as for the Gaussian stochastic model case also here $\mathrm{SNR}_{\mathrm{seg}}$ is improved for the SD-based approaches with respect to the use of Stoch–Lap and Stoch–Lap–U over the whole input range of SNRs. In general, the performance differences are smaller than when a Gaussian distribution is assumed as in Section V-A. We will comment more on this in Section V-D.

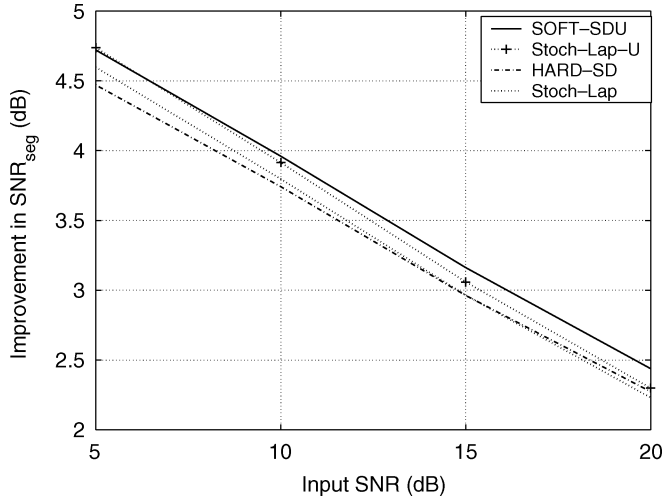


Fig. 8. Performance comparison for Laplace stochastic model versus combined Laplace stochastic/deterministic model for speech signals degraded by F16-cockpit noise in terms of input SNR versus improved $\mathrm{SNR}_{\mathrm{seg}}$.

In terms of $\mathrm{SNR}_{\mathrm{seg}}$, the performance improvement of HARD-SD over the use of a stochastic Laplacian model alone is approximately 0.11 dB. Incorporating the soft decision model between speech absence, the deterministic speech model and the stochastic speech model, i.e., SOFT–SD–U over HARD–SD leads to an additional 0.21 dB improvement. The improvement of SOFT–SD–U over Stoch–Gauss–U is approximately 0.22 dB.

In Fig. 8, similar objective results are shown, but now for signals degraded by F16-fighter cockpit noise. The comparison between SOFT–SD–U and Stoch–Lap–U shows similar performance as for the white noise case. The performance difference between HARD–SD and Stoch–Lap is negligible.

C. PESQ Evaluation

For a further evaluation of the proposed algorithms, we use an extended version of the PESQ measure [11], which predicts the subjective quality of speech signals with good correlation and expresses the quality in a score from 1.0 (worst) up to 4.5 (best). In Fig. 9(a) and (b), we compare PESQ scores for speech signals degraded by white noise and F16-fighter cockpit noise, respectively, when it is assumed that speech is Gaussian distributed under the stochastic speech model. Both SOFT–SD–U and HARD–SD lead to improved PESQ scores with respect to Stoch–Gauss–U and Stoch–Gauss. For signals degraded with white noise SOFT–SD–U and HARD–SD lead to an improvement of approximately 0.19 and 0.1 over Stoch–Gauss–U and Stoch–Gauss, respectively. For signals degraded by F16-fighter cockpit noise, the improvement of Soft–SD–U and HARD–SD over Stoch–Gauss–U and Stoch–Gaus is 0.16 and 0.11, respectively.

In Fig. 10(a) and (b), we compare PESQ scores when it is assumed that speech is Laplacian distributed under the stochastic speech model. For both white noise and F16-fighter cockpit noise, the PESQ difference between HARD–SD and Stoch–Lap is more or less negligible. The PESQ improvement

