INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Effectiveness of near-end speech enhancement under equal-loudness and equal-level constraints

Tudor-Cătălin Zorilă[1], *Sheila Flanagan*[2], *Brian C.J. Moore*[2], *Yannis Stylianou*[1,3]

[1]Toshiba Cambridge Research Laboratory, United Kingdom
[2]Department of Experimental Psychology, University of Cambridge, United Kingdom
[3]Computer Science Department, University of Crete, Heraklion, Greece
{catalin.zorila,yannis.stylianou}@crl.toshiba.co.uk, {saf31,bcjm}@cam.ac.uk

Abstract

Most recently proposed near-end speech enhancement methods have been evaluated with the overall power (RMS) of the speech held constant. While significant intelligibility gains have been reported in various noisy conditions, an equal-RMS constraint may lead to enhancement solutions that increase the loudness of the original speech. Comparable effects might be produced simply by increasing the power of the original speech, which also leads to an increase in loudness. Here we suggest modifying the equal-RMS constraint to one of equal loudness between the original and the modified signals, based on a loudness model for time-varying sounds. Four state-of-the-art speech-in-noise intelligibility enhancement systems were evaluated under the equal-loudness constraint, using intelligibility tests with normal-hearing listeners. Results were compared with those obtained under the equal-RMS constraint. The methods based on spectral shaping and dynamic range compression yielded significant intelligibility gains regardless of the constraint, while for the method without dynamic range compression the intelligibility gain was lower under the equal-loudness than under the equal-RMS constraint.

Index Terms: near-end listening enhancement, equal-loudness constraint, dynamic range compression

1. Introduction

The intelligibility of speech worsens in noisy environments (e.g., train stations, airports, sports arenas) and this can pose a serious communication barrier. Recent research efforts have addressed this issue by exploring the effects of modifications of the clean speech signals (before they are mixed with noise) with the overall RMS power of the signal held constant [1, 2]. This problem is known as near-end listening enhancement or speech intelligibility enhancement (SINE) and the constraint is called the equal-RMS (EQR) constraint. Typical solutions use modifications that are either based on previous intelligibility studies (e.g., simulating the differences between Lombard speech and conversational speech or emphasis of the information-bearing segments of speech) [3–7] or they optimize an objective measure that correlates well with measured intelligibility [8–11]. 'Inclusive' algorithms that would work for both normal-hearing and hearing-impaired listeners or in reverberant conditions are becoming increasingly popular [12, 13].

Most previous evaluations of SINE algorithms have used the EQR constraint. This approach is relatively simple and allows an easy comparison between different algorithms. However, it has the disadvantage that it does not take into account the auditory perception of listeners. Different processing applied to the same speech signal under the EQR constraint may produce markedly different loudness values for each of the modifications, as shown by [14]. Loudness is the perceptual attribute of sounds in terms of which they can be ordered from quiet to loud [15]. In other words, loudness is the subjective impression of the magnitude of sounds. Usually, the SINE modifications result in an increase in loudness, but, in practice, the loudness should be kept within an acceptable range. In particular, it is necessary to avoid excessive loudness for listeners who are close to one or more of the loudspeakers in a public address system. Therefore, it may be more appropriate to evaluate SINE algorithms using an **equal-loudness** (EQL) constraint, whereby the loudness of the speech is the same before and after enhancement processing. Here four state-of-the-art SINE algorithms were evaluated under the EQL constraint, using normal-hearing listeners. The results were compared with similar results obtained using the EQR constraint. To the best of our knowledge, this is the first work assessing the effectiveness of SINE algorithms with the EQL constraint.

An EQL constraint may be also better suited to hearing-impaired listeners than the EQR constraint. It is estimated that 10-15% of the total population worldwide suffers from some form of hearing impairment, a percentage that is expected to increase as the average age of the population increases. Loudness recruitment is a side-effect frequently accompanying the loss of hearing [16]. While sensitivity is reduced for low-level sounds, high-level sounds may be perceived with normal loudness or even with greater-than-normal loudness (hyperacusis), making loudness control very important.

Modeling, measuring and controlling the loudness of sounds has major applications in audio and speech synthesis, broadcasting, hearing instruments and noise control. Mapping sound intensity (physical magnitude) to loudness is a non-trivial cross-disciplinary topic [17–22]. Equating loudness requires both an accurate loudness predictor (model) and a procedure for adjusting the signal level without introducing artifacts [23–28]. Here, we used the loudness model of Glasberg and Moore [22], which has been shown to give accurate predictions of loudness for unprocessed speech and for speech processed using SINE algorithms [14].

