OPEN ACCESS

Short Report

Investigation of test-retest reliability of a new close-to-everyday-life speech test using the Poisson binomial distribution

Untersuchung der Test-Retest-Reliabilität eines neuen alltagsnahen Sprachtests mit der verallgemeinerten Binomialverteilung

Abstract

To enhance the ecological validity of speech intelligibility testing, this study introduces a novel speech test based on a conversation created with synthetic voices embedded in a 3D virtual cafeteria scene. 20 keywords from the conversation at constant signal-to-noise ratios (SNR) were used to assess intelligibility. The test was evaluated with 20 normal-hearing (NH) listeners. Test-retest accuracy was assessed and modeled based on simple and Poisson binomial statistics. Speech reception thresholds (SRT) were –9.7 dB ±2.5 dB SNR (average and standard deviation). After one familiarization run, no further significant training effect was observed. Due to variability in keyword intelligibility at fixed SNR, the "effective" keyword number was found to be n'=32, improving the test-retest reliability of the test compared to a test with 20 keywords of equal intelligibility. Confidence intervals were modelled and confirmed using Bootstrapping.

Zusammenfassung

Zur Erhöhung der ökologischen Validität von Sprachverständlichkeitsmessungen schlägt diese Studie einen neuen Test basierend auf einer Unterhaltung von zwei synthethischen Stimmen vor, welcher eingebettet ist in eine 3-dimensionale virtuelle Cafeteria-Szene. 20 Schlüsselwörter aus der Unterhaltung, dargeboten bei konstantem Signal-Rausch-Verhältnis (SNR), wurden verwendet, um die Sprachverständlichkeit zu erfassen. Der Test wurde evaluiert mit 20 normalhörenden Personen. Die Test-Retest-Genauigkeit wurde ausgewertet und modelliert mit Hilfe von einfacher sowie verallgemeinerter Binomialverteilung. Sprachverständlichkeitsschwellen wurden ermittelt zu –9.7 dB ±2.5 dB SNR (Mittelwert und Standardabweichung). Nach einem Eingewöhnungsdurchgang wurde kein weiterer Trainingseffekt beobachtet. Aufgrund der Variabilität des Schlüsselwortverstehens bei konstantem SNR betrug die „effektive" Schlüsselwortanzahl n'=32 und hat somit die Test-Retest Genauigkeit dieses Tests erhöht im Vergleich zu einem Test mit 20 Schlüsselwörtern bei identischer Verständlichkeit aller Schlüsselwörter. Konfidenzintervalle wurden modelliert und bestätigt mittels Bootstrapping.

Tim Jürgens[1]
Jonas Stohlmann[1]
Fabian Hettler[1]
Jürgen Tchorz[1]
Markus Kallinger[1]
Florian Denk[2]
Hendrik Husstedt[2]

1 Institut für Akustik, Technische Hochschule Lübeck, Lübeck, Germany

2 Deutsches Hörgeräte Institut GmbH, Lübeck, Germany

Introduction

Established speech tests for diagnostics or hearing device evaluation in the laboratory usually do not reflect the complexity of everyday sound environments. Representative room acoustics with several sound sources from different directions, reverberation, and speech and noise that both plausibly fit into the acoustic scene are usually not considered in these tests. However, everyday-life acoustic environments are important for hearing device development to optimize their benefit in situations that their users encounter.

Virtual sound environments (VSE) realized with multichannel loudspeaker systems can simulate such everyday acoustic scenes and offer high reproducibility that is necessary to contrast differences between processing algorithms. However, speech tests used within these VSEs have been mostly either classical matrix tests [1] or

simple standalone sentences [2] which both do not fit naturally into the respective scene and lack the context usually present in everyday conversations.
The first goal of the present study is to introduce a novel speech test that more accurately reflects real-life speech intelligibility tasks than classical speech tests by both, simulating a 3D real-life sound environment and by using speech material that plausibly fits into the scene. The second goal is to evaluate speech intelligibility and model the novel test's reliability in normal-hearing (NH) listeners when measuring at constant signal-to-noise ratios (SNR).

