A FAMILY OF 3GPP-STANDARD NOISE SUPPRESSORS FOR THE AMR CODEC AND THE EVALUATION RESULTS

Masanori Kato, Akihiko Sugiyama, and Masahiro Serizawa

Multimedia Research Laboratories
NEC Corporation
Kawasaki 216-8555, **JAPAN**

ABSTRACT

This paper presents a family of 3GPP-standard noise suppressors and the evaluation results. The family consists of a high-quality version and a low-complexity version. These noise suppressors are based on the MMSE STSA algorithm originally proposed by Ephraim and Malah. To meet the 3GPP requirements with better speech quality, weighted noise estimation, synthesis windowing, and pseudo noise injection are incorporated. Weighted noise estimation enables continuous noise estimation even in the speech period by using a weighted noisy speech. The weight is controlled such that a higher estimated SNR gives a smaller weight. A synthesis window function is applied between inverse transform and overlap-add processing for smooth transition at frame boundaries. Pseudo noise injection, which is not available in the low-complexity version, modifies the spectral gain based on its nonlinearity. The whole family satisfies all the 3GPP requirements. Results of a full set of evaluations specified by 3GPP are presented for the high-quality version.

1. INTRODUCTION

The third generation mobile communication service has already been put in service in some countries. 3GPP (The 3rd Generation Partnership Project) has standardized, for its standard AMR (Adaptive Multi-Rate) codec[1], the AMR noise suppressor. However, a single noise suppressor is not standardized. Instead, the minimum performance requirements for the noise suppressor and the evaluation procedure had been standardized[2]. The operator or the manufacturer can choose whatever noise suppressor algorithm they like. Therefore, 3GPP-standard noise suppressors had been intensively studied with no report of successful development.

Recently, development of three noise suppressors[3]-[5] had been reported at 3GPP. [4] is a high-quality noise suppressor and [5] is its low-complexity counterpart. Their results of evaluations, carried out according to the 3GPP specifications, show that they satisfy all the requirements. These results were submitted for review at 3GPP meetings and have been endorsed at 3GPP SA Plenary meetings[6, 7]. Although [3] was presented outside 3GPP[8], there is no report on [4, 5].

This paper presents a family of 3GPP-standard noise suppressors and the evaluation results. In the next section, the noise suppression algorithms are explained with their complexities. Section 3 presents results of a full set of evaluations specified by 3GPP.

2. NOISE SUPPRESSION ALGORITHMS

High-quality (HQ) and low-complexity (LC) noise suppression algorithms are based on MMSE STSA (Minimum Mean Square Error Short Time Spectral Amplitude) originally proposed by Ephraim and Malah [9]. The noise suppressors incorporate weighted noise estimation [10], synthesis windowing [11], and pseudo noise injection [11]. Noise estimation is carried out using the estimated SNR (Signal-to-Noise Ratio). It enables noise estimation even during the speech section resulting in better tracking capability for nonstationary noise. A synthesis window function is applied between inverse transform and overlap-add processing for smooth transition from a frame to the next by flattening out the gaps at frame boundaries. Pseudo noise injection calculates a noise level that is used to modify the amplitude of the noisy speech and the estimated noise. The injection works as if the estimated SNR were lowered such that a stronger suppression would be applied by a smaller spectral gain. This modification is effective for selectively suppressing medium-amplitude components of the noise based on nonlinearity of the spectral gain characteristics. To reduce the computational complexity, pseudo noise injection is not activated in the LC version.

