DOI: 10.61782/fa.2025.0946

FORUM ACUSTICUM EURONOISE 2025

USING THE MASKING CURVE OF A MASKER SIGNAL TO APPROXIMATE THE SPECTRAL SHAPE OF A TARGET SPEECH SIGNAL*

Pierre Laffitte[1*] **Sankha Subhra Bhattacharjee**[1] **Jesper Rindom Jensen**[1]
Mads Græsbøll Christensen[1]
[1] **Audio Analysis Lab, Aalborg University, Denmark**

ABSTRACT

This work attempts to provide a novel look at speech privacy preservation, by proposing a solution based on sound masking as an alternative or a complement to traditional sound control and noise cancellation methods. We propose a framework for generating a masker signal whose masking curve approximates the spectral shape of the target speech to be masked. The approximation is done by Gradient Descent-based optimization, to minimize the distance between the magnitude spectrum of the target speech and the masker. The results show that the proposed algorithm exhibits the desired effect, reducing the measured annoyance when other metrics are kept constant. Although proper implementation in a real-world system is out of the scope of this paper, it serves to validate the theoretical background proposed here.

Keywords: *psychacoustic annoyance, signal masking, block processing, numerical optimization*

1. INTRODUCTION

Unwanted noise in open-plan offices, healthcare facilities, and residential spaces disrupts concentration, compromises confidentiality, and increases stress. Sound masking technology mitigates these effects by generating a low-level, unobtrusive background sound that conceals conversations and auditory distractions, enhancing privacy and acoustic comfort. It is widely applicable in both controlled sound zones and open acoustic spaces [1–6]. As an alternative, traditional noise control methods, such as active noise control [7, 8] and sound zone control [9, 10], though effective under optimal conditions, rely on precise reference signals or *a priori* information that are often impractical to obtain. In contrast, sound masking leverages the natural masking effect, where one sound renders another less perceptible, without requiring detailed information about the acoustic channels. While most studies evaluate the psychological impact of various fixed masking signals — such as white noise, pink noise, or speech-shaped noise [11] — few explore the creation of a masker tailored to the specific acoustic environment. This work focuses on generating a dynamic masker designed to optimize both privacy and listening comfort by adapting to the characteristics of the speech to be masked.

*__Corresponding author__: pl.laffitte@gmail.com.

2. RELATED WORK

Research on masking sounds for privacy dates back to the 1970s [12, 13]. Early approaches often relied on predefined noise types such as white, pink, babble, or speech-shaped noise [5]. Other sounds, like nature sounds [4,6] or music [6, 14], have been explored, but pink noise remains the most common. To evaluate masker efficiency, metrics like the speech transmission index (STI) [4–6, 15] and articulation index (AI) [16] are frequently used. However, we opted for the short-time objective intelligibility (STOI) [17], designed to better handle time-frequency processing. The key premise of this work is that optimal masking requires tailoring the sound to the speech being masked. While constant noisy sounds are more efficient for masking, they are less tolerable. Psychoacoustic research re-

veals that masking is a frequency-dependent phenomenon [18]. By reducing energy in spectral regions that do not contribute to masking and ensuring enough energy where it does, we can maintain masking efficiency while minimizing annoyance. The Masking Curve [18] describes a signal's masking capability across the frequency spectrum, providing a threshold dB value for each frequency that a given signal can mask. This psychoacoustic approach forms the foundation of our method, balancing intelligibility reduction and listener comfort.

