TOWARDS SPEECH PRIVACY ASSESSMENT FOR VOICE ASSISTANTS: EXPLORING SUBJECTIVE AND OBJECTIVE MEASURES FOR BABBLE NOISE

Anjana Rajasekhar[1]*, Anna Leschanowsky*[1]*, Nils Peters*[2]

[1]*Fraunhofer IIS,*
[2]*Friedrich-Alexander-Universität Erlangen-Nürnberg, International Audio Laboratories*[*]
anjana.rajasekhar@iis.fraunhofer.de

Abstract: The growing prevalence of voice assistants has sparked privacy concerns with respect to content privacy and potential human-based attacks such as eavesdropping which make users feel uncomfortable utilizing them in public. To address these challenges, understanding human privacy perceptions in acoustic environments becomes paramount. This understanding can empower voice assistants to accurately quantify privacy perceptions, adapt conversational patterns, and ultimately enhance human-machine interaction. This study draws inspiration from human-to-human interactions and previous research on acoustic privacy, to quantify privacy perceptions in environments characterized by babble noise. The primary objective is a comprehensive evaluation of both objective and subjective measures to quantitatively capture privacy perceptions in acoustic environments.

1 Introduction

Voice-based devices and virtual assistants have become increasingly popular and significantly influence how tasks are performed and services are utilized. Despite advantages, the rise of privacy concerns regarding these assistants has prompted widespread attention. This paper specifically addresses content privacy, i.e., sensitive linguistic information within audio recordings that is transmitted between humans and devices in loudly spoken form [1]. Thereby, we focus on human-based attacks rather than machine-based attacks, particularly on the potential threat of eavesdropping.

We distinguish two cases: 1) users sharing sensitive information with the machine and 2) the machine vocalizing sensitive information. In the first case, users possess full control over the information they are sharing with devices. To prevent their sensitive information from being eavesdropped, they potentially employ measures such as self-censorship, avoiding the usage of voice assistants in certain environments, or utilizing privacy solutions like voice masks[1] However, in the second case, users have limited control over the sensitive information shared by the device. This lack of control may lead to users avoiding voice assistants altogether - an option unavailable to blind users who rely heavily on text-to-speech technology in the form of screen readers [2]. Secure sound zones using jamming noise allow users to interact privately with voice-based devices but negatively impact the environment through noise pollution as well as user experience [3, 4].

[*]The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.

[1]for instance `https://kck.st/2pvXbr6`, `https://tcrn.ch/3S7IFGS`

 DOI: 10.35096/othr/pub-7088, URN: urn:nbn:de:bvb:898-opus4-70885

Drawing insights from human-to-human conversations, our research aims to contribute to preventing human-based attacks, specifically eavesdropping, in scenarios where users have limited control. Observing that people naturally adjust their communication style based on factors like the acoustic environment, recipient and the type of information [5], context-aware voice-based assistants could similarly tailor their conversation style to safeguard sensitive information from eavesdropping. Previous studies on human-to-human conversations have explored speech privacy in specific contexts, like hospitals and open-plan offices, to understand people's privacy perceptions, and expectations and inform the acoustic design of these environments [6, 7, 8, 9]. One study investigated privacy perception in three distinct noise scenarios revealing that acoustic information significantly impacts how listeners perceive privacy [10]. To the best of our knowledge, we are first to build upon this foundation and explore how the number of background speakers surrounding a hypothetical private conversation influences privacy perceptions utilizing both objective and subjective measures.

2 Experiments

To explore individuals' subjective privacy perceptions and to investigate the appropriateness of current objective measures for speech privacy assessment, sound samples, sourced from the TIMIT dataset [11], are generated as a mix of one target speaker and varying numbers of background speakers at 0 dB signal-to-noise ratio (SNR). To avoid gender-specific influences, the number of background speakers constitutes an equal distribution of male and female speakers.

2.1 Objective Measures

First, we investigate several objective metrics to assess their applicability for measuring speech privacy in the context of voice assistants. Traditional speech privacy standards in the room acoustics field often rely on impulse response measurements, e.g., the speech transmission index (STI) [12], rendering them unsuitable for the usage of voice assistants. However, there is consensus that speech privacy is directly linked to speech intelligibility (SI) [12]. Consequently, speech privacy standards often use SI as an inverse proxy for acoustic privacy [12, 6]. This assumption has led to the development of measures such as the articulation index (AI) and its successor, the speech intelligibility index (SII) [12].