Experimental methods

Participants

Twenty young German native NH listeners (8 female, 12 male, 20–30 years old) participated in the present study. All listeners had pure tone thresholds of ≤20 dB HL at all standard audiometric frequencies between 125 Hz and 8 kHz. Participants provided informed consent before participation. Ethical approval was granted by the Ethics Committee of the University of Applied Sciences Lübeck (approval from 15/05/2023).

Apparatus

A spherical 65-channel coaxial loudspeaker array mounted in an anechoic chamber was used for sound presentation. Participants were seated with their head positioned in the center of the sphere on a height-adjustable chair and used a tablet displaying a graphical user interface (GUI) for giving responses.

Virtual sound environment

A virtual 3D model of the cafeteria of the University campus Lübeck was created using SketchUp 2016, see Figure 1, as two shoe-box-like adjoint rooms, with (length x width x height) 30 m x 20 m x 6 m and 15.7 m x 11 m x 6 m. Acoustic properties of walls, floor, ceiling, and furniture were defined via the RAVEN toolbox [3]. The scene was rendered using 7th order Ambisonics with maxRE decoder. Reverberation times were between 1.3 s for 250 Hz and 0.8 s for 4 kHz. As background noise, the "food court" recording from the ARTE database [4], available in 4th order Ambisonics, was used.

Novel speech test

Google text-to-speech Chirp 3 HD was used to create a dialogue with a male voice ("Fenrir") and a female voice ("Leda"). The dialogue focused on holiday experiences and future travel plans. 20 keywords (structured as pairs) from the sentences were used for testing speech intelligibility and the other words served as carrier structure for passive listening. The dialogue stopped after presentation of a keyword pair (with ongoing background noise) and 10 response alternatives per keyword were displayed on the GUI from which participants needed to select one each. An example sentence is "Vermutlich werden wir in Japan Schlösser besuchen." ("Probably we will visit castles in Japan.") with "Japan" and "Schlösser" being keywords. Participants were instructed to select the most appropriate response and to provide a guess when uncertain (closed-set paradigm). Phoneme distribution of the entire speech material was tested for being representative for German language according to [5]. Selection of presented keywords was fully randomized (i.e., without specific lists). More details about the speech material can be found in [6]; Attachment 1 shows the full dialogue and all response alternatives used. Conduction of the test was done using customized MATLAB scripts.
The scene was first presented to participants at –3 dB SNR for familiarization with the test routine. Afterwards, presentation was done at –8 dB (test) and at –11 dB (test and retest) in randomized order.

Psychometric functions and statistics

Psychometric functions according to [7] were constructed for each participant and for the pooled data of all participants. Paired t-tests were used to assess potential test-retest differences.

Modeling of test reliability

Test reliability to estimate ground-truth intelligibility

95%-confidence intervals were constructed (a) on the assumption of equal intelligibility across all keywords for a given SNR and (b) with considering differences in intelligibility across keywords. For (a), the standard deviation σ_j of intelligibilities around the ground-truth intelligibility p_j in % is given according to the simple binomial distribution as

$$\sigma_j = \sqrt{\frac{p_j(100\% - p_j)}{n}} \qquad (1)$$

with n=20 keywords and j indexing different SNRs (here: j=1 for SNR=–11 dB and j=2 for SNR=–8 dB). 95%-confidence intervals were then constructed from σ_j according to [8] with boundaries of $p_j \pm 1.96 \cdot \sigma_j$ with considering caps at 0% and 100% intelligibility.
For (b) word-specific intelligibilities p_{ji} were evaluated from those keywords that were at least three times presented to the listeners. According to [8] a measure for the diversity of intelligibilities is given as:

$$s_{p_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(p_{ji} - p_j\right)^2} \qquad (2)$$

The Poisson binomial distribution [8] allows now to indicate an "effective" keyword number n_j', i.e., the number of keywords that a fictitious test consisting of equal intel-