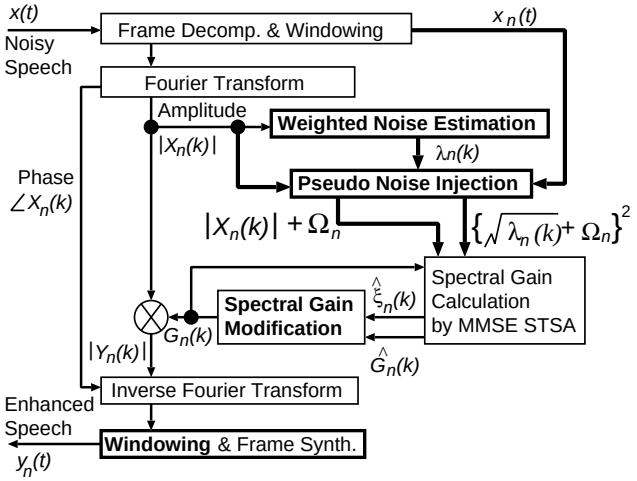

Figure 1: Structure of the high-quality noise suppressor.

Figure 1 shows the structure of the HQ noise suppressor. Functions specific to this noise suppressor are highlighted with bold lines. The input noisy speech, consisting of the desired speech and a noise, is first decomposed into frames of 200 samples with a 40-sample overlap. Each frame of the noisy speech is windowed and mapped onto a frequency domain by a 256-point Fourier transform with 56-sample zero padding. Noise is suppressed independently on the spectral magnitude at each bin by multiplying the spectral gain. It is calculated at each frequency bin from the amplitude $|X_n(k)|$ of the noisy speech and an estimated noise power $\lambda_n(k)$ based on MMSE STSA. A good noise estimate is obtained by weighted noise estimation[10]. The spectral amplitude of the input noisy speech multiplied by the spectral gain is processed by the inverse Fourier Transform with the spectral phase preserved from

Table 1: Average VAF for Japanese Speech Materials.

Condition	W (Noisy Speech Processed by AMR/NS)		X (Clean Speech Processed by AMR)		Y (Noisy Speech Processed by AMR)		Max(X, Y)		Diff. between W and Max(X,Y)	
	VAD1	VAD2	VAD1	VAD2	VAD1	VAD2	VAD1	VAD2	VAD1	VAD2
Overall	0.769	0.788	0.618	0.651	0.817	0.855	0.817	0.855	-5.88 %	-7.84 %
Car Noise	0.631	0.656	0.618	0.651	0.629	0.668	0.629	0.668	**+0.32** %	-1.80 %
Street Noise	0.791	0.822	0.618	0.651	0.858	0.919	0.858	0.919	-7.81 %	-10.6 %
Babble Noise	0.886	0.885	0.618	0.651	0.966	0.979	0.966	0.979	-8.28 %	-9.60 %

Table 2: Average VAF for English Speech Materials.

Condition	W (Noisy Speech Processed by AMR/NS)		X (Clean Speech Processed by AMR)		Y (Noisy Speech Processed by AMR)		Max(X, Y)		Diff. between W and Max(X,Y)	
	VAD1	VAD2	VAD1	VAD2	VAD1	VAD2	VAD1	VAD2	VAD1	VAD2
Overall	0.782	0.800	0.653	0.690	0.831	0.867	0.831	0.867	-5.90 %	-7.73 %
Car Noise	0.650	0.676	0.653	0.690	0.647	0.687	0.653	0.690	-0.46 %	-2.03 %
Street Noise	0.801	0.830	0.653	0.690	0.874	0.931	0.874	0.931	-8.35 %	-10.8 %
Babble Noise	0.893	0.894	0.653	0.690	0.973	0.982	0.973	0.982	-8.22 %	-8.96 %

the noisy speech. Following the inverse Fourier transform, synthesis windowing is applied for smooth transition between frames. After overlap-add processing to synthesize a frame of samples, the time-domain enhanced speech is obtained.

The HQ and the LC versions were implemented, for different applications, on TMS320VC5510 by Texas Instruments [12] and μPD77210 by NEC [13], respectively. They consume 6.98 and 6.28 MIPS with a difference of 11%. Considering that they are 7.52 and 7.43 in WMOPS [4, 5], resulting in 1% difference, the difference in MIPS heavily depends on the chip and how much they are optimized as well as the algorithm itself.

