INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

On the quality and intelligibility of noisy speech processed for near-end listening enhancement

Tudor-Cătălin Zorilă[1], *Yannis Stylianou*[1,2]

[1]Toshiba Cambridge Research Laboratory, United Kingdom
[2]Computer Science Department, University of Crete, Greece
{catalin.zorila,yannis.stylianou}@crl.toshiba.co.uk

Abstract

Most current techniques for near-end speech intelligibility enhancement have focused on processing clean input signals, however, in realistic environments, the input is often noisy. Processing noisy speech for intelligibility enhancement using algorithms developed for clean signals can lower the perceptual quality of the samples when they are listened in quiet. Here we address the quality loss in these conditions by combining noise reduction with a multi-band version of a state-of-the-art intelligibility enhancer for clean speech that is based on spectral shaping and dynamic range compression (SSDRC). Subjective quality and intelligibility assessments with noisy input speech showed that: (a) In quiet near-end conditions, the proposed system outperformed the baseline SSDRC in terms of Mean Opinion Score (MOS); (b) In speech-shaped near-end noise, the proposed system improved the intelligibility of unprocessed speech by a factor larger than three at the lowest tested signal-to-noise ratio (SNR) however, overall, it yielded lower recognition scores than the standard SSDRC.

Index Terms: speech intelligibility enhancement, noisy speech, speech quality

1. Introduction

The intelligibility of speech is decreased when it is presented in a noisy environment. Several techniques have been developed to process speech so that its intelligibility is improved in noise [1]. This type of processing is called “near-end listening enhancement” (NLE) and it has potential applications in public-address systems, telephony or teleconferencing.

There are two major types of NLE systems. Some are applying modifications inspired from previous intelligibility studies (e.g., Lombard speech) [2–7], and others are enhancing the speech using transformations that would optimize some objective intelligibility measures well correlated with human perception [8, 9]. A method based on spectral shaping and dynamic range compression yielded the state-of-the-art performance at the Hurricane Challenge [1]; there, the participants were asked to boost the intelligibility of clean speech in noise under the constraint of equal root mean square (RMS) before and after modifications. Recently, SSDRC was also tested under the constraint of equal-loudness both with normal-hearing and hearing-impaired [10]. The results were similar to those reported earlier.

The standard SSDRC was designed to process clean speech [11] however, in realistic conditions, the input is often noisy. Although typical SNR levels (e.g., 10-20 dB) do not degrade intelligibility, they can have a significant effect on the perceptual quality of speech processed for NLE, particularly in the absence of near-end noise. Denoising techniques can improve the quality in these conditions, but in some applications the increase may not be sufficient. A rather simple noise-tolerant version of SSDRC combined with noise reduction was already proposed in Griffin et al. [12], however the experimental validation was limited to objective metrics only. Here an improved version of noise-tolerant SSDRC was combined with noise reduction and both intelligibility and quality assessments were made subjectively. The proposed approach is denoted multiband SSDRC (MBSSDRC).

The paper is organized as follows. Section 2 first summarizes the standard SSDRC and then details the proposed MBSSDRC system, Section 3 describes the evaluation methodology, the results are discussed in Section 4, and the conclusions are presented in Section 5.

2. Methods

2.1. Standard SSDRC

SSDRC has improved the intelligibility of speech in noise by redistributing energy from the spectro-temporal regions where the SNR was high to those where the SNR was lower. The energy redistribution was performed in a perceptually meaningful way and it had two stages. During the first one (spectral shaping, SS), the SNR at medium and high frequencies was increased by transferring energy from the frequencies below 500 Hz. That was done in three steps. One was increasing the spectral contrast by sharpening the formants, another one was flattening the spectral tilt, and the last one was boosting the frequency range from 1 to 4 kHz. The first two steps were adapted to the voicing nature of the current frame. During the second stage of energy reallocation (dynamic range compression, DRC), the envelope of sounds was statically and dynamically compressed so that its peak-to-RMS values were reduced. Consequently, under constant RMS constraint before and after SSDRC, the level of speech segments more prone to noise masking (fricatives, nasals, and stops) was increased, and that of loud segments (vowels) was decreased. The standard SSDRC was designed to process clean speech. Its simplified block diagram is shown in Fig. 1. The Framing was the short-term signal analysis, DTFT and IDTFT were the direct and inverse Discrete Time Fourier Transform, and OLA was the overlap-and-add waveform synthesis.