The rest of the paper is organized as follows. Section 2 briefly describes the SINE algorithms to be evaluated using the EQL constraint and summarizes the procedure used to equate loudness, Section 3 presents the evaluation methodology and the results, and Section 4 gives a comparison of the present results with previous results obtained using the EQR constraint. Finally, conclusions are presented in Section 5.

http://dx.doi.org/10.21437/Interspeech.2016-594

2. Methods

2.1. Intelligibility enhancement algorithms for evaluation

Four algorithms were selected for evaluation under the EQL constraint. They were chosen based on their high intelligibility gains under the EQR constraint, as reported in earlier studies [1,2,29]. The first three were also jointly evaluated by Zorilă and Stylianou [29], which facilitated a comparison of intelligibility gains under the EQR and EQL constraints. All processing was done using a 16-kHz sample rate and 16-bit resolution.

The first algorithm (SSDRC) was based on the work of Zorilă et al. [4, 7] and used a two-stage energy reallocation strategy. During the first stage (spectral shaping - SS), the speech-to-noise ratio (SNR) at medium and high frequencies was increased by transferring energy from below 500 Hz to higher frequencies. This was implemented by flattening the spectral tilt, sharpening the formants and boosting the mid range (1-4 kHz) energy, the first two operations depending on the voicing nature of the current frame. The second stage applied dynamic range compression (DRC) to the output of the first stage, so as to amplify segments of speech that are more prone to noise masking (fricatives, nasals, and stops), at the expense of reducing the level of segments with higher energy (mostly vowels). As a result of the application of DRC, the overall variations of the time envelope were reduced, as shown in Fig. 1.

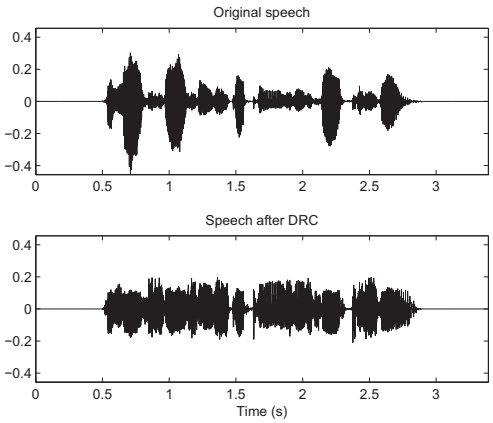


Figure 1: Example of speech before and after application of DRC [4].

The second algorithm (tSER) was suggested by Takou et al. [5] and consisted of a three-stage spectral energy transfer. In one stage, the components below 400 Hz were isolated by lowpass filtering and were passed on unprocessed for combination with the signals from the other stages. In a second stage, the signal was pre-emphasized with a first-order finite impulse response filter that flattened the spectral tilt. The third stage took its input from the second stage and applied a spectral contrast enhancement algorithm resembling the two-tone suppression that occurs in the cochlea [30]. The outputs of these stages were combined after weighting of their magnitudes. No further energy reallocation over time was performed.

The third algorithm (fSERDRC) was based on a more computationally efficient implementation of tSER (denoted fSER) that was combined with the DRC stage of SSDRC [29]. It was shown that SSDRC, tSER and fSERDRC yield similar intelligibility enhancements under the EQR constraint.

The fourth algorithm (SDR) applied modifications derived from an optimization criterion designed to recover the spectral dynamics of speech [10]. The input speech was split into 14 Mel-frequency bands and the power in each band was altered according to mapping functions learned from both speech and noise statistics. The method was shown to considerably improve the recognition rate of speech presented in various stationary noise maskers, outperforming a similar state-of-the-art optimization-based reference system.

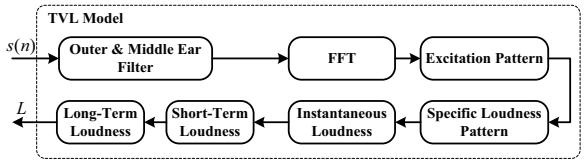


Figure 2: Block diagram of the TVL model [22].