Furthermore, objective SI measures have been thoroughly researched in the field of speech enhancement and speech communication distinguishing between intrusive and non-intrusive [13]. Intrusive measures rely on a reference signal, requiring both clean and noisy speech samples, while non-intrusive measures operate solely on the noisy speech sample [13]. While a clean reference signal is usually not available in real-time applications like voice assistants, intrusive measures can be adapted to non-intrusive ones by first estimating the clean reference signal [14].

In our analysis, we focus on well-established intrusive SI measures as well as on lightweight measures of stationarity. Therefore, we generate samples as described above utilizing ten female and ten male target speakers and varying numbers of background speakers.

The **Speech Intelligibility Index (SII)** is a metric used to quantify and assess the intelligibility of speech in various acoustic environments [15]. It provides a numerical measure of how well speech can be understood or comprehended under specific conditions. The SII is typically expressed as a value between 0 and 1, with 0 indicating poor intelligibility and 1 representing perfect intelligibility. For computation, we rely on the recently released Python implementation of the standard SII protocol [2].

[2] `https://sii.to/programs.html`

The **short-time objective intelligibility (STOI)** measure was primarily developed to assess speech intelligibility after time-frequency weighting, commonly applied in speech enhancement or speech separation techniques [16]. Yet, it was also found appropriate for the evaluation of noisy signals before enhancement. Like SII, it takes on values between 0 and 1 with 1 indicating perfect speech intelligibility. However, STOI was found to perform poorly for modulated noise such as competing speech signals akin to our scenario [17]. To address this limitation, an **extended short-time objective intelligibility (ESTOI)** measure has been proposed [17]. Despite the weaknesses of STOI, we compute both metrics using a Python implementation [3] for completeness.

Spectral flatness (SF) is a metric that characterizes the distribution of energy across different frequencies in a signal [18]. A high SF value indicates a more uniform distribution of energy, resembling a flat spectrum. In situations with fewer background speakers tonal structures in the spectrum are anticipated, leading to lower SF values. In contrast, scenarios with more background speakers and increased noisiness are expected to show an opposite trend. To assess spectral flatness in our experiment, we only rely on the environmental acoustic information, i.e., the mix of background speakers excluding the target speaker.

The **zero-crossing rate** indicates how frequently the signal value crosses the zero axes with noisy sounds tending to have high zero-crossing rates [18]. Here, an increase in the number of background speakers is expected to lead to higher zero-crossing rates. Consistent with the spectral flatness calculation, we only consider the background speaker mix for computing the zero-crossing rate.

2.2 Subjective Evaluation

We conducted a listening test with 12 participants (four females, and eight males) aged 19 to 57. We employed a paired comparison methodology using the webMUSHRA framework [19]. To keep the test completion time manageable, we opted for three distinct levels of background speakers, i.e., 2, 6, and 12. These levels were selected by employing the Automatic Speech Recognition (ASR) system Whisper [20] on the background speaker mix. We used Whisper model base to transcribe the background speaker mix and compared the transcribed output with the corresponding transcripts of audio samples. Due to the background speaker mix, we did not employ Word Error Rate (WER) but the fraction of correctly detected words was found by dividing the number of matching words by the total number of words in the reference transcripts. The three levels were chosen based on their associated significant decrease in word detection accuracy. Finally, each participant was presented with 30 randomized trials of audio samples, each featuring an equal number of male and female background speakers.

Initially, participants were tasked with estimating the number of background speakers in each sample as this aspect may be connected to the overall understanding and perception of the audio sample. Following this, they answered four privacy questions, including three from a previous study [10] and one additional question to complement the existing set. By adding the additional question, we aim to identify correlations and streamline future privacy perception tests by potentially condensing multiple questions into a single inquiry. Participants utilized a 7-point Likert scale (-3 to +3) on a slider to indicate their privacy perception. All rating tasks had to be completed before moving to the next trial.

