



Analysis and Extension of a Near-End Listening Enhancement Method Based on Long-Term Fractile Noise Statistics

Filippo Villani¹, Wai-Yip Chan², Zheng-Hua Tan¹, Jan Østergaard¹, Jesper Jensen^{1,3}

¹Dept. of Electronic Systems, Aalborg University, Denmark

²Dept. Electr. and Computer Eng., Queen's University, Canada

³Oticon A/S, Denmark

fv@es.aau.dk, chan@queensu.ca, zt@es.aau.dk, jo@es.aau.dk, jesj@demant.com

Abstract

This paper addresses the problem of near-end listening enhancement (NELE), where a clean speech signal is modified prior to playback and under an energy constraint to improve intelligibility in noise. We analyze a recently proposed NELE method, optimized using a Speech Intelligibility Index that has been modified to incorporate temporal aspects of the noise via long-term fractile noise statistics. Specifically, we explain the energy allocation strategy adopted by the algorithm, and show that, in contrast to many existing methods, the spectral energy distribution of the modified speech is a function of that of the background noise, but not that of the input speech. Our simulation experiments show that this simple method outperforms well-established spectral shaping NELE methods. In addition, we extend the algorithm by appending an off-the-shelf dynamic range compressor, and show that it performs generally better than state-of-the-art methods for NELE.

Index Terms: near-end listening enhancement, speech intelligibility

1. Introduction

In environments with competing sound sources, speech intelligibility may be severely compromised. To overcome this problem, humans modify their speaking style to produce speech that is easier to understand in the noise background. This is a complex phenomenon known as the Lombard effect [1]. The characteristics of this speaking behavior can vary widely between speakers, but some of the most common features have been identified, e.g., an increase of the overall speech energy, a reduction of the spectral tilt due to an increase of energy in mid- and high-frequency regions, etc. [2–4].

Studies on the Lombard effect have served as inspiration for developing algorithms that attempt to solve the near-end listening enhancement (NELE) problem, i.e., the problem of improving speech intelligibility by modifying an available clean speech signal prior to playback in a known noise background. The heuristics derived from these studies have led to the development of systems that mimic certain features of Lombard speech, resulting in significant gains in speech intelligibility compared to neutrally-produced speech [5–8]. In particular, SSDRC [5] processes clean speech using a set of filters for spectral shaping (SS) and a dynamic range compressor (DRC) that boosts transient speech components at the expense of sonorant and loud components. It is remarkable that SSDRC, despite being agnostic to background noise (see Sec. 2.3), achieves state-of-the-art performance in a wide range of acoustic scenarios [9, 10].

Another class of NELE methods conceptualizes the Lombard effect as an auditory feedback mechanism that adjusts speech production based on the perception of both the envi-

ronment and the speech itself [11]. These methods modify the speech signal to optimize intelligibility metrics by taking the noise background into account, using classic signal processing techniques [12–15] and more recently deep learning [16, 17].

In this paper, we analyze a recently proposed spectral shaping method for NELE [15], named OptFractASII, because it optimizes a modification of the Approximated Speech Intelligibility Index (ASII) [13] which uses fractile statistics of the noise, which better reflects the human ability to glimpse speech in fluctuating noise. In particular, we show that a) the long-term average spectral energy (LTASE) in each sub-band of OptFractASII-processed speech is a function of the background noise spectrum, but *not* of the input speech spectrum; b) the optimal solution of OptFractASII can be explained from a water-filling perspective, by considering the per-sub-band derivative of OptFractASII as the exchange rate at which speech energy is converted into ASII; c) sub-bands for which the ASII exchange rate is too low (typically due to the presence of significant noise power) are allocated zero speech energy.

In a series of simulations, we show that OptFractASII outperforms well-established spectral shaping NELE baselines. Furthermore, we show that by appending a simple off-the-shelf DRC stage that allows the model to shift energy across time, OptFractASII achieves state-of-the-art performance in terms of ESTOI [18] and STGI [19] in a wide range of noise backgrounds, improving over SSDRC [5] as well as more recent deep learning-based systems, such as [16].

2. OptFractASII

In this section, we briefly review OptFractASII [15], a recently proposed spectral shaping method for NELE, before analyzing how it allocates the available speech energy budget to different sub-bands. Finally, we compare spectral characteristics of OptFractASII- and SS-processed speech.

