Signal-to-Noise-Ratio-Aware Dynamic Range Compression in Hearing Aids

May, Tobias; Kowalewski, Borys; Dau, Torsten

Published in:
Trends in Hearing

Link to article, DOI:
10.1177/2331216518790903

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Signal-to-Noise-Ratio-Aware Dynamic
Range Compression in Hearing Aids

Tobias May¹, Borys Kowalewski¹, and Torsten Dau¹

Abstract
Fast-acting dynamic range compression is a level-dependent amplification scheme which aims to restore audibility for hearing-impaired listeners. However, when being applied to noisy speech at positive signal-to-noise ratios (SNRs), the gain function typically changes rapidly over time as it is driven by the short-term fluctuations of the speech signal. This leads to an amplification of the noise components in the speech gaps, which reduces the output SNR and distorts the acoustic properties of the background noise. An adaptive compression scheme is proposed here which utilizes information about the SNR in different frequency channels to adaptively change the characteristics of the compressor. Specifically, fast-acting compression is applied to speech-dominated time-frequency (T-F) units where the SNR is high, while slow-acting compression is used to effectively linearize the processing for noise-dominated T-F units where the SNR is low. A systematic evaluation of this SNR-aware compression scheme showed that the effective compression of speech components embedded in noise was similar to that of a conventional fast-acting system, whereas natural fluctuations in the background noise were preserved in a similar way as when a slow-acting compressor was applied.

Keywords
wide dynamic range compression, signal-to-noise ratio, hearing-aid signal processing

Introduction
One of the primary tasks of a hearing aid is to improve speech recognition through restored audibility (e.g., Jenstad & Souza, 2007; Souza, Boike, Witherell, & Tremblay, 2007; Souza & Turner, 1999). Wide dynamic range compression (WDRC) provides level-dependent amplification. It is therefore capable of improving the audibility of soft speech components while avoiding excessive amplification of high-intensity inputs and the loudness discomfort that would result from it otherwise (e.g., Alexander & Rallapalli, 2017; Villchur, 1973). WDRC is characterized by a number of parameters, such as the attack and release times, compression ratio (CR), compression threshold (CT), and the number of frequency channels. The attack time is usually very short (below 10 ms) such that the compressor can react to a rapid increase in the intensity of the input signal (Alexander & Rallapalli, 2017; Jenstad & Souza, 2005). A compressor is typically classified as fast-acting, with release times shorter than 200 ms, or slow-acting, with release times longer than 200 ms (for a review, see Souza, 2002).

For a maximum audibility benefit, the compression system must be able to follow changes in the speech amplitude on timescales corresponding to the duration of a syllable or even a phoneme. This requires a very-fast-acting system with a release time below about 60 ms (Edwards, 2004). If a longer release time is used, the gain might lag behind the dynamic changes in the speech envelope, leaving low-intensity components underamplified (Jerlvall & Lindblad, 1978; Kuk, 1996). As demonstrated by Braida et al. (1982) and Stone and Moore (1992), the effective compression ratios (ECRs) decrease to only a fraction of the nominal ratios when the release time is too long compared with the rate of the envelope fluctuations in the signal.

¹Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Lyngby, Denmark

Corresponding author:
Tobias May, Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.
Email: tobmay@elektro.dtu.dk

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
Several studies have demonstrated a benefit of fast-acting compression for speech recognition in quiet (Souza & Turner, 1998, 1999; Villechure, 1973). In contrast, Davies-Venn, Souza, Brennan, & Stecker, (2009) found that when audibility was adjusted with linear versus level-dependent amplification using WDRC, the latter was found to be detrimental for speech recognition. This was probably caused by altered level differences between phonemes, distortions of the temporal envelope, or a reduction of the modulation depth of the speech signal (Alexander & Rallapalli, 2017; Gallun & Souza, 2008; Jenstad & Souza, 2005, 2007; Plomp, 1988; Rosen, 1992; Souza & Gallun, 2010; Souza & Gallun, 1996; Stone & Moore, 2003, 2004, 2007, 2008; van Buuren, Festen, & Houtgast, 1999; Walaszek, 2008). Such distortions are typically more pronounced for shorter release times and higher CRs (Alexander & Rallapalli, 2017; Jenstad & Souza, 2005, 2007).