3. PROBLEM FORMULATION

We aim to mask a speech signal $x(n)$ with a masker signal $y(n)$, such that the mixture $x(n) + y(n)$ is unintelligible. To avoid trivial solutions where louder maskers provide better masking, we fix the masker energy $E_y = C$, where C is a constant determined by practical considerations (e.g., desired SNR limits). Since the speech signal $x(n)$ typically has a non-flat spectrum, the masker $y(n)$ does not need uniform energy across all frequencies to effectively mask it. The total masker energy, expressed as:

$$E_y = \sum_{-f_s/2}^{f_s/2} S_{xx}(f) \tag{1}$$

can be minimized by reducing spectral components where little or no masking energy is required. Psychoacoustic research shows that the masking effect is governed by frequency and energy [18]. A masker's energy spectral density $Y(f)$ defines its masked threshold $\nu(f)$ —the minimum energy level required to mask a signal at each frequency. To create a masker $y(n)$ that matches the target speech spectral shape, we iteratively adjust $y(n)$ to minimize the distance between $X(f)$ and $\nu(f)$: This leads to the optimization problem:

$$\min_{y(n)} (\|X(f) - \nu(f)\|)^2 \tag{2}$$

where $\nu(f)$ is nonlinearly dependent on $y(n)$ [18]. To avoid intelligibility leakage, we reset the phase to zero at each frame, simplifying the problem to spectral magnitude only. A block diagram of the process is shown in Fig. 1.

4. PROPOSED METHOD

To achieve this, we choose to generate a control signal $\epsilon(n)$ and shape its spectrum by multiplying it with a

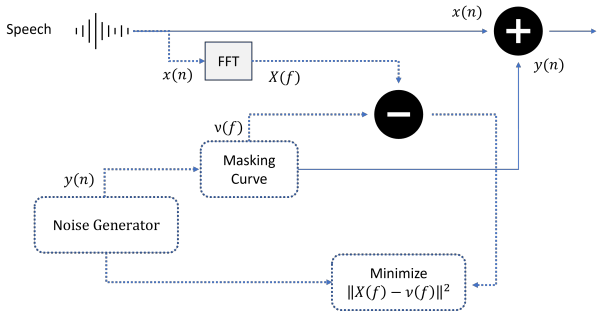


Figure 1: Optimization framework to produce a masker signal, using its masking curve to approximate the speech' spectrum. Dashed lines represents the optimization loop.

weight vector $\vec{W}$ in the frequency domain. The masking curve $\nu(f)$ then depends directly on the resulting time domain signal. The STOI and PA metrics being based on third-octave bands, we choose $\vec{W}$ to have a component along each third octave band, meaning that each filter weight represents an octave band. Our objective is to find the set of filter weights $\vec{W}$ which minimizes the distance between the target speech spectrum and the masked threshold of our masker:

$$\min_{\vec{W}} \sum_{k}^{M} (\|X_k - \nu_k\|)^2 \tag{3}$$

$\tilde{X_k}$ is the k-th third octave band amplitude of the windowed speech frame and ν_k is the masked threshold for the k-th third-octave band with M the total number of bands. When the optimization is finished, the final weight vector $\vec{W}$ is used to filter a raw masker sound, at a power determined by the pre-defined Masker to Signal Ratio (MSR).

4.1 Block processing

We use block-based processing; the target speech to be masked is windowed in the time domain, and each windowed block of signal is then analyzed in the frequency domain. The windowing process uses a classic Hanning window, e.g.,

$$\tilde{x}(x) = \omega(n)x(n) \tag{4}$$

where $\omega(n)$ is the Hanning window function, $x(n)$ the sampled input signal and $\tilde{x}(n)$ the windowed signal frame.

The Hanning window exhibits constant overlap-add property (COLA) thereby allowing reconstruction of the signal in the time-domain after frequency-domain processing [19]. Since this process is block based, each block is adjusted to the desired MSR value, resulting in a fluctuating energy, following the envelope of the input speech maskee.

