Psihološka obzorja / Horizons of Psychology, 17, 4, 33-49 (2008)

Znanstveni empiričnoraziskovalni prispevek

A comparative analysis of different procedures for measuring speech recognition threshold in quiet[#]

Anja Podlesek[1], Luka Komidar[1], Gregor Sočan[1], Boštjan Bajec[1], Valentin Bucik[1], Klas Matija Brenk[1], Jagoda Vatovec[2] and Miha Žargi[2]*
[1]University of Ljubljana, Department of Psychology, Ljubljana, Slovenia
[2]University Medical Centre Ljubljana, Department of Otorhinolaryngology and Cervicofacial Surgery, Ljubljana, Slovenia

Abstract: In Slovenia, the adapted Freiburg Monosyllabic Word Test (FT-SI) has been used to assess the communication function in an audiology patient. To measure the speech recognition threshold (SRT), the ascending procedure that is applied in FT-SI may be very time consuming. The aim of our study was to compare several adaptive procedures with the FT-SI ascending procedure. Based on the analysis of comprehensibility and commonness of stimuli used in FT-SI, the most appropriate words were selected and used in three adaptive procedures: two variants of a descending procedure, both recommended by the ISO 8523-3 standards for measuring an SRT, and the staircase method. On a normal-hearing sample ($N = 36$ in test measurement and $N = 24$ in retest measurement), comparable SRTs were obtained with the adaptive procedures, whereas the FT-SI ascending procedure yielded slightly higher SRTs. When a selected pool of words was used in FT-SI, SRTs became more comparable to the results of the adaptive methods. The study therefore showed that the pool of words used in FT-SI should be revised. Considering relatively short administration time, satisfactory convergent validity, precision and test-retest reliability, the staircase method seems to be the best alternative to the FT-SI ascending procedure.

Key words: speech audiometry, speech recognition threshold, ISO 8523-3 standards, psychophysical methods, adaptive methods

Primerjalna analiza različnih postopkov za merjenje praga govornega razumevanja v tišini

Anja Podlesek[1], Luka Komidar[1], Gregor Sočan[1], Boštjan Bajec[1], Valentin Bucik[1], Klas Matija Brenk[1], Jagoda Vatovec[2] in Miha Žargi[2]
[1]Univerza v Ljubljani, Oddelek za psihologijo, Ljubljana, Slovenija
[2]Univerzitetni klinični center Ljubljana, Klinika za otorinolaringologijo in cervikofacialno kirurgijo, Ljubljana, Slovenija

Povzetek: V Sloveniji za ocenjevanje komunikacijske funkcije avdioloških pacientov uporabljamo prirejen Freiburški test enozložnih besed (FT-SI). Če želimo meriti zgolj prag govornega razumevanja, pa je uporaba naraščajočih serij dražljajev, ki jo uporablja FT-SI, zelo dolgotrajna. Namen naše raziskave

[#]Acknowledgement: This research was supported by Research Agency of Republic of Slovenia ARRS grant L5-6240 and company Slušni aparati Widex d.o.o., Slovenia. Parts of this paper were published in the final report on the study.

[*]Naslov / Address: doc. dr. Anja Podlesek, University of Ljubljana, Faculty of Arts, Department of Psychology, p. p. 580, SI-1001 Ljubljana, Slovenia, e-mail: anja.podlesek@ff.uni-lj.si

je bil primerjati več adaptivnih postopkov z naraščajočim postopkom, uporabljenim v FT-SI. Na osnovi analize razumljivosti in pogostosti dražljajev, uporabljenih v FT-SI, smo izbrali najustreznejše besede in jih uporabili v treh adaptivnih postopkih: v dveh različicah padajočih postopkov, opisanih v standardih ISO 8523-3 za merjenje praga govornega razumevanja, in v metodi stopnic. Na vzorcu normalno slišočih oseb ($N = 36$ v prvem merjenju in $N = 24$ v drugem merjenju) smo z adaptivnimi postopki našli primerljive prage govornega razumevanja, medtem ko so bili pragi, določeni z naraščajočim postopkom, nekoliko višji. Ko smo tudi pri naraščajočem postopku uporabili le izbrane besede, so pragi postali primerljivejši rezultatom adaptivnih postopkov. Raziskava je torej opozorila, da je potrebno bazo besed revidirati. Če upoštevamo kratek čas administracije, zadovoljivo konvergentno veljavnost, natančnost in zanesljivost postopkov v času, se zdi, da je med uporabljenimi postopki metoda stopnic najboljša alternativa naraščajočemu postopku v FT-SI.

