On Optimal Linear Filtering of Speech for Near-End Listening Enhancement

Cees H. Taal, Jesper Jensen, and Arne Leijon

***Abstract*—In this letter the focus is on linear filtering of speech before degradation due to additive background noise. The goal is to design the filter such that the speech intelligibility index (SII) is maximized when the speech is played back in a known noisy environment. Moreover, a power constraint is taken into account to prevent uncomfortable playback levels and deal with loudspeaker constraints. Previous methods use linear approximations of the SII in order to find a closed-form solution. However, as we show, these linear approximations introduce errors in low SNR regions and are therefore suboptimal. In this work we propose a *nonlinear* approximation of the SII which is accurate for all SNRs. Experiments show large intelligibility improvements with the proposed method over the unprocessed noisy speech and better performance than one state-of-the art method.**

***Index Terms*—Near-end enhancement, speech enhancement, speech intelligibility, speech intelligibility index.**

I. Introduction

We consider the situation where a person in a noisy far-end environment wishes to communicate via some communication channel to a person in a noisy near-end environment. The intelligibility of the far-end signal can be affected by background noise from both sides of the communication channel. That is, the noise could originate from the side of the sender (the *far end*) but also from the environment of the receiver (the *near end*). In order to eliminate the negative impact of the far-end noise, one would typically apply a noise-reduction algorithm. Intelligibility degradation due to the near-end noise can be compensated for by pre-processing the recorded far-end speech before playback in the noisy environment. This latter approach, sometimes referred to as near-end listening enhancement [1], is the focus of this work. The common assumptions here are that a clean version of the far-end signal is available (i.e., the potential noise is assumed to be successfully suppressed) and that we have knowledge of the near-end noise statistics [1].

One obvious solution to the near-end listening enhancement problem is to increase the playback level of the speech. However, at a certain point increasing the playback level may not be possible anymore due to loudspeaker limitations or unpleasant playback levels. Therefore many approaches take into account a power constraint and instead of amplification, speech energy is redistributed over time and/or frequency in order to improve speech intelligibility. Many different approaches exist, for example, based on high-pass filtering [2], dynamic range compression [3] or changing the consonant-vowel ratio [4]. However, the majority of these methods do not use any type of mathematical descriptor of speech intelligibility, which makes it difficult to claim any form of optimality.

Fortunately, the effects of additive stationary background noise and linear filtering on speech intelligibility can be well predicted with the speech intelligibility index (SII) [5]. Recently, Sauert and Vary presented a closed-form solution which was optimal in terms of a linear approximation of the SII [1]. However, as we will further explain in this letter, the linear approximation is inaccurate for within-band SNRs smaller than -15 dB. As a consequence, speech energy might be allocated to frequency bands which eventually do not contribute to speech intelligibility. We propose a follow-up to the work of Sauert and Vary by solving the constrained optimization problem based on a nonlinear approximation of the SII which is also accurate for lower within-band SNRs. SII predictions and intelligibility listening tests are performed which indeed show an increase in speech intelligibility.

II. Derivation of Linear Filter

A. Intelligibility Measure

The intelligibility measure which will be used for optimization is based on the standardized SII [5]. We assume that the speech and noise are presented above the threshold in quiet at a comfortable level. Also, effects of masking are excluded from the standard SII procedure (similarly as in [6]). Based on these assumptions the measure can be summarized by the following three stages: (1) First the long-term average spectra of the speech and noise are estimated within critical bands. (2) A within-band SNR is calculated, clipped between -15 and 15 dB followed by normalization to the range of 0 and 1. (3) A weighted average of the normalized within-band SNRs is calculated to obtain one outcome.

