NOISE REDUCTION METHOD FOR WIDEBAND SPEECH CODING

Milan Jelinek[†] *and Redwan Salami*[‡]

[†]University of Sherbrooke, 2500 Boul. Université, Sherbrooke (Quebec), J1H 1K4 Canada
phone: +1 819 821 8000/3893, fax: +1 819 821 7937, email: Milan.Jelinek@Usherbrooke.ca

[‡]VoiceAge Corporation, 750 chemin Lucerne, suite 250, Montreal (Quebec), H3R 2H6 Canada
phone: +1 514 737 4940/239, fax:+1 514 908 2037, email: redwans@voiceage.com

ABSTRACT

We present a new low complexity noise reduction (NR) method based on spectral subtraction and overlap-add analysis/synthesis. A voicing dependent cut-off frequency is introduced, dividing the speech spectrum into two parts. In lower end, the NR gain function varies with frequency bins to minimize distortion at pitch harmonic frequencies while maximizing the suppression between them. In higher end, the gain function is estimated per critical band reducing energy variations. The gain function is further smoothed over time with a smoothing factor adaptive with the actual NR gain to prevent distortion on voiced speech onsets. The NR is as a part of VMR-WB speech codec recently selected as a new 3GPP2 standard for wideband speech applications in cdma2000 3G wireless system.

1. INTRODUCTION

Reducing the level of background noise is very important in many communication systems. For example, mobile phones are used in environments where the communication system needs to operate in the presence of high levels of car noise or street noise. In office applications, such as video-conferencing and hands-free internet applications, the system needs to efficiently cope with office noise. Noise reduction also improves the performance of the speech recognition algorithms increasingly employed in a variety of real environments.

Spectral subtraction is one the most used techniques for noise reduction [1]. Spectral subtraction attempts to estimate the short-time spectral magnitude of speech by subtracting a noise estimation from the noisy speech. The phase of the noisy speech is not processed, based on the assumption that phase distortion is not perceived by the human ear. In practice, spectral subtraction is implemented by forming an SNR-based gain function from the estimates of the noise spectrum and the noisy speech spectrum. This gain function is multiplied by the input spectrum to suppress frequency components with low SNR. The main disadvantage using conventional spectral subtraction algorithms is the resulting musical residual noise consisting of "musical tones" disturbing to the listener as well as the subsequent signal processing algorithms (such as speech coding). The musical tones are mainly due to variance in the spectrum estimates. To solve this problem, spectral smoothing has been suggested, resulting in reduced variance and resolution. Another known method to reduce the musical tones is to use an over-subtraction factor in combination with a spectral floor [2]. This method has the disadvantage of degrading the speech when musical tones are sufficiently reduced.

In the present paper, we introduce a low-complexity NR technique for 50-7000 Hz wideband (WB) speech communication systems, based on spectral subtraction and overlap-add analysis/synthesis. Similarly to the approach used in the EVRC speech codec [3], the amplitude spectrum is divided in critical bands [4] and a gain function based on SNR is computed.

However, in the presented contribution the temporal resolution of the gain function depends on the nature of the speech signal. While the noise energy is always estimated per critical band, the energy of the processed noisy speech frame and the spectral subtraction are performed per frequency bin up to a voicing cut-off frequency. Above this frequency, traditional subtraction following critical bands is used. For high pitched speakers, splitting the processing this way has the advantage of an important distortion reduction of low frequency harmonics and better NR in the valleys between them. At the same time, the smoothing advantage of the per-critical-band subtraction is maintained whenever the signal periodicity or the resolution of the spectral analysis is not high enough.

The NR gain function is smoothed over time using an adaptive smoothing factor inversely related to the actual scaling gain (smoothing is stronger for smaller gains). This approach prevents distortion in high SNR speech segments preceded by low SNR frames, as it is the case for voiced onsets for example.

The described noise reduction algorithm has been used in the Variable-Rate Multimode WB (VMR-WB) speech codec, recently selected as a new 3GPP2 standard for WB speech telephony, streaming, and multimedia messaging services in the cdma2000 third generation wireless system [5].

