Evaluation of a frequency lowering hearing instrument algorithm using a non-inferiority test

Christophe Lesimple, Neil Hockley[*], and Barbara Simon

Bernafon AG, Morgenstrasse 131, 3018 Bern, Switzerland

The primary goal of hearing instrument verification is to demonstrate an improvement on a relevant outcome. It is imprudent to implement an algorithm that improves one outcome while simultaneously degrading another. A traditional test typically uses a superiority hypothesis – H0: New = Conventional and H1: New ≠ Conventional. The absence of statistical significance may be interpreted incorrectly as an absence of clinically relevant differences. An alternative is to start the test with a non-inferiority hypothesis – H0: New < Conventional and H1: New ≥ Conventional. Cross-over designs are often employed because treatment differences are frequently measured within a subject rather than between subjects. Each test period should be long enough for the subject to become acclimatized to each processing change. With these conditions, it is possible to estimate, with the same test, the overall effect of the developed feature and also the period effect. The method of using a cross-over design with a non-inferiority analysis was applied in the testing of a new frequency lowering algorithm. Improved high-frequency functional gain and fricative discrimination was observed. Significant non-inferior SSQ scores between the processing on and off were seen while no period effect was found. These results provide a good approximation of 'real world' acceptance.

INTRODUCTION

Frequency Lowering (FL) algorithms are designed for hearing-impaired people who cannot otherwise obtain benefit from conventional processing (CP) in the high frequencies (HF). The aim of FL processing is to provide improved access to HF cues that would otherwise not be available. Most of the published studies about FL systems are centred on speech recognition and discrimination improvement; however, some of these papers also report the effect of FL systems on sound quality (Simpson *et al.*, 2006; Kuk *et al.*, 2009; Bohnert *et al.*, 2010; Parsa *et al.*, 2013). FL algorithms add artificial signals that may change harmonic ratios, add noise, change timbre, etc., so perceived sound quality may be affected depending on the listener.

Sound quality can be assessed with questionnaires or with perceptual tests. Recent studies, that used questionnaires with different FL algorithms, were unable to show a significant group effect, such as Simpson *et al.* (2006) with the Abbreviated Profile of Hearing Aid Benefit (APHAB) (Cox and Alexander, 1995), Ellis (2012) with the Speech, Spatial and Qualities of Hearing Scale (SSQ) (Gatehouse and Noble, 2004), or Bohnert *et al.* (2010) with self-developed questionnaires. A perceptual test, like the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) design (ITU-R, 2003) used by Parsa *et al.* (2013), investigated subjective ratings with different FL settings and various test stimuli. Test participants rated sound quality

*Corresponding author: nh@bernafon.ch

for speech in noise and music samples with different FL settings. The rating difference was not significant between the CP and FL processing for hearing-impaired adults. Conclusions from these studies might be misinterpreted when the authors report no statistically significant difference between the tested conditions for the following reasons:

1. There is a chance that both processing strategies provide more or less the same perceived sound quality.

2. The sample size might be too small to show an existing difference.

3. The outcomes might not be sensitive to the tested FL algorithm. Sensitivity can be affected by floor and/or ceiling effects, or questions that are not relevant to what is being tested.

These studies are superiority trials that are designed to detect differences between treatments (CPMP, 2000). However, a superiority trial cannot be used to conclude that two treatments, or two processing methods in this case, have the same effect. In order to evaluate if two treatments have the same effect, the Committee for Proprietary Medicinal Products (CPMP) (2000) recommends the use of a non-inferiority hypothesis. To the best of the authors' knowledge, this technique has not been used to evaluate hearing instrument (HI) features to date.

Non-inferiority trial in a cross-over design

This non-inferiority trial seeks to find that a new FL algorithm does not perform worse than the reference CP by more than an acceptable amount, i.e., the non-inferiority margin (M_{NI}). Pocock (2003) and D'Agostino *et al.* (2003) present some key issues that need to be addressed when using a non-inferiority design. Therefore, CP should demonstrate superiority over the unaided condition (a) for the same participant population, (b) with an equivalent HI and (c) for the same outcome measure. These authors also state that each processing strategy needs to be tested for a long enough period of time in order for any processing differences to have a realistic opportunity of being observed.