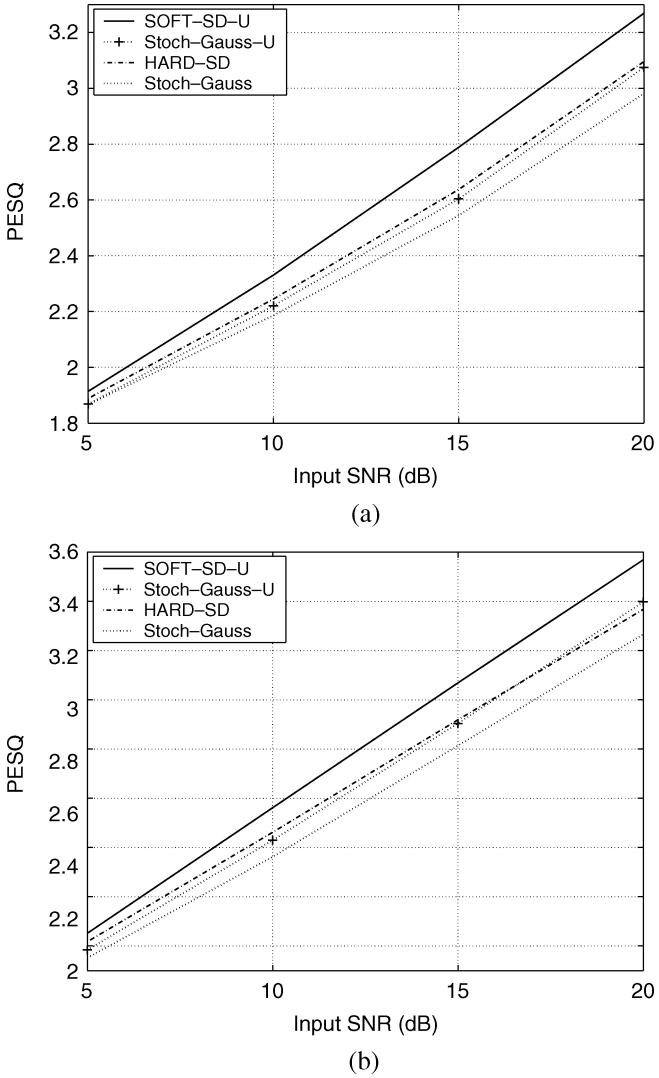


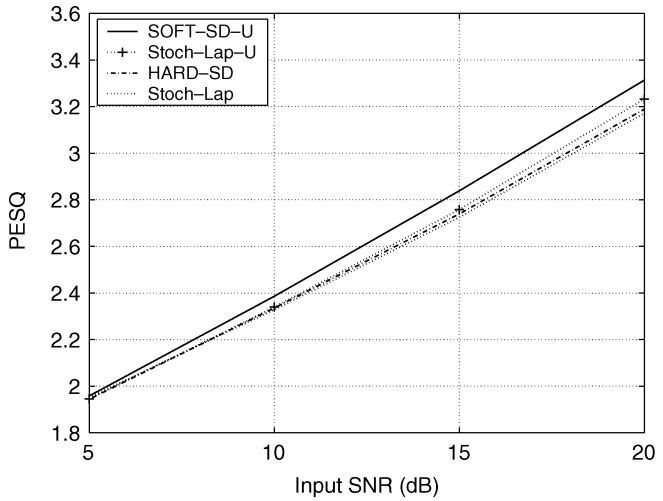
Fig. 9. Performance comparison in terms of PESQ under a Gaussian stochastic model for (a) input signals degraded by white noise and (b) input signals degraded by F16-fighter cockpit noise.

of SOFT–SD–U over Stoch–Lap–U is 0.08 and 0.05 for signals degraded with white noise and F16-fighter cockpit noise, respectively.

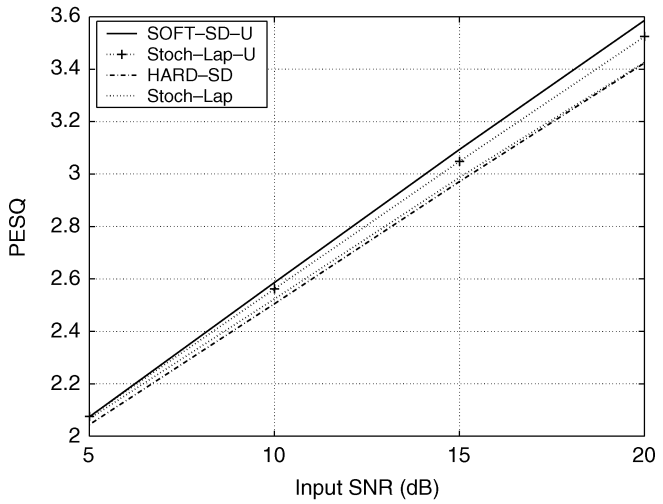
Notice that Figs. 9 and 10 show smaller differences in terms of PESQ score between the several enhancement methods at lower input SNR (e.g., at 5 dB) than at higher input SNR, while in Section V-A and V-B, it is shown that over the whole range of input SNRs, the improvement in terms of $\mathrm{SNR}_{\mathrm{seg}}$ is approximately equal. Although PESQ and $\mathrm{SNR}_{\mathrm{seg}}$ are both quality measures, we cannot expect them to measure the same kind of improvement, since they measure different aspects of quality.