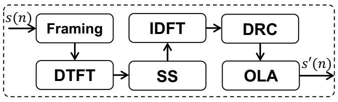

Figure 1: *Block diagram of standard SSDRC [11].*

http://dx.doi.org/10.21437/Interspeech.2017-1225

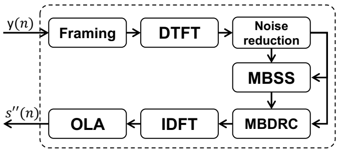

Figure 2: *Block diagram of multiband SSDRC (proposed).*

2.2. Proposed multiband SSDRC

The block diagram is depicted in Fig.2. Apart from the short-term spectral analysis and synthesis blocks, MBSSDRC had a denoising stage, and two multiband expansions of SS and DRC, denoted MBSS and MBDRC, respectively. Both MBSS and MBDRC were applied in the frequency domain and they were adapted to the local SNR estimated during the noise reduction step.

The additive noise model was used for the input signal,

$$y(n) = s(n) + d(n), \tag{1}$$

where $s(n)$ was the clean speech signal and $d(n)$ was the distortion signal.

2.2.1. *Framing and spectral analysis*

Frames of length 20-ms were extracted every 10-ms,

$$y_m(n) = y(n)w(n - mR), \tag{2}$$

where $w(n)$ was the square root Hann window, R was the frame rate and m was the frame index. The DTFT was applied on each frame,

$$Y_m(\omega) = \sum_{n=-\infty}^{\infty} y_m(n)e^{-j\omega n}. \tag{3}$$

Alternatively,

$$Y_m(\omega) = M_y(\omega, m)e^{j\phi_y(\omega,m)}, \tag{4}$$

or similarly,

$$Y_m(\omega) = M_s(\omega, m)e^{j\phi_s(\omega,m)} + M_d(\omega, m)e^{j\phi_d(\omega,m)}, \tag{5}$$

where M_y, M_s and M_d were the magnitude spectra of distorted and clean speech, and of background noise, respectively. ϕ_y, ϕ_s and ϕ_d were the phase spectra.

For SNRs greater than 8 dB it can be shown that the noisy phase ϕ_y is a good estimate of the phase of the clean signal, ϕ_s [13]. In this work the SNR level of the input signal has been fixed to 10 dB, therefore only the magnitude spectrum of the clean speech was estimated, while the phase information was copied from the noisy input.

2.2.2. *Noise reduction*

The optimal log-magnitude spectrum estimator of Ephraim and Malah [14] (LMS) was chosen for denoising since it was shown to outperform other similar noise reduction techniques in terms of speech quality [13]. One function of this stage was to reduce the noise from the input, and the other one was to produce an adaptation signal for the speech intelligibility enhancement that would improve the perceptual quality of the processed speech.

The optimal estimate for the log-magnitude spectrum of clean speech was sought by minimizing the following mean-square error in a statistical sense:

$$J_{LMS} = E\left[\left(\log \hat{M}_s - \log M_s\right)^2\right], \tag{6}$$

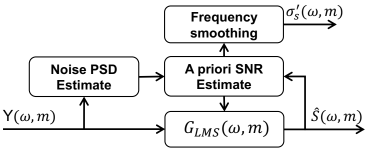

Figure 3: *Simplified block diagram of the noise reduction stage of MBSSDRC.*

where $\log \hat{M}_s$ and $\log M_s$ were the estimated and the true log-spectral magnitudes of clean speech, respectively.