Ključne besede: govorna avdiometrija, prag govornega razumevanja, standardi ISO 8523-3, psihofizikalne metode, adaptivne metode

CC = 2326

Speech audiometry is often used to diagnose the type of hearing loss and to assess the communication function of the patient. Hearing loss and the effectiveness of hearing aids is assessed using the level to which the patient's functioning in everyday life is preserved, and this is often done by examination of speech perception or recognition. In monitoring the effect of the hearing aid, one can study the change in the speech recognition threshold after the implementation of the aid. Both for the patients and the providers of audiometric services, it is desirable to use a measuring method with short administration time. The method should also have proper metric characteristics, i.e. it should be administered objectively and have satisfactory reliability, validity and discriminability.

Various measures are used in speech audiometry (Smoski, 2007). One measure is the speech recognition threshold (SRT), which is the intensity level at which a person can recognize 50% of the words spoken either in quiet environment or in noise. Usually spondees are used to assess this threshold, which is typically not more than 3 (Wilson, Morgan, & Dirks, 1973) to 6 dB (Smoski, 2007) above the average pure-tone thresholds at frequencies 500, 1000 and 2000 Hz. To study the suprathreshold speech discrimination (speech intelligibility), i.e. the ability to understand and repeat words presented at conversational or another suprathreshold level, tests using monosyllabic words are often used where the phonetically balanced lists are presented and the percentage of correctly repeated words at different intensity levels is determined.

In Slovenia, no test has yet been developed for measuring speech recognition threshold. A speech audiogram is most often assessed with the Slovenian adaptation of the German tests developed by Hahlbrock (1953, 1960)––the Freiburg Monosyllabic Word Test and the Freiburg Number Test. Slovenian adaptations were developed in the 1960s (Pompe, 1968). In the first test (FT-SI in the succeeding text), a patient

listens to phonetically balanced columns of 28–29 monosyllabic Slovenian words in a quiet environment. The stimulus intensity level is increased with each column. A speech audiogram representing the percentage of correctly repeated words at each level serves as the basis for estimating the patient's communication function. To derive the audiogram, many columns (and words) have to be presented, which is very time consuming if one only wants to assess the relative gain in speech recognition. The practitioners' need to assess speech recognition in less time, through measuring only speech recognition threshold instead of obtaining the whole audiogram, was the initial motivation of our study. One could perhaps use FT-SI and simply stop measuring when the 50% recognition is achieved. However, several concerns were raised about the length, reliability, and validity of the similar ascending technique in 5 dB steps, which had been recommended by ASHA in 1979 (see ASHA, 1988). The original German version of the Freiburg Test was also criticized for unequal difficulty of the used words (Bangert, 1980; Sukowski, Brand, Wagener, & Kollmeier, 2008). It was thus necessary to conduct an in-depth analysis of the Slovenian version of the Freiburg Test, i.e. FT-SI, and to find a potentially better alternative procedure for measuring an SRT.

Adaptive psychophysical procedures are the first candidates when choosing among different methods for fast, but methodologically sound measurements of sensitivity. An adaptive method which is very simple to use and is thus frequently present in different experimental studies on hearing, is the staircase method or the up-and-down method (Levitt, 1971). In this method, stimulus intensity is varied according to the subject's answer in the preceding trial. Specifically, in a speech recognition test, if the subject recognizes the presented word, stimulus intensity will be decreased. If she/he does not recognize the word, the next stimulus will be presented at a higher intensity. The step between successive intensities is usually held constant. After obtaining from six to eight reversals, the first one is discarded and the threshold is defined as the average of the midpoints of the remaining runs (see Levitt, 1971). The staircase method and its versions proved to yield valid results in several audiometry studies (e.g. Buss, Hall, Grose, & Dev, 2001).