The relative benefit of WDRC versus linear amplification depends on the acoustic condition. When noise is present, the amount of the effective compression and the distortions of the speech envelope seem to be less pronounced compared with the processing of speech in quiet (Rhebergen, Versfeld, & Dreschler, 2009; Souza, Jenstad, & Boike, 2005). Yund and Buckles (1995) studied the impact of multichannel compression on speech recognition in the presence of a fixed-level stationary background noise and found an increased benefit as the signal-to-noise ratio (SNR) decreased. Moreover, Gatehouse, Naylor, and Elberling (2003, 2006) suggested that if the noise is fluctuating with distinct temporal dips, fast-acting compression would provide differential amplification by applying more gain to the low-intensity glimpses of the speech than to the noise peaks, potentially leading to improved intelligibility. This prediction is consistent with recent results from Rhebergen, Maalderink, and Dreschler (2017) and Desloge, Reed, Braida, Perez, and D’Aquila (2017), who established a link between increased speech audibility and improved speech intelligibility when applying fast-acting compression to speech in the presence of fluctuating background noise. On the contrary, compression can negatively affect the output SNR by reducing the speech level and overamplifying portions of the noise occurring in the speech gaps (Alexander & Masterson, 2014; Hagerman & Olofsson, 2004; Naylor & Johannesson, 2009; Rhebergen et al. 2017; Souza et al, 2006). As recently shown by Rhebergen et al (2017), the reduction of the output SNR can be detrimental to speech recognition. Apart from a reduced output SNR, fast-acting compression of mixed sources (e.g., competing talkers or speech in noise) introduces across-signal modulations. Stone and Moore (2007, 2008) demonstrated that this distortion might be detrimental to speech intelligibility, at least when primarily envelope cues are available. Even if the effect on recognition can be small, other perceptual attributes might be affected, such as the perceived noisiness of the sound (e.g., Kuk, 1996; Neuman, Bakke, Mackersie, Hellman, & Levitt, 1998), leading to a perception of reduced overall quality. Therefore, it has been suggested that the compression parameters should be adjusted according to the environment (Kates, 2010; Yund, Simon, & Efron, 1987) to reach the balance point, at which the positive and negative acoustic effects optimally offset each other (Souza, Hoover, & Gallun, 2012).

The hypothesis of the current study was that an optimal hearing-aid compensation strategy should (a) amplify low-level portions of speech, (b) reduce the dynamic range of speech to avoid excessive loudness, (c) avoid amplifying the noise in speech gaps (so-called pumping), and (d) maintain the natural fluctuations in the background noise. To achieve this, an adaptive amplification scheme would be required that selectively changes the characteristics of the compressor in a given time-frequency (T-F) unit depending on whether speech or noise components are dominating. In earlier approaches, such as the K-amp strategy (Killion, Teder, Johnson, & Hanke, 1992) and the dual front-end automatic gain control system (Moore & Glasberg, 1988; Stone, Moore, Alcántara, & Glasberg, 1999), the release time varied according to how long the compression circuit had been activated, which can help to reduce the pumping artifacts. Similar principles have been applied in guided level estimators (Neumann, 2008; Simonsen & Behrens, 2009). Moreover, Lai, Li, Tsai, Chu, and Young (2013) proposed an adaptive WDRC system that adjusted the CR in individual frequency channels depending on the estimated short-term dynamic range. These systems, however, are only sensitive to changes in the overall signal level but do not utilize information related to the presence of the target signal versus the background noise. In the context of binaural WDRC, an adaptive amplification scheme was proposed by Hassager, May, Winberg, and Dau (2017), where knowledge about the acoustic scene in terms of the direct-to-reverberant energy ratio (DRR) was utilized to selectively apply fast-acting compression only to T-F units that are dominated by the direct sound. This direct sound–driven compression scheme, in conjunction with a binaural link, was demonstrated to improve sound source localization and externalization compared with conventional fast-acting compression (Hassager et al., 2017).

