Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing

Tudor-Cătălin Zorilă and Yannis Stylianou

Toshiba Research Europe Ltd., Cambridge Research Laboratory,

208 Cambridge Science Park, Milton Road, Cambridge CB4 0GZ, UK

catalin.zorila@crl.toshiba.co.uk, yannis.stylianou@crl.toshiba.co.uk

Sheila Flanagan and Brian C.J. Moore[a)]

Department of Experimental Psychology, University of Cambridge, Downing Street,

Cambridge CB2 3EB, UK

saf31@cam.ac.uk, bcjm@cam.ac.uk

[a)] Author to whom correspondence should be addressed

Abstract: A model for the loudness of time-varying sounds [B.R. Glasberg and B.C.J. Moore (2012). J. Audio. Eng. Soc. **50**, 331-342] was assessed for its ability to predict the loudness of sentences that were processed to either decrease or increase their dynamic fluctuations. In a paired-comparison task, subjects compared the loudness of unprocessed and processed sentences that had been equalized in: (1) root-mean square (RMS) level; (2) the peak long-term loudness predicted by the model; (3) the mean long-term loudness predicted by the model. Method 2 was most effective in equating the loudness of the original and processed sentences.

PACS numbers: 43.66Cb, 43.66Ba

1. Introduction

There has been considerable interest in recent years in the development of methods of processing speech so as to enhance its intelligibility when background noise and/or reverberation are added after the processing has been applied (Yoo *et al.*, 2007; Zorila *et al.*, 2012; Cooke *et al.*, 2013). Such methods have potential applications in public address systems and in classrooms for use with special populations, such as children with "auditory processing disorder" (Moore *et al.*, 2013). It would be trivial to improve the intelligibility of speech simply by increasing its level, thereby improving the signal-to-noise ratio (SNR). Therefore, processing methods of this type have typically been evaluated under the constraint that the root-mean-square (RMS) level of the speech should be the same before and after processing (Zorila *et al.*, 2012; Cooke *et al.*, 2013). However, what is important in practical applications is that the loudness of the speech should not be increased by the processing; the loudness must be kept within a range that is judged as comfortable by the majority of listeners. Therefore, it may be more appropriate to assess the processing under the constraint that the *loudness* of the speech should be the same before and after processing. Here, we present an evaluation of the accuracy of the loudness model developed by Glasberg and Moore (2002) in equating the loudness of unprocessed and processed speech.

Two types of speech processing were used, both of which have been shown to improve the intelligibility of speech when applied prior to the addition of background noise (Cooke *et al.*, 2013). One method decreased the short-term level fluctuations in the speech (Zorila *et al.*, 2012) while the other increased them (Takou *et al.*, 2013; Zorila and Stylianou, 2014) relative to those of the original speech. The processed signals were therefore thought to provide a strong test of the accuracy of the loudness model. The model used, called the time-varying-loudness (TVL) model (Glasberg and Moore, 2002), takes a time waveform as its input and generates three forms of time-varying loudness: the instantaneous loudness, which is assumed not to be available for conscious perception; the short-term loudness, which is intended to represent the impression of the loudness of a short segment of the sound, for example a syllable in a sentence; and the long-term loudness (LTL), which is intended to represent the overall loudness of a longer sample of the sound, for example a whole sentence.

In this work, it was also assessed whether overall loudness was best predicted by the maximum value of the LTL reached during presentation of a sentence or by the value of the LTL averaged over all times for which the predicted loudness exceeded a certain threshold value.

2. The signal-processing methods

2.1. Spectral shaping with dynamic range compression

The first method was based on spectral shaping combined with dynamic range compression, denoted SSDRC (Zorila *et al.*, 2012). The signal was analyzed in frames and the spectrum in each frame was estimated by discrete time-frequency transform. Spectral peaks (formants) were sharpened and energy was transferred from low frequencies to medium and high frequencies (1-4 kHz), thereby improving the SNR over the frequency range that is most important for intelligibility (ANSI, 1997). Following spectral shaping, dynamic range compression (DRC) was applied to the broadband signal, aiming to amplify the weaker parts of speech that are more prone to noise masking (fricatives, nasals, and stops), while attenuating parts with more energy (vowels). The effect of the DRC was a reduction of the waveform's envelope variations over time, as illustrated in the middle trace of Fig. 1. Hence the processed speech had a smaller dynamic range than the original speech (top trace).

