DELAYLESS SPEECH ENHANCEMENT WITH A VIRTUAL ZERO-PHASE RESPONSE USING A PREDICTION OF PERIODIC SIGNAL COMPONENTS

Kristian Timm Andersen[1,2]*, Thomas Bo Elmedyb*[2] *and Marc Moonen*[1]

[1]KU Leuven, ESAT/STADIUS, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2]Widex A/S, Nymøllevej 6, DK-3540 Lynge, Denmark

ABSTRACT

In this paper, a delayless speech enhancement scheme with zero phase distortion is proposed. It is based on a cascade of adaptive filters that predicts periodic components with a significant auto-correlation for lags larger than a value D. The adaptive filter is positioned at the output of a speech enhancement algorithm, to adjust the phase of the periodic components to the noisy signal, and to remove stochastic signal components with a significant auto-correlation only for lags smaller than D. The stochastic components are enhanced in a separate channel and mixed back together with the periodic components to give an output with no delay or phase distortion compared to the input signal. Such a scheme is useful for low-delay processing where the phase of the signal must be preserved, for instance as a front-end for spatial filtering or when the output is mixed with another source, such as the direct transmitted sound through the vent in an open hearing aid fitting.

Index Terms— low-delay filter bank, zero-phase, real-time prediction, adaptive filters, speech enhancement.

1. INTRODUCTION

In a usual speech enhancement setup, noisy speech is sent through a filter bank, frequency dependent gains are applied and the output is an enhanced speech signal with a certain delay. Different techniques exist for reducing the delay. Filter banks with short analysis or synthesis windows give shorter delays at the cost of a reduced frequency resolution and increased aliasing [1]. A filter bank equalizer, that converts the gains to a linear-phase finite-impulse response (FIR) filter has been proposed as a way to reduce the delay to half of the corresponding discrete Fourier Transform (DFT) filter bank [2]. In addition, a minimum-phase auto-regressive (AR) filter has been proposed that minimizes the delay to a frequency dependent phase shift [3].

In this paper we propose a delayless filtering scheme with a virtual zero-phase response. By delayless virtual zero-phase, we mean that the output signal has the property of zero phase distortion with no delay compared to the input signal, without actually being filtered by a zero-phase filter (which cannot be realized in real-time). The scheme can utilize an existing filter bank with a suitably low delay (approximately a few milliseconds) and is therefore very flexible. The scheme is suitable for any application where a low-delay speech enhancement scheme is required, for instance in a hearing aid or mobile phone. Compared to a minimum-phase filter, a zero-phase filter has no distortion of the phase which can be important when several signal sources are subsequently mixed, for instance when doing spatial filtering.

To illustrate the idea, we consider a periodic signal that is sent through a filter bank with a constant delay of D samples. Since the signal is periodic, the delay through the filter bank bank is equivalent to a phase shift and can be canceled by shifting the phase of the signal output by the phase difference between the input and output of the filter bank. Any gain can be applied to the signal in the filter bank, and because the phase shift cancels the delay, the signal will be identical to a zero-phase filtered signal. The remaining problem would then be how to design a filter that can shift the phase to cancel the delay in the filter bank. However, real world signals are only periodic for a limited time and for this more general problem an adaptive filter is needed that can adapt both magnitude- and phase response to the signal. The problem can be stated as a prediction problem, where the adaptive filter predicts the signal D samples in advance. Signal components with a periodicity shorter than D samples will not be predicted in this way and we will use the term stochastic signals for them.

The proposed method applies speech enhancement to a noisy signal and then filters it with a cascade of adaptive filters that only keeps the periodic signal components phase shifted to be in phase with the noisy input signal. The stochastic signal components are separated from the noisy signal and can be can be added in with a gain factor to the filtered periodic signal since they are in phase with the noisy signal.

