Audio Engineering Society

Conference Paper 1

Presented at the AES 4th International Conference on Audio and Music
Induced Hearing Disorders
2024 May 29–31, Aalborg, Denmark

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Evaluation of Own Voice Perception while Using In-ear Headsets

Miho Takeuchi[1,2], Jeremy Marozeau[2], Randy F. Fela[3], Konstantinos Gkanos[3], and Sidsel M. Nørholm[4]

[1] *FalCom A/S (GN Store Nord), Denmark*
[2] *Hearing Systems, Department of Health Technology, Technical University of Denmark*
[3] *GN Audio A/S (Jabra), Denmark*
[4] *NIRAS A/S, Denmark*

Correspondence should be addressed to Miho Takeuchi (miho.takeuchi@outlook.com)

ABSTRACT

Singing or speaking while using a headset or hearing aid can present significant challenges due to the occlusion effect, alongside variances in processing latency and auditory distortions inherent to these devices. Such challenges are particularly pronounced in activities like singing, where precise control over one's vocal quality is paramount for performance. This study investigates the factors detrimental to the perceptual quality of one's own voice when using in-ear headsets and how each factor influences one's perception of self-voice. Prior literature reviews lead to occlusion, sidetone filters, latency, and distortion as the main factors. Twenty-three assessors participated in a subjective evaluation, which was followed by statistical data analysis. Further investigation was carried out through an assessor screening procedure and text analysis was performed on responses to collected open-ended questions. The study found that occlusion and distortion are the influential factors for the changes in self-voice perception, whereas latency is a determinant. Sidetone filtering becomes a significant factor when occlusion effects are minimized.

1 Introduction

Nowadays, it is common for individuals, regardless of normal or impaired hearing, to experience their own voice through an in-ear headset. One longstanding example are hearing aids, which compensates for hearing loss. As of late, modern in-ear monitors are commonly used for conversations in open spaces such as offices or venues for music performance. These devices can change how naturally we perceive the quality of our own voice, potentially hindering our ability to accurately control our voice to express intended emotions, or achieve the desired sound quality in activities such as during singing. Since in-ear headsets partially block out the airborne acoustic path from the mouth to the ear, the self-voice transmitted externally through the air is attenuated, while self-voice propagating internally through one's body accumulates. The self-voice is altered and is now perceived as muffled and unnatural [1]. However, as depicted in Figure 1, the issue can be mitigated by picking up one's own voice through the headset's microphones and playing back some of the self-voice through the headset, accounting for the losses in the acoustic path. This feature, called sidetone,

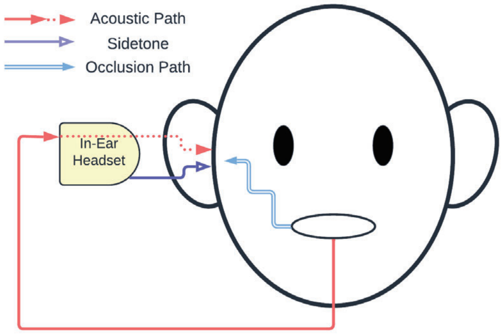

Fig. 1: Acoustical paths for self-voice when using in-ear headsets with sidetone function.

helps reproduce the open-ear experience of one's self-voice as much as possible.

In-ear headsets attenuate higher frequency components in the acoustic path, while the occlusion effect increases the lower frequency content, predominantly between 70–1000 Hz [2]. Therefore, the sidetone is filtered to counteract the changes in frequency content, and active noise cancellation (ANC) is introduced to remove parts of the occlusion effect [3]. The combination of the sidetone filter and the ANC should aim towards an open-ear experience represented by a 0 dB insertion gain (IG), a ratio comparing the sound pressure level in the ear canal between the ear blocked by a headset and the open ear [4].

However, limitations of the in-ear headset's drivers will introduce byproducts such as distortion, and the headset's signal processing will add latency to the system. With these active components in an in-ear headset, it is not a simple task to determine how these different factors influence the perceptual quality of one's voice.