4.2 Optimization

We frame the problem as an unconstrained optimization problem in which we try to minimize the error or distance between a desired variable and an estimated variable. The desired variable is defined as the speech spectrum, $X(f)$, and the estimated variable is the masking curve calculated from the produced masker, $\nu(f)$. The cost function is the l_2-norm of the difference

$$J(\mathbf{W}) = \sum_{k}^{M} (\|X_k - \nu_k\|)^2 \tag{5}$$

which for simplicity is assumed to be convex. To find a global minimum for this function, we use the gradient descent method. Since there is no closed-form solution to the derivative of our error function, we employ a numerical optimization scheme, namely the Finite Difference Method (FDM). In this method, the derivative is approximated by

$$\frac{df}{dx} = \frac{f(x+h) - f(x-h)}{2h} \tag{6}$$

following a rewriting of Taylor's first order decomposition of differentiable functions. For our cost function $J(\mathbf{W})$, this gives the following gradient computation:

$$\Delta = \frac{J(\mathbf{W} + \epsilon) - J(\mathbf{W} - \epsilon)}{2\epsilon} \tag{7}$$

The update equation for weight vector $\mathbf{W}$ is then:

$$\mathbf{W}_{i+1} = \mathbf{W}_i - l_{\text{rate}}\Delta \tag{8}$$

where l_{rate} is a hyper-parameter whose value is chosen empirically.

4.3 Smoothing filter

Once the optimal third-octave band weights are found, we design a filter whose frequency response matches these weights. Applying this filter to the raw masker sound shapes its frequency domain according to the weight values.

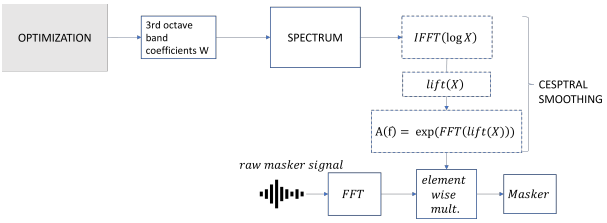


Figure 2: Cepstral windowing method used to smooth the spectrum before converting it to a filter for the raw masker sound

We call this the smoothing filter because we want to avoid a rough masking sound or one that resembles the original speech too closely, as that would compromise intelligibility. To achieve this, we transpose the weights from the third-octave basis to a linear frequency axis, take the cepstrum, and apply a "liftering" filter to retain only the slow spectral variations.

We then return to the frequency domain via the Fourier transform, taking the exponential to reverse the log transformation of the cepstrum. This produces a smoothed spectrogram, which we use to derive the filter for the raw masker sound by taking the inverse Fourier transform.

$$S(f) = \exp(FFT(w * IFFT(\log(FFT(x))))) \tag{9}$$

where w is the "lifter" (i.e. the filter), we apply to the cesptrum to only keep the first few values.

This spectral smoothing technique is called the "cepstral windowing method", for which a graphical description is given in Fig. 2 and is explained in [19].

4.4 Metrics

In this work, we rely on three metrics to evaluate the performance of our framework, which correspond to the two subjective criteria in which we are interested, namely the intelligibility of the speech and the annoyance of the noise.

4.4.1 STOI

Short-term objective intelligence (STOI) is a metric proposed to measure the objective intelligibility of a given speech signal. The subjective evaluation carried out in [17] shows that the STOI objective metric correlates with perceived annoyance. A value of STOI of 0.20 correspond

roughly to a subjective intelligibility score of less than 10%, with a correlation coefficient of 0.95. We noticed from our experimental results that the STOI function does not distinguish very well between two different maskers at low values of intelligibilities, for instance when increasing the volume of the masker. That is, the function exhibits a bit of saturation in the low range (below 0.20), while another metric such as CSII was able to make clearer distinctions between maskers. This is why we chose to also use CSII in addition to STOI.

4.4.2 CSII

Proposed to measure the intelligibility of a speech signal [20], Coherence Speech Intelligibility Index (CSII) computes speech intelligibility, by eq.(14) [20]. According to the paper's finding, a CSII value of 0.10 corresponds roughly to a subjective intelligibility score of less than 10%, with a correlation coefficient of 0.98.