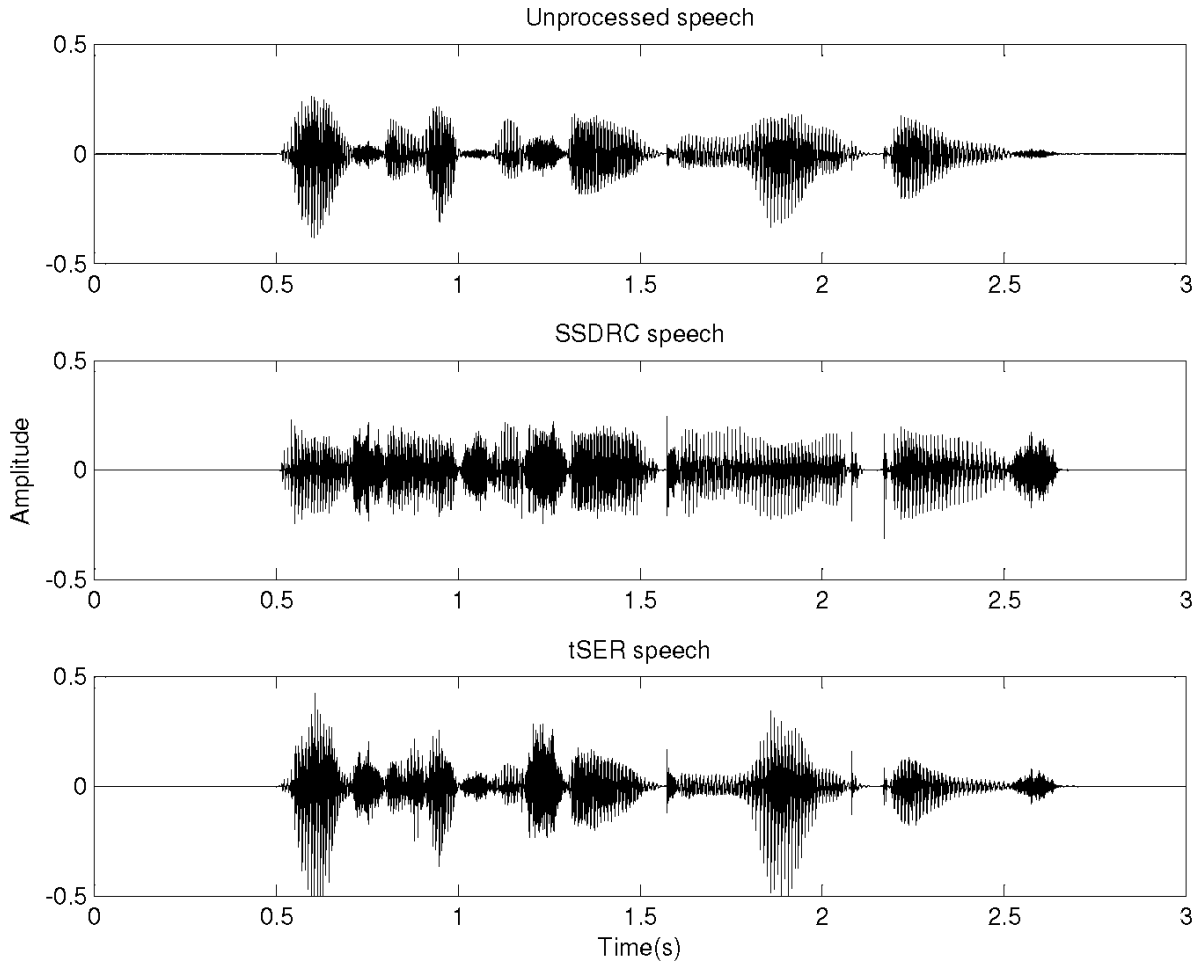

Fig. 1. Waveforms of unprocessed speech (top trace), speech processed using SSDRC (middle trace) and speech processed using tSER. All sentences had the same RMS value. The sentence was “Rice is often served in round bowls”.

2.2. Time-domain spectral energy reallocation

The second method was based on reallocation of energy in frequency using time-domain processing, and is denoted tSER (Takou *et al.*, 2013; Zorila and Stylianou, 2014). This had three processing stages. In one stage, the low-frequency 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 (FIR) 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 (Turicchia and Sarpeshkar,

2005). The outputs of all three stages were combined after weighting of their relative magnitudes. The tSER-processed envelope showed increased envelope fluctuations relative to the original speech, and had a greater dynamic range than the original speech, as illustrated in the bottom trace of Fig. 1.

3. The loudness model

The TVL model used here (Glasberg and Moore, 2002) was an extension of the model for stationary sounds developed by Moore *et al*. (1997). The transfer of sound through the outer and middle ear was modeled using a single FIR filter. Different filters can be used for different sound presentation methods (e.g., free field, diffuse field, or headphone). Here, the diffuse-field option was used, as the stimuli for the experiment were presented using headphones with a diffuse-field response. The version of the model used here was slightly modified to have the middle-ear transfer function given by Glasberg and Moore (2006), as described by Moore (2014).

