# IMPLEMENTATION OF A MULTIPURPOSE NOISE SUPPRESSOR BASED ON A NOVEL SCALABLE FRAMEWORK

*Kazuhiro Yamato, Akihiko Sugiyama, and Masanori Kato*

Media and Information Research Laboratories, NEC Corporation
Kawasaki 211-8666, JAPAN

## ABSTRACT

This paper presents implementation of a multipurpose noise suppressor based on a novel scalable framework for a wide range of applications. Spectral-gain calculation incorporates adaptive control of a spectral-gain floor for an appropriate balance between residual noise and distortion in the enhanced speech. Use of a common spectral gain among channels calculated from a down-mix signal helps reduce computations on gain calculation compared to independent spectral-gain calculation. Subjective evaluation results demonstrate that the difference in the 5-grade mean opinion score with and without the multipurpose noise suppressor in the downlink is improved by as much as 0.5. When it is used as an uplink noise suppressor, the enhanced-speech quality is statistically better than or equal to that with a conventional noise suppressor. Sampling rate vs. MIPS for monaural and stereo implementations, clock rate vs. current, frame size/delay vs. MIPS based on TMS320V5510 are also presented.

***Index Terms***— Speech enhancement, Acoustic noise, Codecs, Distortion

## 1. INTRODUCTION

Noise suppressors are widely used to suppress the background noise and enhance the desired speech in noisy input. The most popular application is cellphone handsets, where it is used as a preprocessor for the encoder in the uplink handset. Another application is voice recorders. The input noisy speech is first encoded and then recorded for efficient use of the memory. A noise suppressor is used, upon playback, after the decoder as a post-processor. This structure enables the listener to enjoy the best personalized balance between the residual noise and the distortion in the enhanced speech.

In the case of voice recorders, the output of the decoder has some speech distortion originating from the encode-decode process, which is called coding distortion. Therefore, the post-processing noise suppressor should minimize additional distortion while keeping comparable level of noise suppression to the preprocessing noise suppressor. This fact results in the need for delicate control of the spectral gain.

There are other differences in the requirements between these applications, namely, the sampling frequency, the number of channels, the frame size for the input and output, the acceptable delay, the total computations, the memory size, implementation platform, and the power consumption. When a wider range of applications, such as handsets for VoIP (voice over Internet protocol) phones, remote conferencing systems, in-car hands-free systems, and preprocessors for speech recognition in noisy environment, are considered, the differences are more widespread. In view of this diversity from a viewpoint of implementation, it is desirable that a single noise suppression algorithm can cover all different requirements by simply changing parameters with no detailed tuning nor adjustment. In the field of mobile communication, noise suppressors with high speech quality have been proposed [1] and their evaluation results have been endorsed by 3GPP (The 3rd Generation Partnership Project). However, these noise suppressors cannot satisfy all the requirements imposed by the applications mentioned earlier under the single framework.

This paper presents implementation of a multipurpose noise suppressor based on a novel scalable framework. In the following section, a scalable framework that can be used for different applications by changing parameter values is explained. Section 4 demonstrates evaluation results for different scalabilities.

## 2. SCALABLE FRAMEWORK

### 2.1. Overall Structure

The scalable framework of the multipurpose noise suppressor has been developed based on the structure in [1]. This is because it provides good balance between the speech quality and the total computations. A block diagram of the multipurpose noise suppressor for two-channel input is depicted in Fig. 1. New functions are highlighted by shaded boxes.

Spectral Gain Modification provides a wide range of balance between the residual noise and distortion in the enhanced speech based on delicately controlled spectral-gain flooring. Down Mix integrates multichannel input signals into their average that is used for spectral-gain calculation for all channels to reduce the number of computations.

