The Journal of the Acoustical Society of America

Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss

--Manuscript Draft--

Manuscript Number:	JASA-00750R1
Full Title:	Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss
Short Title:	Speech intelligibility enhancement
Article Type:	Regular Article
Corresponding Author:	Tudor-Catalin Zorila Toshiba Cambridge Research Laboratory Cambridge, UNITED KINGDOM
First Author:	Tudor-Catalin Zorila
Order of Authors:	Tudor-Catalin Zorila
	Yannis Stylianou
	Sheila Flanagan
	Brian Cecil Joseph Moore
Section/Category:	Speech Communication
Keywords:	speech intelligibility, near-end enhancement, loudness, loudness model
Abstract:	Four algorithms designed to enhance the intelligibility of speech when noise is added after processing were evaluated under the constraint that the speech should have the same loudness before and after processing, as determined using a loudness model. The algorithms applied spectral modifications and two of them included dynamic-range compression. On average, the methods with dynamic-range compression required the least level adjustment to equate loudness for the unprocessed and processed speech. Subjects with normal-hearing (Experiment 1) and mild-to-moderate hearing loss (Experiment 2) were tested using unmodified and enhanced speech presented in speech-shaped noise (SSN) and a competing speaker (CS). The results showed: (a) The algorithms with dynamic-range compression yielded the largest intelligibility gains in both experiments and for both types of background; (b) The algorithms without dynamic-range compression either yielded benefit only with the SSN or yielded no consistent benefit; (c) Speech reception thresholds for unprocessed speech were higher for hearing-impaired than for normal-hearing subjects, by about 2 dB for the SSN and 6 dB for the CS. It is concluded that the enhancement methods incorporating dynamic-range compression can improve intelligibility under the equal-loudness constraint for both normal-hearing and hearing-impaired subjects and for both steady and fluctuating backgrounds.

Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss[*)]

Tudor-Cătălin Zorilă[a)] and Yannis Stylianou[b)]

Toshiba Research Europe Ltd., Cambridge Research Laboratory,

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

Sheila Flanagan and Brian C.J. Moore

Department of Experimental Psychology, University of Cambridge,

Downing Street, Cambridge CB2 3EB, UK

[*)]Some of the data in this paper were presented at Interspeech 2016, San Francisco, USA.

[a)]Emails: catalin.zorila@crl.toshiba.co.uk, ztudorc@gmail.com

[b)]Also at: Department of Computer Science, University of Crete, Heraklion, Greece

Abstract

Four algorithms designed to enhance the intelligibility of speech when noise is added after processing were evaluated under the constraint that the speech should have the same loudness before and after processing, as determined using a loudness model. The algorithms applied spectral modifications and two of them included dynamic-range compression. On average, the methods with dynamic-range compression required the least level adjustment to equate loudness for the unprocessed and processed speech. Subjects with normal-hearing (Experiment 1) and mild-to-moderate hearing loss (Experiment 2) were tested using unmodified and enhanced speech presented in speech-shaped noise (SSN) and a competing speaker (CS). The results showed: (a) The algorithms with dynamic-range compression yielded the largest intelligibility gains in both experiments and for both types of background; (b) The algorithms without dynamic-range compression either yielded benefit only with the SSN or yielded no consistent benefit; (c) Speech reception thresholds for unprocessed speech were higher for hearing-impaired than for normal-hearing subjects, by about 2 dB for the SSN and 6 dB for the CS. It is concluded that the enhancement methods incorporating dynamic-range compression can improve intelligibility under the equal-loudness constraint for both normal-hearing and hearing-impaired subjects and for both steady and fluctuating backgrounds.

I. INTRODUCTION

Several researchers have developed methods of processing speech so as to enhance its intelligibility when background noise and/or reverberation are added after the modifications have been applied (Cooke et al., 2013a). Such methods have potential applications in public-address systems and in classrooms for use with special populations, such as children with "auditory processing disorder". This type of speech processing is called "near-end listening enhancement" (NLE) (Sauert and Vary, 2006). In previous studies of NLE, the unprocessed and processed speech were equated in power, i.e. in root-mean-square (RMS) value. This is referred to here as the equal-RMS (EQR) constraint. However, it is important in practical applications 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. This is especially important for hearing-impaired listeners, who typically experience loudness recruitment and have a smaller dynamic range than for normal-hearing listeners (Steinberg and Gardner, 1937; Fletcher, 1938; Moore et al., 1996). Therefore, it seems more appropriate to assess NLE processing under the constraint that the loudness of the speech should be the same before and after processing. This paper presents an evaluation of four state-of-the-art NLE algorithms under an equal-loudness (EQL) constraint. All four systems have been shown to perform very well under the EQR constraint for normal-hearing (NH) subjects. Both NH