A predecessor to sidetone on in-ear headsets is sidetone on handheld telephones, which was investigated by Appel & Beerends [5]. The main factors influencing the perceptual quality of one's self-voice when using handheld telephones are sidetone latency, sidetone playback level, sidetone distortion, and background noise. An increase in latency, distortion, and background noise causes a deterioration in the mean opinion score (MOS), whereas maintaining the playback level at a moderate level results in the best MOS.

On the contrary, the occlusion effect, a particular concern associated with modern in-ear headsets, can be eliminated, resulting in a notable enhancement in the perceived naturalness of the user's own voice, as previously examined by Liebich & Vary [6]. An implementation of occlusion effect cancellation in hearing devices to supplement a headset's ANC algorithm, commonly known as Feedback ANC (FB ANC), improved how a user perceives the naturalness of their own voice.

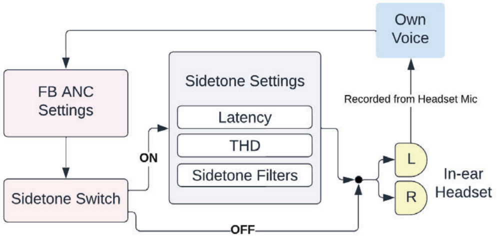

Fig. 2: A schematic of how sidetone & ANC settings of the in-ear headset are controlled.

Not much investigation is conducted on the factors that affect one's perception of self-voice when using in-ear headset settings for users with normal hearing [7]. Therefore, this study aims to answer the following research questions:

- What factors influence the perceptual quality of one's self-voice when using an in-ear headset?
- How does the perceptual quality improve or deteriorate as these factors are adjusted?
- Which factor(s) are the most detrimental in determining the perceptual quality of one's self-voice?

Based on preceding literature and study reviews, it can be hypothesized that the amount of occlusion, side-tone filtering, latency, and distortion might influence the perception of self-voice when using an in-ear headset for users with normal hearing. Within the given timeframe of the project, factors such as background noise and sidetone volume are temporarily left out. Combinations of these four factors will result in different settings on an in-ear headset, hence creating variations on the quality rating of one's self-voice when a user evaluates these scenarios.

This paper is organized as follows: Section 2 explains the audio apparatus used for the study (Section 2.1) and

Table 1: Description of levels of each audio setting and their relations to the identified factors

Setting	Level	Description
FB ANC	Max	Maximum occlusion removal
	Min	Minimal occlusion removal
Sidetone	On	Optimized to 0dB IG with max ANC
	Off	Sidetone is off
Latency	0ms	No latency
	10ms	10ms latency added to system
THD	Min	No distortion
	Max	High amount of distortion of 15%THD

Table 2: All 10 conditions used in the subjective test.

Sidetone	FB ANC	Latency	THD
On	Max	0ms	Min
On	Max	0ms	Max
On	Max	10ms	Min
On	Max	10ms	Max
On	Min	0ms	Min
On	Min	0ms	Max
On	Min	10ms	Min
On	Min	10ms	Max
Off	Max	N/a	N/a
Off	Min	N/a	N/a

the procedure of the subjective test (Section 2.2). The subjective test results are presented in Section 3 and discussed in Section 4, including the application of the assessor screening procedure and its implication on the presented study. Lastly, Section 5 & 6 concludes the study with suggestions for future work.

2 Methods

2.1 Participants

The study involved a total of 23 participants who are GN employees, consisting of 5 women and 18 men, within the age range of 25 to 60 years old. These participants have normal hearing, with English proficiency ranging from conversational to native.

2.2 Apparatus

The participants were equipped with an in-ear headset mock-up similar to the Jabra Elite 85t. The device's signal processing was done on Jabra's external development platform, which adjusts ANC and sidetone with a sampling rate of 48 kHz. Four audio settings were tested: the amount of FB ANC, the addition of sidetone filters, the amount of latency, and total harmonic distortion, THD. The setting levels available for each factor are shown in Table 1, with a description of their technical and perceptual changes. The setting levels are derived from objective measurements and fine-tuned with a pilot test. The ANC and sidetone filters are developed in conjunction with the author and Jabra.