2. PERIODIC SIGNAL ESTIMATION

It is assumed that the noisy speech signal $x(n)$ can be separated into a periodic and a stochastic signal:

$$x(n) = \hat{x}(n) + e(n) \tag{1}$$

where $\hat{x}(n)$ is the estimated periodic signal and $e(n)$ is the stochastic signal. $\hat{x}(n)$ is an all-pass filtered version of $x(n)$ followed by an adaptive filter:

$$\hat{x}(n) = \sum_{k=0}^{K-1} h_k(n) x_A(n-k) \quad (2)$$

where $\boldsymbol{h}(n) = [h_0(n), ..., h_{K-1}(n)]^T$ is a vector of the adaptive filter coefficients and $x_A(n)$ is the output of $x(n)$ filtered by the all-pass filter $A(z)$. In the following, we only consider the situation where the all-pass filter is equal to a straight delay of D samples, i.e. $A(z) = z^{-D}$. The prediction problem consists of estimating the filter coefficients $\boldsymbol{h}(n)$ that minimizes the expected energy of $e(n)$:

$$C(n) = E\{|e(n)|^2\} \quad (3)$$

where $C(n)$ is the cost function to be minimized and $E\{\cdot\}$ denotes the expectation operator. There are many ways to solve this problem, e.g. linear prediction (LP) analysis [4], but in this paper we use a variant of the well-known normalized least-mean square (NLMS) [5] filter due to its low computational complexity and since it does not introduce any further delay in the signal path. Regularization is included by introducing a leakage factor γ and offset a so the final weight update equation is:

$$\boldsymbol{h}(n+1) = (1-\gamma)\boldsymbol{h}(n) + \mu \frac{\boldsymbol{x_D}(n) e(n)}{\boldsymbol{x_D}(n)^T \boldsymbol{x_D}(n) + a} \quad (4)$$

where $\boldsymbol{x_D}(n) = [x(n-D), ...x(n-D-K+1)]^T$ and μ is the step-size.

The periodic signal estimation can be seen inside the dotted box in Figure 1. The auto-correlation function for some examples of signal types can be seen in Figure 2. It is clear from the discussion that the adaptive filter will predict components that have a significant auto-correlation for a lag larger than D. Thus, signal components like voiced speech will be estimated by the filter while components like noise will not. Based on the figures, we make the following observations concerning the adaptive filter : (i) It can adapt to periodic signal components that have a significant auto-correlation for a lag larger than D, (ii) signal components with no significant auto-correlation for a lag larger than D must be removed to minimize the cost funtion $C(n)$ and (iii) the output of the filter $\hat{x}(n)$ must be in phase with $x(n)$ to minimize $C(n)$. In the following, we assume these three observations to be true even though they will only be approximately true for a given implementation of an adaptive filter.

Figure 3 shows an example of a noisy speech signal that is separated into a periodic and a stochastic component.

3. SPEECH ENHANCEMENT WITH ZERO-PHASE RESPONSE

The speech enhancement is performed using a DFT analysis filter bank with an analysis window of length 2D. The gain is

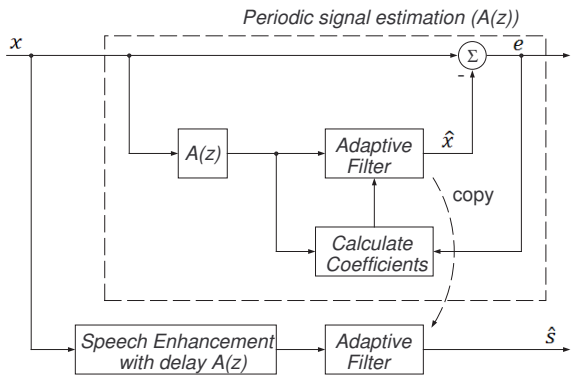

Fig. 1. *The periodic signal is estimated using an adaptive filter after an allpass filter A(z). The adaptive filter is used after a speech enhancement algorithm with a delay $A(z)$ to remove stochastic signal components and adjust the phase of the periodic signal to be in phase with $x(n)$.*