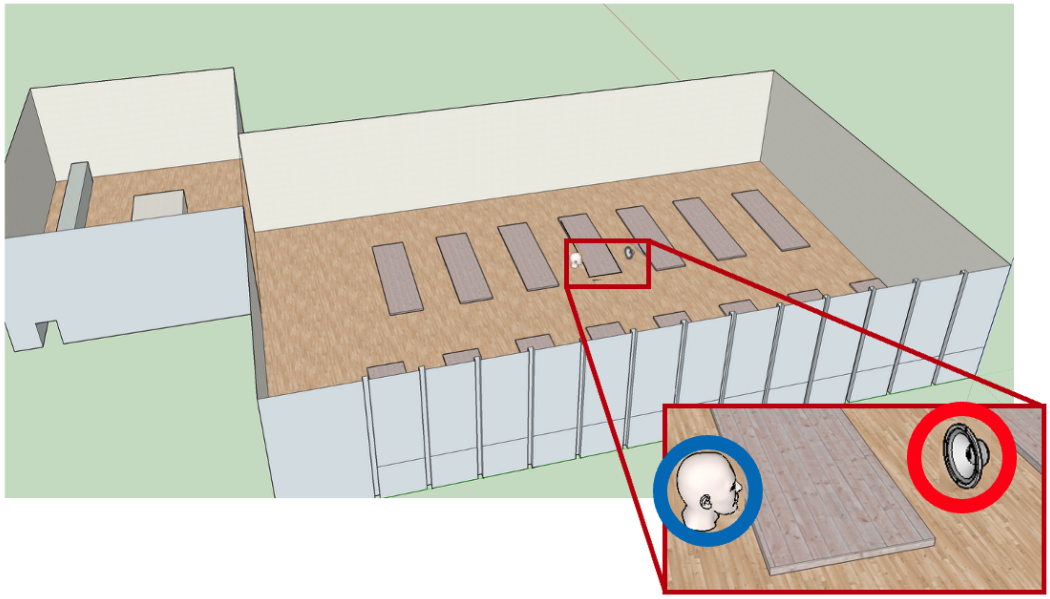

Figure 1: 3D model of the cafeteria created for the present study. The target speech material (red circle) was presented frontally (0°) to the listener (blue circle)

ligibilities across words at a given SNR would have, in order to produce the same 95% confidence intervals as the here-presented test with diverse intelligibilities across words:

$$n'_j = \frac{n^2 p_j(100\% - p_j)}{\sum_{i=1}^{n} p_{ji}(100\% - p_{ji})} = n \cdot \left(1 - \frac{s_{p_j}^2}{p_j(100\% - p_j)}\right)^{-1} \qquad (3)$$

Typically, n' exceeds n; in other words, a test comprising items with varying intelligibility yields narrower confidence intervals and therefore greater measurement precision for speech intelligibility scores. Note that this usually goes alongside a shallower slope of the psychometric function, which renders this test less accurate for adaptive speech reception threshold (SRT) measurements [9]. Using the "effective" keyword number n_j', 95% confidence intervals were constructed assuming binomial distribution according to Eq. (1). Intervals were tested for accuracy with the available data both, with and without compensating for subject-specific intelligibility differences that were extracted from individual psychometric functions.

Test-retest reliability

For hearing device evaluation usually the test reliability to estimate the ground-truth intelligibility is not as important as the test-retest reliability for contrasting two different conditions (for example with and without hearing aids or contrasting two different programs). The present study followed the proposal of [10] to use the method of [11] to estimate 95% confidence intervals for test-retest accuracy. For this, an intelligibility result of a test run is considered as the number of correct responses X. X may be transformed into an intelligibility p by dividing by the number of presented words n. For exploiting Gaussian statistics X is arcsine-transformed to give the random variable:

$$\theta(X, n) = \sin^{-1}\sqrt{\frac{X}{n+1}} + \sin^{-1}\sqrt{\frac{X+1}{n+1}} \qquad (4)$$

This random variable has the advantage of a variance independent of p. Therefore, also the variance of the difference of two values θ_1 and θ_2 given by

$$\sigma^2 = \frac{2}{n+1} \qquad (5)$$

is independent of p. Again, assuming confidence intervals with width $\theta \pm 1.96\sigma$ and backtransformation yields test-retest confidence intervals for the here-proposed novel speech test. These were calculated both, for simple binomial (n=20) and for Poisson binomial statistics with n'. As a sanity check, bootstrapping was done with testing 100 word recognition trials each at fixed SNRs ranging from –22 dB to 5 dB in 0.25 dB steps. Hereby, the (empirically found) average psychometric function of participants was assumed as target function and half of the trials was treated as test and the other half as retest. Simulated data was used to have a much higher number of speech recognition scores (109x100=10,900 in contrast to only 20x3=60 measured scores) and to cover the full range of possible speech intelligibility scores.