2.1. OptFractASII review

As a preparation for the analysis in Sec. 2.2, we briefly review OptFractASII [15]. Let $X(j, l)$ and $V(j, l)$ be the time-frequency representations of a clean speech and noise signal, respectively, obtained using an auditory filterbank with gammatone filters of approximately 1 ERB bandwidth and central frequencies equally spaced on the ERB scale. The indices j and l denote a sub-band- and a time-index, respectively. Let the LTASE within sub-band j be denoted as $\sigma_{X_j}^2 = 1/L_X \sum_{l=1}^{L_X} |X(j, l)|^2$ with L_X being the total number of frames. Let us define the ϕ -fractile noise power as the value $\tilde{\sigma}_{V_j}^2$ for which

$$\text{Prob} \left\{ |V(j, l)|^2 \leq \tilde{\sigma}_{V_j}^2 \right\} = \phi, \quad 0 < \phi < 1, \quad \forall j, l. \quad (1)$$

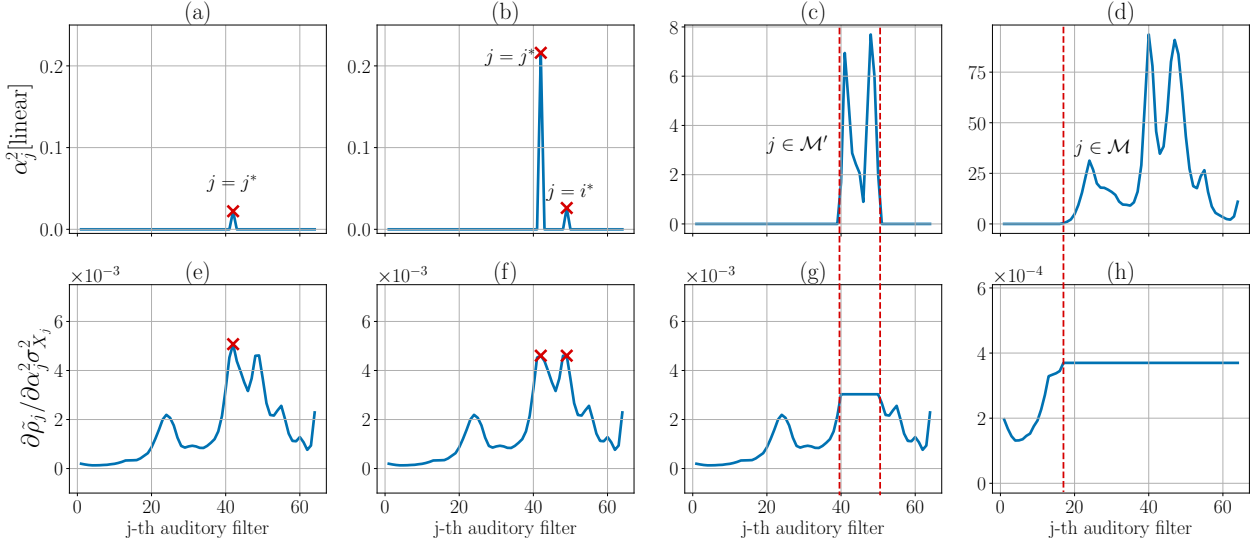


Figure 1: Sub-band gains (a-d) and derivatives of the contribution of the j -th sub-band to ASII with respect to the speech energy (e-h) for different values of the speech energy budget r . The speech energy budget increases from left to right. Note that for visualization purposes, the y-axis scale of subplots (c)-(d) is different from (a)-(b), and (h) is different from (e)-(g).

$\tilde{\sigma}_{V_j}^2$ is evaluated by finding the energy value $|V(j, l)|^2$ for which a fraction ϕ of speech frames in sub-band j have lower energy than the remaining frames.

Let $\tilde{\xi}_j = \sigma_{X_j}^2 / \tilde{\sigma}_{V_j}^2$ denote the within-band fractile SNR. Then, OptFractASII finds the optimal sub-band gain values α_j^2 for the given noise condition by maximizing the ASII computed using the within-band fractile SNR $\tilde{\xi}_j$ rather than the conventional SNR [15], subject to a constraint on the total speech energy available for processing r and a non-negativity constraint on the gain values:

$$\begin{aligned} \alpha_j^2 &= \arg \max_{\alpha_j^2 \geq 0} \sum_{j=1}^J \gamma_j \frac{\alpha_j^2 \tilde{\xi}_j}{1 + \alpha_j^2 \tilde{\xi}_j} \\ \text{s.t. } &\sum_{j=1}^J \alpha_j^2 \sigma_{X_j}^2 = r, \quad \alpha_j^2 \sigma_{X_j}^2 \geq 0, \quad \forall j, \end{aligned} \quad (2)$$

where J is the total number of sub-bands and $\gamma_j > 0$, with $\sum_j \gamma_j = 1$, is the SII band-importance function, which determines the weight of sub-band j to the estimated intelligibility [20].