It was shown in [14] that

$$\hat{M}_s(\omega, m) = G_{LMS}(\omega, m)M_y(\omega, m), \tag{7}$$

or with the noisy phase,

$$\hat{S}(\omega, m) = G_{LMS}(\omega, m)Y(\omega, m), \tag{8}$$

where

$$G_{LMS}(\omega, m) = \frac{\sigma_s(\omega, m)}{1 + \sigma_s(\omega, m)} \exp\left(0.5 \int_{\nu(\omega,m)}^{\infty} \frac{e^{-t}}{t} dt\right). \tag{9}$$

Above,

$$\sigma_s(\omega, m) = \frac{M_s^2(\omega, m)}{M_d^2(\omega, m)}, \tag{10}$$

was the a priori SNR,

$$\nu(\omega, m) = \frac{\sigma_s(\omega, m)}{1 + \sigma_s(\omega, m)}\sigma_y(\omega, m), \tag{11}$$

and

$$\sigma_y(\omega, m) = \frac{M_y^2(\omega, m)}{M_d^2(\omega, m)} \tag{12}$$

was the aposteriori SNR. In this implementation, the upper value of σ_y was limited to 60dB.

The power spectrum density (PSD) of the background was estimated from the first 100-ms of input, assuming that there was no speech activity during that interval. σ_s was updated over time using the decision-directed approach,

$$\sigma_s(m) = \alpha_{nr}\sigma_s(m - 1) + (1 - \alpha_{nr}) \max\left[\sigma_y(m) - 1, \epsilon\right], \tag{13}$$

where the smoothing factor $\alpha_{nr} = \exp\left(-2.2R/T_{\alpha_{nr}}\right)$, R was the frame rate, $T_{\alpha_{nr}}$ was the time constant for the smoothing (2 seconds), and ϵ was -40dB. The adaptation signal for the intelligibility enhancement (σ_s') was computed by passing $\sigma_s(\omega)$ through a zero-phase moving average circuit whose averaging constant was 0.2 at 30 Hz frequency resolution. Simplified block diagram of the noise reduction stage is shown in Fig. 3.

2.2.3. *Multiband spectral shaping*

MBSS had five stages (Fig. 4). The first three had similar functions and implementations as those of standard SS. H_1 sharpened the formants, H_2 flattened the spectral tilt, and H_3 pre-emphasized the mid frequency range. A similar voicing index as in [11] was used to adapt the first two operations. The last two stages of MBSS have performed SNR-adaptive gain thresholding and fixed time smoothing.

Letting

$$G_{ss}'(\omega, m) = H_1(\omega, m)H_2(\omega, m)H_3(\omega, m), \tag{14}$$

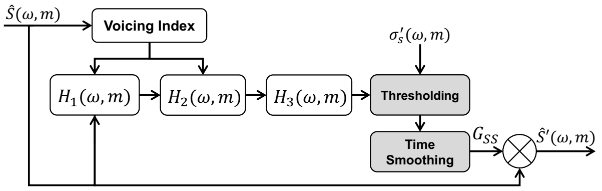

Figure 4: *Block diagram of multiband spectral shaping.*

then after thresholding,

$$G_{ss}(\omega,m) = \begin{cases} G'_{ss}(\omega,m), & if\ \sigma'_s(\omega,m) > \Sigma_{ss} \\ 1, & otherwise \end{cases} . \quad (15)$$

Finally, with time smoothing,

$$G_{ss}(m) = \alpha_{ss} G_{ss}(m-1) + (1-\alpha_{ss}) G_{ss}(m). \quad (16)$$

Above, Σ_{ss} was 0 dB, and the time constant for α_{ss} was set to 100-ms.