Results

Measurements

Figure 2 shows psychometric functions for both individual (grey lines) and group-averaged (black line) data. The average psychometric function was determined by first averaging intelligibility scores across all participants for each SNR and then fitting a logistic function according to [7]. The average psychometric function's SRT was –9.7 dB SNR with an inter-individual standard deviation

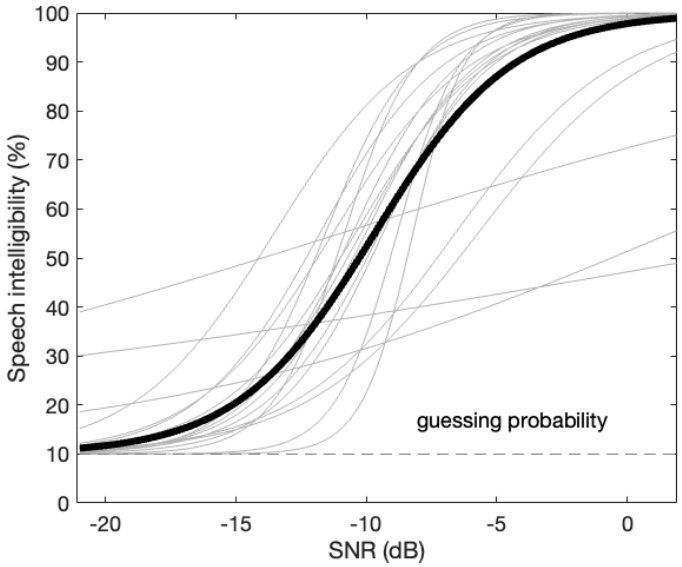

Figure 2: Psychometric functions constructed for each individual participant (thin grey lines) and for the pooled group data of all NH participants (thick black line) for the novel speech test. Individual psychometric functions show similar shapes as the group psychometric function except for 3 outlier curves with much shallower slopes without turning point within the displayed SNR-range.

of ±2.5 dB. The slope of the average psychometric function was 9.5 %-points/dB.

Figure 3 shows individual speech intelligibility scores of test and retest in a Bland-Altman plot, i.e., plotting the difference between test and retest scores against the average between test and retest. A bias of −6 % was observed, meaning that the retest resulted in slightly higher intelligibilities than test. However, a t-test showed that this difference was not statistically significant (p=0.14). Experimentally assessed 95% confidence intervals ranged for this SNR to about ±35% (see Figure 3).

Modeling

For −11 dB SNR, the "effective" keyword number was calculated to be n_1'= 31.5 and for −8 dB SNR to be n_2'=34.1. In the following, a (conservative) average of n'=32 was used. Figure 4 shows 95% confidence intervals for estimation of the ground truth intelligibility according to simple binomial statistics with n=20 (grey lines) and according to Poisson binomial statistics with n'=32 (blue lines). In addition, individual speech intelligibility scores are plotted both as raw data (black circles) and with correction using inter-individual SRT-differences (red crosses), whereas the respective abscissa value was chosen to be the average value across all participants for each condition (−8 dB, −11 dB test and retest). With correction by inter-individual differences 95.0% of measured data was found to be within confidence limits with Poisson statistics, whereas without compensation much less data was within confidence limits.

Figure 5 shows test-retest confidence limits calculated according to [11] with assuming simple (grey lines) and Poisson (blue lines) binomial statistics. Blue circles show simulated test-retest data considering the average psychometric function from Figure 2. 95.3 % of data points were found within confidence limits according to Poisson binomial distribution.

Discussion

This study presents a novel speech test based on keywords extracted from an everyday conversation and embedded in a realistic three-dimensional virtual cafeteria environment. Developing more realistic speech tests that are better representative for everyday life of listeners is a vibrant research field (e.g., [12]). The present study should therefore be regarded as a proof-of-concept demonstrating the feasibility and measurement accuracy of this testing approach.