4.4.3 Annoyance

We consider the Annoyance derived in [21] to measure the degree of annoyance incurred by the masker, given by:

$$PA = N_5(1 + \sqrt{\omega_S^2 + \omega_{FR}^2}) \tag{10}$$

where $\omega_S = (S - 1.75) \times 0.25 \log(N_5 + 10)$ and $\omega_{FR} = \frac{2.18}{N_5^{0.4}}(0.4F + 0.6R)$ and the following variables are defined as follows:

- N_5 relates the perceived loudness of the signal
- S is the sharpness (in acum)
- R measures the roughness (in asper)
- F measures the fluctuation strength (in vascil)

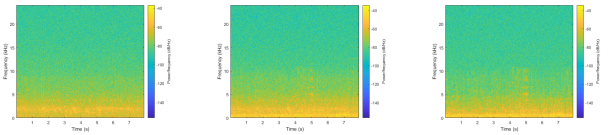
It was shown [21] that this objective annoyance metric correlates with the perceived annoyance.

4.5 Masker signal composition

It was reported in [11] that babble noise is the least annoying type of masker when compared to white noise, pink noise, and speech-shaped noise, so we decided to use it as our masker sound. Additionally, in order to add content/energy in some parts of the spectrum, we added nature sounds to complement the babble sound, which were mostly field recordings from the forest, containing sounds of a river flowing, the rain, and birds chirping. Different combinations of raw sounds were tested, using different mixing proportions, as shown in table 1. Sound a consists of nature sounds and white noise, sound b consists of nature sounds, babble sounds and white noise, sound c consists of babble sounds and white noise, while sound d (noise) was composed of white noise only.

Table 1: Proportion of the raw sounds in the different masking sounds generated and used as input to the optimization process. 'noise' denotes white noise, 'bab' stands for babble noise and 'nat' stands for nature sounds.

sound	noise	bab.1	bab.2	bab.3	nat.1	nat.2	nat.3
a (1)	0.1	0	0	0	0.5	0.5	0.5
b (6)	0.1	0.5	0.5	0.5	0.5	0.5	0.5
c (8)	0.1	0.5	0.5	0.5	0	0	0
d (noise)	1	0	0	0	0	0	0



(a) 'a' (b) 'b' (c) 'c'

Figure 3: Spectra of 'a' (left), 'b' (center), and 'c' (right)

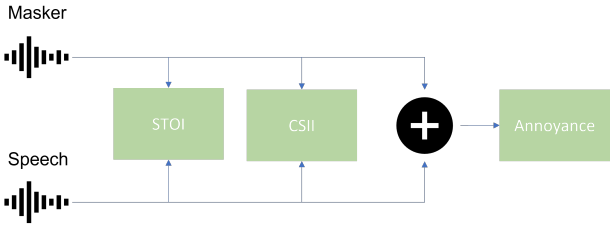


Figure 4: Experimental setup

The spectra of each of the raw masker sounds are given in fig.3; with sound 'a' at the top, sound 'b' second to top, and sound 'c' at the bottom. One can see some sonic patterns in the spectrum; at second 5, for example, between 5 kHz and 10 kHz for sounds 6 and 8. The same pattern is not visible in sound 1, however sound 'a'and sound 'b' seem to share similar content in the 1 kHz-2 kHz frequency band. Similarly, sound 'b' and 8 have similar content in the 0-1 kHz band. sound 'b' seem to be the sound sharing patterns with most other sounds, which is

confirmed by the information provided in Table1. In that sense, sound 'b' is the closest to white noise because it is the "fullest" in frequency.

5. RESULTS

This section presents the measured Intelligibility (STOI and CSII) as well as the annoyance of the masker generated by the proposed methods.

The masker and the input speech are combined to simulate the resulting acoustic signal heard in practice. Annoyance is computed on this signal, while the clean input speech and the masker are used to calculate the intelligibility metrics, as shown in Fig. 4.

All experiments were conducted with a fixed Masker to Signal Ratio (MSR) of -3 dB, meaning the masker power is adjusted at each frame according to the input speech power. This value was chosen as it produced reasonable intelligibility levels.