Solving the problem in (2) leads to the following modified speech LTASE that maximizes fractile ASII [13, 15]:

$$\alpha_j^2 \sigma_{X_j}^2 = \max \left(\frac{\tilde{\sigma}_{V_j} \sqrt{\gamma_j}}{\sqrt{\nu}} - \tilde{\sigma}_{V_j}^2, 0 \right), \quad \forall j, \quad (3)$$

where

$$\frac{1}{\sqrt{\nu}} = \frac{r + \sum_{j \in \mathcal{M}} \tilde{\sigma}_{V_j}^2}{\sum_{j \in \mathcal{M}} \sqrt{\gamma_j} \tilde{\sigma}_{V_j}}, \quad (4)$$

with $\mathcal{M} = \{j : \alpha_j^2 > 0\}$, and where ν is the Lagrange multiplier that ensures that the energy constraint $r = \sum_{j=1}^J \alpha_j^2 \sigma_{X_j}^2$ is obeyed.

2.2. OptFractASII's speech energy allocation strategy

From (3), it follows that the LTASE $\sigma_{Y_j}^2 \triangleq \alpha_j^2 \sigma_{X_j}^2$ of the OptFractASII-processed speech within each band depends on

the input LTASE $\sigma_{X_j}^2$ only through the Lagrange multiplier ν . Hence, $\sigma_{Y_j}^2$ is a function of the energy budget r but *not* the spectral energy distribution $\sigma_{X_j}^2$ of the input speech. In other words, for input speech signals with different LTASE but same total energy available for processing r , OptFractASII would process them to give the same output LTASE, provided the speech signals are presented in the same background noise.

To understand how OptFractASII assigns the available speech energy budget to each sub-band, let us consider the contribution of the j -th sub-band to fractile ASII, see (2):

$$\tilde{\rho}_j = \gamma_j \frac{\alpha_j^2 \tilde{\xi}_j}{\alpha_j^2 \tilde{\xi}_j + 1}. \quad (5)$$

With a certain amount of speech energy $\alpha_j^2 \sigma_{X_j}^2$ allocated to sub-band j , the derivative of $\tilde{\rho}_j$ with respect to the speech energy $\alpha_j^2 \sigma_{X_j}^2$ expresses the exchange rate in sub-band j at which speech energy is converted into fractile ASII:

$$\frac{\partial \tilde{\rho}_j}{\partial (\alpha_j^2 \sigma_{X_j}^2)} = \gamma_j \frac{\tilde{\sigma}_{V_j}^2}{(\alpha_j^2 \sigma_{X_j}^2 + \tilde{\sigma}_{V_j}^2)^2}. \quad (6)$$

We note for later reference that $\partial^2 \tilde{\rho}_j / \partial (\alpha_j^2 \sigma_{X_j}^2)^2 < 0$, which implies that the derivative in (6) is a decreasing function of $\alpha_j^2 \sigma_{X_j}^2$. Considering only the active bands $j \in \mathcal{M}$, (3) reduces to:

$$\alpha_j^2 \sigma_{X_j}^2 = \frac{\tilde{\sigma}_{V_j} \sqrt{\gamma_j}}{\sqrt{\nu}} - \tilde{\sigma}_{V_j}^2, \quad \forall j \in \mathcal{M}. \quad (7)$$

Inserting (7) into (6), we get that, for the active bands, it holds that

$$\frac{\partial \tilde{\rho}_j}{\partial (\alpha_j^2 \sigma_{X_j}^2)} = \nu, \quad \forall j \in \mathcal{M}. \quad (8)$$

Hence, optimal energy allocation is achieved by active bands operating at the same speech energy to ASII exchange rate.

