# On Instrumental Quality Assessment of Speech Enhancement Systems in Three Independent Dimensions

Huajun Yu, Tim Fingscheidt

*TU Braunschweig, Institut für Nachrichtentechnik (IfN), Schleinitzstr. 22, 38106 Braunschweig, Germany*

*Email: {yu, fingscheidt}@ifn.ing.tu-bs.de*

## Introduction

Instrumental quality assessment serves as an important tool for evaluating and developing noise reduction systems. Commonly three measures are of major interest: the quality of the speech component, the level of noise attenuation, and the amount of musical tones, as a specific type of noise distortion. The first two measures can already be obtained in an instrumental manner in many applications [1], see also ITU-T Recommendations P.1100 [2] and P.1110 [3]. Nevertheless, noise distortions such as musical tones are usually still being subjectively evaluated. Recently, a high correlation of the perceived level of musical tones with the log-kurtosis ratio has been observed in [4]. However, this measure is computed from the noisy speech signal and the enhanced speech signal, making this ratio dependent on the level of noise attenuation. In addition, [4] assumes that the noise and speech signals are gamma-distributed in the power spectral domain. Motivated by the idea of an instrumental measure for residual noise distortion being independent of speech distortion and noise attenuation, a modified approach to the log-kurtosis ratio measure is proposed in this paper. Furthermore, two other quantities are presented for evaluating the quality of the speech component and noise attenuation.

In the sequel the instrumental quality assessment framework is briefly recapitulated. The three independent instrumental measures are then described. Finally, the experimental setup and results will be presented.

## Quality Assessment Methodology

In a laboratory setup, we have the clean speech signal $s(n)$ and the background noise signal $n(n)$ with time index $n$ at our disposal. The noisy speech signal $y(n)$ can then be computed as $y(n) = s(n) + n(n)$. The speech enhancement system in our case is a frequency-domain noise reduction system. In the discrete frequency domain, the spectral weights $G(\ell, k)$ are computed from $Y(\ell, k)$, where $\ell$ is the frame index, and $k$ is the frequency bin. The enhanced speech signal can then be estimated as $\hat{S}(\ell, k) = Y(\ell, k) \cdot G(\ell, k)$. Applying the same $G(\ell, k)$ for filtering $S(\ell, k)$ and $N(\ell, k)$ separately, we obtain the filtered speech signal $\tilde{S}(\ell, k)$ and the filtered noise signal $\tilde{N}(\ell, k)$. The time domain signals $\hat{s}(n)$, $\tilde{s}(n)$, and $\tilde{n}(n)$ show then the relationship $\hat{s}(n) = \tilde{s}(n) + \tilde{n}(n)$ due to the linearity property. Note that if no internal access to $G(\ell, k)$ is possible, then signals $\tilde{S}(\ell, k)$ and $\tilde{N}(\ell, k)$ can be obtained following [1] or section 8 in [2, 3].

## Independent Instrumental Measures

Firstly, the quality of the speech component is assessed by the perceptual evaluation of speech quality mean opinion score (PESQ-MOS) [5]. Please note that only the speech *component*, i.e., the filtered clean speech signal $\tilde{s}(n)$ relative to the clean speech signal $s(n)$ is judged with PESQ-MOS. As a second measure, the signal-to-noise ratio (SNR) improvement $\Delta\mathrm{SNR} = \mathrm{SNR}_{\mathrm{out}} - \mathrm{SNR}_{\mathrm{in}}$ in dB is employed to evaluate the effective noise attenuation performance. If both noise and speech components are attenuated, this measure shall only capture the relative improvement. $\mathrm{SNR}_{\mathrm{in}}$ and $\mathrm{SNR}_{\mathrm{out}}$ are computed from $(s(n),\ n(n))$ and $(\tilde{s}(n),\ \tilde{n}(n))$ using ITU-T Recommendation P.56 [6].

Uemura et al. have shown with signals $y(n)$ and $\hat{s}(n)$ that the lower a certain log-kurtosis ratio is, the less musical tones will be perceived [4]. We now propose a modified log-kurtosis ratio

$$\Delta\Psi_{\log} = \log\left(\frac{\Psi_{\tilde{n}}}{\Psi_n}\right), \tag{1}$$