[3] `https://github.com/mpariente/pystoi`

3 Results

3.1 Objective Measures

In Figure 1, we present the STOI, ESTOI and SII values obtained across varying levels of background speakers for 20 different speakers (10 male, 10 female). We do show aggregated scores only as no gender-specific differences could be observed. It's noteworthy that intelligibility methods do not provide a direct prediction of the fraction of words understood but rather a scalar value that is monotonically related to absolute intelligibility. Typically, a logistic mapping is applied based on subjective data to establish this relationship [17]. We refrain from predicting absolute intelligibility, as we did not conduct a subjective intelligibility test and are not concerned with exact intelligibility scores. Instead, we are interested in differences between the number of background speakers and possible correlations with experienced subjective speech privacy. Despite significant differences between STOI estimates and ESTOI and SII estimates, our analysis did not show a significant influence of the number of background speakers on objective intelligibility measures. Overall, STOI scores were consistently higher which is in line with previous research [21, 22].

Furthermore, we investigated measures of stationarity including spectral flatness and zero-crossing rate. Contrary to our expectations, we did not observe the anticipated trend of signals with an increasing number of background speakers to approach spectral flatness values of one or higher zero crossing rates.

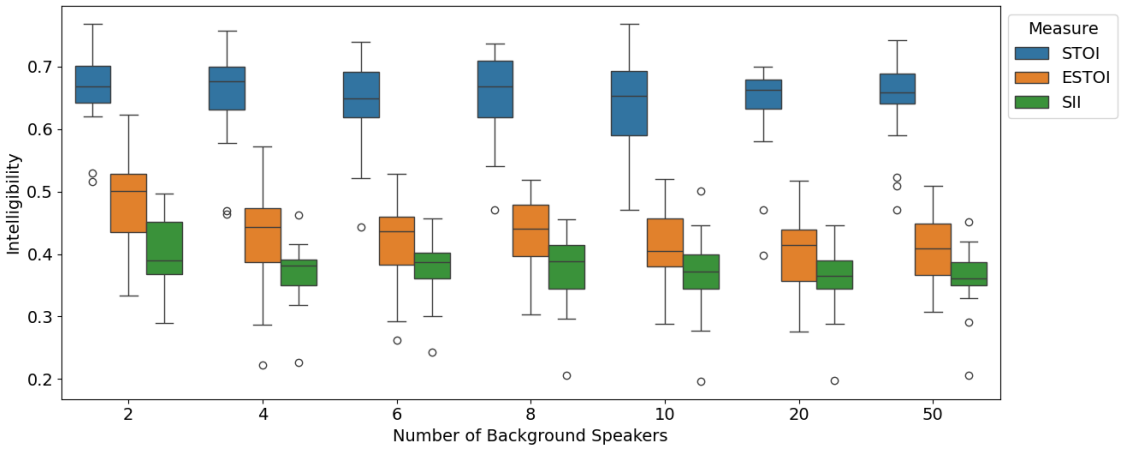

Figure 1 – Intelligibility scores provided for STOI, ESTOI, and SII and computed for 20 mixes (10 male and 10 female target speakers). We show aggregated scores only. Scores do not predict intelligibility per se but are monotonically related to absolute intelligibility [17].

3.2 Subjective Evaluation

Before analyzing participants' privacy perceptions, we assess reliability by checking the consistency of ratings. The principle of transitivity in preferences asserts that if a participant favors scenario A over B and B over C, then they also prefer A over C. In our experiment, participants are consistent within their rating for 78% of trials. As none of the participants showed persistent inconsistency, we did not exclude results for further analysis.

In the listening test, participants were asked to estimate the number of background speakers chosen on a range from 0 to 14. Figure 2 shows that more than 50% accurately guessed samples

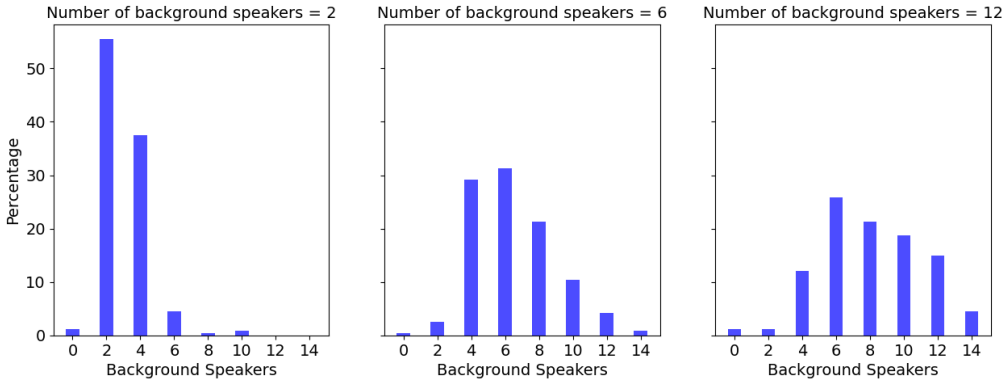

Figure 2 – Participants' Estimates of Background Speakers (original values are 2, 6 and 12) for given Options 0 to 14.