Let us consider output energy distributions for increasing values of r . For this purpose, Fig. 1 shows the values of the sub-band gains and derivatives $\partial \tilde{\rho}_j / \partial (\alpha_j^2 \sigma_{X_j}^2)$ for increasing values

of r . Let us consider the situation where r increases from $r \simeq 0$ to $r = \sum_{j=1}^J \sigma_{X_j}^2$ in successive, infinitesimal quanta of speech energy. The first quantum of speech energy is spent in the sub-band denoted by $j = j^*$, see Fig. 1a, for which $\partial \tilde{\rho}_j / \partial (\alpha_j^2 \sigma_{X_j}^2)$ is maximal (Fig. 1e). This sub-band buys the most ASII with the given quantum of speech energy. For increasing r , its derivative decreases, as $\tilde{\rho}_j$ is a decreasing function of $\alpha_j^2 \sigma_{X_j}^2$, and eventually becomes equal to the derivative of the sub-band, with index $j = i^*$ (Fig. 1b), that has the second highest derivative among all the sub-bands (Fig. 1f). At this point, OptFractASII distributes the successive quanta of speech energy across sub-bands j^* and i^* while keeping $\partial \tilde{\rho}_j / \partial (\alpha_j^2 \sigma_{X_j}^2)$ for $j = j^*, i^*$ constant, until the two derivatives become equal to that of the sub-band with the third best exchange rate. The process continues with OptFractASII distributing energy into all the sub-bands $j \in \mathcal{M}'$ (Fig. 1c), i.e., the active sub-bands, while keeping $\partial \tilde{\rho}_j / \partial (\alpha_j^2 \sigma_{X_j}^2)$ constant for all $j \in \mathcal{M}'$ (Fig. 1g). Fig. 1d shows the optimal gain values α_j^2 , when the speech energy budget is exhausted, $r = \sum_{j=1}^J \sigma_{X_j}^2$ in this case. Fig. 1h verifies that the active sub-bands $j \in \mathcal{M}$ have the same derivative, ν (see (8)). This process can be interpreted as a water-filling process [21], where speech energy is allocated to the active sub-bands by keeping the same “water level” $\partial \tilde{\rho}_j / \partial (\alpha_j^2 \sigma_{X_j}^2)$, i.e., the same speech energy to fractile ASII exchange rate.

From this analysis and thanks to the concavity of the problem in (2) [13], it is clear that the sub-bands that are allocated zero energy are those whose derivatives before processing are less than ν .

We note that the above analysis also applies to OptimalASII [13] if the fractile SNR $\tilde{\xi}_j$ is replaced by the conventional SNR ξ_j , because OptimalASII uses the same mathematical framework as OptFractASII.

2.3. Comparison between OptFractASII and SS

SSDR [5] constitutes the current state-of-the-art for NELE. It is instructive to compare the processing of the spectral shaping (SS) component of SSDRC with that of OptFractASII, because they approach the problem from different angles.

SS is a time-varying system whose filters are inspired by the Lombard speech production process [4]. This system processes speech regardless of the particular noise in which it is to be presented. In fact, its output depends only on the particular spectral shape of the speech to be processed. On the other hand, as noted in Sec. 2.2, OptFractASII only depends on long-term statistics of the noise spectrum and the total speech energy available for processing r , but *not* on the LTASE of the input speech.

To visualize this, Fig. 2 shows the speech LTASE that is output by SS and OptFractASII, when processing two *different* speech signals that are to be presented in the *same* noise background (Fig. 2a-b) and when processing the *same* speech signal that is to be presented in two *different* noise backgrounds (Fig. 2c-d). Fig. 2 verifies our assertion that the output speech LTASE is a function of the input speech LTASE using SS, but of the fractile noise power using OptFractASII. Although the two algorithms appear to be completely different, each of the plots in Fig. 2 shows that both methods reduce the speech spectral tilt. This property, which is imposed by SS to mimic Lombard speech, interestingly appears as an emergent property of OptFractASII, which aims to maximize the speech intelligibility of the processed speech in noise.