Using synthesized speech material for the design of speech tests has also been proposed before (e.g., [13], [14]). Ibelings et al. [13] hereby compared naturalness ratings of Acapela Cloud services and Google TTS for their speech test and found better ratings of the Acapela Cloud female voice, but better ratings of the Google TTS male voice. We found the naturalness of the produced speech material of Google text-to-speech higher both for female and male voices (see [6] for details), possibly due to an update of their speech production models since 2022. Further improvements in quality of speech production can be expected, as the development of these algorithms is ongoing fast.

The present study showed that testing NH listeners with this novel speech material in a 3D scene leads to reasonable psychometric functions that are, on average, slightly lower in SRT than corresponding SRTs of classical speech tests, such as the Oldenburg sentence test (OLSA, [9], SRT=−7.1 dB SNR) with much shallower slope (9.5 %-points/dB compared to 17.1%-points/dB for the OLSA). A shallower slope is advantageous for testing at constant

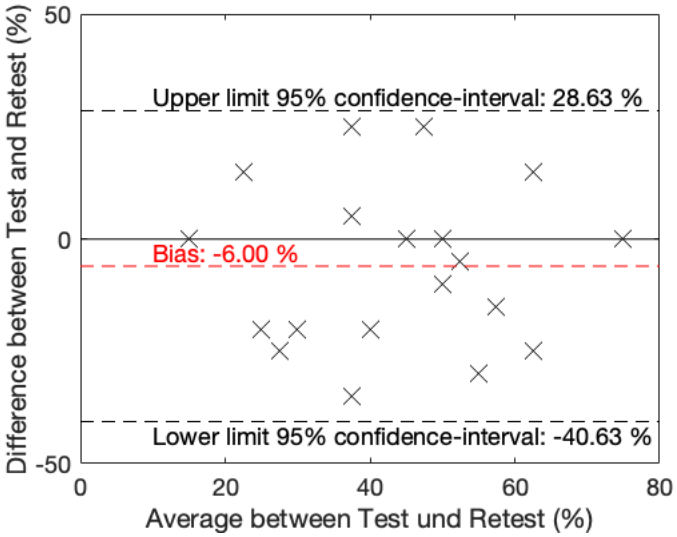

Figure 3: Bland-Altman plot evaluating test-retest accuracy of the novel speech test at –11 dB SNR

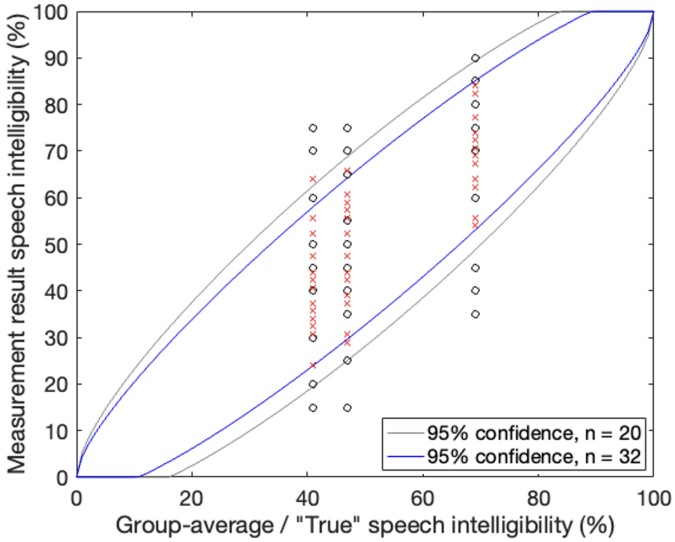

Figure 4: 95%-confidence limits for estimating the ground truth speech intelligibility with one run (20 keywords) of the novel speech test once with simple (n=20) and with Poisson binomial statistics (n'=32). Black circles show raw measured speech intelligibility scores; red crosses show speech intelligibility scores that were corrected for individual performance differences across NH listeners.

SNR, as proposed here, whereas a steeper slope is particularly advantageous for adaptive speech-in-noise testing.