3. EVALUATION BASED ON 3GPP SPECIFICATIONS

Although the HQ and the LC versions both satisfy the 3GPP requirements [4, 5], the evaluation results for the HQ version in combination with the AMR codec for the 3G mobile communications will be presented for space limitation. All items specified by the 3GPP minimum performance requirements [2] were included in the evaluation. Because, at least, two languages are mandatory, Japanese and North American English (English) were selected. The frame size and overlap length were set to 160 and 40 with 56-sample zero padding for a 256 FFT block size to operate with the AMR codec. In the following sections, "AMR" stands for coding and decoding with the AMR codec and "AMR/NS" means that the noise suppressor is applied before "AMR".

3.1. Objective Evaluations

Bit Exactness of the Speech Encoder and Decoder The AMR speech encoder and decoder remain unaltered when the noise suppressor is applied.

Impact on Speech Path Delay An additional algorithmic delay is 5 ms due to a frame overlap of 40 samples for Fourier transform. The processing complexity in WMOPS (Weighted Million Operations Per Second) was evaluated using ETSI basic operators in the C source code. The worst-case count among 720 noisy speech samples was 7.52 WMOPS. The processing delay defined in [2] is calculated with $E * S * P = 50$ as follows:

$$Delay = WMOPS * 20/(E * S * P) = 3.0\ [ms] \quad (1)$$

Since processing delay is 3.0 *ms*, the total additional delay (comprising algorithmic and processing delays) is 8.0 *ms*, satisfying the maximum-of-10 *ms* requirement.

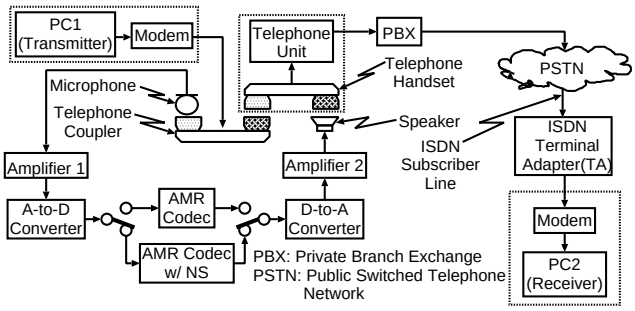

Figure 2: Evaluation Setup for Data Transmission.

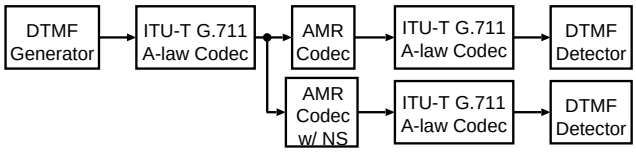

Figure 3: Evaluation Setup for DTMF.

Impact on Channel Activity (Voice Activity Factor Measure) VAF (Voice Activity Factor) measurement for both of the AMR VAD options in the two languages were performed. All the noisy speech materials used in the subjective test in the following section were used in the VAF measurement. Tables 1 and 2 show the average VAF for each of the noise conditions in each of the two languages, respectively. A negative value represents decrease in VAF. The results show that VAF was decreasing for both VAD options in all conditions when NS was active except for the car with VAD1 in Japanese. However, this increase was 0.32% that was considered insignificant at the 3GPP meeting. Therefore, the requirement for channel activity was satisfied.

Interaction with Alternate and Followed by Services To evaluate if there is any impact on data transmission, random binary data of 60 Kbytes in total and 1000 random ASCII characters were transmitted over the public telephone line as in Fig. 2. Transmission of binary data and characters was to evaluate the performance in both the synchronous and the asynchronous modes.

The data from PC1 was converted to the acoustic signal by a telephone coupler and was processed by AMR-NS or AMR. 300

Table 3: Evaluation Results of Data Transmission

Modem Bitrate	Sync. Mode		Async. Mode	
	AMR	AMR/NS	AMR	AMR/NS
1200 bps	No Error	No Error	No Error	No Error
300 bps	No Error	No Error	No Error	No Error

Table 4: Failure Rate for DTMF Detection

AMR Bitrate	AMR	AMR/NS
12.2 kb/s	0.0 % (= 0/256)	0.0 % (= 0/256)
5.9 kb/s	7.4 % (= 19/256)	7.4 % (= 19/256)

and 1200 bps were used as the modem bitrate. The bitrate of the AMR codec was set to 12.2 kbit/s. AMR and AMR-NS were performed by real-time PC software.