ISO 8253-3 standards (1996) for speech audiometry describe two adaptive procedures for measuring an SRT, which were developed by Wilson et al. (1973): (i) the descending procedure using 5 dB steps and (ii) the alternative descending procedure using 2 dB or 5 dB steps. These procedures are simple, rapid, and statistically based procedures for determining the recognition threshold (ASHA, 1988). In both procedures stimulus presentation starts at 20 dB to 30 dB above the average of the subject's pure tone hearing threshold levels at 500 Hz, 1000 Hz and 2000 Hz. Then the speech level is reduced in 5 dB steps. At least two items are presented on each level, until the subject no longer responds correctly to all test items at the specified level. The two descending procedures differ afterwards: (i) In the descending procedure using 5 dB steps, a set of test items (with at least 10 items) is next presented at the level where the subject ceased to respond correctly, and the number of correct

responses is recorded. If the subject scores at least 50% on the set of test items, the intensity is reduced in steps of 5 dB and a new set of test items is presented on each intensity level until the subject scores less than 50% on the set of test items. Usually one level is found to yield somewhat more than 50% and the next lower level somewhat less than 50% recognition (ISO 8253-3, 1996). If the subject scores less than 50% on the set, the level is increased in steps of 5 dB and a new set of test items is presented on each level until the subject scores more than 50% on the set of test items. The SRT is calculated by means of linear interpolation between the lowest level that yielded more than 50% correct responses and the highest level that yielded less than 50% correct responses. (ii) In the alternative descending procedure using 2 dB steps, the descending process is continued in steps of 10 dB until a level is reached at which two consecutive test items are missed. Then the speech level is increased by 10 dB. Two test items are presented at this so-called starting speech level and at each successive 2 dB decrement. This process is continued if at least five out of the first six test items are repeated correctly. If this criterion is not met, the starting speech level is increased by 4 dB to 10 dB. The descending series is terminated when the test subject responds correctly to five of the last six test items presented. The SRT level is calculated according to the Spearman-Kärber method (see ASHA, 1988; ISO 8253-3, 1996).

It is stated in the ISO 8253-3 (1996) standards: "The [descending] procedures are expected to yield comparable results. However, experimental evidence for this is still unavailable." (p. 9)

The aim of our study was to examine convergent validity and reliability of three adaptive methods––the staircase method, the descending procedure, and the alternative descending procedure––and to contrast them with the ascending procedure. The final goal was to develop the computerised adaptive procedure for measuring an SRT in a fast, efficient manner.

Method

Stimuli

The Slovenian version of the Freiburg number test contains only six columns of 10 numbers, whereas the word test (FT-SI) consists of 281 monosyllabic words. To reliably measure SRT, many stimuli are needed. Therefore, the number test could not be used for this purpose. We decided to use the monosyllabic words as stimuli, although this type of speech material is more commonly used for assessing speech intelligibility through a complete audiogram, not for determining a threshold, i.e. a single point on the psychometric curve.

Many of the words in FT-SI are archaic. Although read by a professional speaker, they are sometimes difficult to understand even at the intensities well above

the threshold. We soon realized that the use of existing material is not an optimal option. However, because development of a new base of stimuli and their calibration (construction of the performance-intensity functions for each word) would require an extra and demanding study, we decided that for the purpose of the present study, i.e. the comparison of different procedures for measuring an SRT, the existing verbal material can be used if stimuli with comparable properties are selected. In the adaptive procedures, we wanted to include only words with a similar difficulty level. The rationale for this was the assumption that, because in these procedures the intensity of each stimulus depends on the previous responses, it is important for the responses to depend only on stimulus intensity and not on other characteristics of stimuli, such as word comprehensiveness. Low homogeneity of speech stimuli may compromise the reliability of the measures (ASHA, 1988).

Our first step was to examine the quality of each word. We analysed: (i) the frequency of its use in literary language, (ii) the frequency of its use in colloquial language, and (iii) the clarity of speaker's pronunciation and the word's distinctiveness, i.e. the probability that it may be confused with a different word having a similar phonetic structure.