In this study, the idea of such a scene-aware amplification scheme was studied for acoustic scenes where speech and background noise were presented simultaneously. Specifically, an adaptive amplification system was considered that applied fast-acting compression only to speech-dominated T-F units, while the processing
of noise-dominated T-F units was linearized through a longer release time. The resulting amplification scheme, termed SNR-aware dynamic range compression, was compared with conventional fast- and slow-acting compression systems using three objective metrics based on the ECR as well as relative changes in the modulation spectrum and the broadband SNR.

**System**

The block diagram of the SNR-aware dynamic range compression algorithm is shown in Figure 1. First, the input signal was analyzed by a short-time discrete Fourier transform (STFT). In the acoustic scene analysis stage, a binary decision about speech activity was obtained by applying a threshold criterion to the estimated short-term SNRs in individual frequency channels. This decision was then utilized in the dynamic range compression stage to adaptively adjust the release time of the compressor. Specifically, a short release time was selected if a particular T-F unit was dominated by speech (high SNR), whereas a long release time was used for noise-dominated T-F units (low SNR). Then, a gain function was calculated and applied to the STFT representation of the noisy speech signal. Finally, the output signal was reconstructed using the STFT synthesis stage. All of the individual building blocks are described in detail in the following subsections.

**STFT Analysis**

The input signal was sampled at a rate of 20 kHz and segmented into overlapping frames of 10 ms duration with a shift of 2.5 ms. Each frame was Hann-windowed and zero-padded to a length of 512 samples and a 512-point discrete Fourier transform (DFT) was computed, producing an STFT representation of the input signal (Allen, 1977).

**Speech Detection**

Based on the STFT representation of noisy speech, a binary decision about speech activity was performed for each individual T-F unit. Therefore, the speech power spectral density (PSD) was first obtained in individual DFT bins using the minimum mean-square error estimator by Erkelens, Hendriks, Heusdens, and Jensen (2007). This method relies on an estimate of the noise PSD, which was derived from noisy speech using the algorithm proposed by Hendriks, Heusdens, and Jensen (2010). Both the noisy speech power and the estimated speech PSD were then integrated into seven octave-wide bands, by applying the filterbank described below, and subsequently used to estimate the short-term SNR (Eaton, Brookes, & Naylor, 2013; May, Kowalewski, Fereczkowski, & MacDonald, 2017). Finally, speech activity was detected by applying a threshold to the estimated SNRs in individual T-F units. These thresholds were determined by a training procedure described in the Parameters subsection.

**Filterbank**

The dynamic range compressor operated separately in seven octave-wide bands with center frequencies ranging from 125 Hz to 8 kHz. The octave bands were designed to have rectangular filter weights that were applied to each DFT bin. Given the DFT resolution, the effective filter shape of the individual octave bands was as rectangular as possible. For each octave band, the power of the respective DFT bins was integrated and the magnitude of individual T-F units was returned.

---

**Figure 1.** Block diagram of the SNR-aware compressor consisting of three processing layers: (a) STFT-based analysis and synthesis, (b) acoustic scene analysis, and (c) dynamic range compression. See System section for more details regarding the individual processing steps. ISTFT = inverse short-time discrete Fourier transform; SNR = signal-to-noise ratio; STFT = short-time discrete Fourier transform.
Level Estimation

The magnitude of the individual T-F units was smoothed by a first-order infinite impulse response filter with different time constants associated with attack and release. Given the binary decision about speech activity, two different sets of attack and release time constants were defined for speech-dominated and noise-dominated T-F units: (a) a short attack time of 5 ms and a short release time of 40 ms were used for the speech-dominated T-F units with a high SNR, and (b) a short attack time of 5 ms and a long release time of 2,000 ms were used for the noise-dominated T-F units where the SNR was low. In both cases, a short attack time was chosen to maintain the responsiveness of the compressor to rapid intensity changes, irrespective of whether the dominant signal was speech or noise.

