# FORWARD-BACKWARD DECISION DIRECTED APPROACH FOR SPEECH ENHANCEMENT

*Richard C. Hendriks, Richard Heusdens and Jesper Jensen*

{R.C.Hendriks, R.Heusdens, J.Jensen}@EWI.TUDelft.nl
Delft University of Technology, Dept. of Mediamatics, 2628 CD Delft, The Netherlands

## ABSTRACT

Frequency domain single-channel speech enhancement methods are often defined in terms of the a priori SNR. A widely used method to determine the a priori SNR from noisy speech is the decision directed (DD) approach. An important characteristic of the DD approach is the dependency on previously enhanced frames. This results in biased estimates of the a priori SNR during speech transitions. To overcome this problem we define in this paper a forward-backward DD approach that uses a reversed order of frame processing with a user-definable delay. With the forward-backward DD approach increased subjective and objective performances are obtained. Averaged segmental SNR is increased with more than 0.75 dB and 1.1 dB for speech degraded with white noise at SNRs of 5 dB and 15 dB, respectively.

## 1. INTRODUCTION

Speech conversations held through mobile voice communication systems are often affected by acoustical noise. Since the use of these systems has increased over the years, the interest for noise reduction methods that improve the quality in terms of intelligibility and listeners fatigue has increased as well. For the class of single microphone speech enhancement methods it is typical to assume that the speech signal is uncorrelated with the noise process and that the noise is additive, i.e. $y = x + n$, where $y$ is the noisy speech signal, $x$ the clean speech signal and $n$ the noise process.

Single-channel enhancement techniques are often formulated in the frequency domain, e.g. using the discrete Fourier transformation (DFT). Here clean speech DFT coefficients are estimated by applying a gain function to the noisy speech DFT coefficients. Examples of DFT based enhancement methods derived assuming a Gaussian distribution of clean speech DFT coefficients are the Wiener filter [1] and MMSE-STSA gain function [2]. More recently, gain functions were derived assuming other distributions of the clean speech DFT coefficients, e.g. a Laplacian distribution [3].

Often the gain functions are expressed in terms of the a priori SNR $\xi = \frac{P_{xx}}{P_{nn}}$, where $P_{xx}$ is the power spectrum of the clean speech and $P_{nn}$ the power spectrum of the noise. One method to estimate $\xi$ is the decision directed (DD) approach [2]. The DD approach is well-known because it leads to reduced musical noise. Because the DD approach works in the direction of time we refer to it as the *forward decision directed (FDD) approach*.

The research is supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

With the FDD approach the estimate of $\xi$ for the current frame is dependent on clean speech estimates from the past. Thus, the estimated $\xi$, denoted with $\hat{\xi}$, may be dependent on clean speech estimates from a different speech sound. This leads to biased estimates of $\xi$ and consequently to incorrect noise suppression, especially at the beginning of speech sounds, and moreover to the introduction of echo-artifacts at the offsets of speech sounds. In this paper we present a *backward decision directed* (BDD) approach to overcome this shortcoming of the FDD approach. Instead of using the conventional order of time, we reverse the time index and make the estimate of $\xi$ for the current frame dependent on clean speech estimates from future frames. This implies the need for a (user-definable) delay of several frames. In contrast to FDD, BDD results in less biased estimates of $\xi$ at the beginning of speech sounds and overcomes echo-artifacts at the offsets of speech sounds.

By combining BDD and FDD in a soft decision framework based on a time-adaptive segmentation algorithm for noisy speech, e.g. as in [4], a more efficient use of the noisy speech data is provided and better estimates of $\xi$ at both the start and the end of speech sounds is provided then when solely FDD is used. We will refer to this combination as the combined DD (CDD) approach.