5.1 Hyper parameter selection

5.1.1 Window length

We tested analysis windows of 25 ms, 50 ms, 100 ms, 200 ms, and 400 ms at an MSR of -6 dB for four masker sounds (Table 1). A 100 ms window consistently provided the best balance between low annoyance and low intelligibility. At -6 dB, it produced the lowest annoyance for sounds 6 and 8, the second lowest for white noise (by only 0.3), and a mid-range value for sound 1. At -3 dB, the 100 ms window again gave competitive annoyance levels, remaining close to the lowest values across all sounds. Intelligibility variations were small at -6 dB, staying low (under 0.2 for STOI and 0.10 for CSII), except for sound 1, where the 100 ms window produced the lowest scores—both below the 10% word recognition threshold (as explained in 2). At -3 dB, the 100 ms window again yielded the lowest STOI and CSII scores for sound 1 and remained highly competitive for other sounds. These results lead us to conclude that a 100 ms window is the most suitable for our experiment.

5.1.2 Number of Cepstral Coefficients

The number of cepstral coefficients for spectral smoothing affects the signal reconstruction and thus the masker, by controlling spectrum smoothness. We tested various coefficient counts to assess their impact on intelligibility and annoyance.

Table 2: Effect of window length at MSR of -6dB.

	Sound 'a'			Sound 'b'			Sound 'c'			Noise		
Window	STOI	CSII	PA	STOI	CSII	PA	STOI	CSII	PA	STOI	CSII	PA
25ms	0.15	0.12	99.8	0.10	0.06	93.2	0.11	0.06	87.2	0.10	0.08	99.8
50ms	0.15	0.10	94.1	0.10	0.05	88.4	0.11	0.06	83.1	0.11	0.07	95.6
100ms	0.11	0.05	75.7	0.11	0.07	61.2	0.13	0.08	58.8	0.10	0.07	68.0
200ms	0.14	0.09	73.7	0.11	0.05	66.7	0.11	0.05	64.7	0.11	0.06	68.8
400ms	0.18	0.10	67.9	0.15	0.05	62.9	0.15	0.05	60.3	0.16	0.07	67.7

Table 3: Effect of window length at MSR of -3dB.

	Sound 'a'			Sound 'b'			Sound 'c'			Noise		
Window	STOI	CSII	PA	STOI	CSII	PA	STOI	CSII	PA	STOI	CSII	PA
25ms	0.27	0.45	75.0	0.16	0.13	77.8	0.17	0.13	76.6	0.21	0.36	76.5
50ms	0.18	0.18	80.4	0.16	0.12	66.2	0.16	0.12	66.4	0.18	0.26	67.5
100ms	0.14	0.10	64.8	0.15	0.12	52.1	0.18	0.17	50.6	0.15	0.14	58.1
200ms	0.17	0.14	62.6	0.14	0.09	51.7	0.15	0.08	52.6	0.15	0.11	59.5
400ms	0.24	0.14	54.9	0.25	0.19	50.4	0.24	0.14	49.8	0.23	0.21	57.6

Table 4 show the metrics for three different raw masker sounds. In most cases, intelligibility was only slightly affected, except for the CSII values for sound 'c'. The general trend is that for sounds 6 and 8, higher coefficients lead to higher intelligibility and lower annoyance. For sound 'a', intelligibility remains constant while annoyance decreases with more coefficients. In contrast, for noise, intelligibility decreases and annoyance increases with more coefficients. Based on these results, we selected 250 coefficients as a good compromise between annoyance and intelligibility.

5.2 Evaluation of the proposed optimization method

To validate our method, we evaluated its performance in different iterations to confirm that optimization improved the metrics as expected. We repeated the experiment with four masker sounds (Table1) to ensure consistency. Fig.5 shows the metric evolution for these sounds, with the origin on the x-axis representing 0 iterations (i.e., the raw masking signal as the baseline). For mixes 6, 8 and Noise, optimization decreased intelligibility (STOI and CSII), while mix 1 showed the opposite. However, annoyance decreased significantly for all cases, resulting in more agreeable maskers. This effect likely arises from the algorithm shaping the raw sound by retaining energy where speech is present and omitting regions without signal, producing a quieter and more tolerable masker. Interestingly, for three out of four sounds, the masking power diminished with optimization. We interpret this as the algorithm potentially failing to fully mask speech due to fixed window sizes or frequency-domain errors. If speech is present beyond the frame duration analyzed, it may remain un-

Table 4: Effect of the number of Cepstral coefficients on the measured metrics, at -3 dB.