applied to the signal as a linear-phase FIR filter, since it has a delay D of half the filter length compared to a normal DFT filter bank with a delay of 2D [2]. At the output of the speech enhancement algorithm, the adaptive periodic signal estimation filter is applied. This is seen in the bottom of Figure 1. The adaptive filter removes stochastic signal components and matches the phase of the periodic components with the noisy signal $x(n)$. The output $\hat{s}(n)$ is therefore the periodic components of an enhanced speech signal in phase with the noisy signal. We note that the speech enhancement and the copy of the adaptive filter can be swapped around, if $A(z)$ and the adaptive filter is also swapped around in the periodic signal estimation. This however, would introduce a delay in the adaption of the coefficients since the output of the adaptive filter would have to pass through $A(z)$.

To determine the optimal value of D, we need to consider the consequences for both the adaptive filter and the speech enhancement. Since the adaptive filter estimates signal components with a duration longer than D to be periodic, a small D means that more of the signal will be estimated as periodic and hence will be allowed to pass through the adaptive filter. However, D also determines the delay in the speech enhancement and a small value of D means a low frequency resolution in the filter bank. A large D value therefore means that more noise attenuation can be done in the filter bank while a small D value means that more signal components will be let through the adaptive filter. This is a tradeoff that must be determined in some way. Here, we soften the tradeoff by cascading two adaptive filters with different delay values $D1$ and $D2$ respectively. The first adaptive filter estimates signal components with a duration longer than $D1$ (around 4 to 10ms) and the second filter estimates signal components from the stochastic components e_1 of the first

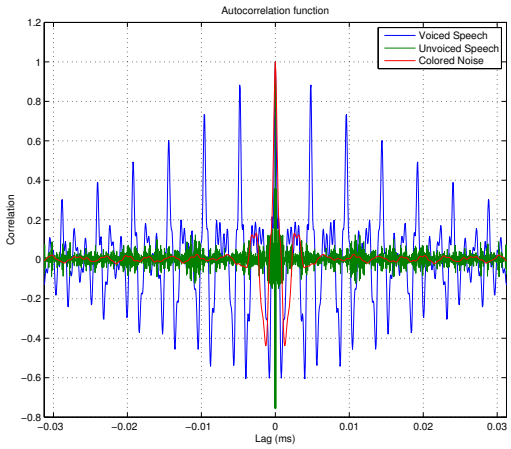

Fig. 2. *Measured auto-correlation function for three signal types: A voiced speech sample, an unvoiced speech sample and white noise that has been filtered to have the same long term spectrum as speech.*

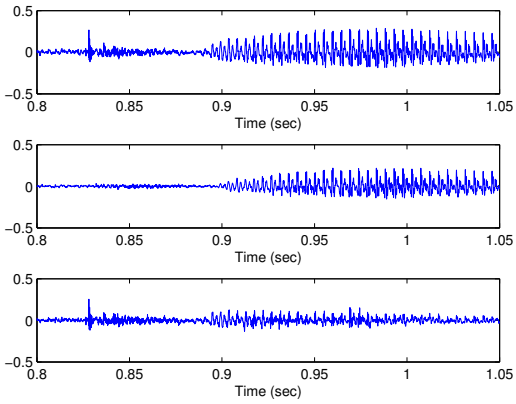

Fig. 3. *A speech signal contaminated with noise is separated into a periodic and a stochastic signal. (Top) Noisy speech signal x. (Middle) Periodic signal $\hat{x}(n)$. (Bottom) Stochastic signal $e(n)$.*