The frequency of the use of words in literary language was determined with FidaPlus corpus of the Slovenian language (FidaPlus, 2007), which contains the information on the frequency of various words used in written documents. The frequency of the use of words in colloquial language was assessed on a sample of 141 students (on average 25 years old) who were given written lists of words in FT-SI. They had to assess, using a 6-point scale (0 – never, 6 – very often), the frequency of occurrence of each word in their everyday life (how often they hear it on TV, radio, use it in spoken language, etc.). The correlation between the frequency of the words' use in literary and colloquial language was .66. The average score (the frequency index) was calculated for each word. A different sample of 44 students (on average 20 years old) participated in measurements of clarity of the words. The words were presented in a large lecture hall for the whole group at the same time and at the usual intensity level of a speaking lecturer, i.e. at approx. 60 dB. After listening to each word, students had to write down what they had heard and at the same time mark a special field if they were not completely certain. For each word, we counted the correctly reproduced words and the marked fields. From the first, we calculated the index of clarity (the proportion of subjects that correctly reproduced the word), and from the second, we calculated the index of certainty (the proportion of those who were certain that their reproduction was correct). The correlation between the indexes of clarity and certainty was .67. The correlation between the frequency of use in the colloquial language and the index of clarity was .35 and the correlation between the frequency of use and the index of certainty was .39.

The words were ranked according to the clarity index. The index value .95 was chosen arbitrarily (there was a clear drop in the index values after the chosen limit) to separate the words of inferior quality from the rest, and that resulted in

selection of 161 out of 281 original words. Next, the average of the three indexes was calculated and 26 words having lowest average were omitted. In the end, 135 words remained in the pool for the adaptive methods (see Podlesek et al., 2007). We compared the phonetic structure of the new pool of words to the old one. We found that in the new pool there was a surplus of the letters—or phonemes (in Slovenian language, a phoneme typically corresponds to a single letter)—*k*, *n*, *v* and *z*, and a lack of letters *d*, *f*, *g* in *p*, but the difference in the amount of letters never exceeded 14%. In the new pool the average length of the words was 3.71 letters ($SD = 0.74$), whereas in the old pool it was 3.57 letters ($SD = 0.73$). It seems that words with a single consonant preceding or following a vowel (CVC) were perceived less clearly than the words with two grouped consonants (CCVC, CVCC, or CCVCC). Nevertheless, the new pool retained the phonetically balanced structure.

Instruments

Computer applications of four procedures for measuring SRTs were developed with MS Visual Basic 6.0: (i) FT-SI, (ii) the staircase procedure, (iii) the descending procedure, and (iv) the alternative descending procedure.

Words, stored as .wav (uncompressed PCM format) files, were presented with a standard personal computer with Creative SB Audigy sound card. The files with speech stimuli were extracted (cut) from FT-SI, where each column of test items was recorded as a single CD track (CD-ROM [SLOG, 2002] was recorded from the original magnetic tape, produced in 1990's). Stimulus files were prepared with Audacity 1.2.3 software. The recording of each word started 100 ms before the beginning of utterance and ended 500 ms afterwards.

The measurements were carried out with words presented binaurally through Sennheiser HD650 headphones. For FT-SI (the ascending procedure) we used the original pool of words (281 words), whereas the pool of words for the adaptive methods consisted of 135 selected words (see Stimuli section). Intensity was controlled by addressing MS Windows' Master and Wave Volume controllers. The generic values of these controllers were assigned exact values in dB SPL by means of the calibration procedure.

Calibration was conducted according to IEC 310 standards, using a Brüel & Kjær Type 4152 artificial ear with coupler DB 843 and the Brüel & Kjær Pulse Labshop 10.0 software. Calibration, as well as the measurements in the experiment, was performed in a laboratory at the Faculty of Arts, Ljubljana, Slovenia. Although the room was quiet and isolated, and the computer that generated the stimuli was placed in the adjacent room, the sound field was still polluted with low frequencies and thus the background noise level was 30±2 dB SPL. For this reason the intensity of the stimuli began at 32 dB SPL. For each stimulus, output intensity at different Master and Wave Volumes was calculated as the average of the 54 peak values over the frequency spectrum of 156–6000 Hz (band-pass filter was used) within the time

interval equal to the duration of the utterance. The final presentation level in dB SPL was obtained after two or three replays of each word at a certain value of volume controllers. For each word, the values of controllers were defined for the output levels between 30 and 99 dB SPL, with 1 dB SPL accuracy.