## 2. BACKWARD DECISION DIRECTED APPROACH

The FDD approach is often used to estimate the a priori SNR $\xi$. Originally in [2] the FDD approach was defined as a linear combination between two equally valid definitions of the a priori SNR,

$$\xi(k,i) = \frac{E\left[|X(k,i)|^2\right]}{P_{nn}(k,i)} \tag{1}$$

and

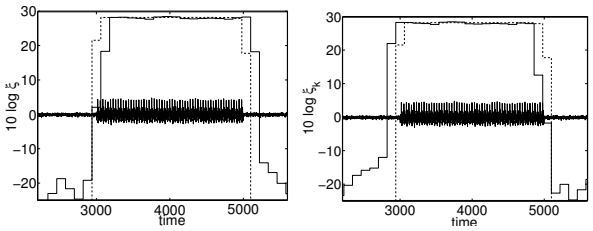
$$\xi(k,i) = E\left[\gamma(k,i) - 1\right], \tag{2}$$

where $|X(k,i)|$ denotes the clean speech amplitude of frame $i$ and frequency bin $k$ and $\gamma(k,i) = \frac{|Y(k,i)|^2}{P_{nn}(k,i)}$ the a posteriori SNR, with $|Y(k,i)|$ the noisy speech amplitude of frame $i$ and frequency bin $k$. With a smoothing factor $\alpha$ that is constrained to be $0 \leq \alpha \leq 1$, the linear combination results in

$$\xi(k,i) = E\left[\alpha \frac{|X(k,i)|^2}{P_{nn}(k,i)} + (1-\alpha)\left[\gamma(k,i) - 1\right]\right]. \tag{3}$$

However, because this expression is hard to implement in practice, approximations were introduced. This led to [2]

$$\hat{\xi}_F(k,i) = \alpha \frac{|\hat{X}(k,i-1)|^2}{P_{nn}(k,i-1)} + (1-\alpha)\max[\gamma(k,i)-1, 0], \tag{4}$$




(a) (b)

Figure 1: *Noisy speech signal with $\xi$ estimated by (a) FDD approach (b) BDD approach (solid) and the true SNR (dashed).*

where $\hat{\xi}_F$ denotes that this is the FDD approach. Among those approximations is the substitution of the expected value of the clean speech amplitude of frame $i$ with its estimate from the previous frame, i.e. frame $i-1$, resulting in a delay in the estimation of $\xi$. Further, the expectation of the a posteriori SNR $\gamma(k,i)$ has been replaced by an estimate based on the periodogram $|Y(k,i)|^2$. In Figure 1a a comparison is shown between the true SNR and the a priori SNR estimated with FDD. Here $\xi$ is estimated for a representative frequency bin of a synthetic speech signal that consists of a noise only, a voiced speech region and again a noise only region. Voiced speech is created by filtering a pulse-train through a time-invariant LPC-synthesis filter whose coefficients were extracted from a speech signal. $\hat{\xi}_F$ is plotted together with the true a priori SNR versus time. As long as the frames $..., i-2, i-1, i$ belong to the same stationary speech region, the delay introduced in (4) leads to highly smoothed, low variance estimates of $\xi(k,i)$ without affecting the estimate badly. However, if there is a transition, i.e. frames $..., i-2, i-1, i$ do not belong to the same stationary region, then the delay in (4) will lead to biased estimates of $\xi(k,i)$ and consequently to too much or too little suppression. During the start of the voiced sound and the second noise only region it can be seen that $\hat{\xi}_F$ is biased with respect to the true SNR. These biased estimates of $\xi$ lead to distorted onsets of speech sounds and echo-like artifacts at speech offsets.