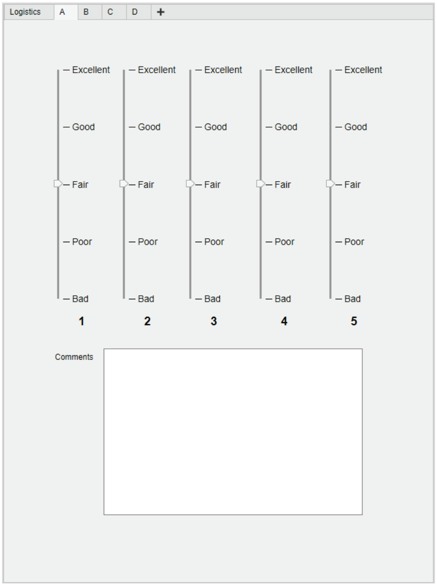

Fig. 3: Rating interface for the subjective test: The 20 sliders for 20 trials are divided into four pages for ease of accessibility.

Figure 2 illustrates the process of varying the parameters to produce different self-voice experiences. It is important to note that if the sidetone is off, no latency, distortion, or filtering effects will be added to the audio being fed back to the headsets. However, it will still be possible to have FB ANC on even though the sidetone is off.

2.3 Subjective Test Procedure

The subjective test employed a full-factorial design of experiments [8], consisting of 10 conditions derived from combinations of FB ANC (2 levels), Latency (2 levels), and THD (2 levels) with sidetone on, as well as FB ANC (2 levels) when sidetone is off (Table 2). Each

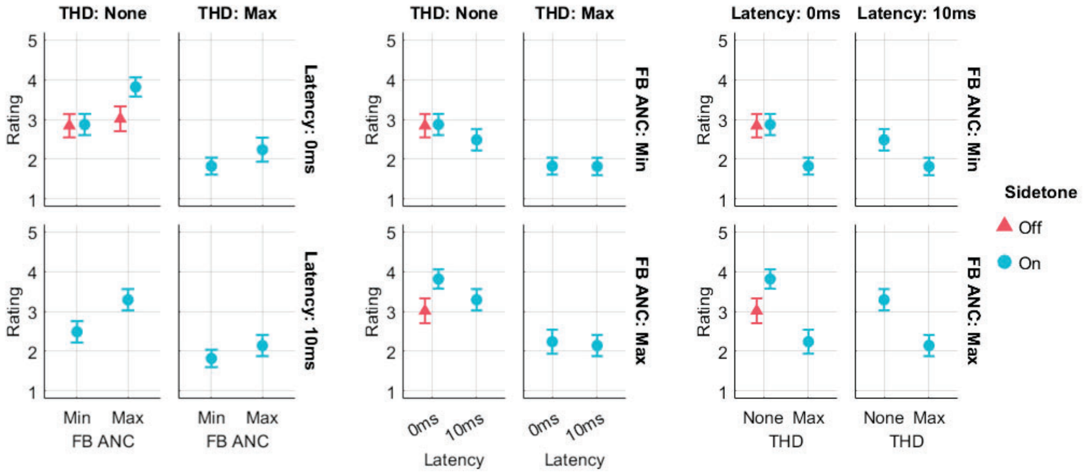

Fig. 4: Mean and CI of self-voice quality ratings for each condition.

condition was repeated twice, resulting in 20 trials per assessor.

The proctor would first set up a random condition. The participant then read a set of Harvard Standardized Sentences [9] at their normal speaking volume while listening to their own voice. The participant was allowed to remove the headsets to compare with the open-ear experience. Therefore, the participant evaluated the overall quality of their own voice on a continuous scale from 1 to 5 (Bad to Excellent) on a corresponding slider on the rating interface shown in Figure 3. The participant was given the option to justify their ratings and state their observations in the comment section.

The collected data from the experiment was analyzed using various statistical techniques, including mean and 95% confidence interval (CI) calculations [10], analysis of variance (ANOVA) [10], assessment of assessor reliability and discrimination ability [11, 12], as well as word frequency analysis of the comments. The relative significance between the F-values of each factor and their corresponding p-values from the ANOVA analysis was examined in detail to evaluate the significance of each factor when perceiving one's voice [10].