Gain Calculation

Given the smoothed level estimation in decibels (dB), a broken-stick gain function was used to derive the respective gains in the individual T-F units. The broken-stick gain function provided a linear gain below the CT and a constant CR above the CT. This gain function was based on the NAL-NL2 prescription (Keidser, Dillon, Flax, Ching, & Brewer, 2011) fitted to the Na standard audiogram corresponding to a flat and moderately sloping hearing loss (Bisgaard, Vlaming, & Dahlquist, 2010) using the settings slow and unilateral. The CTs were derived by measuring the output level of the individual frequency channels in response to stationary speech-shaped noise. The speech-shaped noise had the same long-term average spectrum (LTAS) as the Danish hearing in noise test (HINT) speech material and was normalized to a root mean square–level of 50 dB. The resulting CRs and CTs for the seven octave bands are summarized in Table 1.

Interpolation of Gain Values

The linear gains were interpolated from the channel center frequencies to the DFT frequency axis using a piecewise cubic interpolation to avoid aliasing artifacts. These interpolated gains were subsequently applied to the STFT representation of noisy speech.

STFT Synthesis

After multiplying the gains with the STFT representation of noisy speech, the processed time domain signal was reconstructed by applying an inverse short-time discrete Fourier transform (ISTFT). Specifically, an inverse discrete Fourier transform produced individual time segments that were combined by a weighted overlap-add method (Crochiere, 1980). The weighted overlap-add approach extends the original overlap-add method proposed by Allen (1977) with a synthesis window. A 512-sample tapered cosine window with 39-sample ramps was used as a synthesis window (Grimm, Herzke, Berg, & Hohmann, 2006) to smooth discontinuities at the frame boundaries, which can occur because of temporal aliasing.

Evaluation

Stimuli

Noisy speech was created by mixing clean speech from the Danish HINT (Nielsen & Dau, 2011) with four different types of background noise at seven SNRs (–6, –3, 0, 3, 6, 9, and 12 dB). The following noise types were used: Stationary International Collegium of Rehabilitative Audiology (ICRA)-1 noise and nonstationary ICRA-7 noise representing a six-talker speech babble (Dreschler, Verschuure, Ludvigsen, & Westermann, 2001) as well as car noise and factory noise from the NOISEX database (Varga & Steeneken, 1993). The noise signals were split into two halves of equal size to ensure that there was no overlap between the noise segments used for training the speech detection stage (see Parameters subsection) and evaluation. Following Naylor and Johannesson (2009), the LTAS of all noise types measured in 1/3 octave bands was adjusted to match the LTAS of the Danish HINT corpus.

Each noisy speech mixture consisted of 10 randomly selected HINT sentences from the test lists that were concatenated and mixed with a random noise segment. The noise was normalized to a root mean square–level corresponding to 50 dB while the level of the speech signal was adjusted to yield a predefined SNR. An initial noise-only segment of 250-ms duration was incorporated to ensure that the noise PSD estimator (see Speech Detection subsection) was properly initialized. After processing, this noise-only segment was removed and did not bias the analysis of the objective metrics. For each of the four noise types and seven SNRs, 20 noisy speech mixtures with an average length of 15.5 s were created.

<table>
<thead>
<tr>
<th>Channel center frequency (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CT (dB)</strong></td>
<td>43</td>
<td>43</td>
<td>41</td>
<td>41</td>
<td>37</td>
<td>31</td>
<td>28</td>
</tr>
<tr>
<td><strong>CR</strong></td>
<td>2.2:1</td>
<td>2.2:1</td>
<td>2.2:1</td>
<td>3.0:1</td>
<td>3.5:1</td>
<td>3.3:1</td>
<td>2.5:1</td>
</tr>
</tbody>
</table>

Note. CT = compression threshold; CR = compression ratio.
The ECR was calculated based on the estimated range of input SNRs: following three objective metrics were computed for noise, speech alone, and noisy speech (in the STFT domain). The mixture and then subsequently applied to speech alone, noise, and noisy speech separately. The compression was always estimated based on the noisy speech signal, while inherent fluctuations in the noise-only segments also result in fast changes in the gain function. In contrast, the slow-acting system is able to follow rapid intensity changes of the speech signal, the respective gain functions are shown for a channel center frequency of 2 kHz. The fast-acting system is able to follow rapid intensity changes of the speech signal, while inherent fluctuations in the noise-only segments also result in fast changes in the gain function. In contrast, the slow-acting system only responds to strong onsets and only slowly recovers following the offset of the dominant signal (speech, in this case). Because of the prolonged recovery, the gain remains relatively low after higher intensity segments, leaving other low-level speech components underamplified. The SNR-aware system adaptively switches between fast and slow processing depending on the estimated speech activity. Thus, in speech-active time segments,