The here-proposed speech test shows no significant training effect in NH listeners after one familiarization list at higher SNRs. Possible training effects in HI listeners, however, still need to be investigated. The differences in word intelligibility at a given SNR leads to a higher "efficiency" of the test in comparison with a fictitious test with equal intelligibility for all presented words. This concept is well known from the Freiburg monosyllable test [8] that consists of n'=29 effective words when using a list of 20 words. With n'=32 effective keywords, the present test has a similar or slightly higher number of effective keywords than the Freiburg monosyllable test. Note that not all keyword alternatives were used in the present study, because not all keywords were presented at least 3 times across all listeners. This restriction was necessary to be able to more reliably estimate word-specific intelligibilities needed for the calculation. A larger database similar to that assessed by [8] for the Freiburg monosyllables would be desirable in future studies.

The correction of speech intelligibility scores by individual variability in SRTs results in measured data being within modeled confidence limits, as expected from Poisson binomial statistics, whereas without correction the spread in data is larger. This indicates that the relatively high spread in the raw data is mainly caused by inter-individual variability and not due to inherent test design flaws. Confidence limits of test-retest reliability were confirmed by Bootstrapping and indicate that for 50% test speech intelligibility a retest needs to be beyond 50% ±25% for a statistically significant difference on an individual basis. At the outskirts of the distribution confidence intervals are smaller. Thus, test-retest accuracy is also similar to the Freiburg monosyllable test.

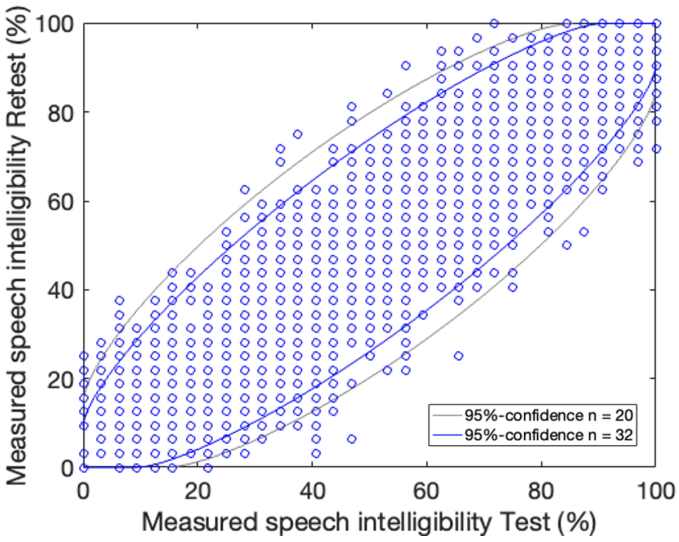

Figure 5: 95%-confidence limits for test-retest accuracy of the novel speech test modelled with simple (n=20, grey lines) and with Poisson (n=32, blue lines) binomial statistics.

Conclusions

A novel speech test with synthetic speech was introduced with the goal of being more representative for everyday life. The test showed no training effect after a familiarization run with NH listeners. Confidence intervals for estimating ground-truth intelligibility and test-retest accuracy were in line with the theory using Poisson binomial statistics with an effective keyword number of n'=32, when using only 20 keywords for testing.

Notes

Conference presentation

This contribution was presented at the 28th Annual Conference of the German Society of Audiology and published as an abstract [15].

Funding

This study was funded by the German Ministry of Research, Technology and Space (grant number: 13FH511KX1).

Use of AI

English language corrections were assisted by ChatGPT (OpenAI, Inc., GPT-5.5, used June 2026). The authors take full responsibility for the content.

Competing interests

The authors declare that they have no competing interests.

