Efficient SNR-Based Subband Post-Processing for Residual Noise Reduction in Speech Enhancement Algorithms

Frederic Mustiere, Martin Bouchard, Miodrag Bolic

{mustiere,bouchard,mbolic}@site.uottawa.ca

uOttawa
L'Université canadienne
Canada's university

Université d'Ottawa • University of Ottawa

Issues in Speech Enhancement

The central issue in speech enhancement consists of the well-known tradeoff between noise reduction and intelligibility.

In practice, a method can rarely consistently improve intelligibility. Usually, practitioners try to at least retain it as much as possible in the noise removal process.

In sensitive applications where intelligibility and naturalness are important (e.g. hearing aid systems), non-aggressive setups for speech enhancement algorithms are privileged, since every day users will difficultly tolerate even isolated artefacts!

The main drawback is thus a larger amount of background leftover noise in the enhanced speech.

Main Objective and Constraints

Assume existing hardware/implementations and limited room for expansion: use existing architecture rather than re-engineer whole system.

► *use postprocessor*, with as simple and efficient implementation as possible (i.e, aim for low computational complexity).

Try and remove surplus background residual noise while retaining the positive features of (pre)enhanced speech (i.e. intelligibility, low distortion, naturalness, etc.)

Proposed Solution Overview

Essentially, scale on a frame-byframe basis the pre-enhanced signals depending on the respective estimated levels of speech and residual noise.

Problem: Even in ideal conditions, such volume-scaling might perceivably and disturbingly modulate the amplitude of the signal.

► *use subband approach*

For simplicity, assume that each subband-domain signal (i.e., each of the decimated signals at the outputs of the filters of the filterbank) are here real-valued and locally viewed as time-domain signals.

Subband gain determination

Proposition: reduce the gain to a single number per band and per frame – i.e., locally reduce it to a scaling factor.

► take advantage of the time/frequency localization of each small frame of data at the output of the decimated filters to formulate some simplifying assumptions, resulting in the application of a fixed gain within one subband over a few consecutive samples.

With:
• $y_m(:,i)$ = pre-enhanced speech vector at subband *m* and frame *i (not necessarily clean!)*, being the sum of:
• $x_m(:,i)$ = clean speech vector, and $r_m(:,i)$ = some residual noise.

Assumptions:
• Speech and noise statistics are fixed over small frames.
• Over *i*th frame, $x_m(:,i)$ and $r_m(:,i)$ are approximately i.i.d. with distributions:

$N(0;\sigma_x(i)^2)$ and $N(0;\sigma_r(i)^2)$

(the sequences can indeed be negative-valued, as opposed to spectral amplitudes in usual frequency-domain processing for example).

With these assumptions, it is easy to show that, for all *k* indexing the subband frame:

$$p(x_m(k,i)\,|\,y_m(:,i)) = N\left(x_m(k,i)\,|\,y_m(k,i)\frac{\sigma_x(i)^2}{\sigma_x(i)^2+\sigma_r(i)^2};\frac{\sigma_x(i)^2\sigma_r(i)^2}{\sigma_x(i)^2+\sigma_r(i)^2}\right)$$

From the above, we can write the conditional expected value of $x_m(:,i)$ in terms of the Signal-to-Residual-Noise-Ratio, denoted here by $SRNR_m(i)$, to obtain the postprocessed enhanced series estimate:

$$\hat{x}_m(:,i) = E(x_m(:,i)\,|\,y_m(:,i)) = \left(1+SRNR_m(i)^{-1}\right)^{-1} y_m(:,i)$$

Practical SRNR approximation

Simple, efficient rule is required to respect the simplicity objective.

Heuristically it was found that satisfactory results can be obtained by using the following rule:

The practical value used to represent the residual noise ratio in each subband is simply taken as the maximum between the fullband estimated SNR and the current subband estimated SNR.

$$SRNR_m(i) \approx \max\left\{SNR_m(i), SNR(i)\right\}$$

From the above equation, we necessarily have:

which is consistent with the effect of the pre-enhancement scheme.

$$SRNR_m(i) \geq SNR_m(i)$$

Performance and Simulations

Speech and noise data:

•20 kHz sampling rate (TIMIT resampled, and NOISEX-92 databases), several speakers, male and female.

•Multiple noise types, at various SNR.

Objective measures:

•Various objective measures (SNR, segmental SNR, intelligibility index (CSII), wideband perceptual speech quality (w-PESQ), composite measures).

Pre-enhancement algorithms:

•Log-spectral MMSE amplitude estimator (LMMSE) [1], both in fullband and subband setups.

•Multi-band spectral subtraction (MSSUB) [2].

Average pre-processed vs. average post-processed scores

	SNR	ASNR	CSII	WPESQ	Covl
LMMSE	12.68	2.79	0.91	1.43	1.61
LMMSE-P	*13.53*	*4.05*	*0.98*	*1.52*	*1.71*
Subband LMMSE	13.16	3.08	0.95	1.47	1.68
Subband LMMSE-P	*13.93*	*4.11*	*0.98*	*1.54*	*1.75*
MSSUB	12.65	2.51	0.94	1.65	1.93
MSSUB-P	*13.08*	*3.49*	*0.98*	*1.77*	*2.07*

(the suffix P indicates that the postprocessor was used)

The post-processor consistently increases the objective scores obtained by the enhancement algorithms.

Most benefits are seen at medium and low SNR, which correspond to situations where improvements are most needed.

Conclusion, Audio Demonstration

A low-complexity add-on to speech enhancement algorithms was proposed.

According to simulation results, the postprocessor can reduce excess of residual noise in enhanced speech without further damaging the remaining speech.

The method is most advantageous when the enhancement algorithm used operates in subbands, in which case the additional complexity is minimal.

Corresponding audio demonstrations are available on-line:

http://www.site.uottawa.ca/bouchard/papers/Eusipco_RNR.zip

References:

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 33, Issue 2, pp. 443–445, April 1985.

[2] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in ICASSP, Orlando, 2002.