Next, details are given for each stage. Let x and ε denote the time-domain signals of the clean speech and noise, respectively. A windowed version of x is denoted by x_m where m denotes the window frame-index. A Hann-window is used with 50% overlap, and 32 ms length. The impulse response of the i^{th} auditory filter is denoted by h_i, where $i \in \{1, \ldots, n\}$ and n is the total number of auditory filters. Subsequently, the energy within one time-frequency (TF) unit is calculated as follows for the clean speech,

$$X_{m,i}^2 = \sum_k |X_m(k)|^2 |H_i(k)|^2, \tag{1}$$

Manuscript received November 23, 2012; revised January 08, 2013; accepted January 08, 2013. Date of publication January 15, 2013; date of current version January 29, 2013. This study was supported by the European Commission within the LISTA Project, FET-Open grant number 256230. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Græsbøll Christensen.

C. H. Taal and A. Leijon are with the Royal Institute of Technology, Sound and Image Processing Laboratory, Stockholm, Sweden (e-mail: taal@kth.se).

J. Jensen is with Oticon A/S, Smørum, Denmark, and also with and Aalborg University, Aalborg, Denmark.

Digital Object Identifier 10.1109/LSP.2013.2240297

where $X_m(k)$ and $H_i(k)$ denote DFTs of x_m and h_i, respectively, with frequency-bin index k. Signals are sampled at 20 kHz where short-time frames are zero-padded to 64 ms before applying the DFT. In total, 64 auditory filters are used where center frequencies are linearly spaced on an equivalent rectangular bandwidth (ERB) scale between 150 and 8500 Hz [7]. Its squared magnitude responses $|H_i(k)|^2$ are chosen as described in [7]. The average energy within one critical band is based on a long-term sample mean over many short-time frames (e.g., several minutes) and is denoted as follows,

$$\sigma_{X_i}^2 = \frac{1}{M}\sum_m X_{m,i}^2, \tag{2}$$

where M equals the total number of short-time frames and similar definitions hold for $\sigma_{\mathcal{E}_i}^2$. Let the SNR within one critical band be denoted by,

$$\xi_i = \frac{\sigma_{X_i}^2}{\sigma_{\mathcal{E}_i}^2}, \tag{3}$$

which is used to calculate an intermediate measure to determine the audibility of the speech in presence of the noise within one band. This SNR is log-transformed, clipped between -15 and $+15$ dB and normalized such that its range is between zero and one, i.e.,

$$d\left(\xi_i\right) = \frac{\max\left(\min\left(10\log_{10}\left(\xi_i\right), 15\right), -15\right)}{30} + \frac{1}{2}, \tag{4}$$

Subsequently, a weighted average is calculated as follows,

$$SII = \sum_i \gamma_i d\left(\xi_i\right), \tag{5}$$

where γ denotes the band-importance function as given in the critical-band SII procedure as found in [5, Table 1]. In summary, this weighting-function reduces the importance of bands with center frequency below 450 Hz and above 4000 Hz. It is expected that (5) is a monotonic increasing function of the intelligibility of the speech in noise [5].

B. Solution to Constrained Optimization Problem

The goal is to maximize the speech intelligibility, i.e., maximize (5), by redistributing the speech energy over the critical bands. Hence, the total energy over all bands remains unchanged. In this work we assume that the statistics do not change over time. Therefore, the derived filter is time-invariant. In practice one could estimate the statistics online, e.g., with a noise-tracker [8], and use a time-varying filter. The same mathematical framework can be used for this time-varying case. Let α_i be a real and non-negative scalar applied to each critical band. Since $\alpha_i \geq 0$ we have the relation $(1/M)\sum_m \left(\alpha_i X_{m,i}\right)^2 = \alpha_i^2\sigma_{X_i}^2$. As a consequence, the constrained problem can be formulated as follows,

$$\begin{aligned} \max \quad & \sum_i \gamma_i d\left(\alpha_i^2\xi_i\right) \\ \text{s.t.} \quad & \sum_i \alpha_i^2\sigma_{X_i}^2 = \sum_i \sigma_{X_i}^2 \\ & \alpha_i^2\sigma_{X_i}^2 \geq 0, \forall i. \end{aligned} \tag{6}$$