subjects and subjects with mild-to-moderate hearing-loss were tested in the present study. It was considered important to test hearing-impaired (HI) subjects, since about 10% of the population in developed countries has some degree of hearing loss (Davis, 1995) and since HI people have marked difficulty in understanding speech when background sounds are present (Moore, 2003). They are therefore likely to have greater difficulty than NH people when listening to public-address systems. People with mild-to-moderate hearing loss often do not use hearing aids, so it was considered appropriate to test the HI subjects without the use of hearing aids, even though some of them were hearing-aid users.

The development of NLE algorithms has been based on a number of approaches. Some approaches are based on fairly basic forms of signal processing such as high-pass filtering followed by amplitude compression (Niederjohn and Grotelueschen, 1976), or signal-to-noise ratio (SNR) recovery (Sauert and Vary, 2006). Spectral contrast enhancement has been used as a possible method of compensating for the reduced frequency selectivity of hearing-impaired people (Baer et al., 1993; Oxenham et al., 2007; Bhattacharya et al., 2011).

Another approach is based on the finding that talkers adaptively change their speaking style from conversational to other more 'specialized' forms in situations where communication is difficult (e.g., when ambient noise is present, or when the listener is not familiar with the language or is hearing impaired). Speakers generally raise their voices

when background noise is present (the Lombard reflex, Lombard (1911)), and speech produced under these conditions ("Lombard" speech) is easier to understand than speech produced in quiet (Dreher and O'Neill, 1957; Summers et al., 1988; Junqua, 1993; Lu and Cooke, 2008). Also, speech that is deliberately spoken clearly is easier to understand than conversational speech for both normal and hearing-impaired listeners (Picheny et al., 1985; Payton et al., 1994; Ferguson and Kewley-Port, 2002) and for non-native listeners (Bradlow and Bent, 2002; Cooke and Lecumberri, 2012). Therefore, one approach is to process conversational speech so as to mimic the effects that occur in "Lombard" speech or clearly spoken speech (Skowronski and Harris, 2006; Zorilă et al., 2012; Takou et al., 2013; Godoy et al., 2014; Jokinen et al., 2014; Erro et al., 2014).

Another approach is to selectively amplify the regions of speech that are thought to carry the most information (Petkov and Kleijn, 2015). For example, Hazan and Simpson (1998) showed that amplifying consonantal regions yielded intelligibility benefits, while Yoo et al. (2007) demonstrated that amplifying transient regions also yielded benefits.

NLE was the topic of the recent "Hurricane Challenge" (Cooke et al., 2013a,b). Participants in the challenge were asked to submit algorithms that would work for both stationary (speech-shaped noise, SSN) and non-stationary (competing speaker, CS) background sounds at different SNRs. All methods were evaluated under the EQR constraint and keyword recognition scores were used to determine the winners. All listeners

had normal hearing. Two of the best performing algorithms in the Hurricane Challenge were compared here under the EQL constraint, using both NH subjects (Experiment 1) and HI subjects with mild-to-moderate hearing loss (Experiment 2).

The rest of the paper is organized as follows. Section II describes how the speech samples were equated in loudness, Section III describes the NLE algorithms, Sections IV and V describe the methods used for Experiments 1 and 2, Section VI presents a discussion of the results, and Section VII gives some conclusions.

II. LOUDNESS EQUALIZATION

The loudness equalization method used here was the same as that described in Zorilă et al. (2016) and was based on the time-varying loudness (TVL) model of Glasberg and Moore (2002). The TVL model is an extension of an earlier model for stationary sounds (Moore et al., 1997). The reader is referred to those papers for details. The TVL model computes two forms of loudness. The short-term loudness represents the loudness of a short segment of sound (e.g., a word). The long-term loudness (LTL) characterizes the overall loudness of a longer segment of sound (e.g., a sentence). Zorilă et al. (2016) showed that the peak value of the LTL for each sentence provided an effective measure for equating loudness across processing conditions, as assessed using loudness matching with normal-hearing listeners. Hence, the peak value of the LTL was used here to equate loudness across processing conditions. The level of each processed sentence was iteratively

adjusted to match the peak LTL value of the same unprocessed sentence. The TVL model with shorter release times for computing the LTL was used in this work (Zorilă et al., 2016).