The paper is organized as follows. In the next section, the overview of the algorithm is given. Section 3 describes the noise energy estimation. In section 4, the NR details and examples are presented. In section 5, the performance of the NR is compared to a WB extension of a well-established NR reference algorithm of EVRC.

2. SYSTEM OVERVIEW

While the algorithm presented in this paper can be used for any application where NR is needed, it has been optimised for use within wideband speech coding systems. To minimize complexity and program memory, many of the parameters it relies upon are thus usually available in a speech encoder.

The Flow chart is outlined in Figure 1. The spectral analysis is performed twice per 20 ms frame using a square root of a Hanning window (which is equivalent to a sine window) with 50% overlap. This window is well suited for overlap-add methods when applied once at spectral analysis stage to obtain the spectrum estimate, and once at the de-noised signal reconstruction before overlap-add. This way, the artefacts introduced through a frequency-domain filtering via the NR gain function are smoothed.

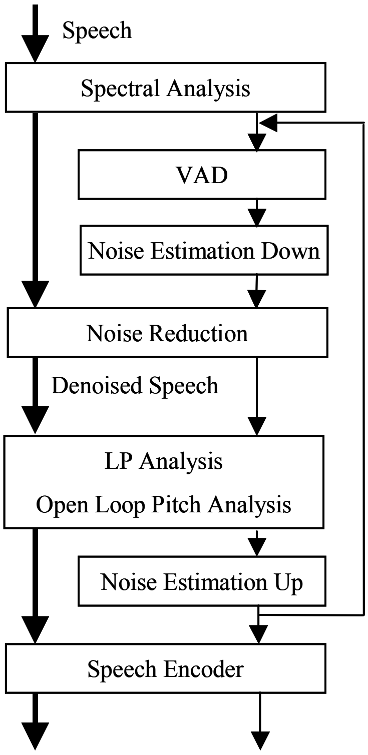

Figure 1: NR system flow chart.

The Voice Activity Detector (VAD) output is then used to control the suppressed noise level. If active speech frame is detected, the gain function is dependent on the SNR for each frequency bin or critical band as mentioned previously. Otherwise, a constant gain function is applied to the whole spectrum.

The noise estimation update is split into two steps for the following reason. Basically, the noise should be updated only on inactive speech frames. In our case, the noise update decision is made based on Linear Prediction (LP) analysis and open loop pitch analysis that are both executed on the de-noised speech signal. The noise can thus be only updated for the next frame. The only exception is when the noise update for the present frame is lower than the previous estimate for some critical band. In this case, the noise can be updated downwards to adjust the estimate in that particular band already before executing NR, independently of speech activity.

A natural question arises why we need to wait for the parameters to make a decision about speech activity, if we already have an estimation of it in the form of the VAD output. The answer is that it is very useful to have the decision about the noise estimation update independent of the VAD output, especially if the parameters for noise update decision are rather insensitive to noise variations. This way, if the noise rises rapidly and VAD erroneously indicates active speech frames, the noise estimate will continue to update and the VAD will not stay locked.

3. NOISE ESTIMATION UPDATE

The parameters used for the noise update decision are pitch stability, voicing, signal non-stationarity, and ratio between 2nd order and 16th order LP residual error energies.

The pitch stability is computed as a pitch estimate difference between several adjacent open loop pitch analyses. The voicing factor corresponds to the normalized correlation of the de-noised, perceptually weighted speech at the estimated open loop pitch period.

The frame non-stationarity assesses the per-critical-band energy variation of the current frame with respect to the long term average, similarly as the spectral deviation used in [3]. In our case it is given by the product over all critical bands of the ratios between the frame energy per critical band, E_{CB}, and the average long term energy per critical band, $\overline{E}_{CB}$, that is

$$\prod_{i=b_{\min}}^{b_{\max}} \frac{\max(\overline{E}_{CB}(i), E_{CB}(i))}{\min(\overline{E}_{CB}(i), E_{CB}(i))}$$

where b_{min} and b_{max} are the minimum and the maximum critical bands respectively.