Participants

Thirty-six otologically normal volunteers, mostly undergraduate students in psychology who were on average 21 years old (SD = 2.3 years), participated in measurements of SRTs. Prior to the study, no participant had any experience with presented stimuli or the measurements of an SRT.

Procedure

After the participant sat down and relaxed, the experimenter collected demographic data and informed her/him about the measurements and the task. ISO 8253-3 (1996) standards on the preparation and instruction of test subjects were followed.

In front of the participant there was a computer screen where the visual signal for the stimulus interval was presented. Namely, a red screen indicated that the word was being presented at that moment or had been recently presented, and a green screen informed the participant that the stimulus would soon be presented and that she/he needed to get ready and pay attention. The screen turned green 1 s before the stimulus started[1].

Stimuli were presented in 5-second intervals. In the mean time, participants had to repeat what they had heard, and the experimenter, who was informed about the correct content of the presented stimulus, clicked a button to save the correctness of the answer.

Every person participated in measurements with all four procedures, first in the measurements with FT-SI ascending procedure, and then in the measurements with the adaptive procedures[2].

The order of the adaptive procedures was varied across participants: Six equal groups of participants were formed and each group was subjected to one of the possible sequences. Due to high absolute and differential sensitivity of the participating

[1]Although an auditory cue is used in the original version of the Freiburg Test, we decided not to use it in order to reduce the possibility of its effect on word recognition. If the intensity of the cue equals the intensity of the stimulus, it is possible that the cue is not heard at low intensities, which may affect the probability of stimulus recognition.

[2]Such an order was chosen to equalize the effect of learning for all the participants. In the ascending procedure, all participants received the same series of words, whereas in the adaptive methods the presentation order of different words was randomized. Therefore it was possible that, in comparison to other procedures, FT-SI would exert the largest systematic effect of learning. Putting FT-SI at the beginning of the experiment therefore seemed a better choice than a randomization of the order of all the procedures.

subjects, fixed steps of 2 dB were used in all procedures and the administration of some procedures had to be modified. In the ascending procedure, the first column of words was presented at the level of 32 dB SPL, as this was the lowest possible suprathreshold level, and then intensity was increased by 2 dB SPL in subsequent columns. If larger step size had been chosen, the percentage of correctly repeated words would have risen too quickly. The staircase method started at 42 dB SPL and steps of 2 dB SPL were used throughout the measurements. The series terminated after 12 reversals, and the average of the last 10 reversals determined the SRT. In the descending procedure, stimulus presentation started at 42 dB SPL[3] and intensity was then reduced in steps of 2 dB SPL – larger step size, such as the one suggested in ISO 8253-3 (1996) standards, might result in the levels that would be lower than the possible minimum (e.g., subjects might correctly repeat both words presented at 34 dB SPL, so that the next pair of words would be played at 29 dB SPL, which was not possible to hear). When the subject no longer responded correctly to both of the items presented at a certain level, a set of 10 test items was presented at the same level, and the intensity was later decreased or increased by 2 dB SPL, until sets with less than or more than 50% correct repetitions were obtained. In the alternative descending procedure, at the start the intensity was 50 dB SPL and was then decreased by 2 dB SPL. When a level was reached at which two consecutive test items were missed at the same level, the speech level was increased by 8 dB. Subsequent decrements had a magnitude of 2 dB and eventual subsequent increments had a magnitude of 8 dB.

In the ascending procedure, stimuli were presented in a predetermined order (as in the original version of FT-SI). Ten lists of stimuli were prepared, with 29 stimuli in the 1st, 4th, 7th and 10th list, and 28 stimuli in the other lists. In each of the adaptive procedures, stimuli were sampled randomly without replacement.

In order to examine the test-retest reliability of different methods, another session of measurements was carried out approximately 9 months later. It was assumed that in the mean time the participants would forget which words were used in the first session. Twenty-four participants were retested, whereas the other 12 did not respond to our request. The procedure used in the first session was repeated, with a different, randomly chosen order of the adaptive methods for each participant. Additionally, after the last adaptive method, the ascending procedure was repeated, but this time stimuli were chosen randomly from the pool of 135 words that were selected to be used in the adaptive methods. This condition was added to compare the results of the ascending procedure using the original FT-SI words with the results of the same procedure using the selected words only (words with higher clarity). If the results were different, this would indicate a possible effect of the presented words on the measured threshold.