III. PROCESSING ALGORITHMS

Speech intelligibility was assessed for four NLE algorithms and for unprocessed speech. There were three rule-based algorithms (Zorilă et al., 2012; Takou et al., 2013; Zorilă and Stylianou, 2015) and one which derived the speech modifications from an optimization criterion (Petkov and Kleijn, 2015). The latter was used only in Experiment 1. Speech spoken by a talker who was listening to SSN (Lombard speech) was assessed in Experiment 2 only. The Lombard speech samples were the same as those used in the Hurricane Challenge (Cooke et al., 2013a).

One algorithm was spectral shaping with dynamic range compression (SSDRC) (Zorilă et al., 2012). This was the winner for 5 out of 6 background conditions in the Hurricane Challenge (Cooke et al., 2013a). It had two processing stages, spectral shaping followed by time-varying amplitude compression. The spectral shaper was frame based and its operation was controlled by a measure of the strength of voicing in the current frame. The spectral shaper transferred energy from components with frequencies below 500 Hz to higher frequencies in such a way that the formants were sharpened, the spectral tilt was flattened, and the SNR in the range 0.5-4 kHz was increased. 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) (Yoo et al., 2007).

The second algorithm was time-domain spectral energy reallocation (tSER) (Takou et al., 2013). This led to good intelligibility gains in the Hurricane Challenge, especially for the CS background (Cooke et al., 2013b). It consisted of three parallel processing stages. In one stage, the components below 400 Hz were passed on unprocessed for combination with signals from the other stages. In the second stage, the signal was pre-emphasized to reduce its spectral tilt, and in the last stage, the spectral contrast of the signal from the previous stage was enhanced using a method resembling two-tone suppression in the cochlea. tSER did not perform energy redistribution over time. For details of the latest implementation of tSER, see Zorilă and Stylianou (2014).

The third algorithm was spectral energy reallocation in the frequency domain followed by dynamic range compression (fSER+DRC) (Zorilă and Stylianou, 2015). The first stage performed processing very similar to that used for tSER, but the processing was frame-based, using 32-ms frames with 50% overlap, and all processing was conducted in the frequency domain. This is more computationally efficient than tSER. The second stage was the same as the DRC stage of SSDRC.

The fourth algorithm was spectral dynamics recovery (SDR) (Petkov and Kleijn, 2015). The signal was split into multiple frequency bands. A distortion measure was used

to characterize the deviation of the dynamics of the noisy speech from the dynamics of speech without noise within each band. This distortion measure was used to derive a parametric relationship between the signal power in a given frequency band before and after addition of the noise. This relationship in turn was used to set the gain in each frequency band so as to preserve the dynamic fluctuations of the speech as much as possible after the addition of noise. Both speech and noise input signals were necessary to train the system. For details, the reader is referred to Petkov and Kleijn (2015).

IV. EXPERIMENT 1: NORMAL-HEARING SUBJECTS

A. Stimuli

The speech and background stimuli used for both experiments were the same as those used for the Hurricane Challenge. The speech consisted of phonetically balanced Harvard sentences (Rothauser et al., 1969) spoken by a native English male and recorded at the Centre for Speech Technology Research, University of Edinburgh, UK (Cooke et al., 2013a). The first 30 Harvard sets (300 sentences in total) were used for the main evaluation, while the 31st set was used for practice. All sentences were generated digitally (16-bit resolution, 16-kHz sampling rate) and had 0.5 s of silence appended at their start and end.

The two background sounds were SSN and CS. The SSN was obtained by filtering white noise so that it had the same average long-term spectrum as the speech produced by

a female reading news-like text. The same female talker was used for the CS. The SSN background was turned on 0.5 s before the start of each sentence and off 0.5 s after the end of each sentence. The CS was gated in a similar way except that its duration was extended so that it stopped at the end of a whole sentence of the CS. Stimuli were presented diotically.

Unprocessed speech was mixed with noise at three SNRs for each background type. These were: −7, −2 and +3 dB for SSN and −19, −12 and −5 dB for CS. For convenience, the SNRs are denoted 'low, 'medium' and 'high', respectively. Because the processed and unprocessed speech were equated for their loudness in quiet, the physical SNR differed across conditions. The changes in level of the processed speech relative to the unprocessed speech needed to equate their loudness are shown in Table I.