with two background speakers. However, only 30% accurately guessed samples with six background speakers, indicating the increased difficulty. Remarkably, when confronted with twelve background speakers, around 25% of the participants wrongly chose six, while nearly 15% correctly selected twelve. As the number of background speakers increases, predicting the exact number becomes notably challenging. Nevertheless, even when participants chose six instead of correctly 12 background speakers, they consistently assigned a lower number to the comparing sample. This suggests that participants were able to discern the increase in background speakers even when they guessed incorrectly.

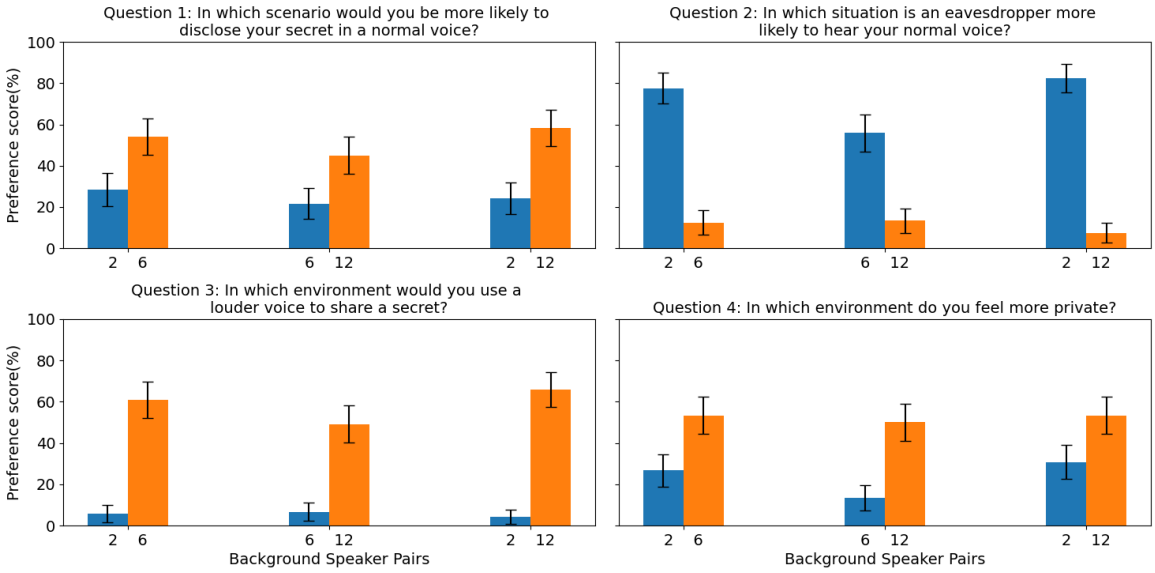

Figure 3 – Preference scores for each question with 95% CI are shown for the pairs (2,6), (6,12), and (2,12). Participants consistently favored higher background speaker levels in Questions 1, 3, and 4, while Question 2 exhibited a preference for lower background speaker levels.

Figure 3 illustrates the preference scores and corresponding 95% confidence intervals for each question concerning two, six and twelve background speakers. The preference scores, expressed as a percentage, are calculated based on the aggregate responses of the twelve participants. It becomes clear that in contrast to the objective measures, participants' preferences are significantly distinct for each comparison across the four questions. Notably, for Questions 1, 3, and 4 (Figure 3), participants consistently favored samples with more background speakers.

Conversely, for Question 2 – 'In which situation is an eavesdropper more likely to hear your normal voice?' – higher preference scores were consistently obtained for samples with fewer background speakers. This inverse relationship of Questions 1, 3, and 4 with Question 2 aligns with our anticipation, considering the reverse wording and focus on the attacker in the second question. To potentially condense questions, we conducted a correlation analysis, using the Pearson correlation coefficient [23]. We found a strong negative correlation only for questions 3 and 2 ($r = -0.7$), and a moderate negative correlation for questions 1 and 2 ($r = -0.43$) and questions 1 and 4 ($r = 0.45$). As most of the questions did not show strong correlations, we can not recommend shortening the questionnaire.