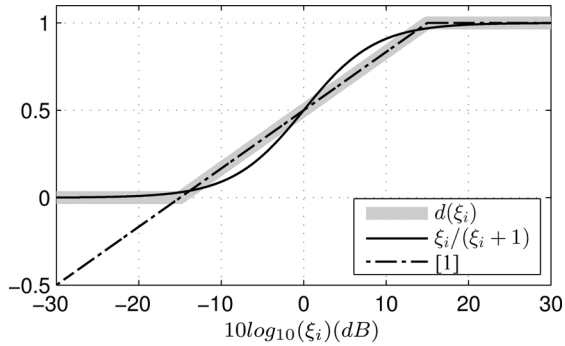

Fig. 1. For mathematical tractability the within-channel intelligibility measure $d(\xi_i)$ is approximated with $\xi_i/(\xi_i+1)$, which is more accurate for lower SNRs compared to the work of Sauert and Vary [1].

In order to find a closed-form solution to this constrained optimization problem Sauert and Vary used an approximation of (4). Under the assumptions as mentioned in Section II-A, the intermediate intelligibility measure per band in the SII, i.e., (4), is approximated by Sauert and Vary as follows [1],

$$d\left(\xi_i\right) \approx \frac{\min\left(10\log_{10}\left(\xi_i\right), 15\right)}{30} + \frac{1}{2}. \tag{7}$$

Hence, the stage where bands are lower-limited to -15 dB is excluded. As an alternative we propose to approximate (4) with the following expression which is mathematically tractable and more accurate for SNRs below -15 dB,

$$d\left(\xi_i\right) \approx \frac{\xi_i}{\xi_i + 1}. \tag{8}$$

Both approximations together with the original intermediate intelligibility, as defined in (4), are shown in Fig. 1. Clearly, the proposed approximation is more accurate for lower SNRs compared to the method from [1]. Note, that the intermediate intelligibility measure scaled by α_i^2, i.e., $\alpha_i^2\xi_i/(\alpha_i^2\xi_i + 1)$, is concave in $\alpha_i^2\sigma_{X_i}^2$. Hence, the weighted average of these concave functions, as in (5), is also concave. We obtain convexity by negation and characterize the problem by the following Lagrangian cost-function,

$$\begin{aligned} J = -\sum_i \gamma_i \frac{\alpha_i^2\xi_i}{\alpha_i^2\xi_i + 1} + \nu\left(\sum_i \alpha_i^2\sigma_{X_i}^2 - r\right) \\ + \sum_i \lambda_i\left(-\alpha_i^2\sigma_{X_i}^2\right), \end{aligned} \tag{9}$$

where $r = \sum_i \sigma_{X_i}^2$ and ν and λ_i are Lagrangian multipliers related to the energy constraint and inequality constraints in (6), respectively. Since our problem is convex and differentiable, any point that satisfies the following Karush-Kuhn-Tucker (KKT) conditions,

$$\begin{aligned} \sum_i \alpha_i^2\sigma_{X_i}^2 &= r \\ \alpha_i^2\sigma_{X_i}^2 &\geq 0, \quad \forall i \\ \lambda_i &\geq 0, \quad \forall i \\ \lambda_i\alpha_i^2\sigma_{X_i}^2 &= 0, \quad \forall i \\ \frac{-2\gamma_i\xi_i\alpha_i}{\left(\xi_i{\alpha_i}^2 + 1\right)^2} + 2\nu\alpha_i\sigma_{X_i}^2 - 2\lambda_i\alpha_i &= 0, \quad \forall i \end{aligned} \tag{10}$$

is guaranteed to be optimal [9]. Solving gives,

$$\alpha_i^2 \sigma_{X_i}^2 = \max\left(\frac{\sigma_{\mathcal{E}_i}\sqrt{\gamma_i}}{\sqrt{\nu}} - \sigma_{\mathcal{E}_i}^2, 0\right), \forall i, \tag{11}$$