B. Subjects

Twenty subjects (8 males) took part. All had audiometric thresholds $\leq$ 20 dB HL for all audiometric frequencies from 0.25 to 8 kHz. Their ages ranged from 19 to 70 years (mean = 26). All were native speakers of English.

C. Procedure

Subjects were tested in a double-walled sound-attenuating booth. Stimuli were presented via Sennheiser HD580 headphones (Wedemark, Germany), which have

approximately a diffuse-field frequency response. The background sound was presented at an equivalent diffuse-field level of 65 dB SPL, so the speech level varied with the SNR. Subjects responded via a Matlab graphical interface. Subjects heard each sentence and background sound only once and were asked to type all of the words they thought they heard on a keyboard. Subjects were encouraged to give their best guess when they were unsure. When the subject indicated that the response was complete, the next stimulus was presented. The test was self-paced and no feedback was given. After every 50 trials, the software encouraged subjects to take a short break. The whole test took roughly one hour to complete.

Stimuli were presented following a Latin square design with factors processing method (5 values) and Harvard set (30 values, six sets per processing method, two for each SNR), making sure that no set was presented more than once. Subjects were tested first with the SSN and then with the CS background, and for each background the SNRs were used in order of increasing difficulty (high SNR first). For the practice set, speech without noise was presented first, followed by speech mixed with SSN and then speech mixed with CS, both at the medium SNR. Subjects were asked to type what they heard. The order of the stimuli for the practice set was the same for all subjects.

The typed answers were first automatically screened to exclude participles such as 'a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', and 'for' and then manually checked to correct obvious

typographical errors. Next, the number of keywords correct was automatically counted. The scores for each processing condition and SNR were averaged across sentence sets and subjects.

D. Results

The average scores are shown in Figs. 1 and 2. Error bars show ± one standard error of the mean. A within-subjects analysis of variance (ANOVA) was performed on the arcsine-transformed data with factors processing method (5 values), background type (2 values) and SNR (3 values). Mauchly's test indicated that the assumption of sphericity was violated for all factors, except for the background type, so the degrees of freedom were adjusted using the Greenhouse-Geisser correction. There were significant main effects of processing method, $F(3.389, 64.386) = 159.8, p < 0.0001$, background type, $F(1, 19) = 209.6, p < 0.0001$, and SNR, $F(1.803, 34.251) = 771.1, p < 0.0001$. There were significant interactions between processing method and background type, $F(2.754, 52.33) = 94.1, p < 0.0001$, processing method and SNR, $F(4.929, 93.646) = 29.6, p < 0.0001$, background type and SNR, $F(1.602, 30.447) = 76.3, p < 0.0001$, and processing method, background type and SNR, $F(5.501, 104.515) = 17.7, p < 0.0001$.

It should be noted that subjects were tested first with the SSN and then with the CS background, so interactions of background type with other factors could, in principle,

partly reflect order effects resulting from practice or fatigue. However, the order of testing processing methods was counterbalanced for each background type, so it seems unlikely that the large and highly significant interaction of background type with processing method resulted simply from an order effect.

Pairwise t-tests (two tailed, with Bonferroni adjustment for multiple comparisons) were used to assess the differences between processing methods for each background type. There were significant differences ($p < 0.05$) for all pairs, except between SSDRC and fSER+DRC (for both backgrounds), and between unprocessed and tSER (only for the CS background). The two processing methods incorporating DRC gave the highest scores for all SNRs, with intelligibility improvements up to 46 and 25 percentage points relative to unprocessed speech, for SSN and CS, respectively, at the low SNR. The tSER method led to small intelligibility improvements for the SSN background, but not for the CS background. The SDR method gave very good scores with the SSN, but it markedly degraded intelligibility with the CS. This outcome was not surprising since SDR was optimized for speech in SSN and many optimization-based procedures are not robust with fluctuating background sounds.

In summary, under the EQL constraint the two processing methods using DRC led to higher intelligibility than for unprocessed speech for both background types, especially for the low SNR. The tSER and SDR methods led to intelligibility improvements with the SSN

background but not with the CS background.

V. EXPERIMENT 2: HEARING-IMPAIRED SUBJECTS

A. Stimuli

The stimuli for Experiment 2 were the same as for Experiment 1, except that processing method SDR was replaced by Lombard speech (Lomb).