4 Discussion

This study investigated objective and subjective measures to assess privacy perceptions in acoustic settings with varying numbers of background speakers. Despite testing established metrics like STOI, ESTOI, SII, spectral flatness, and zero-crossing rate, no significant variations were observed, making them unsuitable for predicting privacy perceptions. This highlights the need for more appropriate objective measures that could predict the level of subjective privacy in acoustic settings. While ASR-based intelligibility measures could be beneficial, they face difficulties in complex acoustic conditions like competing talkers [14]. Based on previous speech privacy standards [12], we assumed a direct link between SI and speech privacy. Yet, we did not conduct a subjective intelligibility test limiting a deeper understanding of the relationship in our setting. Moreover, a recent study found that the dissatisfaction, used as an inverse proxy for speech privacy, was related non-linearly with SI and influenced by additional factors [6]. Therefore, future research could explore measures unrelated to SI such as speaker count estimation as a non-intrusive alternative [24].

Our study represents a pioneering effort as the first to conduct a subjective analysis based on privacy perception and babble noise. The results unveil noteworthy insights, particularly observing a stronger opinion for loudness and attack questions, potentially linked to participants' existing awareness influenced by the Lombard effect [25]. This brings up the question: Are privacy questions more inherently subjective? Additionally, what is the effect of adding environmental sound cues, e.g., a train, with the babble noise on the privacy perception? These questions open a new dimension for future work. We need to analyze whether the combination of environmental cues with babble noise aligns with our study and previous work [10]. Furthermore, the correlation analysis didn't show a strong correlation between Questions 1, 3, and 4 or a strong inverse correlation of all questions with Question 2. This calls for a qualitative approach as the factors influencing correlation are unclear. Therefore, we intend to retain all four questions in future research, as each question provides unique insights into user privacy perceptions. Participants demonstrated awareness of increasing background speakers, despite challenges in precise quantification of background speaker level. In the case of privacy questions, consistent preference for more background speakers was noted in Questions 1, 3, and 4. Notably, Question 2 – 'In which situation is an eavesdropper more likely to hear your normal voice?' – consistently preferred fewer background speakers, revealing nuanced preferences in eavesdropping scenarios. It is important to note, however, that these responses were derived from a paired comparison test, and uncertainties persist regarding individual presentation responses. This uncertainty also paves the way for a valuable avenue in our future research endeavors.

5 Conclusion

As voice assistants continue to be deployed in diverse acoustic settings, understanding user perceptions is crucial for developing systems aligned with expectations and concerns. Our research contributes to this by exploring how individuals perceive privacy in the presence of background speakers. Our subjective evaluation reveals significant differences in user preferences, while various objective measures show no differences. This urges the need for alternative objective measures to capture users' privacy perceptions and allow the development of context-aware voice assistants that effectively safeguard sensitive information from eavesdropping.

References

[1] WILLIAMS, J., K. PIZZI, S. DAS, and P.-G. NOE: *New Challenges for Content Privacy in Speech and Audio*. In *2nd Symposium on Security and Privacy in Speech Communication*, pp. 1–6. 2022. doi:10.21437/SPSC.2022-1.

[2] WILLIAMS, J., J. YAMAGISHI, C. PAUL-GAUTHIER, VALENTINI-BOTINHAO, and J.-F. BONASTRE: *Revisiting Speech Content Privacy*. In *2021 ISCA Symposium on Security and Privacy in Speech Communication*, pp. 42–46. ISCA, 2021. doi:10.21437/SPSC.2021-9.

[3] ZHU, H., X. WANG, Y. JIANG, S. CHANG, and X. WANG: *Secure Voice Interactions With Smart Devices*. *IEEE Transactions on Mobile Computing*, 22(1), pp. 515–526, 2023. doi:10.1109/TMC.2021.3069981.

[4] DONLEY, J., 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. IEEE, Shanghai, 2016. doi:10.1109/ICASSP.2016.7471687.

[5] NISSENBAUM, H.: *Privacy as Contextual Integrity*. *Washington Law Review*, 79, p. 41, 2004.

[6] SATO, H., M. MORIMOTO, S. OHTANI, Y. HOSHINO, and H. SATO: *Subjective evaluation of speech privacy at consulting rooms in hospitals: Relationship between feeling evoked by overhearing speech and word intelligibility score*. *Applied Acoustics*, 124, pp. 38–47, 2017. doi:10.1016/j.apacoust.2017.03.020.