3 Results

The mean of the ratings for each condition are presented in Figure 4, with error bars representing each rating's 95% confidence interval (CI). Each panel represents the average of the same ten conditions plotted over different axes to help the visualization of the effect of each parameter. The left panel shows the effect of FB ANC on the axis. An increase in the average rating of one's self-voice quality can be observed with FB ANC across each condition (THD, Latency). The middle panel shows the effect of latency on the axis. A decrease in the rating with the latency can be seen when no THD was added. The right panel shows the negative effect of THD across all the other conditions (latency and FB ANC). Different colors represent the effect of sidetone. Turning the sidetone on improves the rating only when there is no THD and latency and when the FB ANC is on.

Table 3: ANOVA analysis on FB ANC, latency, THD, and their respective interactions. The interactions of all 3 factors are negligible.

	Factors	F-Value	p-value
3*Individual	FB ANC	33.88	< 0.0001
	Latency	2.8	0.0951
	THD	127.59	< 0.0001
3*Interactions	FB ANC & Latency	0.34	0.5596
	FB ANC & THD	2.71	0.1004
	Latency & THD	1.19	0.2758

The ANOVA analysis of the factors and their interactions are shown in Table 3. Due to the special nature

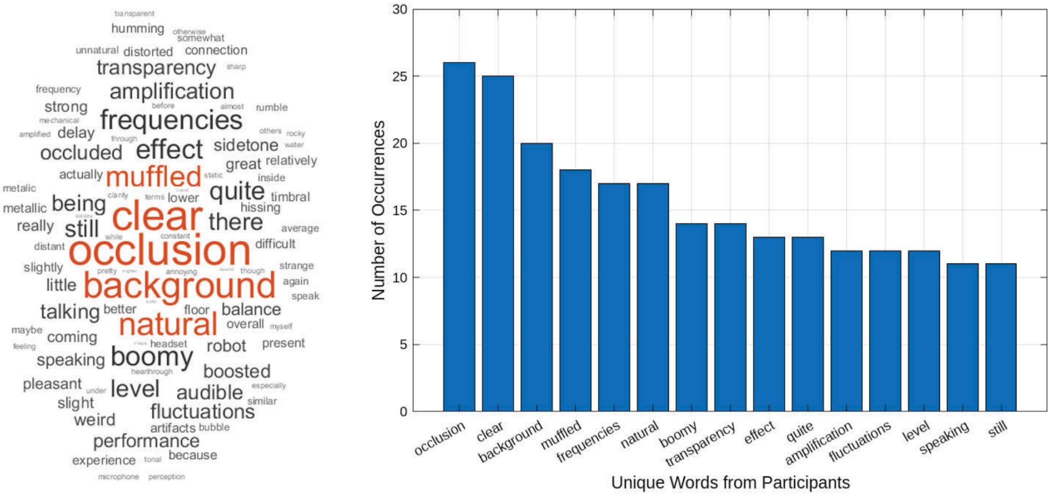

Fig. 5: Visualization of unique word usage of participant comments during the subjective test.

of the sidetone on/off setting, the sample sizes are different when comparing *Sidetone On* and *Sidetone Off*, which may skew the ANOVA test's results. Therefore, ANOVA is computed on the other 3 factors: FB ANC, latency, and THD. From the ANOVA analysis, the low F-values and p-values that exceed the $\alpha = 0.05$ from the interactions suggest a lack of significance between each other. Both FB ANC levels and THD levels indicate a statistically significant difference between each respective factor's two setting levels.

Finally, to support the quantitative relations deduced between one's self-voice quality rating and the factors of interest in this section, a word cloud and a word-frequency histogram were generated from the comments collected from the participants, shown in Figure 5.

4 Discussion

Based on the data analysis and visualization, it can be deduced that an increase in FB ANC and a reduction in THD result in a better self-voice quality rating. The effects of sidetone on one's perception of self-voice are dependent on the FB ANC, but the remaining three factors (FB ANC, Latency, and THD) are independent of one another. The effects of latency are indeterminate.