2.2. Loudness normalization of stimuli

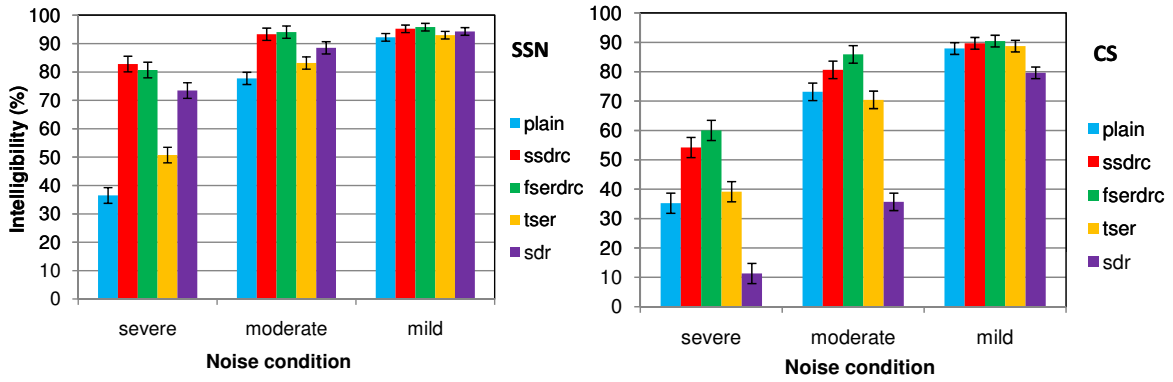
The loudness model of Glasberg and Moore [22] for time-varying sounds (TVL) was used to predict the loudness of the unprocessed and processed speech stimuli (Fig. 2). Firstly, the input signal was passed though a filter that simulates the transfer through the outer and middle ear. Then a multi-resolution spectral analysis was performed on the output, resembling the processing that occurs in the cochlea, and this was used to calculate the excitation pattern evoked by the sound at 1-ms intervals. The excitation pattern simulates the spectral representation of a sound in the cochlea [21]. Specific loudness patterns were calculated by application of a compressive nonlinearity to the excitation pattern [21]. The specific loudness is a kind of loudness density. Instantaneous loudness (not available for conscious perception) was computed by summing the specific loudness values across frequency. The short-term loudness, which represents the loudness of a segment of a sound such as a word in speech or a note in music, was calculated by temporal averaging of the instantaneous loudness values, using a form of averaging representing an automatic gain control, with fast attack and slower release. The long-term loudness (LTL) represents the overall loudness impression of a relatively long sample of the sound (e.g. a sentence), and was calculated from the short-term loudness using an averager with longer time constants. The LTL was used for this work.

Note that all parameters of the model were fixed at standard values, except for the release time of the averager used to calculate the LTL. The exact value of the release time had little effect on the adjustments required to equate the LTL across stimuli. Note also that the model does not take into account the phases of the components, which have a small effect on loudness [31, 32]. However, only one of the enhancement methods used here (tSER) resulted in changes in component phases.

Equating the loudness of stimuli was done as described in [14]. Entire sentences were iteratively rescaled in level until the absolute differences of their peak LTL values were below 0.01 sones. This simple approach has the advantage of not introducing artifacts in the output signal, and was shown to perform well. The TVL model with shorter release times for computing the LTL was used here, as described in [14].

3. Evaluation & Results

The methodology for evaluating the intelligibility benefits of the selected algorithms under the EQL constraint followed the same general guidelines as those used for the Hurricane Chal-



(a) Speech shaped noise.

(b) Competing Speaker type of noise.

Figure 3: Percentage of correctly recognized keywords for the intelligibility assessment under the EQL constraint. The error bars show Fisher's least significant differences (FLSD).

lenge (HC) [1]. The speech signals consisted of the first 30 Harvard sets (300 sentences in total), while the noises were both of stationary (speech-shaped noise - SSN) and non-stationary nature (competing speaker - CS). The Harvard sentences were spoken by a native British English male, and the competing speaker was a woman reading news and Harvard-like sentences. The SSN was generated by passing white noise through a 100-th order finite-impulse filter whose frequency response matched the long-term average spectrum of the CS.

A different approach from the one used for HC was employed to mix the speech and noise samples (Fig. 4). The noise samples were normalized to have an RMS level of -27 dB re full scale. Then, the level of the unprocessed speech was scaled to give three specific SNRs for each noise type. Next, the scaled speech was processed using the SINE algorithms described in Section 2.1 and then equal processed signal was scaled to meet the EQL constraint as described in Section 2.2. The starred branch in Fig. 4 indicates that the target peak LTL value used for loudness normalization came from the unprocessed speech (also denoted as 'plain'). Finally, the loudness-matched enhanced speech and noise signals were added together. The input SNRs were 2 dB larger than the ones used for the HC, i.e. -7, -2 and 3 dB and -19, -12, -5 dB for the SSN and CS maskers, respectively. That was done because we expected a drop in the physical SNR following the loudness equalization. The previous SNRs (for each noise type) were denoted as 'severe', 'moderate' and 'mild', respectively.