N. Coef	Sound 'a'			Sound 'b'			Sound 'c'			Noise		
	STOI	CSII	PA	STOI	CSII	PA	STOI	CSII	PA	STOI	CSII	PA
16	0.14	0.12	67.7	0.15	0.12	56.8	0.18	0.17	54.9	0.17	0.22	54.8
32	0.15	0.11	66.8	0.16	0.12	57.1	0.19	0.17	54.8	0.16	0.16	56.5
64	0.14	0.11	64.7	0.16	0.14	53.2	0.20	0.21	52.3	0.15	0.12	59.3
128	0.13	0.10	65.0	0.17	0.16	52.8	0.20	0.23	51.1	0.15	0.13	58.3
256	0.13	0.10	63.1	0.18	0.18	51.2	0.21	0.28	49.3	0.15	0.13	57.9
512	0.14	0.11	60.5	0.18	0.18	50.1	0.21	0.28	48.8	0.16	0.17	56.4

masked. This suggests a weaker than expected correlation between the masking curve in [17] and the actual intelligibility: Despite producing a theoretically better masker, the algorithm did not achieve lower intelligibility as anticipated.

5.3 Evaluation against the baseline

As a baseline, we measured intelligibility and annoyance metrics for a pink noise masker, used as the control signal for our optimization framework. To ensure a fair comparison, we block-processed the pink noise masker to adjust its energy according to the maskee's power, keeping the SNR constant. Fig. fig. 6 shows the metric evolution for different SNR values. To achieve a CSII of 0.5, we need an MSR of about -2.5 dB, yielding an Annoyance of around 56. The graphs also reveal that while CSII follows SNR monotonously, STOI shows discontinuities in the low-intelligibility range, suggesting saturation. For clearer comparison, Fig. 7 places the baseline and proposed method side-by-side, showing that our method consistently achieves lower annoyance for similar intelligibility at any SNR level. This confirms that our framework produces more efficient and less annoying maskers without compromising intelligibility, a highly successful outcome.

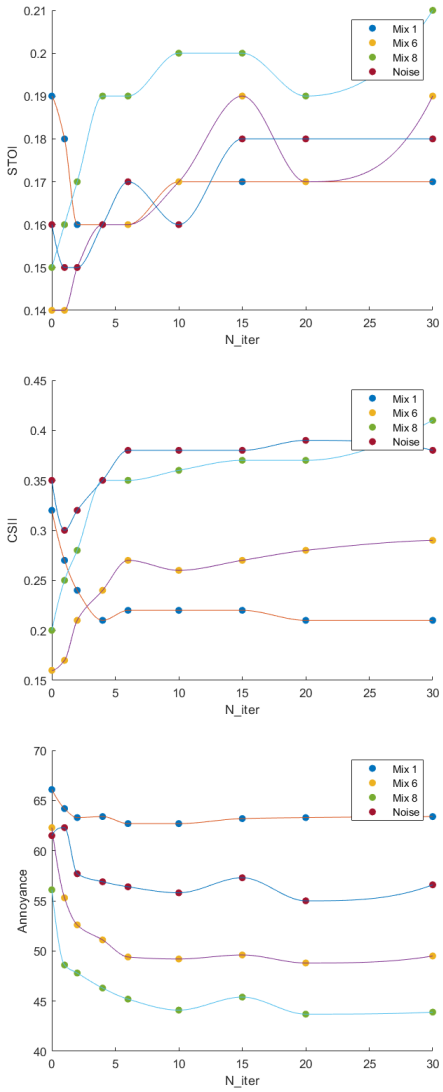


Figure 5: STOI, CSII and Annoyance metrics variation with the number of iterations, for masking sound 1