D. Gaussian Versus Laplace Stochastic Model

In this section, we study the difference in performance between the Gaussian and Laplace stochastic speech model as demonstrated in the objective results in Section V-A, V-B, and V-C and explain the smaller performance difference in terms of $\mathrm{SNR}_{\mathrm{seg}}$ and PESQ between SOFT–SD–U and Stoch–Lap–U



(a)



(b)

Fig. 10. Performance comparison in terms of PESQ under a Laplace stochastic model for (a) input signals degraded by white noise and (b) input signals degraded by F16-fighter cockpit noise.

TABLE II
COMPARISON BETWEEN THE USE OF A GAUSSIAN AND LAPLACE DISTRIBUTION

Speech model	Gaussian model SNR_{seg} (dB)	Laplacian model SNR_{seg} (dB)
Stoch-Gaus/Lap-U	5.0	5.7
SOFT-SD-U	5.9	6.0

than between SOFT–SD–U and Stoch–Gauss–U. To do so, we compare in Table II the average $\mathrm{SNR}_{\mathrm{seg}}$ after enhancement of speech signals that were originally degraded by white noise at an SNR of 10 dB. We see from Table II that the use of a Laplace distribution (Stoch–Lap–U) instead of a Gaussian distribution (Gauss–Stoch–U) for the speech DFT coefficients leads to improved $\mathrm{SNR}_{\mathrm{seg}}$. This is in accordance with the results in [14] where an improvement of approximately 0.5 dB was reported. Moreover, we see from Table II that also the proposed SOFT–SD–U method with the Laplace distribution under the stochastic speech model is slightly better in terms of $\mathrm{SNR}_{\mathrm{seg}}$ as compared to the SD methods where a Gaussian model is used. However, comparing the results for SOFT–SD–U in Table II, we see that the difference between SOFT–SD–U under the two different stochastic models is decreased to approximately 0.1 dB.

Investigation of the Laplace gain function as presented in [4] and experimental analysis given in this section reveal that the 0.7 dB performance improvement of Stoch–Lap over Stoch–Gauss is only partly due to a better speech model, but that there are other beneficial side effects of using the Laplace distribution that lead to performance improvement. More specifically, it can be observed that the better performance is partly connected to the use of the decision-directed approach for estimating the *a priori* SNR. From [4] we know that the gain function under the Laplace distribution applies less suppression than the Wiener gain when the *a posteriori* SNR $\gamma(k,i)$ is high and the *a priori* SNR $\xi(k,i)$ low, a situation that typically arises for speech onsets. The Wiener gain, using the Gaussian distribution, on the other hand does not have this mechanism and will always apply high suppression when $\xi(k,i)$ is low independent on the *a posteriori* SNR. Because the decision-directed approach leads to an underestimated *a priori* SNR at speech onsets [15] due to a dependency on previous frames, the decision-directed based Wiener filter will apply too much suppression on the onsets. The Laplace based gain function, on the other hand, applies less suppression, due to the above described mechanism, and will thus lead to less distorted speech. This effect is visualized in Fig. 11(a), where the SNR per frame after enhancement of a speech signal degraded by white noise at an SNR of 5 dB is shown, together with the original clean speech signal. It is clearly visible that especially at the first half of the speech sound the use of the Laplace distribution leads to improved SNR. This is where the DD approach leads to an underestimation of the *a priori* SNR. In the second half of the speech sound, there is still some improvement, although much smaller because the influence on the *a priori* SNR estimation of the noise only frames preceding the current speech sound decreases as time evolves.