Noisy speech $x^0(t)$ and $x^1(t)$ are divided into frames in Frame Decomp. & Wdw. In channel 0, a frame of speech is transformed into spectral amplitude $|X_n^0(k)|$ and phase $\angle X_n^0(k)$ in Fourier Trans., where $n$ and $k$ represent the indexes to the frame and the frequency bin. An average amplitude $|\bar{X}_n(k)|$ over channels is calculated in Down Mix and provided to Noise Estimation and Gain Calc. A noise power $\lambda_n(k)$ is

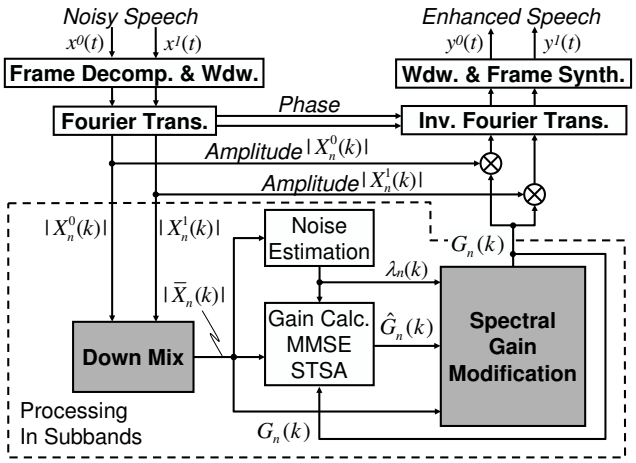



**Fig. 1**. Scalable framework of the noise suppressor.

estimated with the channel-averaged amplitude in Noise estimation. A common spectral gain $\hat{G}_n(k)$ to all channels is calculated with the estimated noise $\lambda_n(k)$ and the average amplitude $|\bar{X}_n(k)|$ in Gain Calc. Spectral Gain Modification imposes a limit on the calculated gain with the estimated noise $\lambda_n(k)$ and the average amplitude $|\bar{X}_n(k)|$ to obtain the final spectral gain $G_n(k)$. It should be noted that Down Mix, Noise Estimation, Gain Calc., and Spectral Gain Modification are all performed in nonuniform subbands for efficiency [1]. The spectral amplitude of the noisy speech in each channel multiplied by $G_n(k)$ is processed by the Inv. Fourier Trans. with the spectral phase preserved from the noisy speech. After overlap-add processing to synthesize a frame of samples in Wdw & Frame Synth., the time-domain enhanced speech is obtained.

### 2.2. Scalability of Speech Quality

Scalability of speech quality is provided by sophisticated contol of the spectral-gain floor that limits the maximum value of the spectral gain. In the case of a post-processing noise suppressor, a large spectral-gain floor is used in speech sections to minimize additional distortion that may lead to fatal degradation in subjective quality. It is set small in nonspeech sections to maximize noise suppression. In non-speech sections, the spectral-gain floor is adaptively controlled based on a long-term average of the output SNR (signal-to-noise ratio) estimate. This control enables sufficient noise suppression and natural transition from the previous speech section simultaneously for better subjective quality. When the noise suppressor is used as a preprocessor for the codec, the spectral-gain floor is fixed to a relatively small value for stronger suppression in speech sections.

Figure 2 illustrates the structure of Spectral Gain Modification. It consists of VAD (voice activity detection), LT-SNR (long-term SNR) Estimation, and Adaptive Gain Flooring. VAD determines a flag $V_n$ representing voice activity

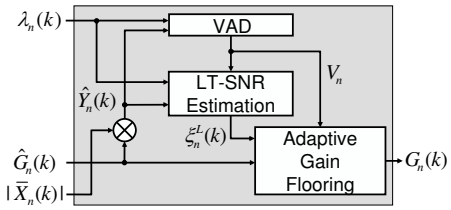



**Fig. 2**. Spectral gain modification.

based on the short-term SNR $\xi_n^S(k)$ that is given by

$$\xi_n^S(k) = \widehat{Y}_n(k)^2/\lambda_n(k) = |\widehat{G}_n(k)\,\overline{X}_n(k)|^2/\lambda_n(k), \quad (1)$$

where $\widehat{Y}_n(k)$, $\lambda_n(k)$, $\widehat{G}_n(k)$, and $\overline{X}_n(k)$ are a tentative output, the estimated noise, the MMSE spectral gain, and the average amplitude of the noisy speech over the channels. $V_n$ is set to 1 when $\xi_n^S(k)$ exceeds a predetermined threshold, and to 0 otherwise. LT-SNR Estimation calculates a long-term SNR $\xi_n^L(k)$ as the ratio of a time-averaged tentative output $\overline{Y}_n(k)$ in speech sections to $\lambda_n(k)$ as

$$\xi_n^L(k) = \overline{Y}_n(k)^2/\lambda_n(k). \quad (2)$$

Adaptive Gain Flooring calculates the final spectral gain $G_n(k)$ as

$$G_n(k) = \begin{cases} \widehat{G}_n(k), & \widehat{G}_n(k) \geq A(V_n, \xi_n^L(k)) \\ A(V_n, \xi_n^L(k)), & \widehat{G}_n(k) < A(V_n, \xi_n^L(k)) \end{cases}, \quad (3)$$