Let us now consider a system where we reverse the processing order of frames, i.e. we reverse the time index. This leads to the following definition of *backward decision directed (BDD) approach*

$$\hat{\xi}_B(k,i) = \alpha \frac{|\hat{X}(k,i+1)|^2}{P_{nn}(k,i+1)} + (1-\alpha)\max[\gamma(k,i)-1, 0], \quad (5)$$

where $\hat{\xi}_B$ denotes that this is the BDD approach, which is dependent on future frames. A necessary assumption for implementation of (5) is an infinite delay. For now we will stick to this assumption, although later on we will show that this assumption can be weakened and that only a finite delay of a few frames is necessary. In Figure 1b we consider the same example, but now $\xi$ is estimated with the BDD approach. The estimate $\xi_B$ at the start of stationary regions is now approximately equal to the true SNR, while the bias in $\xi$ is now moved towards the end of stationary regions. In Figure 2 we compare the FDD and the BDD approach in terms of SNR when combined with an MMSE-STSA [2] enhancement gain function for a natural speech signal degraded with white noise at an SNR of 10 dB. From this example it is obvious that the FDD approach leads to higher SNR values at the end of speech sounds, while the BDD approach

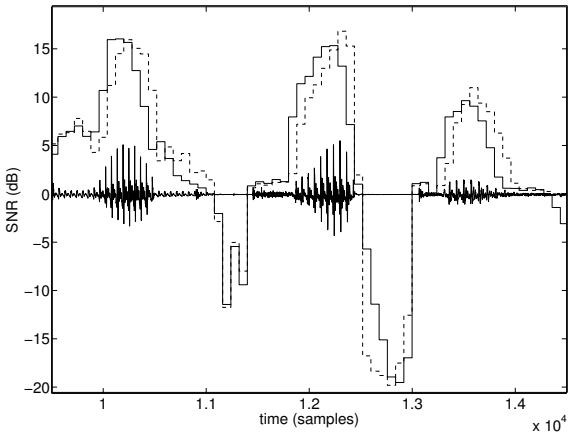



Figure 2: *Comparison of FDD approach (dashed) versus BDD approach (solid) in terms of SNR with $\alpha = 0.98$.*

leads to higher SNR values at the beginning of speech sounds. This suggests to combine the DD approaches in a way such that the advantages of each method may be exploited.

## 3. COMBINED FORWARD-BACKWARD DECISION DIRECTED APPROACH

Figures 1 and 2 indicate that the preference for FDD or BDD depends on the position of a frame within a speech sound. Given that we can identity the position of speech sounds, $\hat{\xi}_B$ and $\hat{\xi}_F$ can be combined into a single estimate $\hat{\xi}(k,i)$ by

$$\hat{\xi}(k,i) \quad = \quad \hat{\xi}_F(k,i)\beta(k,i) + \hat{\xi}_B(k,i)(1-\beta(k,i)), \quad (6)$$

with $\beta(k,i)$, $0 \leq \beta(k,i) \leq 1$. In our scheme the value of $\beta$ varies from frame to frame, dependent on the position of the frame within the speech sound; at the beginning of a speech sound $\hat{\xi}_B$ is preferred, while at the end of stationary regions $\hat{\xi}_F$ is best to use. The length of a speech sound in samples we denote by $N$ and the position of the frame within the speech sound by $n_0$. $N$ and $n_0$ can be determined using a segmentation algorithm for noisy speech as presented in [4]. Obviously, $\beta(k,i)$ depends also on $N$ and $n_0$, but for notational convenience we omit $N$ and $n_0$. In the following we describe three different ways to select $\beta(k,i)$.

The selection $\beta(k,i)$ can be done using training data. First we sample $\beta$ between 0 and 1 in steps of $\delta = 0.05$ and use 6 different signals to compute for each combination of $N$ and $n_0$, denoted by $(N, n_0)$, the amount of improvement in terms of SNR averaged over frames and frequencies for each $\beta$ value. This leads for each $(N, n_0)$ combination, to a curve denoting the average SNR improvement versus $\beta$. Given a frame, $\hat{\xi}_B$ and $\hat{\xi}_F$ can then be combined by selecting that $\beta$ that leads based on the training data to optimal average SNR improvement for the current pair $(N, n_0)$.

In order to obtain an upper bound of the performance we also consider a situation where $\beta$ is optimally determined based on SNR after enhancement using an analysis-by-synthesis approach. This confirmed that $\beta$ is dependent on the location of the frame to be enhanced within its corresponding speech sound.