**Parameters**

The binary decision of speech activity was obtained by thresholding the estimated SNRs in individual T-F units (see Speech Detection subsection). These thresholds were found by maximizing the hit rate minus false alarm rate (H − FA) between the estimated and the true speech activity using a small training set. For this purpose, 10 randomly selected HINT sentences from the training lists were mixed with ICRA-1 and ICRA-7 noise at −5, 0, and 5 dB SNR, producing a training set of $10 \times 2 \times 3 = 60$ noisy speech mixtures. The true speech activity was obtained by applying a threshold criterion of 0 dB to the $a$ priori SNR, which was calculated from the individual speech and noise signals.

The noise PSD estimator by Hendriks et al. (2010) was used with the default parameter set and initialized for each noisy speech mixture by averaging the PSD across the initial noise-only segment of 250 ms. The speech PSD estimator from Erkelens et al. (2007) was configured with the two generalized gamma parameters $\gamma = 1$ and $\nu = 0.6$. Moreover, the smoothing factor $\alpha$ employed by the decision-directed approach corresponded to a time constant of 0.792 s.

**Objective Metrics**

Shadow-filtering (Fredelake, Holube, Schlueter, & Hansen, 2012; Gustafsson, Martin, & Vary, 1996) was employed to investigate the impact of compression on speech, noise, and noisy speech. The compression was always estimated based on the noisy speech signal, while inherent fluctuations in the noise-only segments also result in fast changes in the gain function. In contrast, the slow-acting system only responds to strong onsets and only slowly recovers following the offset of the dominant signal (speech, in this case). Because of the prolonged recovery, the gain remains relatively low after higher intensity segments, leaving other low-level speech components underamplified. The SNR-aware system adaptively switches between fast and slow processing depending on the estimated speech activity. Thus, in speech-active time segments,

The noise PSD estimator by Hendriks et al. (2010) was used with the default parameter set and initialized for each noisy speech mixture by averaging the PSD across the initial noise-only segment of 250 ms. The speech PSD estimator from Erkelens et al. (2007) was configured with the two generalized gamma parameters $\gamma = 1$ and $\nu = 0.6$. Moreover, the smoothing factor $\alpha$ employed by the decision-directed approach corresponded to a time constant of 0.792 s.

**Objective Metrics**

Shadow-filtering (Fredelake, Holube, Schlueter, & Hansen, 2012; Gustafsson, Martin, & Vary, 1996) was employed to investigate the impact of compression on speech, noise, and noisy speech separately. The compression was always estimated based on the noisy speech signal, while inherent fluctuations in the noise-only segments also result in fast changes in the gain function. In contrast, the slow-acting system only responds to strong onsets and only slowly recovers following the offset of the dominant signal (speech, in this case). Because of the prolonged recovery, the gain remains relatively low after higher intensity segments, leaving other low-level speech components underamplified. The SNR-aware system adaptively switches between fast and slow processing depending on the estimated speech activity. Thus, in speech-active time segments,

**Compression Systems**

The following four compression systems were evaluated which all operated in seven octave bands: fast-acting, slow-acting, SNR-aware compression as well as ideal SNR-aware compression based on the $a$ priori SNR. An overview of the respective parameters is given in Table 2. While the conventional fast- and slow-acting compression systems were characterized by the attack and release times, the SNR-aware approach adaptively switched between two sets of attack and release times for speech- and noise-dominated T-F units. The ideal SNR-aware compression system used the true speech activity based on the $a$ priori SNR (see Parameters subsection), rather than the speech activity estimator described in the Speech Detection subsection.