To declare that non-inferiority has been shown, the 95% confidence interval of the difference between both processing systems (FL and CP) should entirely lie above the non-inferiority margin as seen in Fig. 1. Guidelines from the CPMP (2000) also recommend calculating a *p*-value associated with the null hypothesis of inferiority in order to assess the strength of the evidence in favour of non-inferiority.

To reduce the impact of confounding variables and biases (Cox, 2005), a two period cross-over design can be used to assess the processing effect, period effect, and any interaction (Hills and Armitage, 1979). One requirement is that the baseline condition does not change over the two test periods. Thus, it is expected that the test subjects have the same condition at the beginning of each test period. In this design, all participants receive and provide data for both the test FL algorithm and the reference CP. A cross-over design can also be used to test a non-inferiority hypothesis.
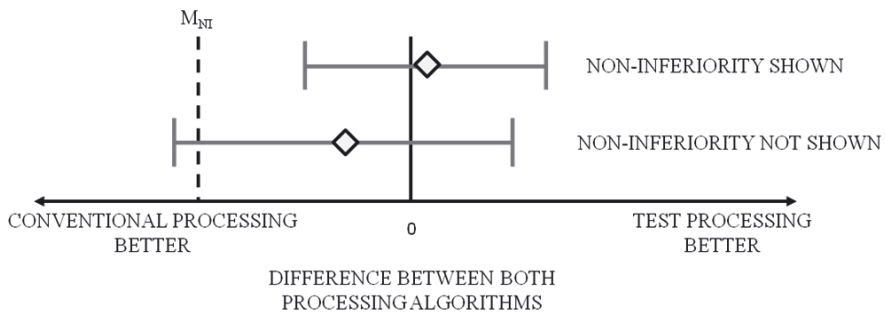
Fig. 1: Schematic presentation of the hypotheses related to a non-inferiority trial. Error bars represent the confidence intervals. In both cases, a superiority hypothesis test would fail to reject the null hypothesis. With a non-inferiority trial in the upper case, it is possible to conclude that the new processing is significantly non-inferior to the reference processing.

Non-inferiority margin determination

The M_{NI} states how close the FL processing must be to the conventional processing. Evidence from previous experiments must show that the CP condition is superior to the unaided one (assay sensitivity) and the test conditions must also be similar for the new trial (constancy assumption). The new processing should also remain superior over the unaided condition by a certain amount (putative comparison). It is reasonable to fix this amount at 50 % of the conventional processing effect over the unaided condition (Jones *et al.*, 1996).

The Speech, Spatial, and Qualities of Hearing Scale (SSQ) questionnaire

The SSQ is a self-report questionnaire divided into three subscales that assess various everyday listening situations. The subject rates their ability to perform in each given listening situation on a scale from 0 to 10 (higher scores always reflect greater ability or less effort). The result is that 'real world' environmental scenarios, with the implemented processing in the HI, can be evaluated.

For this investigation, the qualities subscale of the SSQ was used to investigate various aspects of the perceived sound quality, including: (a) sound quality and naturalness, (b) identification of sound, (c) segregation of sounds, and (d) listening effort. The qualities subscale has shown, in various studies, that CP provides benefit over the unaided condition for adults with mild to severe hearing loss (Noble and Gatehouse, 2006; Jensen *et al.*, 2009; Köbler *et al.*, 2010). The SSQ is available in many languages including German.

Research Question

In this investigation it is hypothesised that the FL algorithm will be judged to provide good sound quality for the hearing-impaired subjects. An improvement over conventional processing is not expected and differences between the FL algorithm and CP should not be clinically relevant. This investigation is designed to show how a non-inferiority test, using the qualities subscale from the SSQ questionnaire, can evaluate the differences in perceived sound quality between the CP and FL processing.