Another approach that we investigate is one where $\beta$ is determined using a predetermined function $f$, with $\beta = f(N, n_0)$.

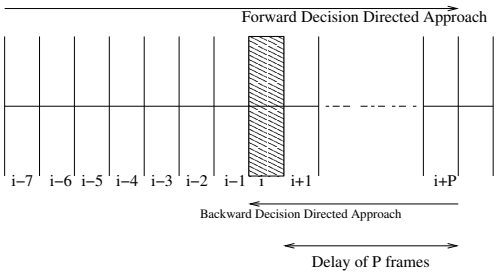



Figure 3: *Procedure for FDD approach with limited delay.*

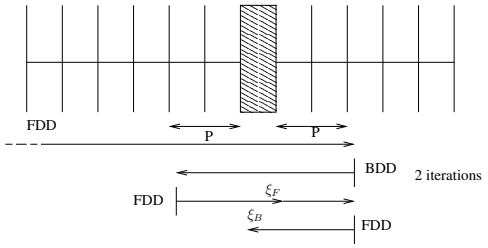



Figure 4: *Procedure for iterative combined forward-backward decision directed approach.*

The shape of $f$ is based on the observation that for frames located at the beginning of stationary regions $\beta$ should be close to zero (here $\hat{\xi}_B$ is typically a better estimate than $\hat{\xi}_F$), while at the end of stationary regions $\beta$ should be close to one. Experimentally, $f(N, n_0)$ is determined as

$$f(N, n_0) = \frac{1}{2}\left(\sin\left(1.5\pi + \pi\frac{n_0 - 1}{N - 1}\right) + 1\right).$$

This approach has an advantage that no training is needed. However, the disadvantage is that the shape of the function $f(N, n_0)$ itself is not adaptive to $(N, n_0)$, which is the case with the training based approach.

The necessary delay for the BDD approach can be limited by using the FDD approach to initialize the BDD approach. The procedure to do this is visualized in Figure 3. First the FDD approach is run up to frame $i + P$ resulting in $\hat{\xi}_F(i + P)$ which can be used to compute an estimate of the clean speech for frame $i+P$. This clean speech estimate $|\hat{X}(k, i+P)|$ can then be used to initialize the BDD approach from frame $i + P$ backwards in time to frame $i$. This procedure limits the delay to $P$ frames.

### 3.1. Iterative Combined Forward-Backward DD Approach

The above described procedure can be further extended with an iterative procedure, such that the DD smoothing process is only applied within stationary speech sounds. Both the FDD and the BDD approach are then run alternately within a surrounding of $2P + 1$ frames around the frame to be enhanced. To apply smoothing only within one speech sound, the restriction is made that this surrounding must stay within the speech sound that corresponds to the current frame. The speech sounds can be identified with a segmentation algorithm [4]. After a couple of iterations the memory of the a priori SNR estimator is mainly dominated by the current speech sound. This in contrast to the FDD approach without iterations [2] where the $\xi$ estimate, because $\alpha \simeq 1$, is largely influenced by preceding speech sounds. Simulation results confirmed that when the number of iterations is larger than 1, the difference between $\hat{\xi}_F$ and $\hat{\xi}_B$ is decreased and that the choice for $\beta$ becomes less sensitive. The procedure for iterative combined forward-backward decision directed (ICDD) approach is described in Figure 4 for a setup with two iterations. First the FDD approach is run up to the frame with index $i + P$, where frame $i$ is the frame to be enhanced. Then the BDD approach is initialized with the clean speech estimate based on the FDD approach and is run down to frame $i-P$. This ends the first iteration. Then the second iteration starts with the forward DD approach, initialized with the clean speech estimate based on the BDD approach of frame $i - P$. During this run, the estimate of the a priori SNR of frame $i$ is then used as representative for $\hat{\xi}_F$. Then as a final step the BDD approach is run for the last time, initialized with the clean speech estimate based on the FDD approach of frame $i + P$, until this run reaches frame $i$. The estimate of the a priori SNR of frame $i$ is then used as representative for $\hat{\xi}_B$.