The processing principle of the four different compression schemes is illustrated in Figure 2 for a speech signal mixed with ICRA-1 noise at 6 dB SNR. Given the noisy speech signal, the respective gain functions are shown for a channel center frequency of 2 kHz. The fast-acting system is able to follow rapid intensity changes of the noisy speech signal, while inherent fluctuations in the noise-only segments also result in fast changes in the gain function. In contrast, the slow-acting system only responds to strong onsets and only slowly recovers following the offset of the dominant signal (speech, in this case). Because of the prolonged recovery, the gain remains relatively low after higher intensity segments, leaving other low-level speech components underamplified. The SNR-aware system adaptively switches between fast and slow processing depending on the estimated speech activity. Thus, in speech-active time segments,

<table>
<thead>
<tr>
<th>Compressor</th>
<th>Attack (ms)</th>
<th>Release (ms)</th>
<th>Speech detection</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast</td>
<td>5</td>
<td>40</td>
<td>Off</td>
<td>–</td>
</tr>
<tr>
<td>Slow</td>
<td>5</td>
<td>2,000</td>
<td>Off</td>
<td>–</td>
</tr>
<tr>
<td>SNR-aware</td>
<td>5/5</td>
<td>40/2,000</td>
<td>On</td>
<td>Estimated SNR</td>
</tr>
<tr>
<td>SNR-aware ideal</td>
<td>5/5</td>
<td>40/2,000</td>
<td>On</td>
<td>$a$ priori SNR</td>
</tr>
</tbody>
</table>

Note. SNR = signal-to-noise ratio.
the SNR-aware system is able to follow rapid intensity changes caused by the short release time, while the use of a long release time for noise-dominant time segments effectively linearizes the processing, which avoids rapid fluctuations in the gain in response to noise-only segments.

**Results**

The ECRs are shown in Figure 3 as a function of the input SNR and the channel center frequency. Each of the four rows represents a different compression scheme, that is, fast-acting (first row), slow-acting (second row), SNR-aware (third row), and ideal SNR-aware compression (fourth row). The left, middle, and right columns show results for the three different signal categories, that is, shadow-filtered speech, shadow-filtered noise, and noisy speech.

As expected, the fast-acting compression system provided the highest ECRs for all three signal categories. For noisy speech (right column), a maximum ECR of up to 2.0 was measured for high frequencies. When using shadow-filtering to analyze the impact of compression on speech and noise separately (left and middle columns), it can be seen that both speech and noise components were compressed, with ECRs of up to 1.6 and 1.3, respectively. The slow-acting compression system did not compress the noise components (with ECRs of 1 and lower) and also provided no compression to the speech components, where the ECR was 1.1 for the entire range of input SNRs. The ECRs of the SNR-aware compressor for the speech components were in a similar range (up to 1.4) as for the fast-acting compressor, while the ECR associated with the noise components was close to 1 (±0.1) for a wide range of input SNRs. Finally, the ECR contours of the SNR-aware and the ideal SNR-aware compressor were very similar to each other for all three signal categories.

Figure 4 shows the relative change in the modulation spectrum (ΔMS) as a function of modulation frequency (ranging from 0.5 to 32 Hz) and the input SNR. Negative values indicate a reduction in modulation depth, while positive values reflect an increase in modulation depth caused by the level-dependent amplification (compression). Again, the four rows represent the different compression schemes (fast-acting, slow-acting, SNR-aware, and ideal SNR-aware compression) and the three columns show results for shadow-filtered speech, shadow-filtered noise, and the noisy speech mixture, respectively.
Fast-acting compression reduced the modulation depth of the shadow-filtered speech signal for modulation frequencies between 0.5 and 8 Hz and this effect increased with increasing SNR. At the same time, the modulation depth of the shadow-filtered noise signal was enhanced with a clear peak around 4 Hz for higher input SNRs. Slow compression did not markedly affect the modulation spectra of the shadow-filtered speech and noise signals. While ΔMS was positive in the range between 0.5 and 8 Hz for the shadow-filtered noise, the individual functions obtained for the different SNRs were fairly flat and did not show any pronounced peak. This coincided with a decreased ECR as already observed in Figure 3. Both SNR-aware systems resembled the conventional fast-acting compressor in terms of ΔMS for shadow-filtered speech. Although modulations around 4 Hz were to some extent enhanced in the shadow-filtered noise, the individual functions were much flatter compared with the fast-acting system and the respective magnitudes were closer to those obtained with the slow-acting compression system.