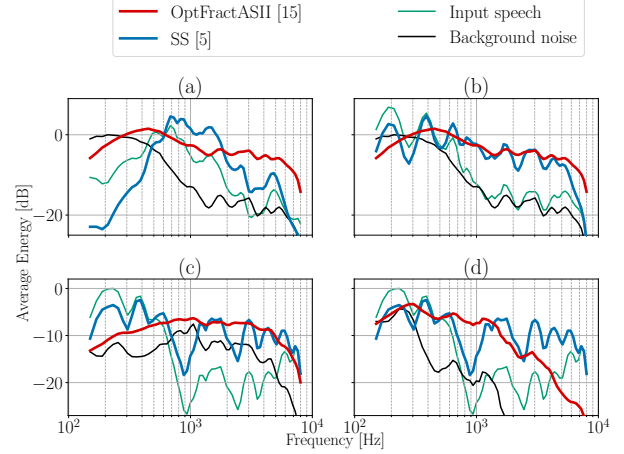


Figure 2: LTASE of OptFractASII- and SS-processed speech for (a) TIMIT and (b) Dantale II speech in the same background noise (SSN at 0dB SNR), and for same AEMST speech in (c) jackhammer and (d) siren noise from UrbanSound8K mixed at 0dB SNR. The plots also show the speech and noise LTASE.

3. Experimental simulations

In this section, we provide two experiments. The first compares method that, like OptFractASII, can modify speech by spectral shaping only. The second compares more advanced methods that can shift energy across frequency and time.

3.1. Experimental setup

To assess performance for different speakers, we use speech signals from three datasets: Dantale II [22], the American English Matrix Sentence Test (AEMST) [23], and DARPA TIMIT [24]. For each dataset, we randomly select speech samples to obtain a corpus of 5 minutes each. The speech signals are then mixed at SNRs between -30 dB and 10 dB with 12 different classes of noise: speech-shaped noise (SSN) whose LTASE matches the average LTASE of one of the three speech datasets, a competing speaker (CS) randomly selected from one of the datasets, and the 10 noise classes of UrbanSound8K [25]. The intelligibility of the processed noisy speech signals is estimated using the spectro-temporal glimpsing index (STGI), which has shown high correlation with speech intelligibility in general, and in particular with speech processed by NELE algorithms [19]. We obtain similar results with ESTOI [18], but these have been excluded due to space limitations.

In every experiment, OptFractASII computes the fractile noise power using $\phi = 0.3$, $r = \sum_{j=1}^J \sigma_{X_j}^2$, and $J = 64$.

3.2. Results - Spectral shaping methods

In this experiment, we want to evaluate NELE methods that aim to impose desired characteristics on the speech spectra. As baselines, we use SEO [6], which was the best performing method of the Hurricane Challenge 1 [26], the spectral shaping component of SSDRC (SS) [5], and OptimalASII [13], since OptFractASII is a modification of this algorithm. Fig. 3 shows the average STGI scores computed on speech signals from the three datasets processed by each method and then mixed with SSN, CS, and UrbanSound8K noise.

In stationary noise such as SSN (top left row of Fig. 3),

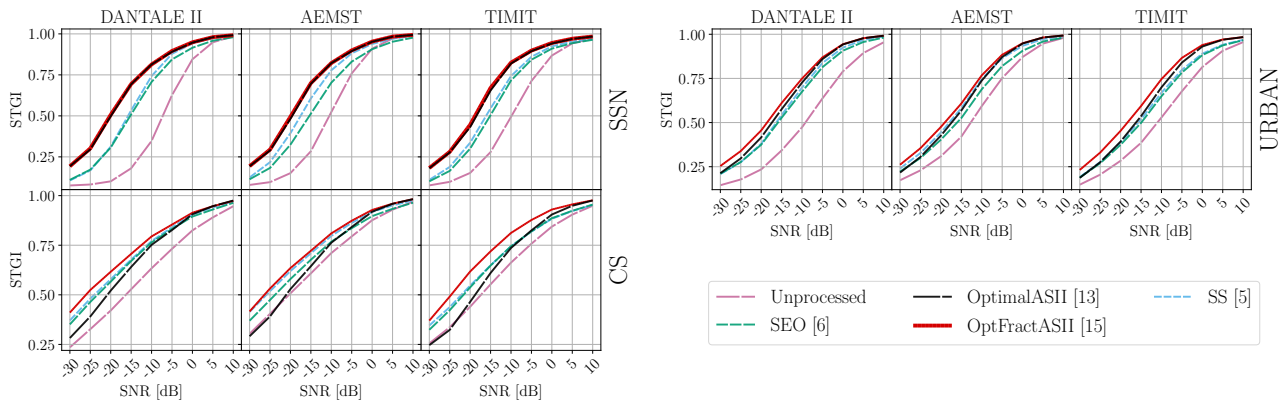


Figure 3: Performance of spectral shaping methods in terms of STGI as a function of input SNR for speech signals from Dantale II, AEMST, and TIMIT. Results for different noise backgrounds are displayed: SSN (top left row), CS (bottom left row), and the average performance over the 10 noise classes of the UrbanSound8K dataset (top right row).