A running estimate of the spectrum of the sound at the output of the FIR filter was obtained by calculating six Fast Fourier Transforms (FFTs) in parallel, using signal segment durations that decreased with increasing center frequency. This was done to give sufficient spectral resolution at low frequencies and sufficient temporal resolution at high frequencies. All FFTs were updated every 1 ms. Each FFT was used to calculate spectral magnitudes over a specific frequency range; values outside that range were discarded. An excitation pattern was calculated from the short-term spectrum at 1-ms intervals, using the same method as described by Moore *et al*. (1997). The next stage was the calculation of the “instantaneous” loudness, which is assumed to be an intervening variable that is not available for conscious perception. The calculation of instantaneous loudness from the excitation pattern was done in the same way as described by Moore *et al*. (1997).

The short-term loudness was calculated from a running average of the instantaneous loudness, using an averaging process resembling the way that a control signal is generated in an automatic gain control (AGC) circuit. The LTL was calculated from the short-term loudness, again using a form of averaging resembling the operation of an AGC circuit, but

with longer time constants. For details, see Glasberg and Moore (2002) and Moore (2014).

In the original version of the TVL model, the release time of the averager used to calculate the LTL had a value of 2000 ms. This relatively long time constant was partly chosen to reflect high-level processes such as memory. Here, we evaluated both the original version of the TVL model and a version in which the release time used to calculate the LTL was shorter, at 200 ms.

When a single sentence is used as input to the model, the predicted LTL builds up over some time, and then stabilizes at roughly a constant value. However, the value still fluctuates to some extent; the fluctuation is greater when the release time constant is shorter, as illustrated in Fig. 2. The question arises as to whether the overall loudness as judged by human listeners is better predicted by the peak value reached by the LTL or by the mean value of the LTL over the time period where its value is reasonably stable. Both approaches were evaluated here.

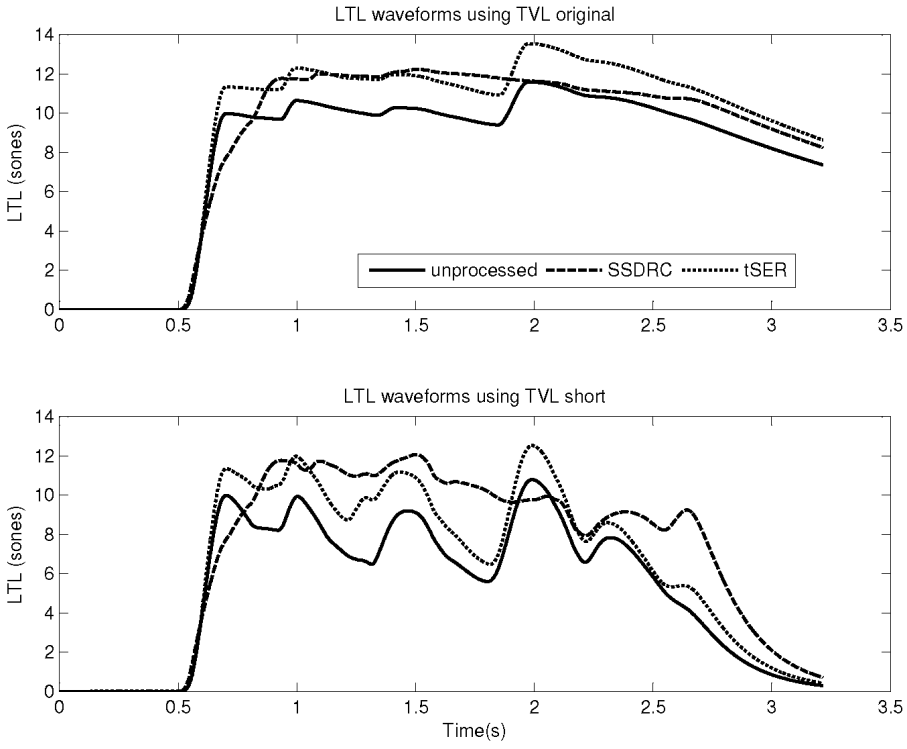

Fig. 2. Long-term loudness as a function of time predicted by the original TVL model (top) and the version of the model with shorter release time (bottom) for the sentence “Rice is often served in round bowls” either unprocessed (solid lines) or processed using SSDRC (dashed lines) or tSER (dotted lines).

4. Loudness comparison experiments

Two experiments were conducted, one using the original release time to calculate the LTL and one using the shorter release time of 200 ms. These are denoted experiments 1 and 2, respectively.

4.1 Subjects

Fifteen subjects (7 male) were tested in experiment 1 and ten subjects (5 male) were tested in experiment 2. All reported having normal hearing and all had audiometric thresholds ≤20 dB HL for all audiometric frequencies from 0.25 to 6 kHz. Their ages ranged from 18 to 70 years for both experiments (mean = 40.3 years for experiment 1 and 40.7 years for experiment 2). Five subjects took part in both experiments. All subjects were native speakers of English.