Table 3 shows that there was no bit error nor character error either with or without NS in both synchronous and asynchronous modes. The noise suppressor has no impact on data transmission.

Interaction with DTMF and Other Signalling Tones Objective evaluations of DTMF transparency were performed. The detailed parameter settings for DTMF are available in [4]. The evaluation system is shown in Figure 3. The DTMF generation, A-law codec and AMR speech codec, and DTMF detection were all performed by software simulations. The evaluation were carried out at 5.9 kbit/s and 12.2 kbit/s, with error free conditions in both modes. Failure rates for the DTMF digits are shown in Table 4. The failure rates are equal for AMR and AMR/NS. The requirement that the latter should not be worse than the other was satisfied.

3.2. Subjective Evaluations

Degradation in Clean Speech Figure 5 shows the results of PC (Paired Comparison) to evaluate degradation in clean speech. The ordinate represents the ratio P showing the preference of AMR/NS over AMR. Satisfaction of $0.45 < P < 0.55$ in all conditions means that degradation in clean speech by AMR/NS is statistically comparable to that by AMR, thus, the requirement is met.

Speech Degradation and Undesirable Effects in the Residual Noise Degradation of speech and undesirable effects in the residual noise were evaluated by 5-grade ACR (Absolute Category Rating). Figure 6 exhibits the results for the car, the street, and the babble noise. Considering the 95% confidence interval, the quality of AMR/NS is better than or comparable to that of AMR. The requirement was satisfied.

Quality Impact of AMR/NS Compared to that of AMR The performance of AMR/NS was evaluated by 7-grade CCR (Comparison Category Rating) in comparison with AMR. Figure 7 depicts the results at bitrates of 12.2 kbit/s and 5.9 kbit/s. Figure 8 shows the results with and without VAD/DTX and different input signal levels at 12.2 kbit/s. A higher score means a higher quality of AMR/NS than that of AMR. When the lower limit of the 95 % confidence interval does not lie in the negative region, AMR/NS has statistically higher quality than AMR. If the upper limit lies in the positive region, the quality of AMR/NS is better than or comparable to that of AMR. Figures 7 and 8 show that the quality of AMR/NS is better than that of AMR in 4 out of 6 conditions independent of the language, the bitrate, the input signal level, and the use of VAD/DTX. In addition, the quality of AMR/NS is better than or comparable to that of AMR for all conditions. Therefore, the requirements are met.

Subjective SNR Improvement Figure 4 depicts the subjective SNRI calculated from the CCR evaluation results. In either language, the SNRI is higher than 6dB in more than 2 conditions. Moreover, the SNRI is over 4dB in more than 2 conditions in the remaining 4 conditions. Therefore, the requirements are met.

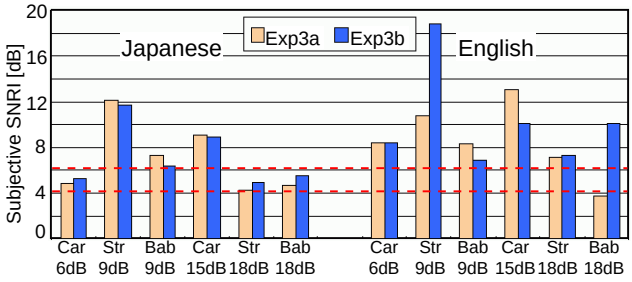

Figure 4: Subjective SNR Improvement.