The ratio between $E(2)$ and $E(16)$, the LP residual energies after 2nd order and 16th order analysis, reflects the fact that to represent a signal spectral envelope, a higher order of LP is generally needed for speech signal than for noise. In other words, the difference between $E(2)$ and $E(16)$ is supposed to be lower for noise than it is for active speech.

The noise estimate is updated if none of these parameters indicate an active speech frame, i.e. a quite conservative approach is taken. Further, a hangover is added to further diminish the risk of updating the noise on an active speech frame.

4. NOISE REDUCTION

Noise reduction is applied on the signal domain and de-noised signal is then reconstructed using overlap and add. The reduction is performed by scaling the spectrum using a scaling function limited between g_{min} (corresponding to –14 dB) and 1 and derived from the frequency dependent SNR.

4.1 Noise Reduction Gain Function

The scaling function is computed as a function of SNR and given by

$$g_s(f) = \sqrt{k_s\ SNR(f) + c_s}$$

$SNR(f)$ is defined as a ratio between the current frame energy $E_f(f)$ and the estimated noise energy $E_n(f)$, f being a discrete frequency. The scaling function $g_s(f)$ is bounded by

$E_{min} \leq (g_s)^2 \leq 1$. The values of k_s and c_s are determined such as $(g_s)^2 = E_{min}$ for $SNR = 1$, and $(g_s)^2 = 1$ for $SNR = 45$. That is, for $SNR = 1$ and lower, the noise reduction is limited to –14 dB, and no noise suppression is performed if the ratio of frame and noise energies is 45 or higher.

The noise energy is always estimated over critical bands, i.e. $E_n(f)$ is constant inside each critical band. The current frame energy estimation is however dependent on the voicing cut-off frequency – below that frequency, the energy is estimated per frequency bin and above that frequency, the energy is estimated per critical band. Consequently, the gain function varies with each frequency bin up to the cut-off frequency, but it is constant over critical bands above that frequency.

This new feature allows for preserving the energy at frequencies near to harmonics preventing distortion while strongly reducing the noise between the harmonics. In practice, this feature can be exploited only for voiced signals and, depending on the frequency resolution of the frequency analysis employed, for signals with relatively short pitch period. However, these are precisely the signals where the noise between harmonics is most perceptible.

An example of the effect on the lower part (0-1500 Hz) of a high pitched speech spectrum can be seen in Figure 2 for 10 dB SNR car noise. In the upper plot, per critical band processing of the whole spectrum for a noisy speech (pale curve) is compared to the spectrum of the corresponding original clean speech (dark curve). In the lower plot, the cut-off frequency logic is used. It can be seen that the noise is better suppressed in the valleys between the harmonics and that they remain generally better preserved.

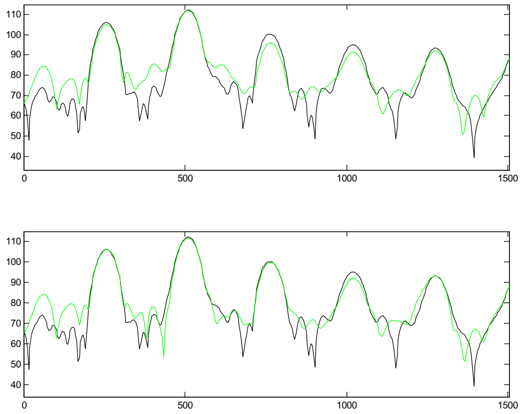

Figure 2: Example of the effect of per bin processing in low frequencies of high pitched voiced spectrum.

The actual scaling applied to the spectrum is performed using a smoothed scaling gain updated in every frequency analysis as

$$\bar{g}_s(f) = \alpha(f)\,\bar{g}_s(f) + (1 - \alpha(f))\,g_s(f),$$

where the smoothing factor α is inversely related to the gain and given by

$$\alpha(f) = 1 - g_s(f)$$

That is, the smoothing is stronger for smaller gains g_s. Temporal smoothing of the gains prevents audible energy oscillations, especially in higher part of the spectrum. Controlling the smoothing using α prevents distortion in high SNR speech segments preceded by low SNR frames, as it is the case for voiced onsets for example, This is illustrated in Figure 3. The upper plot shows about 1 second of the original clean speech. In the 2nd plot, the smoothing is done with constant value of $\alpha = 0.9$. In the lower plot, α is adaptive as described above.