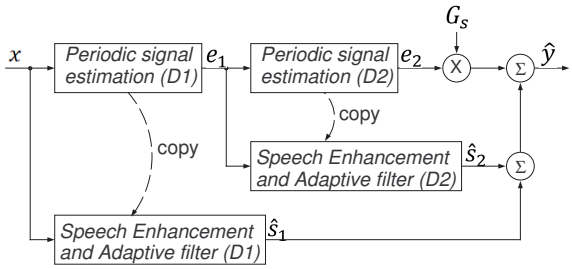

Fig. 4. *Two cascaded adaptive filters. The output from the two speech enhancement blocks can be summed with the enhanced stochastic signal since all signals are in phase.*

filter with a duration longer than $D2$ (less than 4ms). This is illustrated in Figure 4. The stochastic components e_2 of the second adaptive filter comprises all components that cannot be estimated by any of the two filters. It will be dominated by noise, transient signals and onsets like short bursts and plosives in speech. Since it consists of components with a significant auto-correlation only for lags smaller than D1 and D2, it is known that the power spectral density of these components will be relatively flat compared to the periodic components. Therefore, we justify only applying a simple scaling with gain G_s to the stochastic signal e_2 to perform a speech enhancement on the stochastic signal components without adding any delay or phase distortion. Finally the two outputs of the speech enhancement blocks with a virtual zero-phase response $\hat{s}_1$ and $\hat{s}_2$ are added to the enhanced stochastic signal to give the output $\hat{y}$.

4. EXPERIMENTAL RESULTS

Eight speech samples sampled at 16kHz were mixed with babble noise, generated by adding 12 speech samples, at various signal-to-noise (SNR) ratios and given as input to the proposed system. For the first adaptive filter, the used values were $D1 = 63$ samples, $\mu = 0.05$, $\gamma = 2 \cdot 10^{-3}$, $a = 0.05$, $K = 128$. For the second adaptive filter, the same values were used except $D2 = 15$, $\mu = 0.25$, $K = 64$. The parameters for the adaptive filter were chosen to minimize audible distortion. For the speech enhancement systems, the DFT analysis filter bank were designed using a Hann window of length $2 \cdot D1$ and $2 \cdot D2$ respectively. The gain was calculated as the Wiener gain [6], which was smoothed over time using the decision-directed a priori SNR [7] with smoothing parameter $\alpha = .97$ and the noise was estimated using an unbiased MMSE-based method [8]. The gain is used to design a linear-phase filter with a delay equal to $D1$ and $D2$ respectively [2]. The gain for the stochastic signal G_s is also calculated as a smoothed Wiener gain. Since the stochastic signal mainly consists of high frequency components, the gain is calculated based on a 1st order high-pass filtered version of the stochastic signal with a cutoff frequency of 4kHz. PESQ results can be seen in Figure 5. For comparison, a reference speech enhancement system s_{Ref} was also tested, equivalent to the speech enhancement used in $D1$ in the proposed scheme without an adaptive filter. The figure shows the results for the reference s_{Ref}, the output of the speech enhancement and adaptive filters $\hat{s}_1 + \hat{s}_2$ and the total output $\hat{y}$. It is seen that for low SNR values they perform very similarly. For higher SNR values, $\hat{s}_1 + \hat{s}_2$ performs significantly worse, which illustrates that

transient information becomes more important for the speech quality at higher SNR values. s_{Ref} and $\hat{y}$ have very similar performance, which illustrates that the proposed scheme is successful in performing speech enhancement without any significant degradation in the speech quality. Informal listening tests confirmed that they have a similar overall quality although there is some difference in the quality of the artifacts that are present. The reference scheme exhibits some musical noise which is a well-known phenomenon when the noise cannot be estimated perfectly. The proposed scheme actually has a somewhat lower degree of musical noise compared to the reference system. From listening tests it was found that the adaptive filter attenuates musical noise making the speech more pleasant to listen to. It is noted, that if the step-size μ is too large, it can also introduce certain phase-related artifacts when adapting to the signal which can make the speech sound unnatural. Due to this phenomenon, μ was set to a relatively small value. Using a larger step-size would enable the adaptive filter to estimate more of the signal as periodic components, at the price of more artifacts.