the advantage of OptFractASII and OptimalASII over the other two algorithms is evident for all the SNRs and speech datasets. In fluctuating noise, such as CS (bottom left row of Fig. 3), OptFractASII is still the best performing method in every condition. In urban noise (top right row of Fig. 3), which is the most diverse class of noise in this experiment in terms of noise spectral shapes and temporal fluctuations, OptFractASII is uniformly better than all the baseline methods. Pairwise comparisons (t-test with Bonferroni correction) show that STGI and ESTOI scores obtained with OptFractASII are significantly ($p < 0.05$) higher than for the baseline methods, with very few exceptions: no significant difference ($p > 0.05$) could be found between OptFractASII and OptimalASII in SSN (SNR = -30 dB) and in CS and UrbanSound8K (SNR > 5 dB). This is to be expected, as OptFractASII’s gains are designed to be similar to OptimalASII under these conditions [15].

3.3. Results - Methods for time-frequency energy allocation

The purpose of this experiment is to show the importance of allowing speech energy to be shifted not only across frequency, but also across time. To do this, we appended the DRC stage used by SSDRC to OptFractASII (resulting system called OptFractASII+DRC) and compared it to the plain OptFractASII [15]. We compared performance against two more baselines: SSDRC [5], and a recent neural model [16], here referred to as NELE-GAN, which uses spectrograms of the clean speech and noise to produce a modified speech signal that jointly optimizes several speech intelligibility and quality metrics.

Fig. 4 shows STGI scores for each of the noise classes presented in Sec. 3.1. The results are average pooled over the three speech datasets, and the SNRs are grouped into three ranges (low if $-30 \text{ dB} \leq \text{SNR} < -15 \text{ dB}$, mid if $-15 \text{ dB} \leq \text{SNR} < 0 \text{ dB}$, and high if $0 \text{ dB} \leq \text{SNR} \leq 10 \text{ dB}$). From the figure, it is clear that STGI increases when we append a DRC to OptFractASII. This demonstrates the beneficial effect on speech intelligibility provided by systems that redistribute energy across time, by, e.g., amplifying low-energy temporal regions at the expense of high-energy ones. Furthermore, OptFractASII+DRC, despite using an off-the-shelf DRC stage, performed better than or as good as SSDRC, and definitely better than NELE-GAN, in every of the conditions shown in Fig. 4. STGI and ESTOI scores obtained with OptFractASII+DRC are statistically significantly ($p < 0.05$) better than those of the baselines. For the

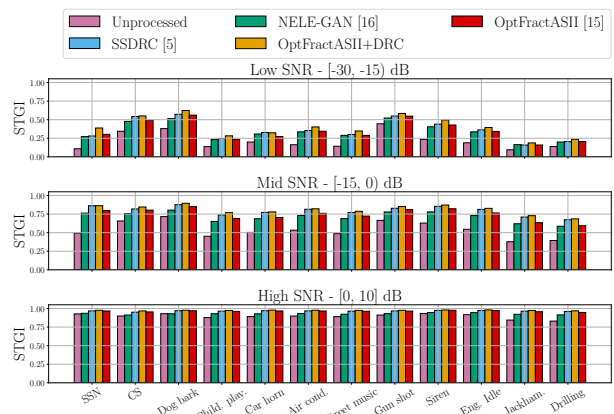


Figure 4: Performance of NELE methods in terms of STGI in SSN, CS, and UrbanSound8K’s noise classes at different SNRs. The STGI scores are computed as the average across the three speech datasets. OptFractASII is included as baseline to judge the effect of the DRC-stage in OptFractASII+DRC.

car horn class at low and mid SNRs, and for SSN and air conditioner at mid SNRs, no statistical difference ($p > 0.05$) with respect to SSDRC could be found.