where $\Psi_{\tilde{n}}$ and $\Psi_n$ are a kurtosis related to the *filtered noise* signal and to the *noise* signal, respectively. We use $\Delta\Psi_{\log}$ to quantify noise distortion (such as, e.g., musical tones). Different from [4], where $|N(\ell, k)|^2$ are assumed to be gamma-distributed in the power spectral domain, no such assumption is needed here. In the theory of higher-order statistics, the kurtosis $\Psi_x$ of a random variable $x$ is defined as $\Psi_x = \frac{E\{[x-\mu]^4\}}{(E\{[x-\mu]^2\})^2}$, where $E\{\cdot\}$ is the expectation operator and $\mu = E\{x\}$. Similar to this definition, an *instantaneous kurtosis* of *log-squared amplitude* noise DFT coefficients for each frame $\ell$ can be computed as

$$\Psi_n(\ell) = \frac{\frac{1}{K}\sum_{k=1}^{K}\left[10\log(|N(\ell,k)|^2) - \overline{10\log(|N(\ell,k)|^2)}\right]^4}{\left(\frac{1}{K}\sum_{k=1}^{K}\left[10\log(|N(\ell,k)|^2) - \overline{10\log(|N(\ell,k)|^2)}\right]^2\right)^2}, \tag{2}$$

with $\overline{10\log(|N(\ell,k)|^2)} = \frac{1}{K}\sum_{k=1}^{K} 10\log(|N(\ell,k)|^2)$ and $K$ being the DFT length.

The kurtosis $\Psi_{\tilde{n}}(\ell)$ can straightforwardly be computed by replacing $N(\ell, k)$ with $\tilde{N}(\ell, k)$ in (2). The two terms $\Psi_{\tilde{n}}$ and $\Psi_n$ will then be calculated as

$$\Psi_{\tilde{n}} = \frac{1}{L}\sum_{\ell=1}^{L}\Psi_{\tilde{n}}(\ell),\ \Psi_n = \frac{1}{L}\sum_{\ell=1}^{L}\Psi_n(\ell), \tag{3}$$

where $L$ indicates the frame length. Finally, $\Delta\Psi_{\log}$ can be computed with $\Psi_{\tilde{n}}$ and $\Psi_n$ without any assumption about probability distribution functions.

Furthermore, it can be shown that the three instrumental measures PESQ-MOS, $\Delta$SNR, and $\Delta\Psi_{\log}$ are independent from each other.

## Experimental Setup and Results

Our experiments are performed with automotive noises. Eight clean speech signals (four male and four female), and 12 in-car background noise signals, each with a length of 8s, are taken from the NTT database [7] and the ETSI background noise database [8], respectively. The $\mathrm{SNR}_{\mathrm{in}}$ values are chosen from $-5$ dB to 20 dB with a step-size of 5 dB. Signals are windowed by a Hann window of length 512 samples, followed by a DFT

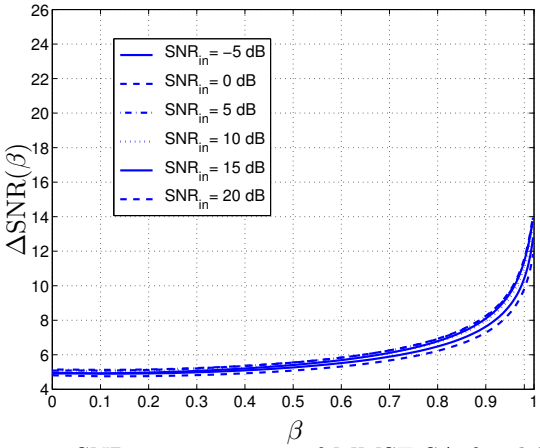


Figure 1: SNR improvement of MMSE-SA for different input SNR levels

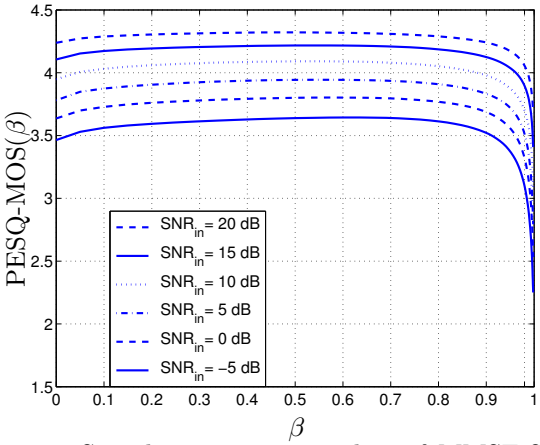


Figure 2: Speech component quality of MMSE-SA for different input SNR levels

with length $K = 512$ and a frame shift of 50%. All signals are sampled with 16 kHz. Ephraim and Malah's MMSE-SA estimator [9] will be evaluted with the *a priori* SNR $\xi(\ell,k)$ being estimated by the decision-directed (DD) approach [9]