where ν is chosen such that the energy constraint is satisfied,

$$\frac{1}{\sqrt{\nu}} = \frac{r + \sum_{i\in\mathcal{M}} \sigma_{\mathcal{E}_i}^2}{\sum_{i\in\mathcal{M}} \sqrt{\gamma_i}\sigma_{\mathcal{E}_i}}, \tag{12}$$

and where $\mathcal{M} = \left\{i \in \{1,\ldots,n\} : \alpha_i^2 > 0\right\}$ denotes the set of critical band indices for which the optimal α_i^2 is positive. Since the set $\mathcal{M}$ depends on α_i^2, the Lagrange multiplier ν is also dependent on α_i^2. In order to cope with this recursive dependency, the optimal value of ν may be found by evaluating (11) for a range of ν-values or, e.g., using a bi-section method [9] such that the energy constraint is satisfied.

III. Experimental Evaluation

The proposed method (PROP) is compared with the unprocessed noisy speech (UN) and a baseline method by Sauert and Vary [1] (SAU) by means of SII predictions and an intelligibility listening test. For a fair comparison we use the exact same implementation for both methods as explained in Section II-A, where only the applied α_i differs.

A. SII Predictions

SII predictions as described in Section II-A are calculated for PROP, SAU and UN for four different noise types (babble, car, white, speech-shaped) and SNRs in the range between -30 and -5 dB in steps of 2 dB. Per condition, a total number of 100 sentences are randomly selected from the Timit database [10] to calculate an average SII. The results are shown in Fig. 2. The proposed method shows better performance than the baseline method for the lower-SNR range for almost all noise types. For the higher SNR-range the proposed method shows similar performance as the baseline method. This is expected since the main difference between the two methods, as was explained in the previous section, occurs when frequency bands drop below -15 dB, which will only happen at lower global SNRs. Inspection of the car noise condition reveals that most bands are still above -15 dB which explains the same performance. Both algorithms result in a large intelligibility improvement compared to the unprocessed noisy speech except for the white-noise condition, where the baseline method actually decreases speech intelligibility for $\mathrm{SNRs} < -15$ dB.

B. Listening Experiment

Sixteen Dutch-speaking subjects listened to sentences from the Dutch Matrix-test [11]. The material consists of 5-word sentences spoken by a female speaker, which are of the grammatical form name-verb-numeral-adjective-noun (e.g., Ingrid owns six old jackets), as proposed by Hagerman [12]. Each word in the sentence is picked randomly from a list of 10 possible words. The subject had access to the closed set of words by means of a 10-by-5 matrix on a computer screen. The task of the listener is to select via a graphical user interface the understood words. Sentences are played only once.

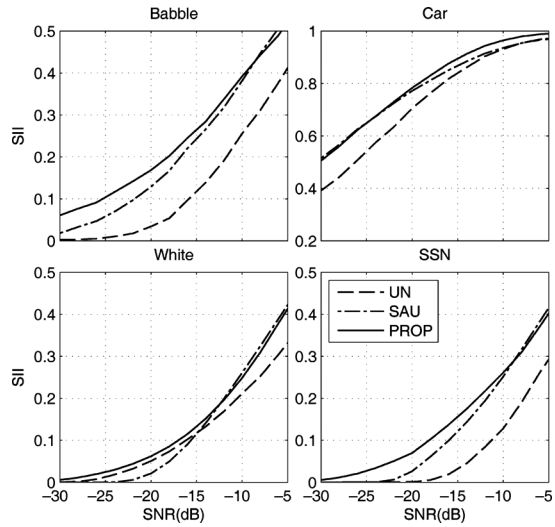

Fig. 2. SII predictions for the proposed method (PROP), the baseline method by Sauert *et al.* [1] (SAU) and the unprocessed noisy speech (UN).