2.2.4. Multiband dynamic range compression

MBDRC had three stages. First, the full-band (time domain) DRC gains were computed at the frame rate, then SNR-adaptive gain thresholding was applied in each frequency band, and finally the gains were smoothed over frames (Fig. 5). The first stage had four steps. One was reconstructing the waveform from the spectral coefficients; another one was applying full wave rectification on the previous waveform, extracting 30-ms frames at every 30-ms, picking the maximum value in each frame, and re-sampling the sequence of maxima at the original frame rate of MBSSDRC; at a third stage, the level of the envelope $e(m)$ computed previously was statically compressed using the strategy described in [11], yielding the gains $g(m)$; finally, the dynamics of $g(m)$ were enhanced to have a faster response to on-sets and a slower one to off-sets,

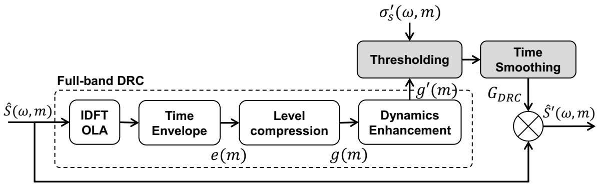

Figure 5: *Block diagram of multiband DRC.*

$$g'(m) = \begin{cases} a_r g'(m-1) + (1-a_r) g(m), & g(m) < g'(m-1) \\ a_a g'(m-1) + (1-a_a) g(m), & g(m) \geq g'(m-1) \end{cases} , \quad (17)$$

where a_r and a_a corresponded to 40-ms and 1-ms release and attack time constants, respectively.

The MBDRC gains were applied only to those bands that exceeded a threshold SNR,

$$G_{drc}(\omega,m) = \begin{cases} g'(m), & if\ \sigma'_s(\omega,m) > \Sigma_{drc} \\ 1, & otherwise \end{cases} , \quad (18)$$

where Σ_{drc} was 10 dB. The last step was time smoothing,

$$G_{drc}(m) = \alpha_{drc} G_{drc}(m-1) + (1-\alpha_{drc}) G_{drc}(m). \quad (19)$$

The time constant for α_{drc} was 5-ms. The waveforms and spectrograms of noisy speech processed using the standard SSDRC or the proposed MBSSDRC are depicted in Figs. 6-7.

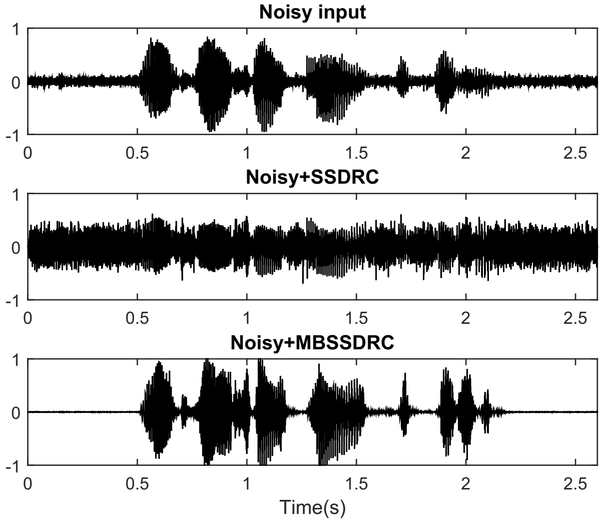

Figure 6: *Example of noisy speech (speech shaped noise, 10 dB SNR) processed using the standard SSDRC (middle panel) or the MBSSDRC (lower panel).*

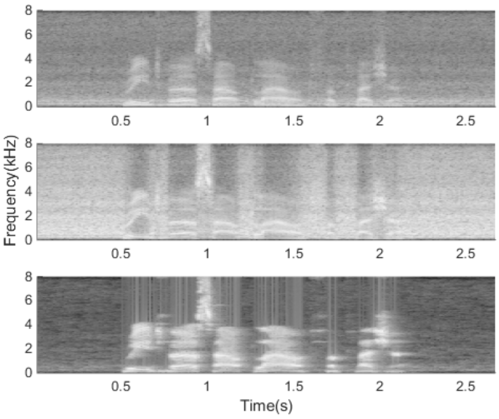

Figure 7: *Spectrograms of signals in Fig. 6. The upper panel is for the unprocessed noisy speech, the middle one is for SSDRC, and the lower one is for MBSSDRC.*