The significance of FB ANC on the influence of one's self-voice is likely due to the form factor of in-ear headsets giving an underlying impression of occlusion issues. This allows participants to pay closer attention to such issues during their evaluation process, suggested by how "occlusion" is the most mentioned word amongst participant comments in Figure 5. On the other hand, how THD levels influence one's perception of self-voice is likely due to the alienation of one's distorted voice, as participants described with negative words such as "distorted", "fluctuation", "strange", and "weird" for the cases where THD is maximized.

Table 4: ANOVA analysis between FB ANC and sidetone and their interaction.

Factors	F-Value	p-value
FB ANC	16.32	0.0001
Sidetone	9.03	0.003
FB ANC & Sidetone Interaction	7.72	0.006

Sidetone's dependency on FB ANC levels might be explained by performing ANOVA analysis on an isolated group of data points, removing data points corresponding to conditions with 10 ms latency or maximum THD (Table 2). Table 4 shows the ANOVA results of the isolated analysis between FB ANC and sidetone, where both factors have significant F-values, and low p-values. Moreover, the interactions between FB ANC and Sidetone also have a fairly significant F-value accompanied by a p-value below the threshold. Therefore,

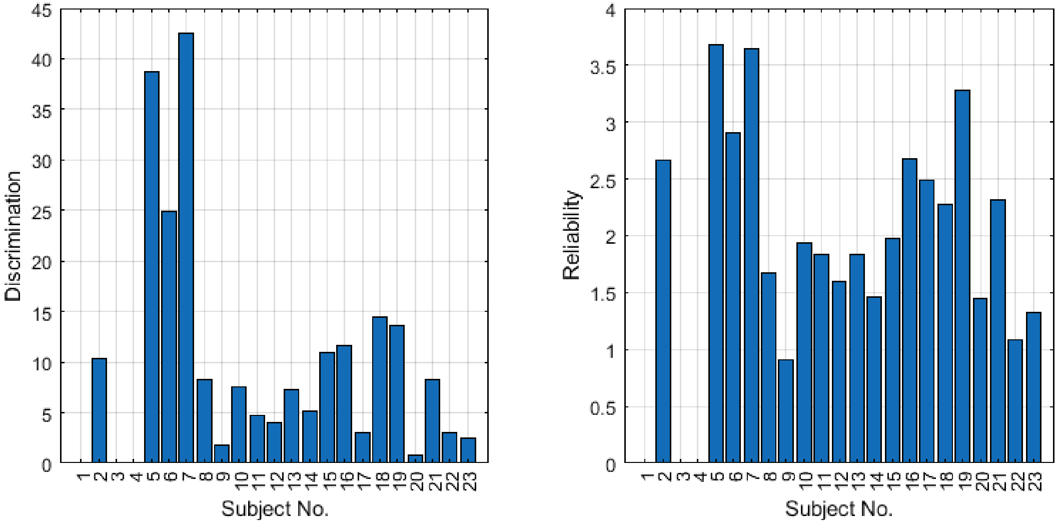

Fig. 6: Reliability and Discrimination metrics for all participants.

the ANOVA analysis suggests that both FB ANC and Sidetone have a significant influence on the perception of one's voice, but also indicates that their interaction with one another have influence on perception as well. This is also suggested by feedback received by participants specifically noting that they feel neutral about a completely passive headset with no sidetone only thinking that their voice is slightly muffled. Switching on sidetone only increases the volume of their own voice slightly, while having occlusion effects remain.

Even though the effect of latency appeared to be indeterminate, there were a few participants that mentioned words and phrases such as "delay", "echo", and "light early reflection", indicating that they were able to perceive the latency changes with critical listening. This suggests a listening skill gap between the 23 participants, where some are casual and untrained listeners and some were critical listeners. Assessor screening is therefore instigated. Participant's performance metrics were evaluated using eGauge, a set of screening methods recommended by ITU-R [11, 13, 14, 15].