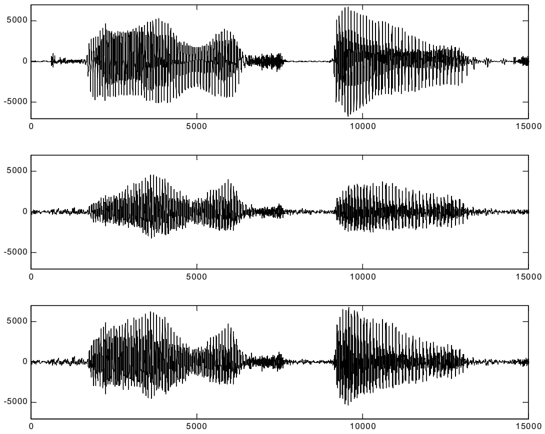

Figure 3: Example of the influence of adaptive gain scaling on a voiced onset.

4.2 Voicing Cut-off Frequency Estimation

For the purpose of the described method, the estimation of the cut-off frequency does not need a high precision. It is based on the voicing parameter and given by

$$f_c = 0.00017118\,e^{17.9772\,r_x}$$

where r_x is the voicing. This exponential characteristics has been found experimentally by first estimating the cut-off frequency with high precision using similar methods as employed in low rate parametric speech coding. These methods are based on successively high-pass filtering of the perceptually weighted speech signal and tracking the drop of the normalized correlation value below a threshold. The cut-off frequencies found were then plotted as a function of the mean normalized correlation corresponding to each of them. The dependency was close to an exponential and the characteristics above have been found through least mean squares approximation.

The cut-off frequency f_c is bounded by $325\text{ Hz} \leq f_c \leq 3700\text{ Hz}$. The number of critical bands having an upper frequency not exceeding f_c is then processed by frequency bin. The lower bound means that if f_c is lower than 325 Hz, all the spectrum is processed by critical bands. The upper bounds limits the per bin processing to the first 17 critical bands only.

Interestingly, even if a Spectral Distortion (SD) measure cannot properly assess all the advantages of the described processing, the SD for the method when the cut-off frequency estimation depends on the voicing has been lower than for any fixed cut-off frequency, including the extreme cases of all spectrum processed per frequency bin and all spectrum processed per frequency band. The SD (evaluated following critical bands) is shown in Figure 4. The upper curve represents the SD as a function of a (fixed) number of critical bands where per-bin NR is used (0 meaning that all spectrum has been processed per critical band and 20 meaning that first 20 critical bands have been all processed per frequency bin.). It can be seen that the SD for all those fixed cut-off frequencies is always above the SD when cut-off frequency varies with the voicing (lower constant curve).

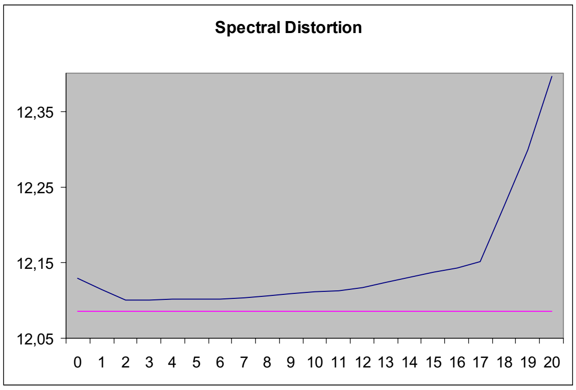

Figure 4: Spectral distortion as a function of a fixed cut-off frequency expressed in number of critical bands.

5. IMPLEMENTATION AND PERFORMANCE

The presented NR has been implemented as a part of VMR-WB speech codec. VMR-WB can operate in one of 4 modes. Modes 0, 1 and 2 are specific to cdma2000 system with mode 0 providing highest quality and mode 2 operating at lowest average bit rate. Mode 3 is designed for direct, transcoding free operation with 3GPP/ITU AMR-WB speech coding standard [6].