[7] CLAMP, P. J., D. G. GRANT, D. A. ZAPALA, and D. B. HAWKINS: *How private is your consultation? Acoustic and audiological measures of speech privacy in the otolaryngology clinic*. *European Archives of Oto-Rhino-Laryngology*, 268(1), pp. 143–146, 2011. doi:10.1007/s00405-010-1342-8.

[8] CAVANAUGH, W. J. and G. C. TOCCI: *Speech privacy in healthcare buildings:review of early studies and current procedures for analysis*. *The Journal of the Acoustical Society of America*, 123(5_Supplement), pp. 3193–3193, 2008. doi:10.1121/1.2933328.

[9] UTAMI, S. S., J. SARWONO, N. A. ROCHMADI, and N. SUHERI: *Speech privacy and intelligibility in open-plan offices as an impact of sound-field diffuseness*. Inter-noise 2014, Melbourne, Australia, November, 2014.

[10] LESCHANOWSKY, A., S. DAS, T. BÄCKSTRÖM, and P. P. ZARAZAGA: *Perception of privacy measured in the crowd – paired comparison on the effect of background noises*. INTERSPEECH 2020, Shanghai, China, October 25–29, 2020.

[11] GAROFOLO, J., L. LAMEL, W. FISHER, J. FISCUS, D. PALLET, N. DAHLGREN, and V. ZUE: *Timit acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium, 1992.

[12] STOUT, J.: *Speech privacy standards*. Cambridge Sound Management, Inc, 2015.

[13] FENG, Y. and F. CHEN: *Nonintrusive objective measurement of speech intelligibility: A review of methodology*. *Biomedical Signal Processing and Control*, 71, p. 103204, 2022.

[14] KARBASI, M. and D. KOLOSSA: *Asr-based speech intelligibility prediction: A review*. *Hearing Research*, p. 108606, 2022.

[15] *ANSI/ASA S3.5-1997 (R2017) - Methods for Calculation of the Speech Intelligibility Index*. https://webstore.ansi.org/standards/asa/ansiasas31997r2017, 1997.

[16] TAAL, C. H., R. C. HENDRIKS, R. HEUSDENS, and J. JENSEN: *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217. 2010. doi:10.1109/ICASSP.2010.5495701.

[17] JENSEN, J. and C. H. TAAL: *An algorithm for predicting the intelligibility of speech masked by modulated noise maskers*. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11), pp. 2009–2022, 2016. doi:10.1109/TASLP.2016.2585878.

[18] PEETERS, G.: *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. Tech. Rep., Icram, 2004.

[19] SCHOEFFLER, M., F.-R. STÖTER, B. EDLER, and J. HERRE: *Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (MUSHRA)*. In *1st Web Audio Conference*, pp. 1–6. 2015.

[20] RADFORD, A., J. KIM, T. XU, G. BROCKMAN, K. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. In *Proceedings of the 40th International Conference on Machine Learning*. 2023.

[21] BRAMSLØW, L., G. NAITHANI, A. HAFEZ, T. BARKER, N. H. PONTOPPIDAN, and T. VIRTANEN: *Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm*. *The Journal of the Acoustical Society of America*, 144(1), pp. 172–185, 2018. doi:10.1121/1.5045322.

[22] LÓPEZ-ESPEJO, I., A. EDRAKI, W.-Y. CHAN, Z.-H. TAN, and J. JENSEN: *On the deficiency of intelligibility metrics as proxies for subjective intelligibility*. *Speech Communication*, 150, pp. 9–22, 2023. doi:10.1016/j.specom.2023.04.001.

[23] *SPSS tutorials: Pearson correlation*. Kent State University, 2023.

[24] STÖTER, F.-R., S. CHAKRABARTY, B. EDLER, and E. A. HABETS: *Classification vs. regression in supervised learning for single channel speaker count estimation*. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 436–440. IEEE, 2018.

[25] GIGUÈRE, C., C. LAROCHE, E. BRAULT, J.-C. STE-MARIE, M. BROSSEAU-VILLENEUVE, B. PHILIPPON, and V. VAILLANCOURT: *Quantifying the lombard effect in different background noises*. *The Journal of the Acoustical Society of America*, 120(5), pp. 3378–3378, 2006.