[3]We wanted to measure the SRT without prior knowledge of pure-tone thresholds and at the same time did not want to start the measurements too far away from the threshold, for example at the level of 50 dB HL as suggested by Martin and Stauffer (1975). The level of 42 dB SPL seemed a reasonable choice, especially because this level was also chosen as the starting level in the staircase method. Recognition of words was 100% at the chosen level in all the participants.

Results and discussion

With two-way mixed-design analyses of variance we first examined if the SRT and procedure duration differed among the procedures and if they were affected by the order in which the procedures were administered. Greenhouse-Geisser correction was used when necessary. Results for four procedures (the ordinary version of FT-SI and the three adaptive methods) were included in calculations presented below.

In both measurements, the procedure was the only factor that significantly affected the duration of measurements (see Table 1). To define an SRT in the first measurement, 114.6 stimuli had to be presented on average with the ascending procedure (SD = 24.9), 44.4 stimuli (SD = 4.4) with the staircase method, 47.5 stimuli (SD = 13.6) with the descending procedure, and 32.2 stimuli (SD = 4.2) with the alternative descending procedure. In the second measurement, the average number of stimuli presented before the completion of the procedure was 30.8 (SD = 5.4) for the staircase method, 52.8 (SD = 18.2) for the descending procedure, 37.5 (SD = 5.2) for the alternative descending procedure, and 126.8 (SD = 18.5) for the ascending procedure. Overall, the adaptive methods were completed two to three times quicker than the FT-SI ascending procedure.

Table 1. *Summary of ANOVAs of thresholds and durations in two measurements*

	F	df_1	df_2	p	MSE	$\eta^2_{partial}$
1st measurement – threshold						
procedure	3.42	2.14	64.18	.036	1.96	.10
order	0.30	5.00	30.00	.907	6.51	.05
procedure × order	1.61	10.70	64.18	.123	1.96	.21
2nd measurement – threshold						
procedure	30.28	3.00	54.00	.000	1.01	.63
order	1.97	5.00	18.00	.132	3.67	.35
procedure × order	0.93	15.00	54.00	.535	1.01	.21
1st measurement – duration						
procedure	198.23	1.46	43.93	.000	501.62	.87
order	0.46	5.00	30.00	.806	166.56	.07
procedure × order	0.58	7.32	43.93	.776	501.62	.09
2nd measurement – duration						
procedure	30.28	3.00	54.00	.000	245.20	.63
order	0.17	5.00	18.00	.969	272.35	.05
procedure × order	1.03	15.00	54.00	.443	245.20	.22

Overall, the thresholds obtained with different procedures were slightly lower in the second measurement than in the first one, but the difference between the results of the first and the second measurement was not statistically significant,