In VMR-WB, all internal processing including NR algorithm is done at 12.8 kHz sampling frequency. The NR spectral analysis uses windows length of 256 samples giving the spectral resolution of 50 Hz. As a consequence, splitting the spectrum for per-bin and per-band processing has been used only for pitch frequencies higher than about 110 Hz.

The algorithm performance has been evaluated by a formal MOS test against the reference NR system, a WB extension of the EVRC NR algorithm. A preliminary version of the VMR-WB codec has been used in the test, equipped with the reference NR and with the NR presented in this paper. The MOS test results are summarized in Figure 5 for 10 dB SNR car noise, 20 dB SNR car noise, 15 dB SNR street noise and 20 dB SNR office noise. It is important to note that in the 20 dB SNR cases, the bistreams have been corrupted by 2% of erased frames and by 2% of frames suffering from half-rate reduction (i.e. discarding 50% of selected bits).

It was possible to keep the complexity of the NR quite low, given the fact that most of the used parameters are available in a speech codec. The per-frequency-bin processing adds only a tiny increase because the noise estimation and hence the denominator of the SNR computation, is still done per critical band. The complexity (accounting for the cut-off frequency estimation, NR and de-noised signal reconstruction) has been evaluated to about 2.2 WMOPS using an automated WMOPS counter.

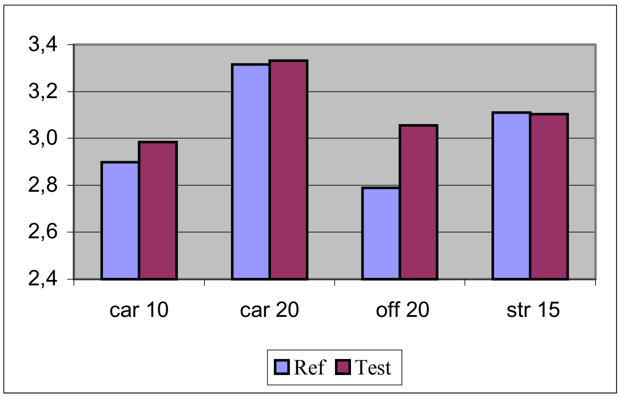

Figure 5: Comparison of MOS scores for VRM-WB codec, mode 0, equipped with the reference and the tested NR.

6. CONCLUSION

We have presented a new noise reduction method for WB speech coding. The main features consist in a NR processing depending on a cut-off frequency and in an adaptive smoothing of the NR gain function. The cut-off frequency is a function of the frame voicing. Below the frequency, per-bin NR is performed. Above the frequency, NR is done per critical band. It has been shown that this approach gives better results than processing the whole spectrum in the same way. The NR algorithm has been implemented in the VMR-WB speech codec. Its performance has been shown to be superior to the WB extension of the established EVRC NR reference.

ACKNOWLEDGEMENTS

The authors wish to thank Bruno Bessette for his comments to the algorithm design.

REFERENCES

[1] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proc. IEEE ICASSP*, Washington, DC, USA, Apr. 1979, pp. 208–211.

[3] 3GPP2 C.S0014-0 “Enhanced Variable Rate Codec (EVRC) Service Option for Wideband Spread Spectrum Communication Systems”, 3GPP2 Technical Specification, Dec. 1999.

[4] J. D. Johnston, “Transform coding of audio signal using perceptual noise criteria,” *IEEE J. Select. Areas Commun*., vol. 6, pp. 314–323, Feb. 1988.

[5] M. Jelinek, et al, “Advances in source-controlled variable bit rate wideband speech coding,” in *Porc SWIM - Lectures by Masters in Speech Processing*, Maui, HI, USA, Jan., 2004.

[6] B. Bessette, et al, “The Adaptive Multi-Rate Wideband Speech Codec (AMR-WB),” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no 8, pp. 620-636, Nov. 2002.