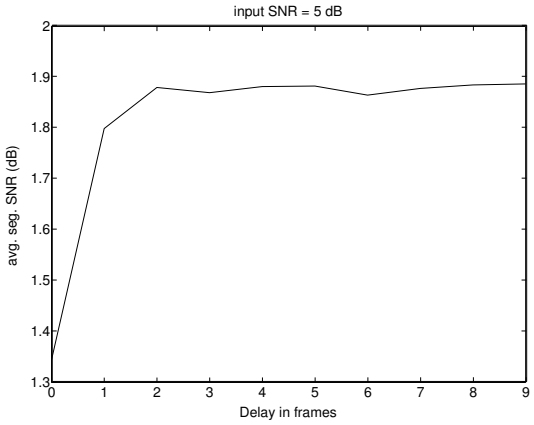



Figure 5: *Performance of the CDD approach in terms of segmental SNR in dB versus adjusted delay in frames.*

## 4. EXPERIMENTAL RESULTS

In this section we evaluate the ideas presented in Section 2 and 3 by means of objective and subjective experiments. As enhancement method we use in all experiments MMSE-STSA [2]. Furthermore, we use in all DD based methods a smoothing factor $\alpha = 0.98$, as proposed in [2]. Signals are degraded with white noise. Noise statistics are measured during silence and assumed to be stationary. The frame sizes are chosen to be 160 samples long with 50 percent overlap and a sampling frequency of 8 kHz. Computation of the $(N, n_0)$ pairs is done using a segmentation algorithm as presented in [4].

As a first experiment we evaluate the influence of the delay limitation on the CDD approach. This is done by observing the influence of the amount of allowed delay on the average segmental SNR. In this experiment, the framework proposed in Figure 3 was used. Figure 5 shows the results in terms of average segmental SNR for the CDD approach versus the delay $P$ in frames, averaged over eight different speakers. Figure 5 shows that a delay of two or three frames is already enough, and that using more delay does not further increase the performance.

Figure 6 shows a comparison between ICDD and FDD in terms

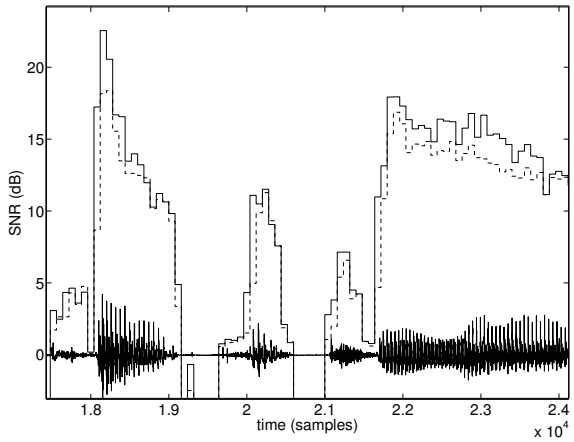



Figure 6: *Comparison between FDD (dashed) and CDD approach (solid) in terms of SNR per frame versus time with* $\alpha = 0.98$*, P=3, 1 iteration and input SNR of 15 dB.*

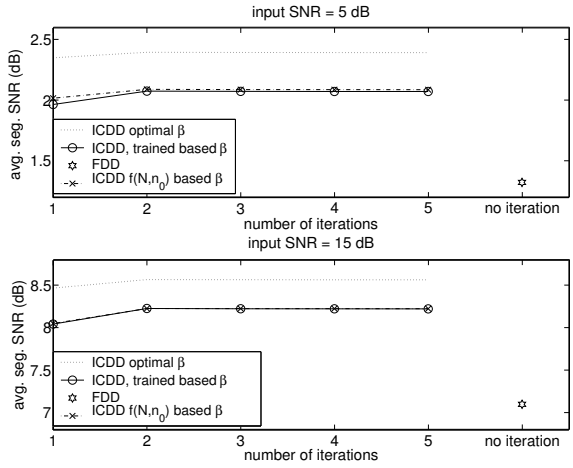



Figure 7: *Comparison in terms of segmental SNR between ICDD with P=3 frames delay versus the number of iterations and FDD.*