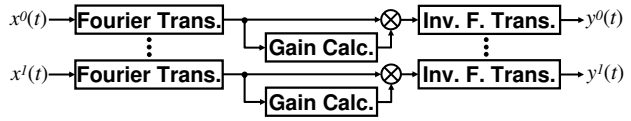
$$A(V_n, \xi_n^L(k)) = G_{fl}^1 V_n + (1 - V_n)\,\tilde{A}(\xi_n^L(k)). \quad (4)$$

$A(V_n, \xi_n^L(k))$ is an adaptive spectral-gain floor that is a function of the VAD flag $V_n$ and the long-term SNR $\xi_n^L(k)$. It keeps a large value $G_{fl}^1$ in speech sections when $V_n = 1$. In non-speech sections with $V_n = 0$, it is controlled by a monotonically decreasing function $\tilde{A}(\xi_n^L(k))$ of $\xi_n^L(k)$ with an upper limit $G_{fl}^1$ and a lower limit $G_{fl}^0$. This monotonically-decreasing nature guarantees a high spectral-gain floor for low SNRs to enable smooth transition from the previous speech section where considerable residual noise exists. When the SNR is high, the residual noise in speech sections is negligible and smaller residual noise in nonspeech sections is desirable with a low spectral-gain floor. It should be noted that a special case of this noise suppressor with $V_n = 1$ and a small value of $G_{fl}^1$ reduces to that in [1], providing scalability of speech quality.

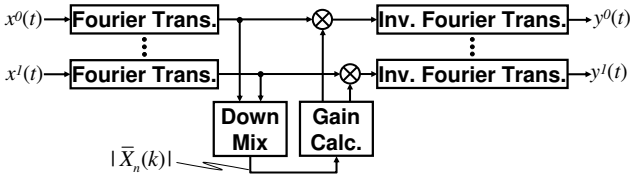
### 2.3. Channel Scalability

Common gain-calculation to all channels is introduced for computational efficiency in the case of multichannel noise suppression. This is illustrated in Fig. 3 (b). The efficient structure shares the same spectral gain among channels that is calculated from a down-mixed signal

$$|\overline{X}_n(k)| = \frac{1}{N_{ch}} \sum_{j=1}^{N_{ch}} |X_n^j(k)|, \quad (5)$$




(a) Straightforward structure with independent gains




(b) Efficient structure with a common gain

**Fig. 3**. Straightforward and efficient multichannel structures.

in contrast to the straitforward structure in Fig. 3 (a) with independent gains. $N_{ch}$ is the number of channels. [1] indicates that the fast Fourier transform (FFT), spectral gain calculation, and the inverse FFT take $1.2$, $0.8$, and $1.2$ MIPS in an implementation for an 8kHz sampled input. Therefore, reduction $R(N_{ch})$ in MIPS is approximated by

$$R(N_{ch}) \approx \frac{0.8 * (N_{ch} - 1)}{(1.2 + 1.2 + 0.8) * N_{ch}} = 0.25 * (1 - 1/N_{ch}). \quad (6)$$

$R(N_{ch})$ is $13\%$ for a stereo input and approaches $25\%$ as $N_{ch}$ increases.

### 2.4. Sampling-Rate Scalability

Because Down Mix, Noise Estimation, Gain Calc., and Spectral Gain Modification are all performed in nonuniform subbands, a subband table that defines subband decomposition of frequency bins should be prepared for each sampling rate. A subband index $i$ for a frequency bin index $k$ is calculated so that the bandwidth of a subband keeps almost the same bandwidth independent of the sampling frequency. An example of the subband table can be found as Tab. 1 in [1]. By switching the subband tables depending on the sampling rate, the spectral gain can be calculated by a single algorithm.