Two thresholds were used – the first to weed out unreliable and non-discriminating participants, the second to distinguish the listening skills amongst the reliable participants. The first critical threshold of 1 was applied to both the discrimination and reliability metrics to separate the participants into two groups – a reliable and discriminating participant versus an unreliable or not-discriminating participant (see Figure 7). As Subjects 1, 3, 4, 9, and 20 are unreliable or have poor discrimination skills, their ratings are removed to generate a new mean and CI for the 10 setting's quality ratings, as shown in Figure 7. There is a slight improvement in discriminating the quality for sidetone on/off settings and latency settings, but the CI on these settings cloud the significance of these trends due to the smaller data set.

Observing the general trend of reliability and discrimination metrics in Figure 6, the critical threshold is brought up further with a discrimination threshold of 5 and reliability threshold of 2 for optimal separability between critical listeners and moderately skilled listeners. All reliable participants are re-distributed into 2 groups, in which the two groups then undergo a new set of mean and CI calculations, resulting in Figure 8.

FB ANC and THD show the same trends, but the critical listeners were able to discriminate latency changes better when THD is minimized. The same cannot be said for latency changes under maximum THD – as THD effects are significantly overpowering latency effects, as shown by a significant F-value from the ANOVA analysis.

Despite being able to clean up the raw data with assessor screenings, a familiarization procedure should

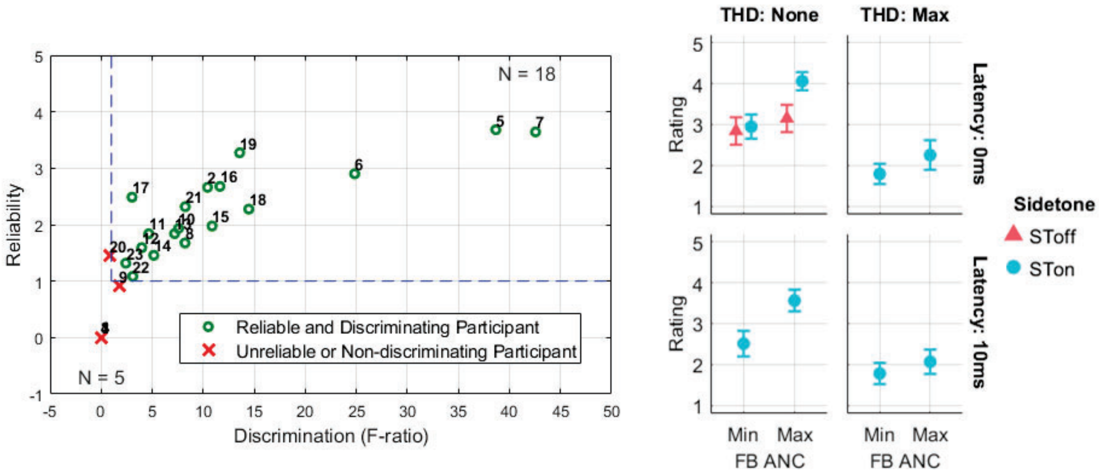

Fig. 7: Categorization of reliable to unreliable participants, and the resulting Mean and CI amongst reliable participants.

have been conducted before the beginning of each test to provide a better rating framework for all participants. Participants should be allowed to experience 2–3 settings that range from good to bad quality, so that their discrimination abilities can be standardized; therefore, having a data set with a clear distinction between the different audio settings. Regardless, as the general trends of how the investigated factors influence self-voice remain similar, the experiment still shows reliable results despite the gaps in listening skills.

5 Limitations and Future Works

Limited options for a sidetone filter may have hindered the representation of the effects of sidetone. The limitation of this study is that the sidetone filter is being optimized to 0dB IG under maximum FB ANC settings. However, if ANC is turned off and no longer aids in reducing low-frequency components of occlusion, it requires a new sidetone filter as using the old filter will result in a non-zero IG. The filter needs to counter both the occlusion issues and assist the high-frequency components in the acoustic path to achieve 0 dB IG. This may help with bigger findings on how sidetone filtering affects the perception of one's self-voice. An investigation on a variety of simple filter shapes, such as low-pass or high-pass filters, might also be a benefactor for further investigation.