Signals are sampled at 20 kHz and degraded with two noise types: speech shaped noise (SSN) and babble noise. Again PROP, SAU and UN are included in the experiment. SNRs of $[-20 \quad -17 \quad -14]$ and $[-11 \quad -8 \quad -5]$ are chosen for the processed speech (SAU, PROP) and unprocessed speech, respectively, to have roughly similar speech intelligibility for the two conditions. For each condition the listener is presented with 5, 5-word sentences through headphones in a silent room where each sentence was used only once. As a consequence, each subject listened to a total of

$$(5 \text{ sentences} \times 2 \text{ noise types} \times 3 \text{ algorithms} \times 3 \text{ SNRs}) = 90 \text{ sentences}$$

in total. The order of the sentences was randomized. The score per user and for one condition was obtained by the average percentage of correct words.

The results are shown in Fig. 3 which illustrates the average-user intelligibility scores with standard errors for both noise types. It is clear from the results that both the proposed algorithm and the baseline method results in a large improvement in intelligibility compared to the unprocessed noisy speech. Rough estimates of the 50% speech reception thresholds (SRTs) are obtained with linear interpolation, which indicate SRT improvements of 7–10 dB depending on noise and algorithm type. From the results we also see that the proposed method results in more intelligible speech (approximate increase of 10%) over the baseline method in all SSN conditions and one babble noise condition. For the highest two SNRs with the babble noise the proposed and baseline method show similar performance. Note that these listening test results are in line with the SII predictions in Fig. 2.

A 3-way analysis of variance (ANOVA) with factors noise-type (babble, SSN), SNR $(-20, -17, -14)$ and algorithm type (PROP, SAU) shows that the overall improvement of PROP compared to SAU is statistically significant ($\mathrm{p} = 0.002$).

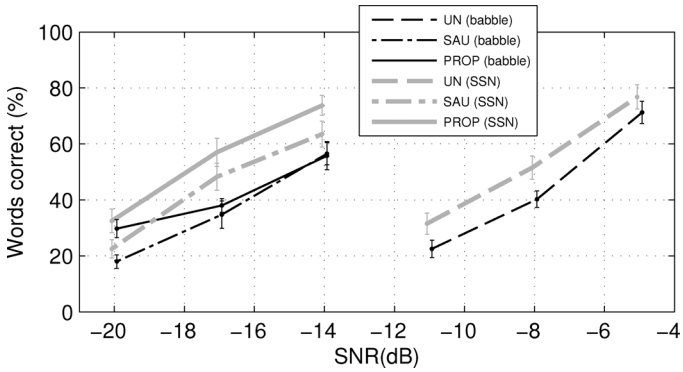

Fig. 3. Average-user intelligibility scores with standard errors for the proposed method (PROP), the baseline method by Sauert and Vary [1] (SAU) and the unprocessed noisy speech (UN).

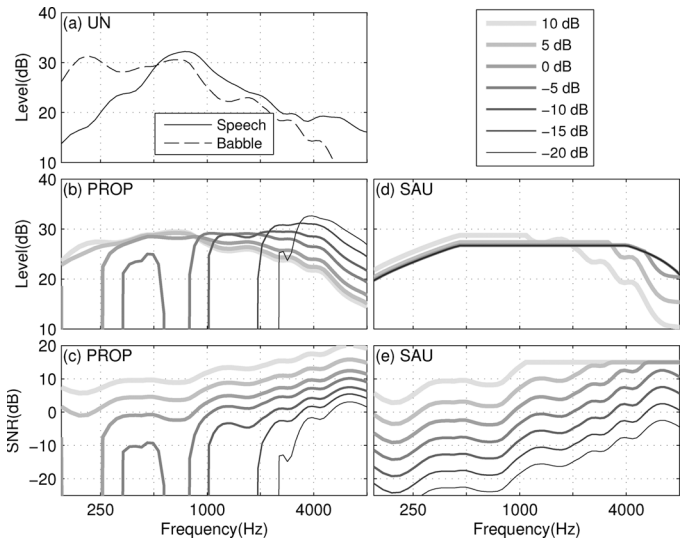

Fig. 4. (a) Spectra for unprocessed speech (UN) and babble noise. (b) Resulting speech spectra and (c) resulting per-band SNRs for proposed method (PROP) for different global SNRs. (d, e) Processed speech spectra and per-band SNRs for Sauert and Vary [1] (SAU).