### 2.5. Frame-Size Scalability

Some applications such as cellphone handsets require a specific frame size along with an overlap which are determined in the communication standard. Others are more flexible to the settings of these parameters. One of the most common settings for 8kHz sampling is a frame size of $128$ with $50\%$ overlap. A larger frame size increases the delay and a deeper overlap necessitates heavier computation because the overlapped samples are processed twice. When the delay requirement is critical such as in two-way communications, a small frame size should be used. Otherwise, the frame size could be large efficiency. The block size for the Fourier transform is set larger than or equal to the frame size. For higher sampling frequencies, the frame size increases accordingly.


**Fig. 4**. Subjective-evaluations setup for post-processing.

## 3. IMPLEMENTATION

The multipurpose noise suppressor has been implemented on DSPs, Embedded Processors, and PCs. The DSPs include TMS320C55x by Texas Instruments, $\mu$PD7205x and $\mu$PD7721x by NEC. The embedded processors are represented by Xscale and StrongARM by Intel with the help of Windows CE Operating System (OS). As PC implementations, IA32 (Intel Architecture 32bit) on Windows and Linux and PowerPC by IBM and Motorola on Mac OS are available.

## 4. EVALUATION

### 4.1. Scalability of Speech Quality

Subjective evaluations were performed with the setup shown in Fig. 4. The algorithm for the encoder and the decoder was AMR [2]. Car, street and babble noise were used following the 3GPP standard [3]. SNRs were set to $6$, $12$, and $18$dB for the car noise, and $9$, $15$, and $21$dB otherwise. $24$ subjects at ages between $20$ and $40$ were asked to score the enhanced speech based on Absolute Category Rating (ACR) with a 5-grade mean opinion score (MOS) [4]. The enhanced-speech quality in the downlink with and without the multipurpose (post-processing) noise suppressor was evaluated with PSI-CELP [5] noise suppressoror the multipurpose noise suppressor as the uplink (preprocessing) noise suppressor. Figure 5 illustrates the evaluation results. The height of the bar represents the average score and the vertical line at the top of the bar exhibits the $95\%$ confidence interval.

When PSI-CELP was used as the uplink noise suppressor for the car noise with an $18$dB SNR, use of the multipurpose noise suppressor in the downlink improved the score by about $0.5$ with a statistically significant difference as shown by a dashed oval. There was no statistically significant degradation in any of the evaluated conditions. It has been demonstrated that the multipurpose noise suppressor is effective for post-processing to suppress noise with coding distortion when insufficient noise suppression is provided by preprocessing.

For evaluation as a preprocessing noise suppressor, the enhanced speech was directly compared to that by the conventional noise suppressor [1]. $22$ subjects at ages between $20$ and $40$ were asked to score the enhanced speech based on Comparison Category Rating (CCR) with a 7-grade comparison MOS (CMOS) [4]. SNRs were set to $6$ and $15$ dB for the car noise, and $9$ and $18$dB otherwise. Note that a noise suppressor is not used in the downlink and the input speech is not distorted. The results are shown in Fig. 6. The height of the bar from the $0$ line represents the average CCR. A positive score means superiority of the multipurpose noise suppressor.

The multipurpose noise suppressor achieved $0.4$ point higher

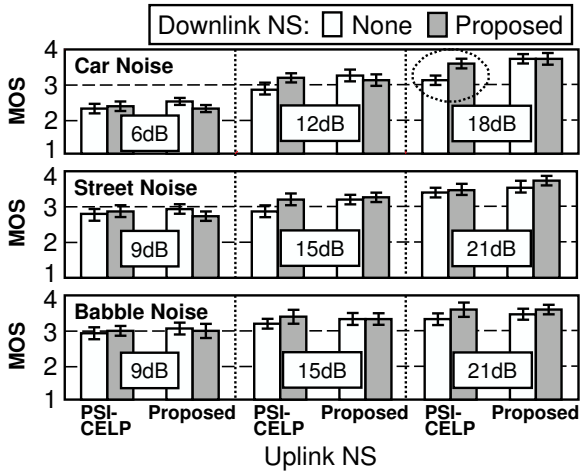



**Fig. 5**. Subjective evaluation results for post-processing.

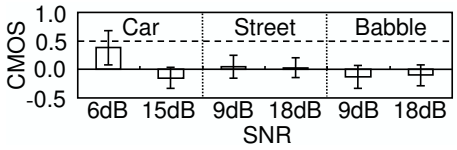



**Fig. 6**. Subjective evaluation results for preprocessing.

score than [1] for the car noise with a 6dB SNR. For other conditions, enhanced-speech quality is statistically comparable. Therefore, the performance of the multipurpose noise suppressor for undistorted speech is higher than or equal to that of [1], which had obtained endorsement by 3GPP.