$$\xi'(\ell,k) = \beta \cdot \frac{|\hat{S}(\ell-1,k)|^2}{\hat{\phi}_{NN}(\ell-1,k)} + (1-\beta)\cdot P[\gamma(\ell,k)-1], \quad (4)$$
$$\xi(\ell,k) = \max\{\xi'(\ell,k),\ \xi_{\min}\},$$

with a smoothing factor $\beta$, the enhanced speech signal of the previous frame $\hat{S}(\ell-1,k)$, the *a posteriori* SNR $\gamma(\ell,k) = \frac{|Y(\ell,k)|^2}{\hat{\phi}_{NN}(\ell,k)}$, and $\xi_{\min} = -15\,\text{dB}$. The estimated noise power spectrum $\hat{\phi}_{NN}(\ell,k)$ is computed via minimum statistics [10]. In this paper, we choose $0 \le \beta \le 0.998$. The corresponding results of PESQ-MOS($\beta$), $\Delta$SNR($\beta$), and $\Delta\Psi_{\log}(\beta)$ are shown in Figs. 1-3. It can be observed that with $\beta$ in (4) being increased noise attenuation in terms of SNR improvement increases exponentially. Therefore, concerning noise attenuation a large $\beta$ close to unity is of interest. However, transient distortions for the speech component will unfortunately occur with $\beta$ being chosen close to unity, which leads to a strong smoothing for the $\xi(\ell,k)$ estimation. This phenomenon can be well analyzed by the PESQ-MOS measures of the speech component in Fig. 2 for different $\text{SNR}_{\text{in}}$ levels. It can be observed that with $\beta \ge 0.7$ PESQ-MOS scores decrease monotonically. The most cited benefit of employing the DD approach is to reduce musical tones by setting $\beta$ close to unity, which carries out a smoothing procedure obtaining a more consistent estimate of $\xi(\ell,k)$. The $\Delta\Psi_{\log}(\beta)$ measures in Fig. 3 reflect this conclusion: When $\beta$ is increased from 0 to 1, $\Delta\Psi_{\log}$ accordingly will decrease meaning musical tones will be attenuated.

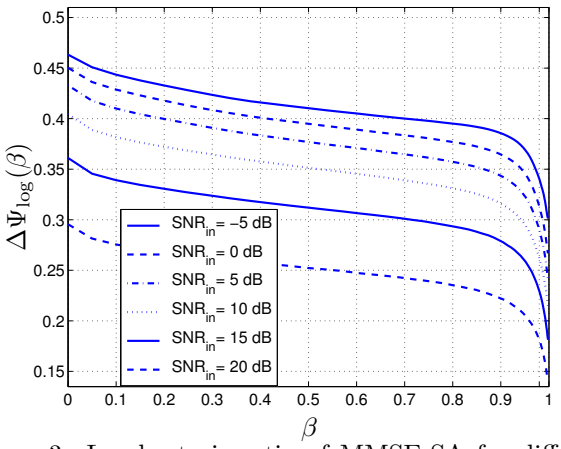


Figure 3: Log-kurtosis ratio of MMSE-SA for different input SNR levels

It can also be noted that $\Delta\Psi_{\log}$ increases with the decreasing $\text{SNR}_{\text{in}}$ levels. This shows, that more musical tones will be generated especially for low $\text{SNR}_{\text{in}}$ levels. This phenomenon can be confirmed by informal listening tests.

## Conclusions

We have presented three independent instrumental quality measures for evaluating speech enhancement systems. Among them a log-kurtosis ratio measure for noise distortion in terms of musical tones is proposed. All measures serve to optimize parametrization of a speech enhancement system.

## References

[1] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards Objective Quality Assessment of Speech Enhancement Systems in A Black Box Approach," in *Proc. of ICASSP'08*, Las Vegas, NV, Apr. 2008, pp. 273–276.

[2] "ITU-T Recommendation P.1100, Narrow-Band Hands-Free Communication in Motor Vehicles," 2008.

[3] "ITU-T Recommendation P.1110, Wideband Hands-Free Communication in Motor Vehicles," 2009.

[4] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic Optimization Scheme of Spectral Subtraction based on Musical Noise Assessment via Higher-Order Statistics," in *Proc. of IWAENC'08*, Seattle, WA, Sep. 2008.

[5] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," 2005.

[6] "ITU-T Recommendation P.56, Objective Measurement of Active Speech Level," 1993.

[7] NTT Advanced Technology Corporation, "Multi-Lingual Speech Database for Telephonometry," 1994.

[8] "ETSI EG 202 391-1, Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation technique and Background Noise Database," 2008.

[9] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 9, no. 5, pp. 504–512, July 2001.