4.2 Procedure

A paired-comparison procedure was used. Ten Harvard sentences (Rothauser *et al.*, 1969) were used, spoken by a man. On each trial, the same sentence was presented twice in succession, once unprocessed and once processed with one of the two methods (either SSDRC or tSER). The order of the unprocessed and processed sentences was random with the constraint that the unprocessed sentence occurred equally often in the first and second positions. The unprocessed sentence had an overall diffuse-field equivalent level of 65 dB SPL (its level and spectrum at the eardrum were the same as would be produced if the sound were presented in a diffuse field with a level of 65 dB SPL at the position corresponding to the center of the listener's head). The two sentences within a trial were equalized either in RMS level, in the peak LTL predicted by the TVL model, or in the mean LTL predicted by the TVL model (see below for details of the equalization procedure). The subject was asked to use a slider on a screen, controlled by a computer mouse, to indicate whether the first or the second sentence was louder and by how much. The scale ranged from –3 (sentence 1 much louder) to +3 (sentence 2 much louder). A slider setting of 0 indicated that the two sentences were equal in loudness. The scale was continuous. All ten sentences were used with each equalization method and processing method. Pairs of sentences for the different

equalization methods (3 types) and different speech-processing methods (2 types) were interleaved and presented in an order that was different for each subject, with the constraint that the same sentence was never presented twice in succession.

When the unprocessed sentence was judged as louder than the processed sentence, any non-zero response was scored as a negative number. Conversely, when the processed sentence was judged as louder than the unprocessed sentence, any non-zero response was scored as a positive number. The coded responses for each processing method were averaged across all sentences. For simplicity, the result is called the "mean score." The equalization method that led to a mean score closest to zero was deemed to be the method that gave the most accurate loudness equalization.

4.3 Equalization of the original and processed speech

For each unprocessed sentence, the following were calculated: (1) the RMS level; (2) the peak LTL predicted by the TVL model; (3) the mean LTL predicted by the TVL model averaged across all values of the LTL that were above 1 sone. For experiment 1, the overall amplitude of a given processed sentence was iteratively scaled until either: (1) the RMS level was matched to that of the same unprocessed sentence; (2) the peak LTL matched that of the same unprocessed sentence; (3) the mean LTL matched that of the same unprocessed sentence. This was done separately for each sentence. The resulting scaled amplitudes were those used in experiment 1. For experiment 2, the amplitudes of all sentences were scaled either so that the peak LTL was 10 sones (corresponding to the average peak LTL for the unprocessed sentences before scaling) or so that the mean LTL was 7 sones (corresponding to the mean of the mean LTL of the unprocessed sentences before scaling).

In experiment 1, for peak LTL equalization, the level of the SSDRC-processed speech was reduced, on average, by 0.2 dB, and that of the tSER-processed speech was reduced by 2.6 dB, relative to the levels required for equal RMS. For mean LTL equalization, the level of the SSDRC-processed speech was reduced, on average, by 1.8 dB, while that for tSER-processed speech was reduced by 3.2 dB. Although the mean reduction was very small for peak LTL equalization and SSDRC-processed speech, the change in level varied across

sentences from −1.7 to 2.8 dB (standard deviation = 1.4 dB). Thus, equating the RMS level of individual unprocessed and SSDRC-processed sentences would probably not lead to equal loudness for all sentences.

In experiment 2, for peak LTL equalization, the level of the SSDRC-processed speech was reduced, on average, by 0.9 dB, and that of the tSER-processed speech was reduced by 2.6 dB, relative to the levels required for equal RMS. For mean LTL equalization, the level of the SSDRC-processed speech was reduced, on average, by 4.1 dB, while that for tSER-processed speech was reduced by 2.9 dB.

It should be noted that the level reductions described above are not solely a result of differences in the temporal properties of the unprocessed and processed speech; they result at least partly from spectral differences between the unprocessed and processed speech. For a fixed RMS level, both types of processing result in a reduction of low-frequency energy and an increase of medium- and high-frequency energy. The medium and high frequencies contribute more to loudness than the low frequencies, so the spectral changes result in an increase in loudness. This point is discussed in more detail later.

4.4 Stimulus generation and presentation

Stimuli were generated digitally (16-bit resolution, 16-kHz sampling rate) and presented via Sennheiser HD580 headphones (Wedemark, Germany), which have approximately a diffuse-field frequency response. Subjects were seated in a sound-attenuating chamber. They responded using a computer mouse, as described above. No feedback was given.

5. Results

5.1 Experiment 1 (longer release time)

The mean scores for experiment 1, averaged across subjects, are shown in Fig. 3. For sentences equated in RMS level (left pair of bars), the values were positive, by 0.22 scale units for original versus SSDRC and 0.43 scale units for original versus tSER. This means that, at equal RMS level, speech processed using either SSDRC or tSER was louder than the original speech. For sentences equated in peak LTL, the mean score was just above zero

(0.05) for original versus SSDRC and below zero (–0.33) for original versus tSER. Thus, when equated for peak LTL, the processed sentences were well matched in loudness to the original sentences for SSDRC and were slightly less loud for tSER. For sentences equated in mean LTL, the mean score was –0.30 for original versus SSDRC and –0.47 for original versus tSER. Thus, when equated for mean LTL, the tSER-processed sentences were somewhat less loud than the original sentences. Averaged across processing methods, the mean scores were 0.32 for RMS equalization, –0.14 for peak LTL equalization, and –0.39 for mean LTL equalization.