### 4.2. Channel Scalability and Sampling-Rate Scalability

Figure 7 (a) shows the sampling rate vs. a measured computations in MIPS on TMS320V5510. The solid and dashed lines exhibit a stereo and a monaural implementations. The ratio of stereo MIPS to monaural MIPS is approximately $1.8$ at any sampling rate. This reduction agrees with (6) for $N_{ch} = 2$.

In Fig. 7 (b), clock frequency vs. current on TMS320V5510 is demonstrated for $8$ and $44.1$kHz sampled input signals. During measurement, the DSP performed only noise suppression. The current linearly changes with the clock frequency independent of the input-signal sampling frequency. When the power consumption is critical and a fraction of the computational power on the DSP is needed, reduction in clock frequency is useful.

### 4.3. Frame-Size Scalability

Table 1 illustrates examples of different frame-parameter values, algorithmic delay, and computations in MIPS for $8$ and $16$kHz sampling. As was discussed in 2.5, it is clearly shown that a small frame size and shallow overlap provides a short delay at the expense of total computations.

## 5. CONCLUSION

Implementation of a multipurpose noise suppressor based on a novel scalable framework has been presented. The framework provides scalability of speech quality by adaptive control of a spectral-gain floor. Channel scalability and sampling-rate scalability are achieved by calculation of a single gain based on a down-mix signal for all channels. Subjective evaluation results have demonstrated that the noise suppressor is useful for post-processing when the preprocessing noise suppressor does not provide sufficiently small residual noise. As a preprocessing noise suppressor, a positive CCR score of $0.4$ has been obtained in a low-SNR car environment in comparison with the conventional 3GPP noise suppressor.

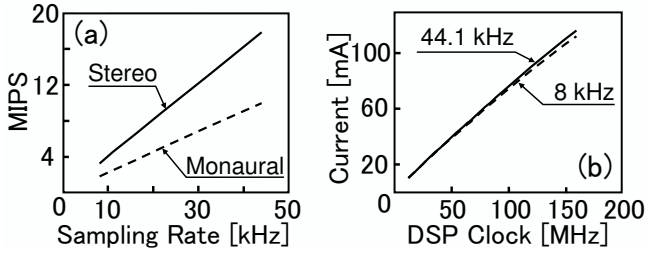



**Fig. 7**. (a) Sampling Rate vs. MIPS. (b) Clock Frequency vs. Current.

**Table 1**. Frame size, delay, and computational load.

| $F_s$ kHz | FFT Size | Frame Size | Overlap Size | Algorithmic Delay | MIPS |
|---|---|---|---|---|---|
| 8 | 256 | 160 | 40 | 5 ms | 3.2 |
| 8 | 512 | 256 | 256 | 30 ms | 1.8 |
| 16 | 512 | 256 | 256 | 15 ms | 3.6 |


## ACKNOWLEDGMENT

The authors would like to thank Dr. Osamu Hoshuyama of Media and Information Research Laboratories, NEC Corporation for his comments and encouragement.


## 6. REFERENCES


[1] M. Kato and A. Sugiyama, "A Low-Complexity Noise Suppressor with Nonuniform Subbands and a Frequency-Domain Highpass Filter," ICASSP'06, pp.473-476, May 2006.

[2] "Digital Cellular Telecommunication System (Phase 2+); Adaptive Multi-Rate (AMR); Speech Processing Functions General Description," 3GPP TS 06.71 Release 98, Jun. 1999.

[3] "Minimum Performance Requirements for Noise Suppresser Application to the AMR Speech Encoder," 3GPP TS 06.77 V8.1.1, Apr. 2001.

[4] ITU-T COM 12 "Methods for Subjective Determination of Transmission Quality," Recommendation P.800, Aug. 1996.

[5] T. Oya, H. Suda, and T. Miki, "Pitch Synchronous Innovation CELP (PSI-CELP) - PDC Half Rate Speech CODEC", IEICE Technical Report, RCS93–78, Nov. 1993.