3. Evaluation & Results

Two subjective evaluations were conducted independently to assess the intelligibility (Experiment A) and the quality (Experiment B) of speech processed using MBSSDRC. The clean speech sounds were the same as for the Hurricane Challenge [1], i.e., recordings of phonetically balanced Harvard sentences [15] uttered by a native English male. The input noisy speech sounds were formed by mixing the previous samples with a stationary type of noise; real street noise extracted from the Aurora-4 database [16] was added for Experiment A, and speech-shaped noise (SSN) was used for Experiment B. The input SNR level in both cases was 10 dB. The sampling rate of all stimuli was 16 kHz.

Three speech styles were tested in Experiment A. One was unprocessed noisy speech (NSY_PLAIN), and the other two were speech processed using either MBSSDRC (NSY_MBSSDRC) or using noise reduction plus the standard

SSDRC (NSY_NR+SSDRC). The processed and unprocessed stimuli were equated in RMS and they were listened to in a SSN background at -9 and -4 dB output SNR. The SSN was generated by filtering white noise so that it had the long-term average spectra of a female talker [1].

Four speech styles were tested in Experiment B. One was unprocessed clean speech (CLN_PLAIN), another one was noisy speech processed using standard SSDRC (NSY_SSDRC), and the last two were NSY_MBSSDRC and NSY_NR+SSDRC. For this experiment, all stimuli were equated in loudness as it was suggested in [17] and they were presented in quiet conditions (no additional background noise was added after processing).

Twenty participants (9 males) were tested in Experiment A, and another ten (5 males) took part in Experiment B. Their ages ranged from 19 to 49 yr (mean=23). All were native British English speakers having normal hearing, and they were paid for participation. The evaluations were conducted using headphones in sound proof booths at the Center for Speech Technology Research (University of Edinburgh, UK). Subjects responded via a Matlab graphical interface for both tests. For Experiment A, they heard each stimulus only once and were asked to type what they heard on a keyboard. The stimuli order followed a Latin square design with factors processing method and Harvard set number, so that no set was presented more than once per participant. Each Harvard set had 10 sentences. For Experiment B, participants were asked to rate the quality of each sample on a five points scale (1 was the lowest and 5 was the highest). They were able to listen to each stimulus more than once. The stimuli order followed a Latin square design with the processing method as the only factor. The same Harvard sentences (set number 4) were repeated for all methods at Experiment B.

The average percentage of correctly recognized keywords across participants for Experiment A is depicted in Fig. 8, and the average quality scores for Experiment B are depicted in Fig. 9.

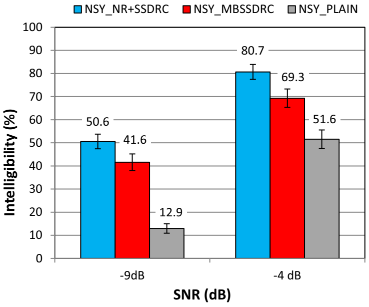

Figure 8: *Average keyword recognition rates for the intelligibility experiment (speech-shaped noise background).*

4. Discussion

Repeated-measures analysis of variance (ANOVA) was performed on the scores from Experiments A with factors processing method (3) and SNR (2). There were significant main effects of processing method ($F(2, 38) = 146$, $p < 0.001$) and SNR ($F(1, 19) = 222$, $p < 0.001$). Pairwise t-tests (two tailed, with Bonferroni adjustment for multiple comparisons) were used to check the differences between processing methods. All pairs had significant differences ($p < 0.001$). Another ANOVA was performed for Experiment B with the processing method (4) as the only factor. The effect of processing method was significant ($F(3, 27) = 212$, $p < 0.001$). Similarly, all pairs of processing methods had significant differences ($p < 0.01$).