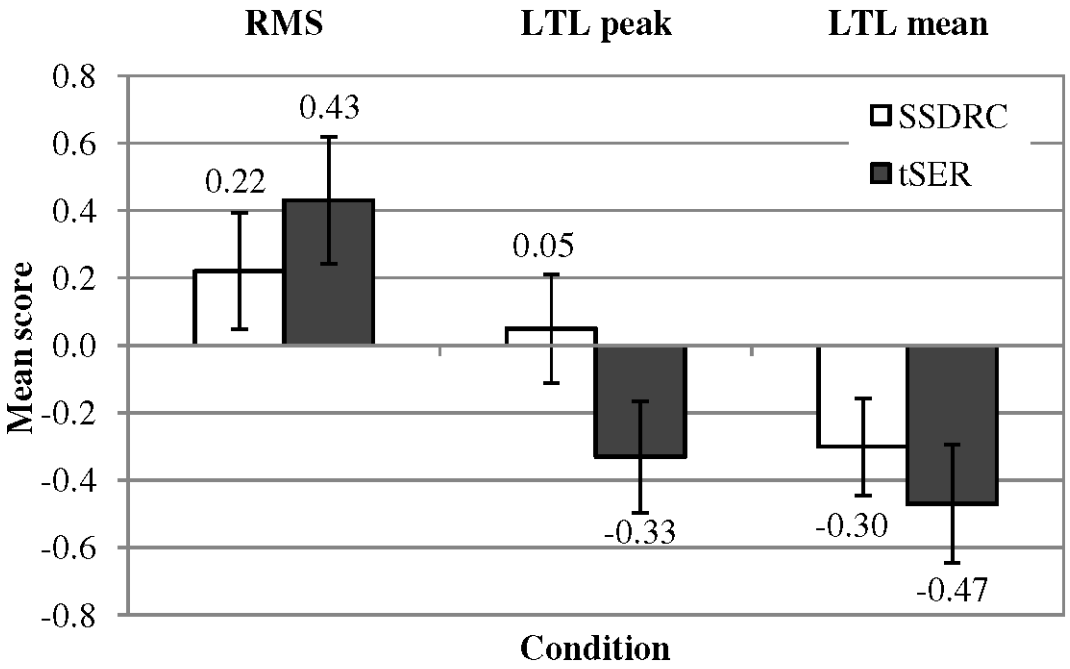

Fig. 3. Results of experiment 1 (LTL calculated with using the original TVL model with the longer release time) showing mean ratings of the loudness of processed speech relative to that of unprocessed speech for two types of processing (SSDRC, open bars, and tSER, shaded bars) when the unprocessed and processed speech were equated in terms of: (1) RMS level (left pair of bars); (2) the peak value of the LTL (middle pair of bars); (3) the mean value of the LTL (right pair of bars). Error bars show ±1 standard error.

A two-way repeated-measures analysis of variance (ANOVA) was conducted on the scores with factors equalization method and type of processing. Mauchly's test indicated that

the assumption of sphericity was violated for the factor equalization method and for the interaction of equalisation method with type of processing so the degrees of freedom were adjusted using the Greenhouse-Geisser correction. There was a significant main effect of equalization method: $F(1.07,14.94) = 23.7$, $p < 0.001$. There was no significant effect of type of processing ($p > 0.05$), but there was a significant interaction of equalization method and type of processing: $F(1.31, 18.38) = 9.63$, $p < 0.005$.

A series of *t*-tests (two-tailed) was conducted to assess whether the mean score for each equalization method and processing method was significantly different from zero. For RMS equalization, the mean for tSER was significantly above zero ($t(14) = 2.25$, $p = 0.041$), but the mean for SSDRC was not. For peak LTL equalization, the means did not differ significantly from zero for either processing method ($p > 0.05$). For mean LTL equalization, the mean for tSER processing was significantly below zero ($t(14) = 2.69$, $p = 0.018$), while the mean for SSDRC processing did not differ significantly from zero. Overall, it can be concluded that equalization based on the peak LTL was the best method for equating the loudness of the unprocessed and processed sentences.