Finally, the input/output SNR analysis for the four compression schemes and a linear reference condition (dashed line) is shown in Figure 5. All tested compression systems led to a reduction in the output SNR, which was most pronounced at higher input SNRs. The fast-acting compressor reduced the output SNR by up to 4.8 dB, while the slow-acting system was closest to the linear reference condition. The SNR-aware compressor produced a consistently higher output SNR than the fast-acting system over the complete range of input SNRs. This benefit was about 2 dB at higher input SNRs and was very similar for the SNR-aware and the ideal SNR-aware compressors.

In general, the objective metrics computed for the SNR-aware and the ideal SNR-aware compressor were very similar, suggesting that the accuracy of the SNR estimator was sufficiently high. The performance of the speech detection algorithm is summarized in Table 3 in terms of the hit rate (H), the false alarm rate (FA), and the $H/C_0$ for different frequency channels. While the $H/FA$ was not higher than 34.7 % for the lowest two frequency channels, performance increased up to 59.0 % at higher center frequencies.
The analysis of ΔMS indicated that distortions of the speech components are an inevitable consequence of fast-acting compression. A rapidly changing gain function reduces the temporal contrasts of the speech components which, in turn, reduces the modulation power. This is also reflected in the ECRs, which are highest for the fast-acting compression scheme. As pointed out by Villchur (1989), the reduction in modulation power is not necessarily detrimental, as long as it coincides with an improvement in speech audibility. At the same time, fast-acting compression increases the modulation depth of noise signal components at positive SNRs. As shown in Figure 4, the largest increase was

![Figure 4](image)

**Figure 4.** Relative change in modulation spectra (ΔMS) caused by fast-acting (first row), slow-acting (second row), SNR-aware (third row), and ideal SNR-aware compression (fourth row) as a function of the modulation frequency and the input SNR. Results were averaged across all four noise types. The black dashed line indicates the zero line while the left, middle, and right columns show results for shadow-filtered speech, shadow-filtered noise, and noisy speech, respectively. SNR = signal-to-noise ratio.

![Figure 5](image)

**Figure 5.** Input/output SNR analysis for the four different compression schemes and a linear system averaged across all four noise types. SNR = signal-to-noise ratio.

**Table 3.** Performance Analysis of the Binary Speech Detection Algorithm in Terms of H, FA, and H − FA in Percentage as a Function of the Channel Center Frequency Averaged Across All Noise Types and SNRs.

<table>
<thead>
<tr>
<th>Channel center frequency (Hz)</th>
<th>125</th>
<th>250</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>8000</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>53.1</td>
<td>55.2</td>
<td>67.5</td>
<td>72.0</td>
<td>74.0</td>
<td>73.2</td>
<td>81.3</td>
</tr>
<tr>
<td>FA</td>
<td>18.9</td>
<td>20.5</td>
<td>13.5</td>
<td>15.9</td>
<td>18.3</td>
<td>21.8</td>
<td>22.3</td>
</tr>
<tr>
<td>H − FA</td>
<td>34.2</td>
<td>34.7</td>
<td>54.0</td>
<td>56.1</td>
<td>55.7</td>
<td>51.4</td>
<td>59.0</td>
</tr>
</tbody>
</table>

Note. H = hit rate; FA = false alarm rate; H − FA = hit rate minus false alarm rate; SNR = signal-to-noise ratio.
found around the 4-Hz region, which corresponds to the

typical maximum in the speech modulation spectrum
(e.g., Plomp, 1983; Souza & Gallun, 2010). This results

from the compressor gain following short-term fluctua-
tions in the intensity of the dominating speech signal,

which disrupts the natural fluctuations in the back-
ground noise. As a consequence, the glimpses of noise

that are cyclically amplified because of the increased gain
during the speech pauses may lead to a sensation of

pumping and increased overall noisiness (Neuman

et al., 1998). Such processing thus is likely to decrease the

SNR in the modulation domain, which has been

proposed to be detrimental for speech intelligibility

(Jørgensen & Dau, 2011; Jørgensen, Ewert, & Dau,

2013). Furthermore, the long-term level of the noise is

increased at the output of the compressor, causing a

reduced output SNR (Naylor & Johannesson, 2009). In

contrast, slow-acting compression avoids the amplifica-
tion of the noise components. As shown in Figure 2, the

changes in the gain function of the slow-acting system
do not follow the fluctuations of speech very closely.