Post-hoc analysis is performed for six pairwise comparisons between the two algorithm types for each SNR (two-sided t-tests with Bonferroni-corrected p-values). From this analyses we found that the difference in performance was significant for [babble noise, $\mathrm{SNR} = -20$] ($\mathrm{p} = 0.004$) and [SSN, $\mathrm{SNR} = -14$] ($\mathrm{p} = 0.004$).

IV. Filter Properties

To explain the difference in performance between the two algorithms it is of interest to investigate their properties as a function of SNR. In Fig. 4 several processed critical-band speech-spectra are shown for both methods with babble noise and SNRs between -20 and 10 dB (see caption for exact subplot descriptions).

One difference between the two methods can be clearly observed from subplot (c) and (e) which show per-band SNRs. Note that with the proposed method no per-band SNR falls below -15 dB while the baseline method is still spending energy in bands with SNRs below -15 dB. In the latter case this energy is lost, since it does not contribute to speech intelligibility according to the SII. This is probably the main reason for better performance at lower SNRs with the proposed method as was observed in both the listening test and the SII-predictions.

Also note in Fig. 4 that with the baseline method the shape of the speech becomes independent of noise type and SNR when the per-band SNRs fall below 15 dB. In fact, the bandpass shape of the speech, as observed in subplot (d), equals the band-importance function γ_i in (5). This result was also found by [13] where shaping the speech spectrum as the band-importance functions should have a positive effect on speech intelligibility. Although the proposed method shows similar behavior for higher SNRs, we conclude that this is not optimal for lower SNRs. In fact, SII-predictions for white noise, as shown in Fig. 2, indicate that this approach could even decrease speech intelligibility.

V. Conclusions

A new linear filter was proposed to optimize the intelligibility of speech in noise for the near-end listener. This was accomplished by redistributing the speech energy over frequency such that an approximation of the speech intelligibility index (SII) was maximized. SII predictions and intelligibility listening test experiments show large intelligibility improvements with the proposed method and better performance than one state-of-the art baseline method. In contrast to the baseline method, the proposed method sets certain frequency bands to zero when they do not contribute to intelligibility anymore.

References

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

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

[3] T. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on power recovery and dynamic range compression," in *Proc. EUSIPCO*, 2012, pp. 2075–2079.

[4] S. Gordon-Salant, "Recognition of natural and time/intensity altered cvs by young and elderly subjects with normal hearing," *J. Acoust. Soc. Amer.*, vol. 80, no. 6, pp. 1599–1607, 1986.

[5] ANSI, "Methods for Calculation of the Speech Intelligibility Index" ANSI, New York, 1997, S3.5-1997.

[6] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, 2010.

[7] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 9, pp. 1292–1304, 2005.

[8] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *IEEE Int. Conf. Acoust., Speech Signal Processing*, 2010, pp. 4266–4269.

[9] S. Boyd and L. Vandenberghe, "KKT optimality conditions," in *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004, ch. 5.5.3, pp. 243–246.

[10] J. Garofolo, Timit: Acoustic-Phonetic Continuous Speech Corpus NIST, Boulder, CO, USA, 1993.

[11] J. Koopman, R. Houben, W. A. Dreschler, and J. Verschuure, "Development of a speech in noise test (matrix)," in *8th EFAS Congr., 10th DGA Congr.*, Heidelberg, Germany, Jun. 2007.

[12] B. Hagerman, "Sentences for testing speech intelligibility in noise," *Scandinavian Audiology*, vol. 11, no. 2, pp. 79–87, 1982.

[13] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Amer.*, vol. 43, no. 1, pp. 81–86, 1968.