6. CONCLUSION

We have introduced a framework which optimizes the level of a masker sound at different frequency bins, in order to reduce the total energy of the masker by concentrating it only in frequency regions where it is needed. This is achieved by comparing the masking curve of the masker to the target speech maskee's spectrum, and finding a filter to shape the masker's spectrum so that the masking curve matches the maskee's spectrum. Using some intelligibility metrics as well as annoyance metrics from the literature (CSII, STOI, PA), our experimental results show that the masker signal produced by this framework exhibits lower annoyance at similar levels of intelligibility. Therefore this can find practical use in the design of masking systems for speech privacy, or at least guide the design of such systems, as used in office spaces or in cars for example.

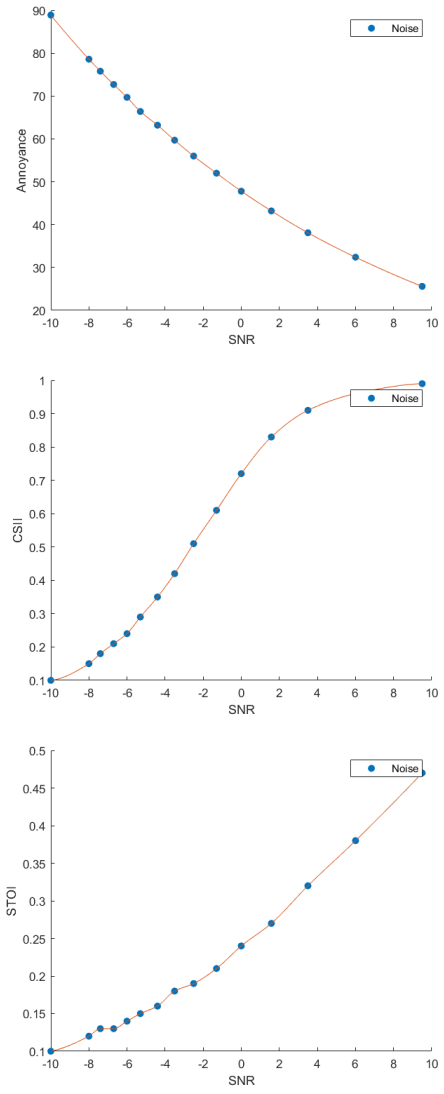


Figure 6: Evolution of STOI, CSII and Annoyance values for a pink noise masker vs. SNR.

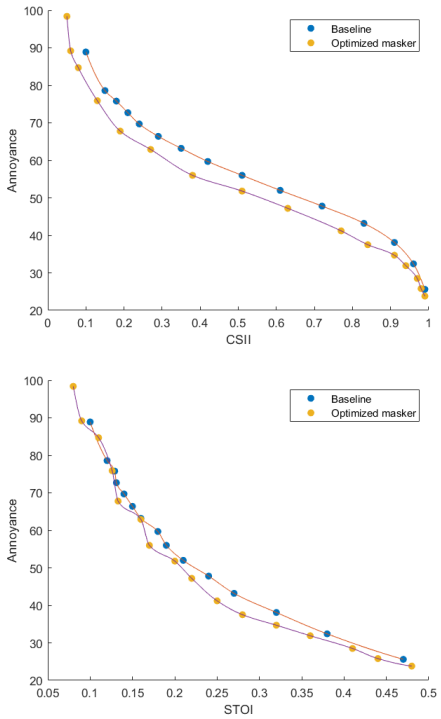


Figure 7: Comparison of baseline masker and optimized masker performance by looking at their corresponding Annoyance and SNR levels in function of CSII and STOI.

7. REFERENCES

[1] J. Donley, C. Ritz, and W. B. Kleijn, "Improving speech privacy in personal sound zones," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 311–315, 2016.