MATERIAL AND METHOD

A cross-over trial with a three-week period was judged to be sufficient for the subjects to become acclimatized to each processing scheme. This time period should be long enough for each subject to experience most of the situations or environments described in the SSQ questionnaire. Two groups were created with two different experimental sequences. The Sequence A group received the FL algorithm in the first period, whereas the Sequence B group received the CP. After three weeks, the processing type was switched so that the Sequence A group received CP and the Sequence B group received the FL algorithm. Group allocation was done using minimization of the following predictive factors: (a) high-frequency hearing loss and (b) participant amplification experience.

Participants

Fourteen subjects between 41 and 79 years of age (average = 64) with a bilateral sensorineural hearing loss took part in this trial. Twelve of them were experienced HI users who had previously used the same compression scheme. The other two subjects were first time users. There were thirteen males and one female. The high-frequency hearing loss was defined by the high-frequency average (HFA) of the air conducted thresholds at 4, 6, and 8 kHz. Ten of the fourteen participants had a HFA that was greater than 70 dB HL. Figure 2 shows the average hearing thresholds for all participants.

Test hearing instrument

The FL algorithm was employed in a commercially available receiver-in-the-ear (RITE) HI. The appropriate acoustic coupling was selected for each subject as recommended by the fitting software. The same instrument was used for both experimental periods. Only the FL algorithm was enabled or disabled during the trial. The gain, compression factors, and automatic features were identical over both periods.

To minimize the placebo effect that is commonly found in HI evaluations (Dawes *et al.*, 2013), the participants were unaware of which algorithm they were trying at any instance in time and of the specific intent of the tested feature.
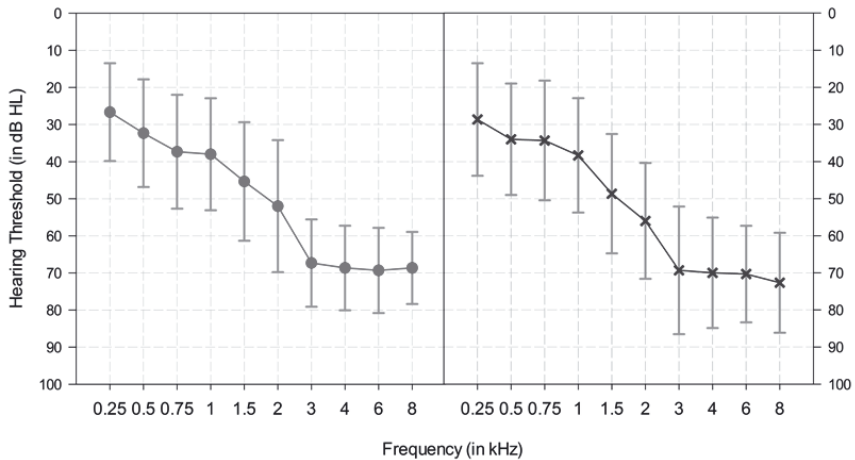
Fig. 2: Average hearing-threshold levels for the fourteen trial participants at each air conducted audiometric frequency. Error bars show the standard deviation.

Non-inferiority determination

The non-inferiority margin determination was based on previous internal pilot investigations. Figure 3 shows what was considered in the determination of the M_{NI}.
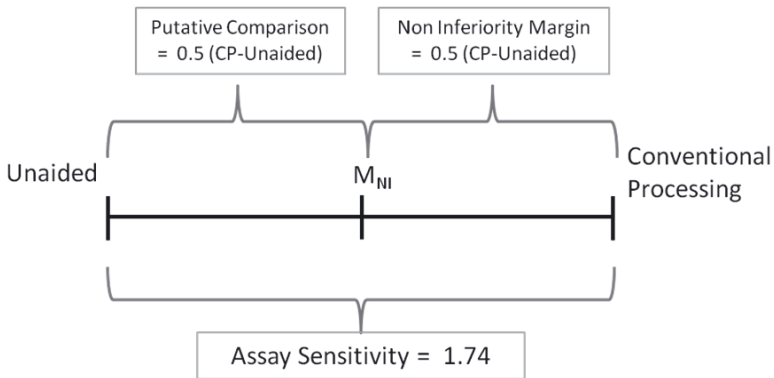

Fig. 3: Considerations in the determination of the M_{NI}. (a) The assay sensitivity will be based on an already known CP effect over the unaided condition. (b) The putative comparison will control the effectiveness of the tested processing.