To support our discussion of the aforementioned described mechanism, we show in Fig. 11(b) experimental results averaged over 24 different speech signals degraded by white noise at an SNR of 10 dB. We compare in terms of $\mathrm{SNR}_{\mathrm{seg}}$, enhancement using Stoch–Lap with Stoch–Gauss, while the *a priori* SNR was estimated with both the DD approach and the maximum likelihood approach [1], [16]. With the maximum likelihood approach, the *a priori* SNR is computed based on an averaged noisy speech power spectrum over the current and the two last frames. The latter approach leads in general to more musical noise than the DD approach; however, it has a smaller dependency on previous frames. Fig. 11(b) shows that the Laplace distribution still leads to somewhat better performance, but that by elimination of the dependency and consequently the previously described mechanism, the performance gain of the Laplace distribution over the Gaussian distribution is decreased from 0.7 to 0.15 dB. Moreover, this mechanism also explains why the improvements of the SD methods lead to a relatively smaller improvement when the Laplace distribution is used. Specifically, one advantage of the deterministic model is the independence of the *a priori* SNR estimation and, therefore, it is independent of

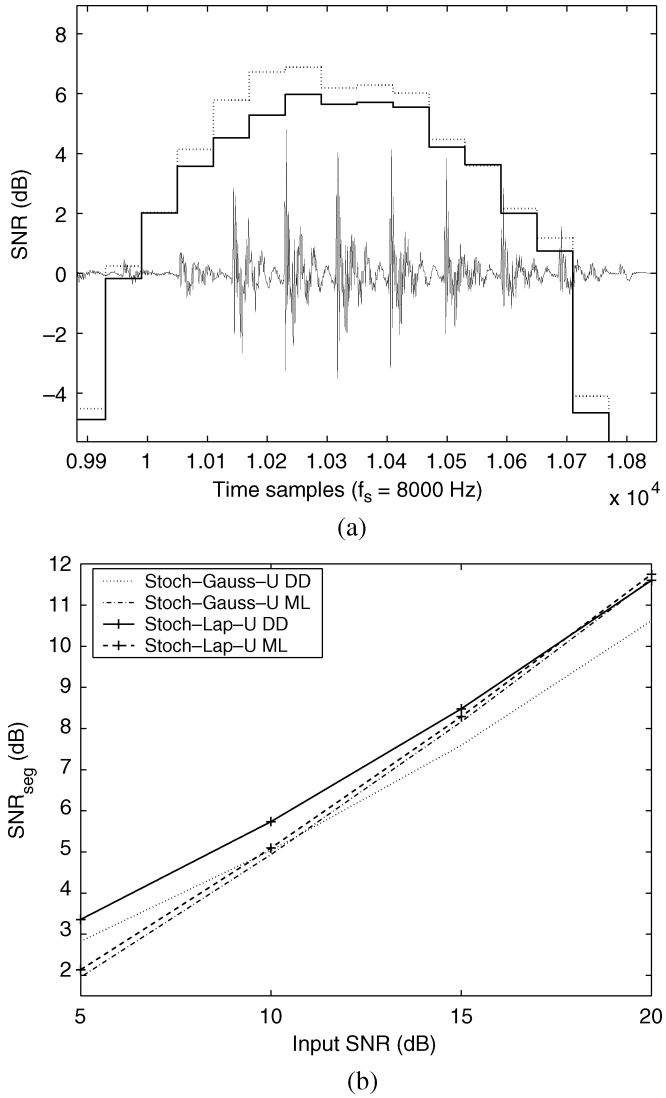


Fig. 11. Performance comparison in terms of SNR over time between (a) Stoch–Lap–U (dotted) and Stoch–Gauss–U (solid) and (b) Stoch–Lap–U versus Stoch–Gauss–U when ξ is estimated with both the decision-directed and the maximum likelihood approach.

the use of a decision-directed approach. This overcomes, similarly as with the Laplace gain function, an over-suppression at the start of stationary speech sounds and explains why combining the Laplace model with the deterministic model leads to a relatively smaller improvement than combining the Gaussian distribution as a stochastic model with the deterministic speech model.