[2] J. Donley, C. Ritz, and W. B. Kleijn, "Multizone soundfield reproduction with privacy- and quality-based speech masking filters," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1041–1055, 2018.

[3] D. Wallace and J. Cheer, "Optimisation of personal audio systems for intelligibility contrast," 05 2018.

[4] J. Carlsson, "Evaluation of masking sounds in an existing open-plan office master's thesis in the master's programme in sound and vibration," tech. rep.

[5] V. Hongisto, "Effect of sound masking on workers in an open office," tech. rep.

[6] A. Haapakangas, E. Kankkunen, V. Hongisto, P. Virjonen, D. Oliva, and E. Keskinen, "Effects of five speech masking sounds on performance and acoustic satisfaction. implications for open-plan offices," *Acta Acustica united with Acustica*, vol. 97, pp. 641–655, 7 2011.

[7] W. Zhu, L. Luo, J. Sun, and M. Christensen, "A new variable step size algorithm based hybrid active noise

control system for gaussian noise with impulsive interference," in *2020 IEEE 6th International Conference on Computer and Communications, ICCC 2020*, (United States), pp. 1072–1076, IEEE, Dec. 2020. Publisher Copyright: © 2020 IEEE.; 6th IEEE International Conference on Computer and Communications, ICCC 2020 ; Conference date: 11-12-2020 Through 14-12-2020.

[8] R. Serizel, M. Moonen, J. Wouters, and S. Jensen, "A speech distortion weighting based approach to integrated active noise control and noise reduction in hearing aids," *Signal Processing*, vol. 93, p. 2440–2452, Sept. 2013.

[9] S. Bhattacharjee, L. Shi, G. Ping, X. Shen, and M. Christensen, "Study and design of robust personal sound zones with vast using low rank rirs," in *International Conference on Acoustics, Speech, and Signal Processing*, (United States), IEEE Signal Processing Society, May 2023.

[10] T. Lee, L. Shi, J. Nielsen, and M. Christensen, "Fast generation of sound zones using variable span trade-off filters in the dft-domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 363–378, 2021.

[11] D. Wallace, "Practical audio system design for private speech reproduction thesis for the degree of doctor of philosophy," tech. rep., 2020.

[12] R. Waller, "Office acoustics—effect of background noise," *Applied Acoustics*, vol. 2, no. 2, pp. 121–130, 1969.

[13] A. C. C. Warnock, "Acoustical privacy in the landscaped office," *Journal of the Acoustical Society of America*, vol. 53, pp. 1535–1543, 1973.

[14] S. J. Schlittmeier and J. Hellbrück, "Background music as noise abatement in open-plan offices: A laboratory study on performance effects and subjective preferences," *Applied Cognitive Psychology*, vol. 23, pp. 684–697, 2009.

[15] Y. Zhang, D. Ou, and S. Kang, "The effects of masking sound and signal-to-noise ratio on work performance in chinese open-plan offices," *Applied Acoustics*, vol. 172, 1 2021.

[16] F. Zarei, J. Lee, R. Mackenzie, and V. L. Men, "Evaluation of the uniformity of sound-masking systems in an open-plan office," *Applied Acoustics*, vol. 186, 1 2022.

[17] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 4214 – 4217, 04 2010.

[18] S. V. D. Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *Eurasip Journal on Applied Signal Processing*, vol. 2005, pp. 1292–1304, 6 2005.

[19] J. O. Smith, *Spectral Audio Signal Processing*. `http`:http://ccrma.stanford.edu/ jos/sasp//–`ccrma.stanford.edu/~jos/sasp/`, accessed 2023. online book, 2011 edition.

[20] J. M. Kates and K. H. Arehart, "Coherence and the speech intelligibility index," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2224–2237, 4 2005.

[21] U. Widmann, "A psychoacoustic annoyance concept for application in sound quality," *The Journal of the Acoustical Society of America*, vol. 101, pp. 3078–3078, 05 1997.