4. Conclusions

In this paper we analyzed the energy allocation strategy adopted by OptFractASII, a recently proposed method for NELE [15] which uses long-term noise statistics to optimally maximize an approximation of the Speech Intelligibility Index. We showed that the long-term average spectral energy of OptFractASII’s output depends exclusively on background noise statistics, in contrast to the state-of-the-art model SSDRC. We also explained the energy allocation strategy adopted by OptFractASII using a water-filling interpretation. In simulation experiments, we showed that OptFractASII outperforms other well-established NELE spectral shaping systems. In addition, OptFractASII performed generally better than SSDRC and other baselines across a wide range of noise conditions, speakers, and SNRs, when an off-the-shelf dynamic range compression stage is appended.

5. Acknowledgment

This work is partly funded by a grant from the William Demant Foundation.

The authors would like to thank Yannis Stylianou for providing an implementation of SSDRC.

6. References

- [1] É. Lombard, "Le signe de l'élévation de la voix," in *Annales des Maladies de l'Oreille, du Larynx du Nez et du Pharynx*, vol. 37, 1911, pp. 101–119.
- [2] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.
- [3] L. Youyi and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, pp. 3261–3275, 11 2008.
- [4] J.-C. Junqua, "The lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.
- [5] T.-C. Zorila, 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.
- [6] R. Takou, N. Seiyama, and A. Imai, "Improvement of speech intelligibility by reallocation of spectral energy," in *Proc. Interspeech*, 2013, pp. 3605–3607.
- [7] M. Koutsogiannaki and Y. Stylianou, "Modulation enhancement of temporal envelopes for increasing speech intelligibility in noise," in *Proc. Interspeech*, 2016, pp. 2508–2512.
- [8] C. Chermaz and S. King, "A sound engineering approach to near end listening enhancement," in *Proc. Interspeech*, 2020, pp. 1356–1360.
- [9] C. Chermaz, C. Valentini-Botinhao, H. Schepker, and S. King, "Evaluating near end listening enhancement algorithms in realistic environments," in *Proc. Interspeech*, 2019, pp. 1373–1377.
- [10] J. Rennie, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-enhancing speech modifications - the hurricane challenge 2.0," in *Proc. Interspeech*, 2020, pp. 1341–1345.
- [11] H. Brumm and S. Zollinger, "The evolution of the lombard effect: 100 years of psychoacoustic research," *Behaviour*, pp. 1173–1198, 2011.
- [12] B. Sauert and P. Vary, "Near end listening enhancement: Speech intelligibility improvement in noisy environments," in *Proc. ICASSP. IEEE*, 2006, pp. 493–496.
- [13] C. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, 2013.
- [14] A. Fuglsig, J. Jensen, Z.-H. Tan, L. Bertelsen, J. Lindof, and J. Østergaard, "Minimum processing near-end listening enhancement," *IEEE/ACM Trans. Audio Speech and Lang. Proc.*, vol. 31, pp. 2233–2245, 2023.
- [15] F. Villani, W.-Y. Chan, Z.-H. Tan, J. Østergaard, and J. Jensen, "Near-end listening enhancement using a noise-robust linear time-invariant filter," in *Proc. IWAENC*, 2024, pp. 444–448.
- [16] H. Li and J. Yamagishi, "Multi-metric optimization using generative adversarial networks for near-end speech intelligibility enhancement," *IEEE/ACM Trans. Audio Speech and Lang. Proc.*, vol. 29, pp. 3000–3011, 2021.
- [17] M. Shifas, T.-C. Zorila, and Y. Stylianou, "End-to-end neural based modification of noisy speech for speech-in-noise intelligibility improvement," *IEEE/ACM Trans. Audio Speech and Lang. Proc.*, vol. 30, pp. 162–173, 2022.
- [18] J. Jensen and C. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Acoust., Speech, Sig. Proc.*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [19] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction," in *Proc. Interspeech*, 2021, pp. 206–210.
- [20] ANSI, *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America, 1997.
- [21] J. Proakis and M. Salehi, *Digital communications*. McGraw Hill, 2001.
- [22] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, 2003.
- [23] B. Kollmeier, A. Warzybok, S. Hochmuth, M. Zokoll, V. Uslar, T. Brand, and K. Wagener, "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *Int. J. Audiol.*, vol. 54, no. sup2, pp. 3–16, 2015.
- [24] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa Timit acoustic-phonetic continuous speech corpus cd-rom {TIMIT}," 1993.
- [25] J. Salamon, C. Jacoby, and J. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. on Multimedia*, 2014, pp. 1041–1044.
- [26] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane Challenge," in *Proc. Interspeech*, 2013, pp. 3552–3556.