It was expected that the HI subjects would have more difficulty than the NH subjects, so the SNRs for unprocessed speech were increased to −4, 1 and 6 dB and −9, −6, 0 dB for the SSN and CS backgrounds, respectively. The SNRs are again denoted 'low', 'medium' and 'high'.

B. Subjects

Ten subjects (8 males) were tested. Their ages ranged from 28 to 70 years (mean = 64). All were native speakers of English. The average audiogram for the left and right ears is shown in Fig. 3. Subjects had mild-to-moderate hearing loss (thresholds between 20 and 60 dB HL) over the frequency range that is most important for speech intelligibility (0.5-4 kHz). Half (5) of the subjects did not use hearing aids. All subjects listened unaided. Stimuli were presented diotically, using the same headphones as for Experiment 1.

C. Procedure

The procedure was the same as for Experiment 1. The changes in level required to

equate the loudness of the processed and unprocessed sentences are shown in Table I.

D. Results

Average scores across subjects are shown in Figs. 4 and 5. A within-subjects ANOVA was performed on the arcsine-transformed data with factors processing method (5 values), background type (2 values) and SNR (3 values). Mauchly's test indicated that the assumption of sphericity was violated for all factors, except for background type, so the degrees of freedom were adjusted using the Greenhouse-Geisser correction. There were significant effects of processing method, $F(1.662, 14.961) = 34.6, p < 0.0001$, background type, $F(1, 9) = 17.2, p < 0.05$, and SNR, $F(1.093, 9.84) = 55.9, p < 0.0001$. There were significant interactions between processing method and SNR,
$F(3.392, 30.524) = 9.6, p < 0.0001$, background type and SNR,
$F(1.828, 16.456) = 9.6, p < 0.05$ and processing method, background type and SNR,
$F(3.443, 30.991) = 3.9, p < 0.001$. There was no significant interaction between processing method and background type ($p > 0.05$).

Pairwise t-tests (two tailed, with Bonferroni adjustment for multiple comparisons) comparing processing methods for each background type indicated significant difference ($p < 0.05$) for all pairs, except for SSDRC versus fSER+DRC, Lomb versus unprocessed, Lomb versus tSER and tSER versus unprocessed. The non-significant pairs were the same for the two backgrounds. SSDRC and fSER+DRC gave the highest scores for all SNRs and

both background types, with improvements relative to unprocessed speech up to 37 and 17 percentage points for the SSN and CS backgrounds, respectively, at the low SNR. Lombard speech led to lower intelligibility than unprocessed speech for the low and medium SNRs.

To assess whether the performance of the HI subjects was affected by the amount of their hearing loss, we derived a composite measure of performance for each HI subject by averaging scores for the low and medium SNRs; the high SNR was not included since some subjects scored close to ceiling for this SNR. This was done separately for each background type and separately for the unprocessed and SSDRC-processed speech. Composite scores were arcsine transformed for statistical analysis. We quantified the amount of hearing loss as the mean audiometric threshold for the better ear at 2 and 4 kHz (designated $HL_{2,4}$), since this has been shown to be highly correlated with the ability to understand speech in noise (Smoorenburg, 1992).

For the SSN background, there was a strong correlation between the composite score for unprocessed speech and $HL_{2,4}$ ($r = -0.81$, $p = 0.005$), indicating that greater hearing loss at 2 and 4 kHz was associated with poorer understanding of speech in noise, consistent with the finding of Smoorenburg (1992). The benefit from SSDRC processing (i.e., the difference between composite scores for the SSDRC-processed and unprocessed speech) for the SSN background was positively correlated with $HL_{2,4}$, but the correlation just failed to reach significance ($r = 0.59$, $p = 0.07$). Thus, there was a trend for subjects with greater

hearing loss to obtain more benefit from the SSDRC processing. For the CS background there was again a negative correlation between the composite score for unprocessed speech and $HL_{2,4}$, but the correlation just failed to reach significance ($r = -0.59$, $p = 0.07$). The benefit from SSDRC processing for the CS background was not significantly correlated with $HL_{2,4}$ ($r = 0.29$, $p = 0.41$).