Based on the observation of a 1.74 score improvement on the SSQ qualities subscale from a previous pilot study, it was appropriate to set the M_{NI} to 50% of the historical effect (improvement from unaided to CP). Due to the fact that the same RITE HIs with the same compression scheme and fitting rationale were used in both trials, it is assumed that the assay sensitivity is obtained and that the constancy assumption is held. The M_{NI} for this trial can be set to a 0.87 SSQ score degradation.

RESULTS

Each participant filled out an SSQ questionnaire after each three-week test period. Mean scores are shown for both processing types and for each qualities subscale attributes in Fig. 4.
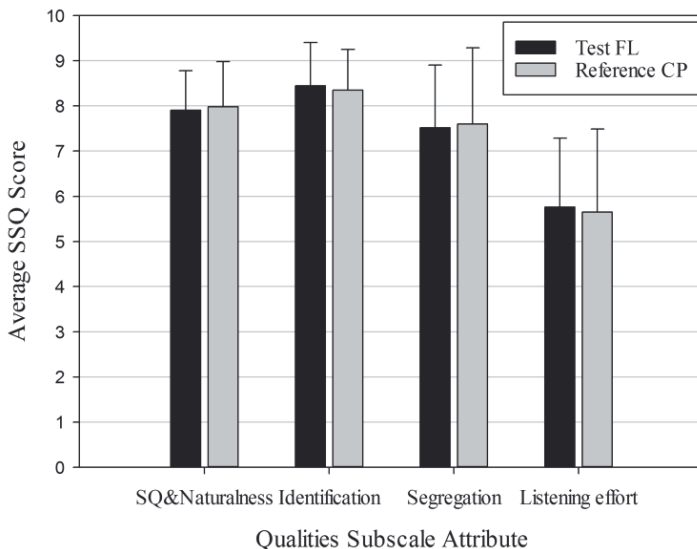

Fig. 4: Average qualities scores from the SSQ within sound quality and naturalness, identification of sound, segregation of sounds, and listening effort (n=14). Results with the HIs with the FL algorithm are in black and the results with CP are in grey. Error bars show one standard deviation.

Based on the rating difference between both processing schemes, it is possible to compute the mean difference and the 95% confidence interval for this outcome. To conclude that FL processing is significantly non-inferior to the CP, the lower boundary of the 95% confidence interval should be higher than the M_{NI}. Under the null hypothesis, the data are distributed with the M_{NI} as means and will follow a *t*-distribution with N-2 degrees of freedom. The *p*-value is derived from these data and all the findings are shown in Table 1.

Qualities Attributes	Non-Inferiority Test Null hypothesis (H0): Difference $\leq$ - 0.87				Reject H0?
	Difference (T - R)	95 % CI of the difference		p-value	Non-Inferior?
		Lower Bound	Upper Bound		
SQ & Naturalness					
FL-CP	-0.08	-0.43	0.27	0.001	Yes
Identification of Sound					
FL-CP	0.08	-0.31	0.47	0.007	Yes
Segregation of Sounds					
FL-CP	-0.08	-0.43	0.26	0.001	Yes
Listening Effort					
FL-CP	0.11	-0.91	1.13	0.096	No

Table 1: Non-inferiority test results for the different attributes from the qualities subscale of the SSQ. The difference between FL processing and CP imply that positive values are in favour for the tested algorithm.

The qualities subscale outcomes show that the developed FL is significantly non-inferior to the CP for the following attributes: sound quality and naturalness, identification of sound, and segregation of sounds. For the listening effort attribute, the null hypothesis could not be rejected as the confidence-interval lower boundary is smaller than the M_{NI}.