5.2 *Experiment 2 (shorter release time)*

The mean scores averaged across subjects are shown in Fig. 4. The pattern of the results is similar to that for experiment 1. For sentences equated in RMS level (left pair of bars), the values were positive, by 0.28 scale units for original versus SSDRC and 0.48 scale units for original versus tSER. Thus, at equal RMS level, speech processed using either SSDRC or tSER was louder than the original speech. For sentences equated in peak LTL, the mean score was very close to zero (–0.01) for original versus SSDRC and slightly below zero (–0.24) for original versus tSER. Thus, when equated for peak LTL, the original and processed sentences were reasonably well matched in loudness. For sentences equated in mean LTL, the mean score was –0.64 for original versus SSDRC and –0.24 for original versus tSER. Thus, when equated for mean LTL, the SSDRC-processed sentences were markedly softer than the original sentences, while the tSER-processed sentences were slightly softer. Averaged across processing methods, the mean scores were 0.38 for RMS equalization, –0.125 for peak LTL

equalization, and –0.44 for mean LTL equalization.

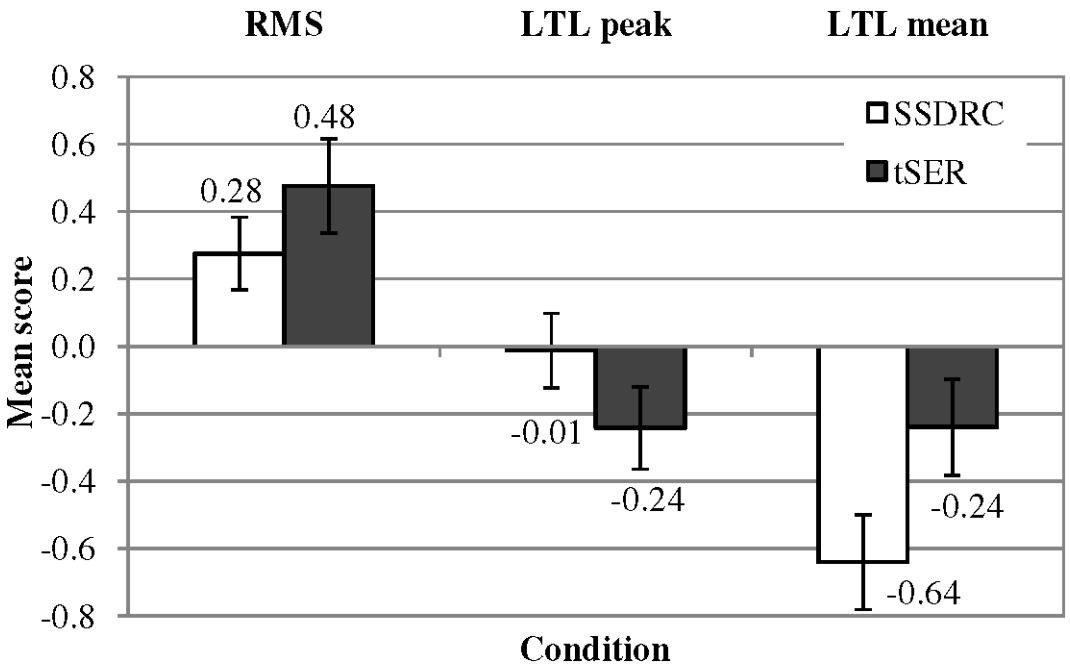

Fig. 4. As Fig. 3, but showing the results for experiment 2, which used the TVL model with a shorter release time for calculating the LTL.

A two-way repeated-measures ANOVA was conducted on the scores with factors equalization method and type of processing. Mauchly's test indicated that the assumption of sphericity was violated for the factor equalization method so the degrees of freedom were adjusted using the Greenhouse-Geisser correction. There was a significant main effect of equalization method: $F(1.094,9.825) = 34.8$, $p < 0.001$. There was no significant effect of type of processing ($p > 0.05$), but there was a significant interaction of equalization method and type of processing: $F(2, 18) = 17.5$, $p < 0.001$.

A series of *t*-tests (two-tailed) was conducted to assess whether the mean score for each equalization method and processing method was significantly different from zero. For RMS equalization, the means for both SSDRC and tSER were significantly above zero ($t(9) > 2.55$, $p = 0.031$). For peak LTL equalization, the means did not differ significantly from zero for either processing method ($p > 0.05$). For mean LTL equalization, the mean for SSDRC processing was significantly below zero ($t(9) = 4.55$, $p = 0.0014$), while the mean for tSER processing did not differ significantly from zero. Overall, it can be concluded that

equalization based on the peak LTL was the best method for equating the loudness of the unprocessed and processed sentences.

6. Discussion

The results showed that for speech processed to either increase or decrease its dynamic range, equalization of the processed and unprocessed speech based on the peak LTL predicted by the TVL model led to more accurate equalization of loudness as perceived by human listeners than equalization based on the mean value of the LTL or the RMS level. This was true for both values of the release time constants used to calculate the LTL.