Unrelated-samples t-tests (two-tailed) were conducted to assess whether there was a significant difference between the composite intelligibility scores for the users and the non-users of hearing aids (HA) for each background. For the SSN background, the scores did not differ significantly ($t(43.54) = 1.56, p = 0.12$), but they did for the CS background ($t(47.94) = 2.01, p = 0.05$), being higher (better) for the non-users than for the users. The average values of $HL_{2,4}$ for the non-users and the users of HA were 39 and 47.5 dB, respectively, the difference reflecting the fact that the non-user group included more subjects with small values of $HL_{2,4}$. The difference in speech scores across the two groups can probably be accounted for by the difference in $HL_{2,4}$ values. The improvement produced by SSDRC was not significantly different for the two groups for either background.

In summary, under the EQL constraint the pattern of results was similar to that obtained for the NH subjects in Experiment 1. The two processing methods using DRC led to higher intelligibility than for unprocessed speech for both background types, especially

at the low SNR. The tSER method led to small intelligibility improvements with the SSN background but not with the CS background. The Lombard speech did not give higher intelligibility than for unprocessed speech for either background type.

VI. DISCUSSION

The average intelligibility scores are plotted against the physical SNRs after loudness equalization for both experiments in Figs. 6-9. Second-order polynomials were fitted to the data to ease visual comparison of scores for the different processing methods. Note that the fits are perfect, because three parameters were used to fit three points for each method. The figures show that the ranking of processing methods as depicted in Figs. 1-5 was preserved. In other words, under both EQR and EQL constraints, SSDRC and fSER+DRC gave clear intelligibility enhancements for speech in both background types, whereas the other methods either yielded benefits only in the SSN, or yielded no clear benefits.

It is noteworthy that Lombard speech gave significantly better scores than unprocessed speech in the Hurricane Challenge under the EQR constraint when tested using NH subjects (Cooke et al., 2013a,b). This is consistent with what is shown in Figs. 8 and 9 for the HI subjects tested here; when the scores were plotted against physical SNR, the scores for the Lombard speech were higher than those for unprocessed speech. However, under the EQL constraint, the Lombard speech did not yield higher scores than the unprocessed speech. Evidently, the reduction in level required to equate the loudness of

the Lombard speech to that of the unprocessed speech offset the beneficial effect observed under the EQR constraint. This shows that different outcomes can be obtained under the EQR and EQL constraints.

Speech reception thresholds (SRTs) were estimated from the fitted polynomial functions to assess the differences between the two groups of subjects. The SRT50 is defined as the SNR at which a subject achieved 50% keywords correct. This was estimated only for the unprocessed speech, as scores did not fall as low as 50% correct for some of the processing conditions. For the NH subjects, the average SRT50s were -5.7 dB and -16.7 dB for the SSN and CS backgrounds, respectively. The average SRT50s for the HI subjects were -4 dB and -10.6 dB. Hence, the increase in SRT associated with hearing loss was 1.7 dB for the SSN and 6.1 dB for the CS. The larger effect of hearing loss for the CS than for the SSN is consistent with results in the literature using both speech and artificial sounds as the fluctuating background (Duquesnoy, 1983; Baer and Moore, 1994; Peters et al., 1998; Moore, 2003).

Although the SRT is often defined as the SNR required for 50% correct words or sentences, in practical situations higher scores would be needed to achieve effective communication. Hence, SRT80 values corresponding to 80% keywords correct were also calculated. This had the added benefit of allowing SRT values to be determined for all conditions, with only minor extrapolation. The fitted polynomial functions were used to

determine the SRT80 values, with extrapolation where needed. For the NH subjects, the SRT80 values for unprocessed speech were -1.6 dB for the SSN and -9.9 dB for the CS. Hence, performance was considerably better with the fluctuating background. For the HI subjects, the SRT80 values for unprocessed speech were 0.2 dB for the SSN and -0.2 dB for the CS. Evidently, the HI subjects obtained little benefit from the fluctuations in the CS background, which is consistent with previous studies using both speech and non-speech fluctuating backgrounds (Festen and Plomp, 1990; Peters et al., 1998). SRT80 values for the unprocessed stimuli were 1.8 dB and 9.7 dB lower for the NH than for the HI subjects for the SSN and CS backgrounds, respectively.

The changes in SRT80 for each processing condition relative to the SRT80 value for unprocessed speech are shown in Figs. 10-11 for the NH and HI subjects, respectively. The two processing methods incorporating DRC gave improvements in SRT80 that were 3 dB or more for both subject groups and both types of background. Remarkably, the improvements for the HI group were between 6 and 7 dB for the CS background. This indicates that SSDRC and fSER+DRC processing can markedly decrease the SNR required for adequate intelligibility for HI subjects.