4. CONCLUSION

A family of 3GPP-standard noise suppressors and the evaluation results have been presented. Depending on the computational restriction, these noise suppressors incorporate weighted noise estimation, synthesis windowing, and pseudo noise injection in the MMSE STSA algorithm to achieve high quality. Results of a full set of evaluations specified by 3GPP have confirmed that the high-quality version of the family satisfies all the 3GPP requirements. The LC version also meets all the 3GPP requirements.[5]

5. REFERENCES

[1] "Digital cellular telecommunications system (Phase 2+); Adaptive Multi-Rate (AMR); speech processing functions; General description," 3GPP TS 06.71 Release 98.

[2] "Minimum performance requirements for noise suppresser application to the AMR speech encoder," 3GPP TS 06.77 V8.1.1, Apr. 2001.

[3] "Test results of Mitsubishi AMR-NS solution based on TS 26.077," 3GPP Tdoc S4-020251, May 2002.

[4] "Test results of NEC AMR-NS solution based on TS 26.077," 3GPP Tdoc S4-020415, July 2002.

[5] "Test results of NEC low-complexity AMR-NS solution based on TS 26.077," 3GPP Tdoc S4-020417, July 2002.

[6] "TSG SA WG4 status report at TSG SA#16," TSGS#16-020221, Jun. 2002.

[7] "TSG SA WG4 status report at TSG SA#17," TSGS#16-020431, Sep. 2002.

[8] S. Furuta and S. Takahashi, "A noise suppressor for the AMR speech codec and evaluation test results based on 3GPP specifications," Proc. IEEE Workshop on Speech Coding, pp.159-161, Oct. 2002.

[9] 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, Dec. 1984.

[10] M. Kato, A. Sugiyama and M. Serizawa, "Noise Suppression with High Speech Quality Based on Weighted Noise Estimation and MMSE STSA," *Proc. IWAENC2001*, pp.183-186, Sep. 2001.

[11] A. Sugiyama, T. P. Hua, M. Kato, M. Serizawa, "Noise suppression with synthesis windowing and pseudo noise injection," Proc. of ICASSP'02, pp. 545-548, May 2002.

[12] "TMS320VC5510 Fixed-Point Digital Signal Processor DATA Manual," SPRS076C, Texas Instruments, May 2002.

[13] "μPD77210,77213 Data Sheet," U15203EJ-3V0DS00, NEC, Nov. 2001.

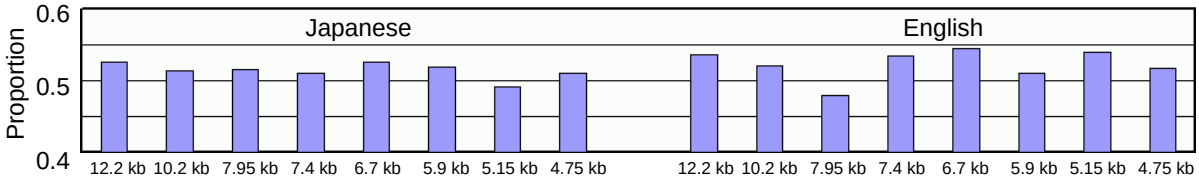

Figure 5: Paired Comparison Results.