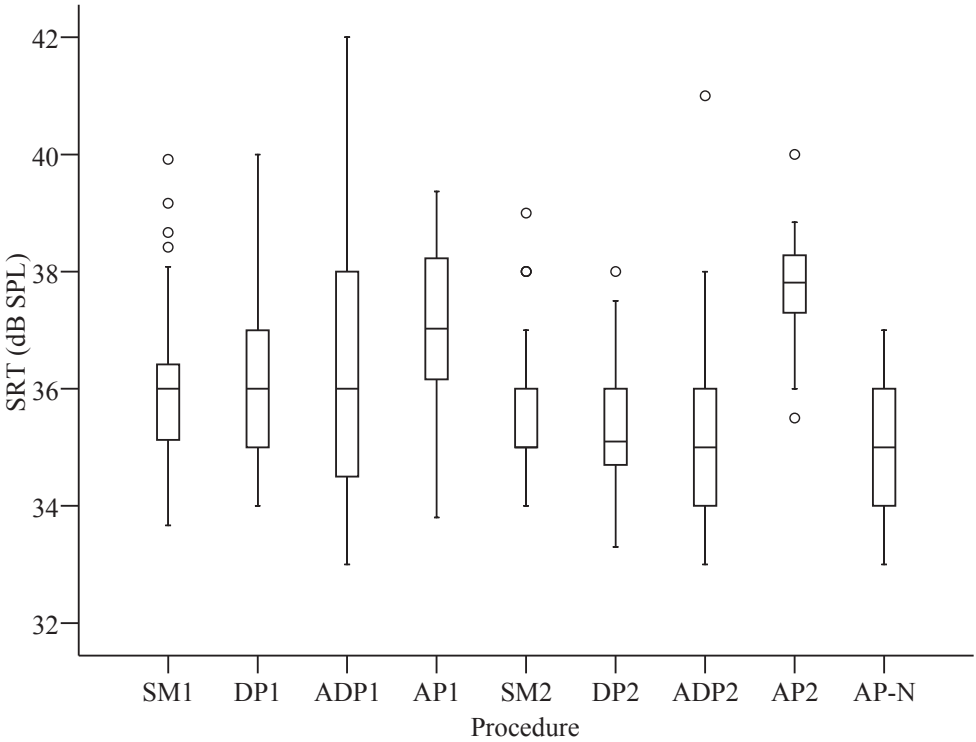

Figure 1. Speech recognition thresholds, as measured with different procedures: SM –staircase method, DP – descending procedure, ADP – alternative descending procedure, AP – the ascending procedure, AP-N – ascending procedure with a new pool of words. The numbers in the names of the procedures represent the measurement order: 1 – the first measurement ($N = 36$), 2 – the second, retest measurement ($N = 24$). Boxes show the interquartile range and whiskers show the absolute range with exception of outliers, represented by circles.

Hotelling's $T = .036$, $F(1, 23) = 0.826$, $p = .373$. In both measurements, procedure was the only factor that significantly affected the SRT (see Table 1). Figure 1 shows the SRTs obtained with different procedures. The adaptive methods yielded comparable results (in the first measurement: staircase method $M = 36.1$ dB SPL, the descending procedure $M = 36.2$ dB SPL, the alternative descending procedure $M = 36.1$ dB SPL; in the second measurement: staircase method $M = 35.6$ dB SPL, the descending procedure $M = 35.4$ dB SPL, the alternative descending procedure $M = 35.5$ dB SPL), whereas the ascending procedure gave the average threshold that was somewhat higher ($M = 36.9$ dB SPL in the first measurement and $M = 37.8$ dB SPL in the second measurement). In both measurements, the largest dispersion of individual results was obtained with the alternative descending procedure. Among the adaptive methods, the staircase method gave least inter-individually variable results.

The higher threshold in the ascending procedure compared to the ones in the adaptive methods could be explained in two ways. First, contrary to the adaptive methods, only ascending series were presented in this procedure. In obtaining various measures, the ascending thresholds are usually higher than the descending ones (Robinson & Koenigs, 1979; Wall, Davis, & Myers, 1984), primarily because of the habituation error and hysteresis (see Gescheider, 1997). Perhaps, due to the influence of top-down processes (attention, expectations), subjects follow descending series much easier than ascending series. In descending series, they expect subsequent stimuli to be audible, so they are more motivated to listen and pay more attention to the interval in which the stimulus is to be presented. All the adaptive procedures contained descending series, whereas the FT-SI ascending procedure did not, which might result in lower thresholds for the first ones. After the first measurement, we examined the reversals in the staircase method, and the reversals in the descending series were indeed lower (M = 35.4 dB) than reversals in the ascending series (M = 36.9 dB). Moreover, the average reversal intensity in the ascending series resembled the threshold obtained with the FT-SI ascending procedure. Thus, it seemed that the difference between the results of different procedures might be a consequence of the type of stimulus series used for estimating a threshold. Second, it is possible that the thresholds were higher in the FT-SI ascending procedure because the whole pool of words was used, so the words with higher difficulty were presented also. This might have resulted in a lower percentage of correctly repeated words and consequently in a higher threshold. To see if this explanation is correct, we repeated the ascending procedure at the end of the second measurement, but this time using only the selected words, i.e. words with similar difficulty (with high frequency of use, clearly pronounced words and words distinct from other phonetically similar words). The use of selected words in the ascending procedure resulted in a threshold that was similar to the ones obtained with the adaptive methods (see Figure 1) and much lower than the one obtained with the ordinary version of this test (with archaic and rare words included). This implies that the difference between the thresholds obtained with the ordinary version of the FT-SI ascending procedure and the thresholds obtained with the adaptive methods can principally be attributed to different stimuli used in the measurements. The selection of words therefore plays a very important role in measuring an SRT. When selecting words for use in the adaptive methods, a relatively large number of words with certain phonemes had to be dropped from the pool (e.g., in the recordings, the phoneme p was sometimes underemphasized by the speaker, and the correct reproduction of the word was therefore less probable; the exclusion of such words finally resulted in a small under-representation of the phoneme p). This indicates the need for revision of the recorded materials for future audiometry measurements. Words which are infrequent in everyday communication or easily confounded with phonetically similar words should be excluded from the lists.