VII. CONCLUSIONS

Several NLE algorithms were evaluated under the constraint of equal loudness. Subjects with normal hearing and with mild-to-moderate hearing loss were tested, using

steady speech-shaped noise (SSN) and a competing speaker (CS) as backgrounds. Two algorithms incorporating dynamic range compression led to substantial intelligibility improvements for both subject groups and both types of background. The algorithms that did not incorporate dynamic-range compression either led to improved intelligibility only for the SSN background or led to no clear benefit for either background. For the hearing-impaired group, Lombard speech led to improved intelligibility when compared to unprocessed speech at the same physical RMS level, but not when compared at equal loudness, showing that the equal-level and equal-loudness constraints can lead to different outcomes.

Acknowledgments

The authors would like to thank the LISTA Project for the evaluation corpus and Petko Petkov for providing the SDR stimuli for Experiment 1. The work of author BCJM was supported by the Engineering and Physical Sciences Research Council, UK (grant number RG78536). We thank Michael Stone and an anonymous reviewer for helpful comments.

REFERENCES

Baer, T., and Moore, B.C.J., **(1994)**. "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," J. Acoust. Soc. Am., **95**, 2277-2280.

Baer, T., Moore, B.C.J., and Gatehouse, S., **(1993)**. "Spectral contrast enhancement of

speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev., **30**, 49–72.

Bhattacharya, A., Vandali, A., and Zeng, F.G., **(2011)**. "Combined spectral and temporal enhancement to improve cochlear-implant speech perception," J. Acoust. Soc. Am., **130**, 2951–2960.

Bradlow, A., and Bent, T., **(2002)**. "The clear speech effect for non-native listeners," J. Acoust. Soc. Am., **112**, 272–284.

Cooke, M., and Lecumberri, M., **(2012)**. "The intelligibility of Lombard speech for non-native listeners," J. Acoust. Soc. Am., **132**, 1120–1129.

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

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

Davis, A., **(1995)**. *Hearing in Adults* (Whurr, London), 1–1026.

Dreher, J.J., and O'Neill, J.J., **(1957)**. "Effects of ambient noise on speaker intelligibility for words and phrases," J. Acoust. Soc. Am., **29**, 1320–1323.

Duquesnoy, A.J., **(1983)**. "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons," J. Acoust. Soc. Am., **74**, 739–743.

Erro, D., Zorilă, T.C., and Stylianou, Y., **(2014)**. "Enhancing the intelligibility of statistically generated synthetic speech by means of noise-independent modifications," IEEE/ACM Trans. Audio Speech Lang. Process., **22**, 2101–2111.

Ferguson, S.H., and Kewley-Port, D., **(2002)**. "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am., **112**, 259–271.

Festen, J.M., and Plomp, R., **(1990)**. "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am., **88**, 1725–1736.

Fletcher, H., **(1938)**. "Loudness, masking and their relation to the hearing process and the problem of noise measurement," J. Acoust. Soc. Am., **9**, 275–293.

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.

Godoy, E., Koutsogiannaki, M., and Stylianou, Y., **(2014)**. "Approaching speech intelligibility enhancement with inspiration from Lombard and clear speaking styles," Comput. Speech Lang., **28**, 629–647.

Hazan, V., and Simpson, A., **(1998)**. "The effects of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," Speech Commun., **24**, 211-226.

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

Junqua, J.C., **(1993)**. "The Lombard reflex and its role on human listeners and automatic speech recognizers," J. Acoust. Soc. Am., **93**, 510–524.

Lombard, É., **(1911)**. "Le signe de l'élévation de la voix," Ann. Malad. Oreille., **37**, 101–119.

Lu, Y., and Cooke, M., **(2008)**. "Speech production modifications produced by competing talkers, babble, and stationary noise," J. Acoust. Soc. Am., **124**, 3261–3275.

Moore, B.C.J., Wojtczak, M., and Vickers, D.A., **(1996)**. "Effect of loudness recruitment on the perception of amplitude modulation," J. Acoust. Soc. Am., **100**, 481–489.

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., **(2003)**. "Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms," Speech Commun., **41**, 81–91.

Niederjohn, R.J., and Grotelueschen, J.H., **(1976)**. "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," IEEE Trans. Acoust. Speech Signal Process., **24**, 277–282.