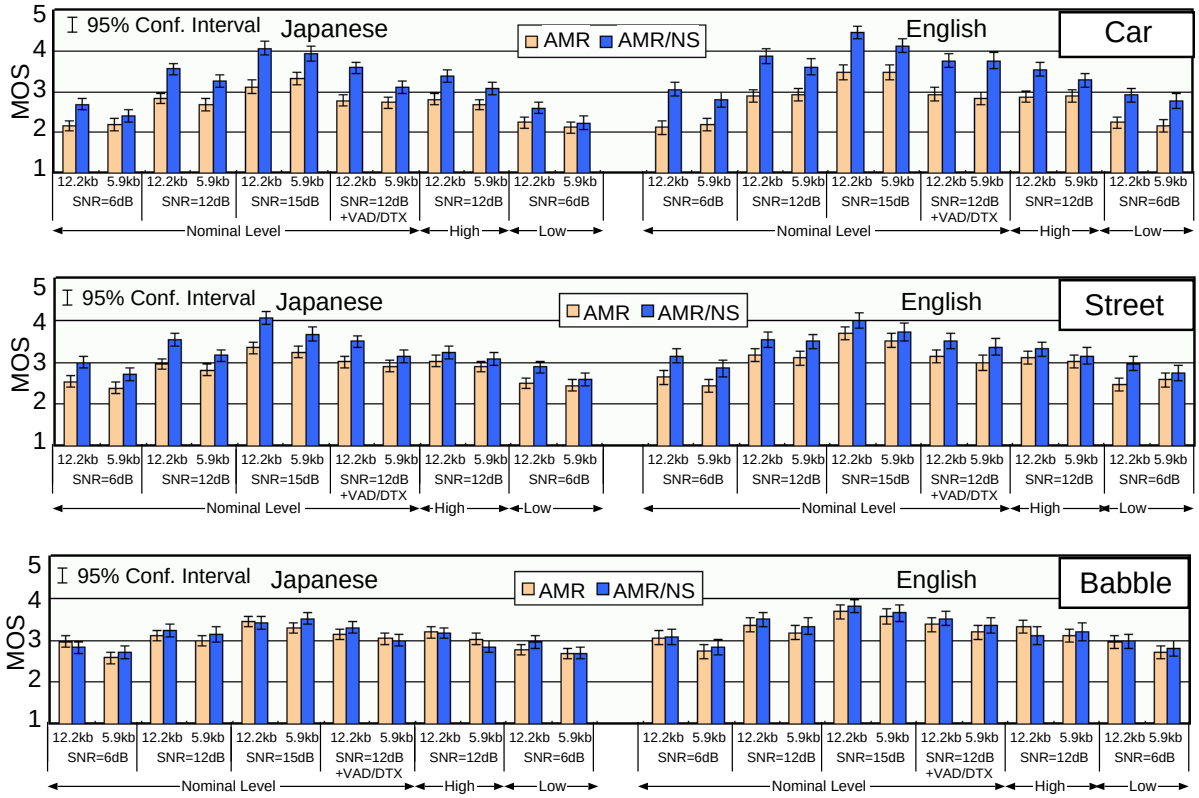

Figure 6: ACR Results.

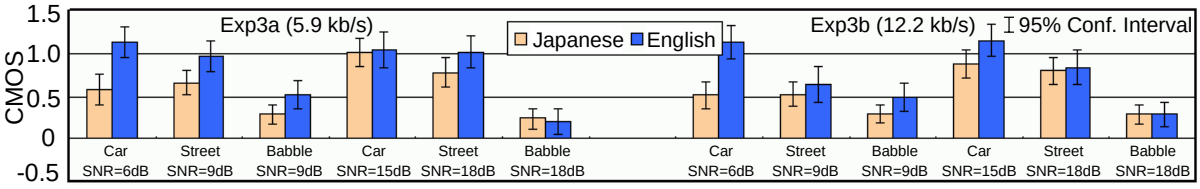

Figure 7: CCR Results (I).

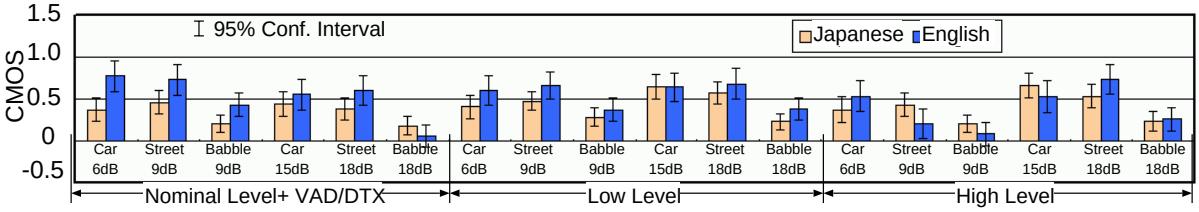

Figure 8: CCR Results(II).