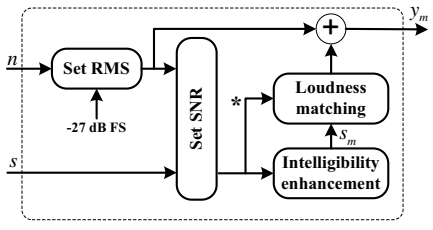


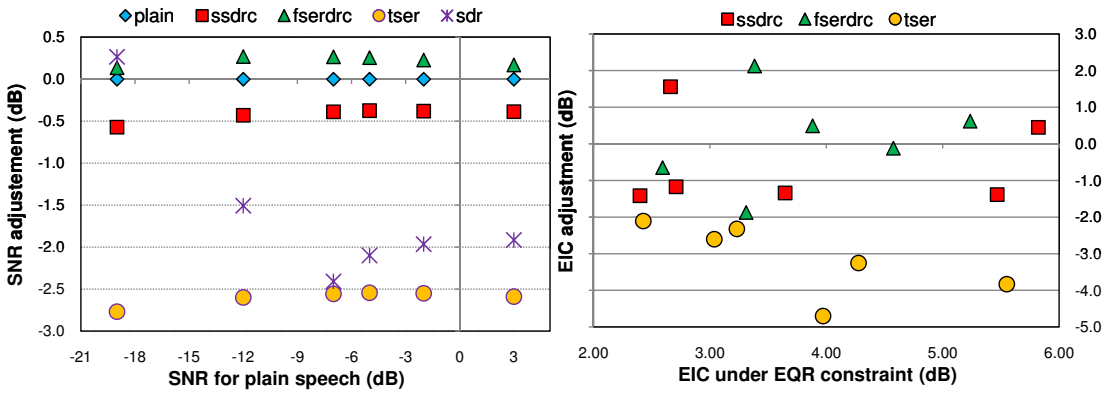
Figure 4: Diagram of method of scaling level to meet the equal-loudness constraint.

Twenty subjects took part in the evaluation, all having normal audiograms for all audiometric frequencies from 0.25 to 8 kHz. The listening test was conducted in a sound-proof room at the Department of Experimental Psychology, University of Cambridge, UK. Stimuli were presented via Sennheiser HD580 headphones at an equivalent input level of 65 dB SPL. Each sentence was presented no more than once and subjects were asked to type what they heard. The test was self-paced, controlled via a Matlab graphical interface, and took roughly one hour to complete.

Fig. 3 shows the percentage of correctly recognized keywords averaged across all subjects for each SNR, noise type and processing condition. Repeated-measures analysis of variance (ANOVA) was performed on the arcsine-transformed scores with factors processing condition (5), noise type (2) and SNR (3). There were significant main effects of processing condition $F(4, 76) = 159.0, p < 0.0001$, noise type $F(1, 19) = 235.7, p < 0.0001$ and SNR $F(2, 38) = 807.8, p < 0.0001$. Only SSDRC and fSERDRC yielded significant higher scores than for plain speech ($p < 0.05$) for both noise types. Intelligibility gains were larger with the stationary masker than with the fluctuating masker, and the intelligibility gains were greatest for the lowest (severe) SNR. SDR yielded intelligibility gains with the SSN but led to reduced intelligibility with the CS.

4. Discussion

The results showed significant intelligibility gains for speech processed by either SSDRC or fSERDRC, but little or no gain for tSER. Since it was previously established that fSER and tSER yield similar recognition rates under the EQR constraint [29], the drop in performance of tSER under the EQL constraint can be attributed to the lack of a dynamic range compression stage. A more in-depth view of the effects of holding the loudness constant can be obtained by plotting the average SNR adjustments needed to obtain equal loudness (Fig. 5a). The adjustment is relative to the EQR case. It can be seen that, on average, the level of SSDRC samples was reduced by roughly 0.5 dB, and was slightly increased for fSERDRC, explaining the similar intelligibility gains for the two, as shown in Fig. 3, but also the slightly better performance for the latter with the CS background. The level of tSER-processed speech was decreased by more than 2.5 dB with the EQL constraint (explaining the lower intelligibility scores), thus indicating a much severe need to control the loudness of speech processed by this system. SDR processing yielded significant intelligibility gains with the SSN, despite an average level reduction of 2 dB. However, SDR processing decreased intelligibility with



(a) Mean SNR adjustment resulting from equating loudness.