One issue that arises is whether loudness equalization could be performed equally well using a model based on the long-term-average spectra of the stimuli. To assess this, we calculated the long-term spectra across the ten sentences for unprocessed, SSDRC-processed and tSER-processed stimuli with equal RMS levels, and with levels adjusted to give equal peak LTL or equal mean LTL (for the longer time constant only). We then used the loudness model of Moore *et al*. (1997) for stationary sounds, with the modified middle-ear transfer function described by Glasberg and Moore (2006), to predict the loudness in each case. The results are summarized in Table 1. The predicted loudness values for the unprocessed stimuli are all the same, because no adjustment of level was applied for these stimuli. When RMS levels were equalized, the model for stationary sounds predicted that both types of processed stimuli would be louder than the unprocessed stimuli, as found in the data. Equalization based on the mean LTL led to a predicted loudness level that was 1.5 phons higher for SSDRC-processed speech than for unprocessed speech and was almost the same for tSER-processed speech and unprocessed speech. For equalization based on the peak LTL, which led to the most accurate equalization of loudness in our experimental data, the loudness model for stationary sounds did not predict a constant loudness across processing conditions. In particular, the predicted loudness level for SSDRC-processed speech was 2.8 phons higher than for the unprocessed speech and the predicted loudness level for tSET-processed speech was 1 phon higher than for unprocessed speech. We conclude that the dynamic aspects of the stimuli did influence their loudness and that there are benefits in using the model for time-

varying sounds to equalize the loudness of unprocessed and processed speech.

The results of the experiments are consistent with earlier results showing that the LTL predicted by the TVL model could give accurate predictions of the loudness of speech that had been subjected to multi-channel amplitude compression of the type that is often used in broadcasting (Moore *et al.*, 2003). It is also consistent with the results of Rennies *et al.* (2013), obtained using both loudness matching and categorical loudness scaling, which showed that the LTL gave reasonably accurate predictions of the loudness of a variety of speech-like signals (including speech-shaped noise, unprocessed speech, and speech that was subjected to filtering, reverberation, and amplitude compression and expansion), whereas the predictions were not as accurate when based on the short-term loudness derived using the TVL model or other loudness models (Chalupper and Fastl, 2002; Rennies *et al.*, 2009).

Table 2 summarizes the results of the two experiments, showing the mean adjustments in level relative to equal RMS required to equalize the peak LTL or the mean LTL and the mean rating obtained for each equalization method and type of processing. For SSDRC, the correlation between the level adjustments and the ratings for the combined results of the two experiments was 0.97 ($p < 0.05$). The best-fitting linear regression line was

$$\text{Rating} = 0.212(\text{Level adjustment}) + 0.18 \quad (1)$$

This implies that the rating would be 0, i.e. the SSDRC-processed and unprocessed sentences would be equally loud, when the RMS level of the SSDRC-processed sentences was reduced by 0.8 dB. The level reduction based on equalizing the peak LTL was 0.2 dB with the original release time constant and 0.9 dB with the shorter time constant, both of which are reasonably close to the “ideal” value of 0.8 dB. The level reduction based on equalizing the mean LTL was 1.8 dB with the original release time constant and 4.1 dB with the shorter time constant. This last value is markedly larger than the “ideal” value, suggesting better performance with the original release time.

For tSER, the correlation between the level adjustments and the ratings for the combined results of the two experiments was 0.99 ($p < 0.05$). The best-fitting linear regression line was

$$\text{Rating} = 0.274(\text{Level adjustment}) + 0.455 \quad (2)$$

This implies that the rating would be 0, i.e. the tSER-processed and unprocessed sentences would be equally loud when the RMS level of the tSER-processed sentences was reduced by 1.7 dB. The level reduction based on equalizing the peak LTL was 2.6 dB with both the original release time constant and the shorter time constant, reasonably close to the "ideal" value. The level reduction based on equalizing the mean LTL was 3.2 dB with the original release time constant and 2.9 dB with the shorter time constant, both somewhat larger than the "ideal" value.

Overall, the results suggest that the level adjustments based on matching the peak value of the LTL were somewhat closer to the adjustments required to actually match the loudness of the unprocessed and processed speech than level adjustments based on the mean value of the LTL. This was the case using both the original release time constant and the shorter time constant. Thus, level adjustments based on matching the peak LTL seem to be preferable.

7. Summary and conclusions

Speech processing to enhance its intelligibility when noise is added after processing can either increase or decrease the speech dynamic range, depending on the method of processing, and can also change the average spectral shape of the speech. These changes can alter the loudness of the speech when the overall RMS level is held constant. This paper assessed the effectiveness of three methods in equating the loudness of unprocessed and processed speech, for two methods of speech processing, one that decreased the dynamic range (SSDRC) and one that increased it (tSER). The original and processed speech were equated in terms of: (1) RMS level; (2) the peak LTL predicted by the TVL model; (3) the mean LTL predicted by the TVL model. Two versions of the TVL model were used, one with the original longer release time for calculating the LTL (experiment 1) and the other with a shorter release time (experiment 2).