Some dismissed factors are also worthy of investigation. Similar to telephony sidetone [5], sidetone volume is part of the research as listening to one's own voice at an altered level can cause an altered perception in one's own voice similar to Lombard Effects – a quieter sidetone volume may force one to talk louder, and a louder sidetone volume might cause one to start whispering, despite instructing a subject to speak at a normal effort [16, 17]. As participants have commented about the level quite a few times (Figure 5), further investigation on this factor should be considered.

A broader range of latency values should be investigated, due to the insignificant effects seen with the latency levels in this study. The disruption of comb filtering effects might give some insight into how users respond to self-voice latency [18]. As this study investigates the in-ear headset form factor, an expanded study on a range of form factors, such as on-ear and over-ear headsets, can be considered to fully encompass the sidetone performances of all types of contemporary headsets.

The procedure of the subjective test can be further improved with the addition of a familiarization stage, allowing all the participants to have a standardized rating range, and providing better quality data sets for analysis.

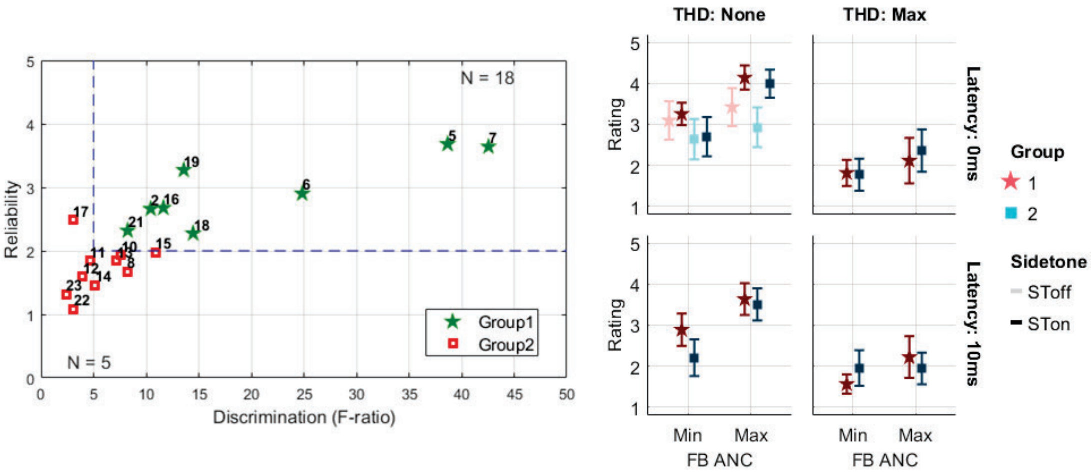

Fig. 8: Further categorization of reliable participants. Group 1 - critical listener; Group 2 - lesser skilled listener.

Lastly, the study can lead to the development of an objective perceptual self-voice quality measure for headset sidetone levels based on the factors investigated, which can aid in predicting the quality of one's self-voice for an unknown headset completed with objective acoustical measurements.

6 Conclusion

The present study investigated how several acoustical factors including sidetone settings, latency levels, feedback ANC levels, and total harmonic distortion influence the perception of one's own voice when wearing in-ear headsets. Instrumental measurements were conducted to define the level settings as well as design the experiment. A subjective listening test was conducted using a 5-scale continuous quality rating to assess the own voice perception. Twenty-three assessors participated in the study. The procedure aims to evaluate the impact of acoustical factors on the perception of one's own voice and understand the extent of their influence on quality.

Referring back to the research questions posed in Section 1, the study suggests that occlusion levels, sidetone, and distortion levels influence the perception of self-voice the most when using an in-ear headset and that the interactions between sidetone and occlusion levels are influential to one's self-voice perception as well. As occlusion is reduced by increasing the strength of FB ANC, how one perceives their self-voice improves; As distortion increases, one's perception of their self-voice deteriorates. When less occlusion is experienced by a person, the existence of sidetone becomes more and more detrimental to the quality of self-voice. Out of the four factors investigated, the effects of latency are the most uncertain.