(b) EIC change resulting from equating loudness (both SSN and CS noise types).

Figure 5: Impact of applying the EQL constraint on the SNR and on the intelligibility gains obtained using the EQR constraint.

the CS. This result is not surprising since SDR was originally designed for stationary maskers. The results for SDR processing are consistent with other studies showing low performance of optimization-based algorithms for speech presented in fluctuating noise [1].

Although the average SNR adjustments in Fig. 5a may indicate that there is no meaningful difference between the EQR and EQL constraints when the processing chain includes DRC, there are cases when the adjustments were larger. The box-and-whisker chart of SNR corrections for EQL across the whole evaluation set reveals such examples (Fig. 6); some sentences required scaling by more than $\pm$ 4 dB.

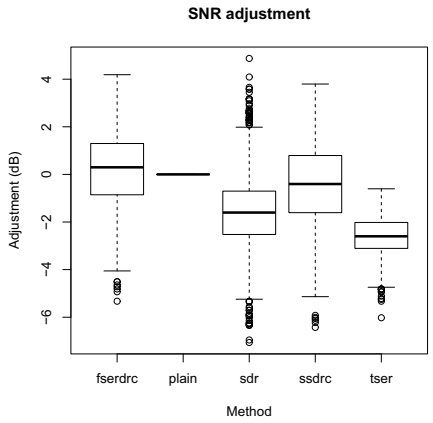


Figure 6: Box-and-whisker chart of SNR adjustments averaged across all noise conditions and across all sentences. The lower and upper whiskers show the minimum and maximum values of the adjustments, the lower and upper box boundaries show the first and third quantiles, and the thick lines show median values.

Figure 5b shows the relationship between the intelligibility gains yielded under the two constraints for the first three algorithms. EIC stands for equivalent intensity change, and is a metric indicating how much the unprocessed (plain) speech would have to be amplified (positive EIC) or attenuated (negative EIC) to achieve the intelligibility scores of the SINE systems tested here. EICs were used to rank the systems participating in the HC [1].

The same logistic mappings of intelligibility and SNR as described in [1] were used to compute EIC values for the results presented here. In Fig. 5b, the abscissa represents the EICs data obtained using the EQR constraint [29], while the ordinate shows the EIC change for a given type of processing resulting from application of the EQL constraint as opposed to the EQR constraint. As expected, the resulting values for all noise types and conditions correlated well with the SNR adjustments shown in Fig. 5a. On average, the EIC for SSDRC dropped by 0.6 dB, the EIC for fSERDRC increased by 0.1 dB, and the EIC for tSER dropped by 3 dB. No data obtained under the EQR constraint were available to assess SDR in a similar way. Comparing the EIC difference for plain speech during the test here and the one obtained with the EQR constraint, an average of 1.9 dB was obtained, which is reasonably close to the expected 2 dB value.

Overall, the benefits obtained using SINE processing with DRC did not differ significantly under the EQL constraint and the EQR constraint. However, significantly reduced intelligibility gains (equivalent to an EIC of 3 dB) were obtained under the EQL constraint with the algorithm that did not include DRC. A further evaluation is planned to assess the effectiveness of the SINE algorithms with hearing-impaired listeners under the EQL constraint.

5. Conclusions

Four modern speech-in-noise near-end intelligibility enhancement algorithms were evaluated by replacing the typical equal-RMS (EQR) constraint with an equal-loudness (EQL) constraint. The latter can serve to prevent the algorithms from leading to an excessive loudness for some listeners. The results showed that the intelligibility gains for the algorithms employing dynamic range compression (DRC) did not differ significantly under the EQR and EQL constraints. However, the performance of an algorithm without DRC was worse under the EQL than under the EQR constraint by an amount equivalent to a change in signal-to-noise ratio of more than 3 dB.

6. Acknowledgements

The authors would like to thank Petko Petkov for providing the SDR stimuli used for the evaluation.

7. References

[1] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Commun.*, vol. 55, pp. 572–585, 2013.

[2] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.

[3] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by highpass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 4, pp. 277–282, 1976.

[4] C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012, pp. 635–638.

[5] R. Takou, N. Seiyama, and A. Imai, "Improvement of speech intelligibility by reallocation of spectral energy," in *Proc. Interspeech*, 2013, pp. 3605–3607.