Attachments

Available from https://doi.org/10.3205/zaud000092

1. zaud000092_Attachment1.pdf (232 KB)
   Speech material (Stohlmann J [6])

References

1. Jansen T, Hartog L, Oetting D, Hohmann V, Kayser H. Benefit of Hearing-Aid Amplification and Signal Enhancement for Speech Reception in Complex Listening Situations. Trends Hear. 2024;28:23312165241271407. DOI: 10.1177/23312165241271407
2. Oreinos C, Buchholz JM. Evaluation of Loudspeaker-Based Virtual Sound Environments for Testing Directional Hearing Aids. J Am Acad Audiol. 2016;27(7):541-56. DOI: 10.3766/jaaa.15094
3. Schröder D, Vorländer M. RAVEN: A real-time framework for the auralization of interactive virtual environments. In: European Acoustics Association, editor. Proceedings of the Forum Acusticum, Aalborg, Denmark. Aalborg, DK: European Acoustics Association; 2011. p. 1541-6.
4. Weisser A, Buchholz J, Oreinos C, Badajoz Davilla J, Galloway J, Beechey T, Keidser G. The Ambisonic Recordings of Typical Environments (ARTE) Database. Acta Acust utd Acust. 2019;105(4):695-713. DOI: 10.3813/AAA.919349
5. Kohler KJ. Einführung in die Phonetik des Deutschen. Berlin: E. Schmidt Verlag; 1995.
6. Stohlmann J. Development and Modeling of a Speech Intelligibility Test in a Simulated Everyday Environment Using Virtual Acoustics and Synthetic Speech [BSc thesis]. Technische Hochschule Lübeck; 2025.
7. Jürgens T, Brand T. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model". J Acoust Soc Am. 2009;126(5):2635-48. DOI: 10.1121/1.3224721
8. Holube I, Winkler A, Nolte-Holube R. Modellierung der Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung – Hat der Freiburger Einsilbertest 29 Wörter pro Liste? Z Audiol. 2018;57(1):6-17.

9. Wagener K, Brand T, Kollmeier B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. Z Audiol. 1999;38(3):86-95.

10. Holube I, Winkler A, Nolte-Holube R. Modellierung und Verifizierung der Test-Retest-Reliabilität des Freiburger Einsilbertests in Ruhe mit der verallgemeinerten Binomialverteilung. Z Audiol. 2020;2:1-25.

11. Thornton A, Raffin M. Speech-discrimination scores modeled as a binomial variable. J Speech Hear Res. 1978;21(3):507-18. DOI: 10.1044/jshr.2103.507

12. Andersen P, Bramsløw L, Pedersen A, Dau T, Kressner A. Using more realistic speech material to enhance ecological validity in the Everyday Conversational Danish Sentence Test. J Acoust Soc Am. 2026;159(4):3062-74. DOI: 10.1121/10.0043241

13. Ibelings S, Brand T, Holube I. Speech Recognition and Listening Effort of Meaningful Sentences Using Synthetic Speech. Trends Hear. 2022;26:23312165221130656. DOI: 10.1177/23312165221130656

14. Ibelings S, Brand T, Ruigendijk E, Holube I. Development of a Phrase-Based Speech-Recognition Test Using Synthetic Speech. Trends Hear. 2024;28:23312165241261490. DOI: 10.1177/23312165241261490

15. Jürgens T, Stohlmann J, Hettler F, Tchorz J, Kallinger M, Denk F, Husstedt H. Untersuchung der Test-Retest-Reliabilität eines neuen alltagsnahen Sprachtests mit der verallgemeinerten Binomialverteilung. In: Deutsche Gesellschaft für Audiologie e. V., editor. 28. Jahrestagung der Deutschen Gesellschaft für Audiologie. Oldenburg, 04.-06.03.2026. Düsseldorf: German Medical Science GMS Publishing House; 2026. Doc081. DOI: 10.3205/26dga081

Corresponding author:
Tim Jürgens
Institut für Akustik, Technische Hochschule Lübeck, Mönkhofer Weg 239, 23562 Lübeck, Germany
tim.juergens@th-luebeck.de

Please cite as
Jürgens T, Stohlmann J, Hettler F, Tchorz J, Kallinger M, Denk F, Husstedt H. Investigation of test-retest reliability of a new close-to-everyday-life speech test using the Poisson binomial distribution. GMS Z Audiol (Audiol Acoust). 2026;8:Doc15. DOI: 10.3205/zaud000092, URN: urn:nbn:de:0183-zaud0000923

This article is freely available from
https://doi.org/10.3205/zaud000092

Published: *2026-07-22*