Therefore, distortions in the modulation spectrum of
the noise components, as shown in Figure 4, are of

much smaller magnitude. This leads to a more

linear behavior in terms of the input/output SNR

analysis. However, a slow-acting system does not pro-

vide any substantial compression to the speech signal

components.

The SNR-aware compression scheme appears to com-
bine the desired properties of the two conventional sys-
tems. The analysis of the ECR suggests that the effective

compression of speech embedded in noise, as provided

by the SNR-aware system, is very similar to the one

obtained with conventional fast-acting compression.

This behavior should be advantageous, as it is linked

to improved audibility (Alexander & Rallapalli, 2017).

At the same time, the fluctuations in the gain function
become much slower when speech is absent, which

avoids the amplification of noise-only segments and

increases the output SNR relative to that obtained with

fast-acting compression. This is also reflected in the

ECRs associated with the noise components, which clo-

sely resemble the behavior of the slow-acting compres-

sor. Thus, the SNR-aware compression scheme

maintains the acoustic properties of the background

noise similar to slow-acting compression while applying

fast-acting compression to the speech signal components.

Preserving the modulation fidelity of the background

noise may facilitate the target-background segregation,

improve the perceived quality of the acoustic scene, and

aid speech recognition in adverse conditions.

The SNR-aware compression scheme utilizes an esti-
mation of the short-term SNR to detect speech-
dominated T-F units. The estimation accuracy of this

speech detection stage, as reflected by the H – FA,

was as high as 59% and generally in a similar range as

the speech detector used in the DRR–aware compression

scheme (Hassager et al., 2017). Instead of using the

output of the speech detection stage directly for noise

reduction, the binary classification of speech activity

was used to adaptively select different time constants

for speech and noise components. Thus, estimation

errors in the speech detection stage do not introduce

clearly audible artifacts, and only limit the effective com-

pression of speech components. In a binaural setup with

two hearing aids, the estimation of speech activity could

be further improved by spatial cues (May, van de Par, &

Kohlrausch, 2011), which would allow the application of

fast-acting compression to speech-dominated T-F units

corresponding to a target source at a specific spatial

location.

Conclusion

This study presented a scene-aware amplification strat-

ey that adaptively changes the characteristics of the

compressor depending on the estimated speech activity

in individual T-F units. Specifically, fast-acting compres-

sion was applied to speech-dominated T-F units where

the SNR was high, while slow-acting compression was

performed for noise-dominated T-F units with a low

SNR. A systematic analysis using three technical metrics

showed that this SNR-aware compression scheme

achieved similar ECRs compared with conventional

fast-acting compression, while the natural fluctuations

in the background noise were preserved in a similar

way as processing the noise components with a conven-
tional slow-acting system. Future work will quantify the

subjective benefit of the SNR-aware compression scheme

by performing behavioral listening tests.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with

respect to the research, authorship, and/or publication of this

article.

Funding

The authors disclosed receipt of the following financial support

for the research, authorship, and/or publication of this article:

This research was supported by the Technical University of

Denmark and funding from Sonova AG (Stäfa, Switzerland).

Note

1. The speech detection performance of the first two frequency

channels was relatively poor, probably because of the high

temporal resolution of 10 ms which limited the frequency

resolution and caused the first two octave filters to be

based on only up to five discrete Fourier transform bins.

As a consequence, the estimated signal-to-noise ratio was

less reliable, which limited the speech detection
performance. Thus, the signal-to-noise ratio estimate of the third channel (500 Hz) was used for the first two channels, for which nevertheless individual thresholds were found as described in the Parameters subsection.

References