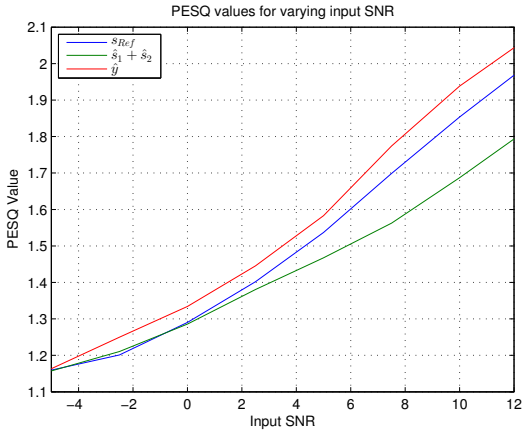

Fig. 5. *PESQ values as a function of input SNR.*

Figure 6 shows the cross-correlation function between the noisy input signal x and the output of the proposed scheme $\hat{y}$. It is seen that the function has a maximum at lag zero, which shows that the proposed scheme has a delay of zero samples, and that it is approximately symmetric, which shows that the phase has not been distorted. The bottom plot in Figure 6 shows the waveform of a voiced speech signal for x, $\hat{y}$ and the reference scheme s_{Ref}. It is seen that the phase matches between x and $\hat{y}$ while the gain of the signals are different. s_{Ref} is delayed compared to x, but since it is a linear-phase scheme it also has no phase distortion.

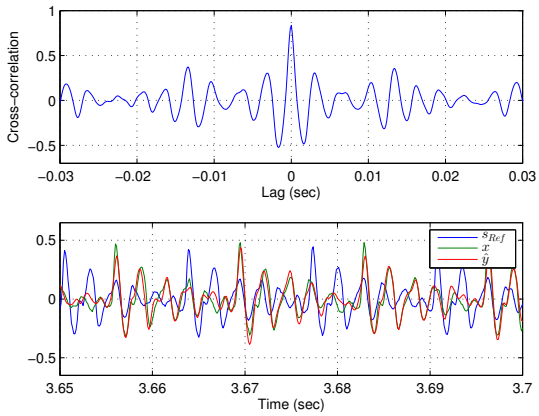

Fig. 6. *(Top) The cross-correlation function between the noisy input signal x and the virtual zero-phase output signal $\hat{y}$. (Bottom) Waveform of inputs and outputs. It is seen that the proposed scheme matches the phase of the noisy input.*

5. CONCLUSION

A delayless speech enhancement scheme with virtual zero-phase response has been proposed, which allows a noisy speech signal to be processed with a frequency dependent gain without modifying the phase or delaying the signal. This is done by predicting periodic signal components and phase shifting them to match the input signal. Stochastic components that cannot be predicted are not phase shifted. It has been shown that the proposed method results in similar speech quality to a reference linear-phase speech enhancement scheme and that the proposed scheme has somewhat less musical noise. It has also been demonstrated by a practical example that the output has approximately zero phase distortion compared to the input signal. Further research includes investigating other prediction methods than the NLMS method used in this paper.

6. REFERENCES

[1] D. Hermann, E. Chau, R.D. Dony, and S.M. Areibi, "Window based prototype filter design for highly oversampled filter banks in audio applications," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, April 2007, vol. 2, pp. II–405–II–408.

[2] H. W. Löllmann and P. Vary, "Generalized filter-bank equalizer for noise reduction with reduced signal delay," *Proc. of European Conf. on Speech Communication and Technology (Interspeech, Lisbon, Portugal)*, Sep. 2005.

[3] H. W. Löllmann and P. Vary, "Low delay filter-banks

for speech and audio processing," in *Speech and Audio Processing in Adverse Environments*, Eberhard Hänsler and Gerhard Schmidt, Eds., Signals and Communication Technology, pp. 13–61. Springer Berlin Heidelberg, 2008.

[4] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Institute of Electrical and Electronics Engineers, New York, 2000, Originally published: New York : Macmillan, 1993.

[5] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall information and system sciences series. Prentice Hall, 2002.

[6] P. Scalart and J.V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Acoustics, Speech and Signal Processing, 1996. ICASSP 1996. IEEE International Conference on*, May 1996, vol. 2, pp. 629–632 vol. 2.

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

[8] T. Gerkmann and R.C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 4, pp. 1383–1393, May 2012.