[6] E. Jokinen, M. Takanen, M. Vainio, and P. Alku, "An adaptive post-filtering method producing an artificial lombard-like effect for intelligibility enhancement of narrowband telephone speech," *Comput. Speech Lang.*, vol. 28, pp. 619–628, 2014.

[7] C. Zorilă, Y. Stylianou, T. Ishihara, and M. Akamine, "Near and far field speech-in-noise intelligibility improvements based on a time-frequency energy reallocation approach," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2016, in press.

[8] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommunikation*, 2010.

[9] C. Valentini-Botinhaoa, J. Yamagishia, S. King, and R. Maia, "Intelligibility enhancement of hmm-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 665–686, 2014.

[10] P. Petkov and W. Kleijn, "Spectral dynamics recovery for enhanced speech intelligibility in noise," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 327–338, 2015.

[11] H. Schepker and J. Rennies, "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 138, no. 5, pp. 2692–2706, 2015.

[12] A. Jemaa, N. Mechergui, G. Courtois, A. Mudry, S. Djaziri-Larbi, M. Turki, H. Lissek, and M. Jaidane, "Intelligibility enhancement of vocal announcements for public address systems: a design for all through a presbycusis pre-compensation filter," in *Proc. Interspeech*, 2015, pp. 70–74.

[13] J. Rennies, A. Volgenandt, H. Schepker, and S. Doclo, "Model-based adaptive pre-processing of speech for enhanced intelligibility in noise and reverberation," in *Proc. Interspeech*, 2015, pp. 2619–2620.

[14] C. Zorilă, Y. Stylianou, S. Flanagan, and B.C.J. Moore, "Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing," *J. Acoust. Soc. Am. Express Letters*, 2016, in press.

[15] B.C.J. Moore, *An introduction to the psychology of hearing (6th Ed.)*. Leiden: Brill, 2013.

[16] ——, *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues (2nd Ed.)*. Chichester: Wiley, 2007.

[17] J. Steinberg, "The loudness of a sound and its physical stimulus," *Phys. Rev.*, no. 26, pp. 507–523, 1925.

[18] H. Fletcher and W. Munson, "Loudness, its definition, measurement and calculation," *J. Acoust. Soc. Am.*, vol. 5, pp. 82–108, 1933.

[19] ——, "Relation between loudness and masking," *J. Acoust. Soc. Am.*, vol. 9, no. 1, pp. 1–10, 1937.

[20] E. Zwicker and B. Scharf, "A model of loudness summation," *Psychol. Rev.*, vol. 72, pp. 3–26, 1965.

[21] B.C.J. Moore, B.R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 2015.

[22] B.R. Glasberg and B.C.J. Moore, "A model of loudness applicable to time-varying sounds," *J. Audio Eng. Soc.*, vol. 50, no. 5, pp. 331–342, 2002.

[23] E. Torick, R. Allen, and B. Bauer, "Automatic control of loudness level," *IEEE Trans. on Broadcasting*, vol. BC-14, no. 4, pp. 143–146, 1968.

[24] R. Aarts, "A comparison of some loudness measures for loudspeaker listening tests," *J. Audio Eng. Soc.*, vol. 40, no. 3, pp. 142–146, 1992.

[25] E. Vickers, "Automatic long-term loudness and dynamics matching," in *111th AES Convention*, 2001.

[26] S. Klar and G. Spikofski, "On levelling and loudness problems at television and radio broadcast studios," in *112th AES Convention*, 2002.

[27] E. Skovenborg and S. Nielsen, "Evaluation of different loudness models with music and speech material," in *117th Convention of AES*, 2004.

[28] I. Yanushevskaya, C. Gobl, and A. Chasaide, "Voice quality in affect cueing: does loudness matter?" *Frontiers in Psychology*, vol. 4, 2013.

[29] C. Zorilă and Y. Stylianou, "A fast algorithm for improved intelligibility of speech-in-noise based on frequency and time domain energy reallocation," *Proc. Interspeech*, pp. 60–64, 2015.

[30] L. Turicchia and R. Sarpeshkar, "A bio-inspired companding strategy for spectral enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 2, pp. 243–253, 2005.

[31] H. Gockel, B.C.J. Moore, and R.D. Patterson, "Influence of component phase on the loudness of complex tones," *Acta Acust. united Ac.*, vol. 88, pp. 369–377, 2002.

[32] H. Gockel, B.C.J. Moore, R.D. Patterson, and R. Meddis, "Louder sounds can produce less forward masking: Effects of component phase in complex tones," *J. Acoust. Soc. Am.*, vol. 114, pp. 978–990, 2003.