The results were similar for the two experiments. When equated in RMS level, the processed speech was judged as louder than the unprocessed speech for both SSDRC and tSER; the difference was significant for tSER in experiment 1 and for both SSDRC and tSER

in experiment 2. When equated in peak LTL, the loudness of the processed speech did not differ significantly from that of the unprocessed speech for either processing method. When equated in mean LTL, the processed speech was judged as softer than the unprocessed speech; the difference was significant for tSER in experiment 1 and for SSDRC in experiment 2. It is concluded that the method based on the peak LTL is effective in equating the loudness of processed and unprocessed speech for processing that either decreases or increases the dynamic range of the speech.

Acknowledgments

We thank two reviewers for helpful comments on an earlier version of this paper.

References

ANSI (**1997**). *ANSI S3.5-1997. Methods for the calculation of the speech intelligibility index* (American National Standards Institute, New York).

Chalupper, J., and Fastl, H. (**2002**). "Dynamic loudness model (DLM) for normal and hearing impaired listeners," Acta Acust. united Ac. **88**, 378-386.

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (**2013**). "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proceedings of Interspeech* (Lyon, France), pp. 3552-3556.

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio Eng. Soc. **50**, 331-342.

Glasberg, B. R., and Moore, B. C. J. (**2006**). "Prediction of absolute thresholds and equal-loudness contours using a modified loudness model," J. Acoust. Soc. Am. **120**, 585-588.

Moore, B. C. J. (**2014**). "Development and current status of the "Cambridge" loudness models," Trends Hear. **18**, 1-29.

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). "A model for the prediction of thresholds, loudness and partial loudness," J. Audio Eng. Soc. **45**, 224-240.

Moore, B. C. J., Glasberg, B. R., and Stone, M. A. (**2003**). "Why are commercials so loud? - Perception and modeling of the loudness of amplitude-compressed speech," J. Audio Eng. Soc. **51**, 1123-1132.

Moore, D. R., Rosen, S., Bamiou, D. E., Campbell, N. G., and Sirimanna, T. (**2013**). "Evolving concepts of developmental auditory processing disorder (APD): a British Society of Audiology APD special interest group 'white paper'," Int. J. Audiol. **52**, 3-13.

Rennies, J., Holube, I., and Verhey, J. L. (**2013**). "Loudness of speech and speech-like signals," Acta Acust. united Ac. **99**, 268-282.

Rennies, J., Verhey, J. L., Chalupper, J., and Fastl, H. (**2009**). "Modeling temporal effects of spectral loudness summation," Acta Acust. united Ac. **95**, 1112-1122.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225-246.

Takou, R., Seiyama, N., and Imai, A. (**2013**). "Improvement of speech intelligibility by reallocation of spectral energy," in *Proceedings of Interspeech* (Lyon, France), pp. 3605-3607.

Turicchia, L., and Sarpeshkar, R. (**2005**). "A bio-inspired companding strategy for spectral enhancement," IEEE Trans. Speech. Audio Proc. **13**, 243-253.

Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C. C., Durrant, J. D., Kovacyk, K., Shaiman, S. (**2007**). "Speech signal modification to increase intelligibility in noisy environments," J. Acoust. Soc. Am. **122**, 1138-1149.

Zorila, C., Kandia, V., and Stylianou, Y. (**2012**). "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proceedings of Interspeech* (Portland, OR, USA), pp. 635-638.

Zorila, C., and Stylianou, Y. (**2014**). "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," in *Proceedings of Interspeech* (Singapore), pp. 2050-2054.

Table 1. Loudness calculated using a model for stationary sounds based on the long-term average spectrum, with various forms of equalization across processing method.

	Unprocessed			SSDRC			tSER		
Equalization method	RMS	Peak LTL	Mean LTL	RMS	Peak LTL	Mean LTL	RMS	Peak LTL	Mean LTL
Loudness, sones	20.0	20.0	20.0	27.9	24.5	22.3	24.7	21.4	20.7
Loudness level, phons	83.2	83.2	83.2	87.8	86.0	84.7	86.1	84.2	83.7

Table 2. Summary of the results of experiment 1 (original release time) and experiment 2 (shorter release time), showing the average level adjustments (relative to equal RMS) required to equate the peak value of the LTL and the mean value of the LTL, together with the mean loudness ratings.

	RMS		LTL peak		LTL mean	
	SSDRC	tSER	SSDRC	tSER	SSDRC	tSER
Level adjustment re RMS Original, dB	0	0	–0.2	–2.6	–1.8	–3.2
Mean rating	0.22	0.43	0.05	–0.33	–0.30	–0.47
Level adjustment re RMS Shorter, dB	0	0	–0.9	–2.6	–4.1	–2.9
Mean rating	0.28	0.48	–0.01	–0.24	–0.64	–0.24