Oxenham, A.J., Simonson, A.M., Turicchia, L., and Sarpeshkar, R., **(2007)**. "Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing," J. Acoust. Soc. Am., **121**, 1709–1716.

Payton, K.L., Uchanski, R.M., and Braida, L.D., **(1994)**. "Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing," J. Acoust. Soc. Am., **95**, 1581–1592.

Peters, R.W., Moore, B.C.J., and Baer, T., **(1998)**. "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am., **103**, 577-587.

Petkov, P., and Kleijn, W., **(2015)**. "Spectral dynamics recovery for enhanced speech intelligibility in noise," IEEE/ACM Trans. Audio Speech Lang. Process., **23**, 327–338.

Picheny, M.A., Durlach, N.I., and Braida, L.D., **(1985)**. "Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech," J. Speech Hear. Res., **28**, 96–103.

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

Sauert, B., and Vary, P., **(2006)**. "Near end listening enhancement: Speech intelligibility improvement in noisy environment," in *Proc. ICASSP*, 493–496.

Schepker, H., and Rennies, J., **(2015)**. "Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index," J. Acoust. Soc. Am., **138**, 2692–2706.

Skowronski, M.D., and Harris, J.G., **(2006)**. "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," Speech Commun., **48**, 549–558.

Smoorenburg, G.F., **(1992)**. "Speech reception in quiet and in noisy conditions by

individuals with noise-induced hearing loss in relation to their tone audiogram," J. Acoust. Soc. Am., **91**, 421–437.

Steinberg, J.C., and Gardner, M.B., **(1937)**. "The dependence of hearing impairment on sound intensity," J. Acoust. Soc. Am., **9**, 11–23.

Summers, W.V., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I., and Stokes, M.A., **(1993)**. "Effects of noise on speech production: acoustic and perceptual analyses," J. Acoust. Soc. Am., **84**, 917–928.

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

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

Zorilă, T.C., and Stylianou, Y., **(2014)**. "On spectral and time domain energy reallocation for speech-in-noise intelligibility enhancement," in *Proc. Interspeech* (Singapore), 2050–2054.

Zorilă, T.C., and Stylianou, Y., **(2015)**. "A fast algorithm for improved intelligibility of

speech-in-noise based on frequency and time domain energy reallocation," in *Proc. Interspeech* (Dresden, Germany), 60–64.

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

Zorilă, T.C., Stylianou, Y., Flanagan, S., and Moore, B.C.J., **(2016)**. "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., **140**, 402–408.

Tables

TABLE I. Mean, minimum, maximum and standard deviation (Std) values of the changes in level needed to equate the loudness of the processed sentences to that of the unprocessed sentences for Experiments 1 and 2.

Processing	Mean (dB)	Min (dB)	Max (dB)	Std (dB)
SSDRC	-0.4	-6.4	3.8	1.6
fSER+DRC	0.2	-5.3	4.2	1.6
tSER	-2.6	-6	-0.1	0.8
SDR	-1.6	-7.1	4.9	1.6
Lomb	-2.3	-7.1	2.1	1.4

Figure Captions

FIG. 1. Average keyword recognition scores for Experiment 1 for the steady speech-shaped noise (SSN) background for each SNR. Error bars show ± one standard error.

FIG. 2. As Fig. 1 but for the competing speech (CS) background.

FIG. 3. Average audiograms for the left and right ears of the subjects participating in Experiment 2.

FIG. 4. Average keyword recognition scores for Experiment 2 for the steady speech-shaped noise (SSN) background for each SNR. Error bars show ± one standard error.

FIG. 5. As Fig. 4 but for the CS background.

FIG. 6. Intelligibility as a function of physical SNR after loudness equalization for Experiment 1 and the SSN background. Data points were fitted using second-order polynomials for visual guidance.

FIG. 7. As Fig. 6 but for the CS background.

FIG. 8. Intelligibility as a function of physical SNR after loudness equalization for

Experiment 2 and the SSN background. Data points were fitted using second-order polynomials for visual guidance.

FIG. 9. As Fig. 8 but for the CS background.

FIG. 10. Differences between SRT80 values for processed speech and unprocessed speech for normal-hearing subjects (Experiment 1).

FIG. 11. Differences between SRT80 values for processed speech and unprocessed speech for hearing-impaired subjects (Experiment 2).