7 Acknowledgments

This project was conducted during the first author's master thesis project at the Danmarks Tekniske Universitet (DTU), sponsored by GN. The authors would also like to thank the participants at GN for their time, and Professor Cumhur Erkut for his feedback during the presentation of this project.

References

[1] Reinfeldt, S., Östli, P., Håkansson, B., and Stenfelt, S., "Hearing One's Own Voice During Phoneme Vocalization—Transmission by Air and Bone Conduction," *The Journal of the Acoustical Society of America*, 128(2), pp. 751–762, 2010, doi:10.1121/1.3458855.

[2] Pörschmann, C., "Influences of bone conduction and air conduction on the sound of one's own

voice," *Acta Acustica united with Acustica*, 86(6), pp. 1038–1045, 2000.

[3] Hu, S., Rajamani, R., and Yu, X., "Active Noise Control for Selective Cancellation of External Disturbances," in *Proceedings of the 2011 American Control Conference*, pp. 4737–4742, IEEE, 2011, doi:10.1109/ACC.2011.5991142.

[4] Baumfield, A., Hickson, L., and McPherson, B., "Performance of Assistive Listening Devices using Insertion Gain Measures," *Scandinavian Audiology*, 22(1), pp. 43–46, 1993, doi:10.3109/01050399309046017.

[5] Appel, R. and Beerends, J. G., "On The Quality of Hearing One's Own Voice," *Journal of the Audio Engineering Society*, 50(4), pp. 237–248, 2002.

[6] Liebich, S. and Vary, P., "Occlusion Effect Cancellation in Headphones and Hearing Devices—The Sister of Active Noise Cancellation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, pp. 35–48, 2021, doi:10.1109/TASLP.2021.3130966.

[7] Hengen, J., Hammarström, I. L., and Stenfelt, S., "Perception of One's Own Voice After Hearing-Aid Fitting for Naive Hearing-Aid Users and Hearing-Aid Refitting for Experienced Hearing-Aid Users," *Trends in Hearing*, 24, p. 2331216520932467, 2020, doi:10.1177/2331216520932467.

[8] Oehlert, G. W., *A First Course in Design and Analysis of Experiments*, Freeman, 2010.

[9] Rothauser, E., "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, 17(3), pp. 225–246, 1969.

[10] Walpole, R. E., *Introduction to Statistics*, Prentice Hall International, 1997.

[11] ITU-R BS.2300, "Methods for Assessor Screening," Standard, International Telecommunication Union, Geneva, Switzerland, 2014.

[12] Schlich, P., "GRAPES: A method and a SAS® Program for Graphical Representations of Assessor Performances," *Journal of sensory studies*, 9(2), pp. 157–169, 1994, doi:10.1111/j.1745-459X.1994.tb00238.x.

[13] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation: Theory, Method and Application*, John Wiley & Sons, 2007.

[14] Lorho, G., Le Ray, G., and Zacharov, N., "eGauge — A Measure of Assessor Expertise in Audio Quality Evaluations," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, Audio Engineering Society, 2010.

[15] Fela, R. F., Zacharov, N., and Forchhammer, S., "Assessor Selection Process for Perceptual Quality Evaluation of 360 Audiovisual Content," *Journal of the Audio Engineering Society*, 70(10), pp. 824–842, 2022, doi:https://doi.org/10.17743/jaes.2022.0037.

[16] Lane, H. and Tranel, B., "The Lombard Sign and the Role of Hearing in Speech," *Journal of speech and hearing research*, 14(4), pp. 677–709, 1971, doi:10.1044/JSHR.1404.677.

[17] Chong, H. J., Choi, J. H., and Lee, S. S., "Does the Perception of Own Voice Affect Our Behavior?" *Journal of Voice*, 2022, ISSN 0892-1997, doi:https://doi.org/10.1016/j.jvoice.2022.02.003.

[18] Jekosch, U. and Möller, S., "Speech Quality and the E-model," *The Journal of the Acoustical Society of America*, 105(2), pp. 974–974, 1999, ISSN 0001-4966, doi:10.1121/1.425325.