CONCLUSION

Showing client benefit for a newly developed algorithm is the desired outcome for the verification of many HI features. However, when the signal is manipulated, it seems also important to assess how the perceived sound quality might be affected. The use of a non-inferiority hypothesis is probably the only way to show that the sound quality is not significantly degraded with a new algorithm. For the FL algorithm evaluated in this investigation, it was possible to address this concern by using the qualities subscale from the SSQ questionnaire with a non-inferiority test. Three out of the four attributes from the qualities subscale were significantly non-inferior. Based on these outcomes it is expected that this FL algorithm will provide a perceived sound quality that is comparable to the CP and that this will result in good acceptance by the HI wearer.

REFERENCES

Bohnert, A., Nyffeler, M., and Keilmann, A. (**2010**). "Advantages of a non-linear frequency compression algorithm in noise," Eur. Arch. Otorhinolaryngol., **267**, 1045-1053.

Cox, R.M. (**2005**). "Evidence-based practice in provision of amplification," J Am Acad Audiol. 2005, **16**, 419-38.

Cox, R.M., and Alexander, G.C. (**1995**). "The abbreviated profile of hearing aid benefit," Ear. Hearing, **16**, 176-186.

CPMP: Committee for Proprietary Medicinal Products (**2000**). "Points to consider on switching between superiority and non-inferiority," European Medicines Agency (EMEA), CPMP/EWP/482/99.

D'Agostino, R.B. Sr., Massaro, J.M., and Sullivan, L.M. (**2003**). "Non-inferiority trials: design concepts and issues – the encounters of academic consultants in statistics," Stat. Med., **22**, 169-186.

Dawes, P., Hopkins, R., and Munro, K.J. (**2013**). "Placebo effects in hearing-aid trials are reliable," Int. J. Audiol., **52**, 472-477.

Ellis, R.J. (**2012**) "Benefit and predictors of outcome from frequency compression hearing aid use," PhD thesis, University of Manchester, UK.

Gatehouse, S., and Noble, W. (**2004**). "The speech, spatial and qualities of hearing scale (SSQ)," Int. J. Audiol., **43**, 85-99.

Hills, M., and Armitage, P. (**1979**). "The two-period cross-over clinical trial," Br. J. Clin. Pharmacol., **8**, 7-20.

ITU-R (**2003**). *Recommendation BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems.* International Telecommunications Union, Geneva, Switzerland.

Jensen, N.S., Akeroyd, M.A., Noble, W., and Naylor, G. (**2009**). "The Speech, Spatial and Qualities of Hearing scale (SSQ) as a benefit measure," NCRAR conference on *The Ear-Brain System: Approaches to the Study and Treatment of Hearing Loss*, Portland, October 2009 (poster).

Jones, B., Jarvis, P., Lewis, J.A., and Ebbutt, A.F. (**1996**). "Trials to assess equivalence: the importance of rigorous methods," Brit. Med. J., **313**, 36-39.

Köbler, S., Lindblad, A.C., Olofsson, A., and Hagerman, B. (**2010**). "Successful and unsuccessful users of bilateral amplification: differences and similarities in binaural performance," Int. J. Audiol., **49**, 613-627.

Kuk, F., Keenan, D., Korhonen, P., and Lau, CC. (**2009**). "Efficacy of linear frequency transposition on consonant identification in quiet and in noise," J. Am. Acad. Audiol., **20**, 465-479.

Noble, W., and Gatehouse, S. (**2006**). "Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing Scale (SSQ)," Int. J. Audiol., **45**, 172-181.

Parsa, V., Scollie, S., Glista, D., and Seelisch, A. (**2013**). "Nonlinear frequency compression: effects on sound quality ratings of speech and music," Trends Amplif., **17**, 54-68.

Pocock, S.J. (**2003**). "The pros and cons of noninferiority trials," Fundam. Clin. Pharmacol., **17**, 483-490.

Simpson, A., Hersbach, A.A., and McDermott, H.J. (**2006**). "Frequency-compression outcomes in listeners with steeply sloping audiograms," Int. J. Audiol., **45**, 619-629.