of SNR over time, together with the clean speech signal. The signal under consideration originates from a female speaker at an input SNR of 15 dB. The ICDD approach was implemented as demonstrated in Figure 4 with P=3 and using one iteration. The determination of $\beta$ is done by computing for each frame the $(N, n_0)$ pair, followed by selection of $\beta$ using the method described above based on training data. As expected, Figure 6 demonstrates that in terms of SNR the ICDD approach performs better than the FDD approach, especially during the start of each speech sound.

Figure 7 shows a comparison in terms of average segmental SNR versus the number of iterations for input SNRs of 5 and 15 dB between ICDD with $\beta$ selected using $(N, n_0)$ pairs based on training data, ICDD with $\beta$ selected using $\beta = f(N, n_0)$, ICDD with $\beta$ chosen optimally given the clean speech signal, and the FDD approach. The delay was chosen as $P = 3$, according to experimental results shown in Figure 5. The results are averaged over 8 speech signals. From Figure 7 it follows that most improvement is gained when going from one to two iterations, and that selection of $\beta$ using the function $\beta = f(N, n_0)$ gives approximately the same performance as when based on trained data. Furthermore, when $\beta$ can be ideally adapted to the underlying speech signal, given knowledge of the clean speech, an extra improvement of maximum 0.3 dB can be gained. The improvement of ICDD with two or more iterations over FDD is 0.75 dB and 1.1 dB for respectively input SNRs of 5 dB and 15 dB.

For subjective performance evaluation an informal OAB listening test was performed with 6 participants, the authors not included. In this test the listeners were presented first the original noise free signal followed by two different enhanced signals in randomized order. Each series was repeated 4 times. Both signals were enhanced using the MMSE-STSA estimator where one was combined with FDD approach and one with ICDD approach with $P = 3$ and 2 iterations. Selection of $\beta$ was performed using the function $\beta = f(N, n_0)$. In this listening test we used white noise at input SNRs of 15 dB and 5 dB. The relative preference of the ICDD approach over FDD approach was 68% for both input SNRs of 5 dB and 15 dB. A statistical Wilcoxon signed rank test, revealed that for both input SNRs the difference between the two methods was significant at a significance level of $p = 0.025$. The preference for ICDD is mainly due to the better estimation of the a priori SNR resulting in less distortions at the start of stationary sounds.

## 5. CONCLUSIONS

A backward decision directed (BDD) approach has been presented. This approach overcomes the introduction of distortions at the start of speech sounds and is based on a time-reversed processing order of frames. Consequently the estimation of the a priori SNR with BDD is dependent on future frames. Using a soft-decision framework the forward decision directed (FDD) and BDD approach can be combined, which leads to less biased estimates of $\xi$ at the beginning of speech sounds and overcomes echo-artifacts at offsets of speech sounds. Furthermore a limited delay BDD approach is presented, which makes it possible to reduce the delay to a few frames. Objective experiments demonstrated improvements of more than 7 dB of local SNR and improvements of more than 0.75 dB and 1.1 dB average segmental SNR for input SNRs of 5 dB and 15 dB respectively. Informal listening tests show a preference for the proposed method.

## 6. REFERENCES

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

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

[3] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using Laplacian speech priors," in *Int. Workshop on Acoustic, Echo and Noise Control*, September 2003, pp. 87–90.

[4] R. C. Hendriks, R. Heusdens, and J. Jensen, "Adaptive time segmentation of noisy speech for improved speech enhancement," in *IEEE Int. Conf. Acoust., Speech,Signal Processing*, March 2005, vol. 1, pp. 153–156.