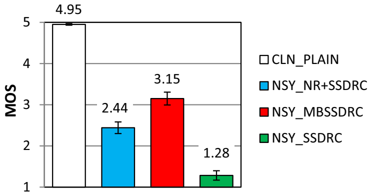

Figure 9: *Average MOS scores for the quality experiment (quiet listening conditions).*

In Experiment A, as expected, NSY_NR+SSDRC yielded the highest intelligibility gains, followed by NSY_MBSSDRC. One possible explanation for this result could be that the SNR thresholds for MBSSDRC were set too high, so only part of the speech cues had their audibility enhanced when they were presented in noise. These thresholds are however important for improving the perceptual quality in the absence of near-end noise, as the results from Experiment B seem to indicate. There, NSY_MBSSDRC had an important MOS benefit over NSY_NR+SSDRC. As the input noises used for Experiments A and B were different, direct comparisons between the intelligibility and the quality scores cannot be made. However, a more recent quality test under the same type of input noise as for Experiment A, confirmed the ranking depicted in Fig. 9 up to a scaling factor.

The trade-off between the speech understanding in noise and the speech quality in quiet will be explored in a future work. One possible improvement of MBSSDRC could be to jointly adapt it to the noise at the transmitter and receiver's end; that would optimize the intelligibility when the noise level at the receiver is strong, and would improve the quality when the noise at the receiver does not impair speech understanding. A preliminary experiment with a fluctuating type of near-end noise (competing speaker) showed that neither standard SSDRC nor MBSSDRC had any significant effects on the intelligibility of noisy input speech. This could be another improvement to explore in the future.

5. Conclusions

An extension of a state-of-the-art speech in noise intelligibility enhancement system based on spectral shaping and dynamic range compression (originally designed to work with clean signals) was suggested in this paper. The aim was to improve the perceptual quality of the processed speech presented in a quiet listening environment when the input signal is noisy. Noise reduction and multiband intelligibility enhancement adapted to the local SNR were suggested in this context. The quality and the intelligibility of the processed noisy speech were evaluated in two separate trials. It was found that: (a) In quiet listening conditions, the proposed system improved the MOS of the baseline SSDRC system; (b) In speech-shaped near-end noise, the proposed system boosted the intelligibility of the unprocessed speech by a factor larger than three at the lowest tested SNR, however it yielded lower intelligibility gains than the standard SSDRC.

6. 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 Communication*, no. 55, pp. 572–585, 2013.

[2] J. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Commun.*, vol. 20, pp. 13–22, 1996.

[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] T. 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] 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.

[6] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from lombard and clear speaking styles," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 629–647, 2014.

[7] 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.

[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] 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.

[10] T. Zorilă, Y. Stylianou, S. Flanagan, and B. Moore, "Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 189–196, 2017.

[11] T. 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.*, vol. 24, no. 10, pp. 1808–1818, 2016.

[12] A. Griffin, T. Zorilă, and Y. Stylianou, "Improved face-to-face communication using noise reduction and speech intelligibility enhancement," in *Proc. ICASSP*, 2015, pp. 5103–5107.

[13] P. Loizou, *Speech Enhancement: Theory and Practice (second edition).* CRC Press, 2013.

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE T. Acoust. Speech*, vol. 33, no. 2, pp. 443–445, 1985.

[15] E. Rothauser, W. Chapman, N. Guttman, H. Silbiger, M. Hecker, G. Urbanek, K. Nordby, and M. Weinstock, "Recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[16] N. Parihar and J. Pircone, "Distributed speech recognition front-end large vocabulary continuous speech recognition evaluation," Tech. Rep., 2002.

[17] T. Zorilă, Y. Stylianou, S. Flanagan, and B. 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.*, vol. 140, no. 1, pp. 402–408, 2016.