VI. Conclusion

In this paper, we proposed the use of a combined stochastic-deterministic speech model for DFT-domain based speech enhancement. Under the deterministic speech model, clean speech DFT coefficients are modeled as a complex exponential across time. Using the combined speech model, we derived an MMSE estimator for clean speech where speech presence uncertainty is taken into account. The use of this estimator leads to less suppression of voiced speech sounds and less muffled speech than when a stochastic speech model alone is used. Although

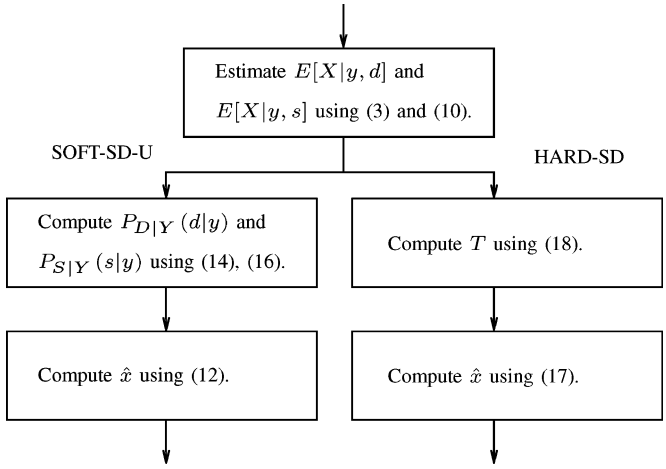


Fig. 12. Block-scheme of proposed algorithms.

the presented method is general and can be extended to be used with other distributions under the stochastic representation, we demonstrated the use of the combined stochastic-deterministic speech model using the Gaussian and Laplace distributions. Objective experiments showed that the use of the proposed MMSE estimator leads to improvements over the use of a stochastic speech model alone. Moreover, evaluation with PESQ demonstrated an improvement in speech quality. Further, we presented a discussion on the performance difference between the use of the Gaussian and the Laplace distribution under the stochastic speech model.

Appendix

In Fig. 12, we outline the steps of the proposed algorithms. These steps should be performed for each and every DFT coefficient $Y(k,i)$ of the noisy input signal.

References

[1] 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.

[2] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 3, pp. 197–210, Jun. 1978.

[3] 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.

[4] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sep. 2005.

[5] W. B. Kleijn and K. K. Paliwaf, *Speech Coding and Synthesis*. New York: Elsevier, 1995.

[6] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust, Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

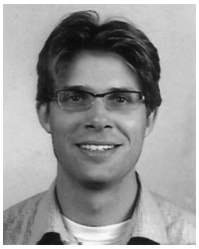
[7] B. D. Rao and K. S. Arun, "Model based processing of signals: a state space approach," *Proc. IEEE*, vol. 80, no. 2, pp. 283–307, Feb. 1992.

[8] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation for improved speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 6, Nov. 2006, to be published.

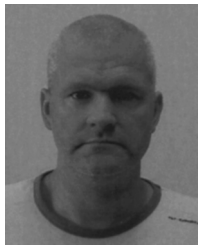
[10] S. K. Kay, *Fundamentals of Statistical Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1998, vol. 2.

[11] J. G. Beerends, "Extending p.862 PESQ for assessing speech intelligibility," *White Contribution COM 12-C2 to ITU-T Study Group 12*, Oct. 2004.
[12] J. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Piscataway, NJ: IEEE Press, 2000.
[13] G. Dewey, *Relative Frequency of English Speech Sounds*. Cambridge: Harvard Univ. Press, 1923.
[14] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Proc. Int. Workshop Acoust., Echo and Noise Control (IWAENC)*, Sep. 2003, pp. 87–90.
[15] R. C. Hendriks, R. Heusdens, and J. Jensen, "Forward–backward decision directed approach for speech enhancement," in *Proc. Int. Workshop Acoust. Echo and Noise Control (IWAENC)*, Sep. 2005, pp. 109–112.
[16] 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.



Richard C. Hendriks received the B.Sc. and M.Sc. degrees, both in electrical engineering, from Delft University of Technology, Delft, The Netherlands, in 2001 and 2003, respectively. He is currently pursuing the Ph.D. degree in the Department of Mediamatics, Delft University of Technology.

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, 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 Center for PersonKommunikation (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.