Usually the slope of the recognition curve is lower for monosyllabic words than for spondees or sentence tests (Brand & Kollmeier, 2002). It is common to use

spondees in measuring an SRT, because the estimation of the threshold can be more reliable (exact) with the steeper slope of the psychometric curve. One could assume that the measurement error in our study was probably somewhat larger because we used monosyllabic words as stimuli than it would have been had we used the bisyllabic words or even more structured stimuli. However, in our study, the slope of the recognition curves was quite high. Table 2 shows standard deviation of the momentary SRTs in different methods. In the staircase method, the momentary SRT was calculated for each direction of the series as the average of the upper and the lower threshold determined in that series direction. In the ascending and the descending procedure, standard deviation was calculated as the ratio of the difference between the lowest level that yielded more than 50% correct responses and the highest level that yielded less than 50% correct responses and the difference between z-value corresponding to the proportion of correct responses at both levels, respectively. In the alternative descending procedure, standard deviation was estimated as the ratio of the difference between the threshold level determined by the Spearman-Kärber method and the level of the lowest intensity reached and the difference between the proportion of correct responses at the threshold level (50%) and at the lowest intensity. The latter was estimated to be 0 if six out of the last six presented words were repeated incorrectly or .167 if five out of the last six presented words were repeated incorrectly.

Table 2. *Descriptive statistics for standard deviations of the recognition curve*

Procedure	*Me*	Min	Max	Predicted 90% transition interval
SM1	1.57	0.85	3.17	5.2
SM2	1.34	0.50	2.44	4.4
DP1	1.51	0.00	7.89	5.0
DP2	1.46	0.00	7.89	4.8
ADP1	5.17	1.00	7.24	17.0
ADP2	4.65	1.00	7.24	15.3
AP1	1.77	0.80	4.85	5.8
AP2	1.97	0.82	7.94	6.5
AP-N	1.59	0.52	4.85	5.2

Note. SM – the staircase method, DP – descending procedure, ADP – alternative descending procedure, AP – ascending procedure, AP-N – ascending procedure with a new pool of words. The numbers in the names of the procedures represent the measurement order: 1 – the first measurement ($N = 36$), 2 – the second measurement (retest; $N = 24$).

The lower the standard deviation in Table 2, the steeper the psychometric curve, which indicates a more precise measurement of the SRT. We see that the staircase method and the descending procedure offered the most precise estimation of the SRT, whereas the alternative descending procedure yielded the least precise

(reliable) SRT estimate. In the alternative descending procedure the estimation of *SD* of the recognition curve is not straightforward. Because only two stimuli are presented at each intensity level, it is difficult to estimate the proportion of recognition. If the subject fails to satisfy the criterion of correctly repeating at least one out of the last six stimuli, we estimate that at the intensity of the last presented stimuli the proportion of recognition is between 0 and 0.167. However, such a proportion could as well be reached at a higher intensity, but the recognition is not measured with a big enough precision to find that out. Thus, it seems that the alternative descending procedure allows only for a rough approximation of an SRT. According to the criterion of intra-individual reliability (precision) of the SRT estimation, the staircase method and the descending procedure have the advantage over the other two procedures for measuring an SRT.

The slope of the discrimination function was around 20–25% per dB (this estimation is based on the results obtained with the descending procedure), which is comparable to the slopes for the sentence tests (see Brand & Kollmeier, 2002), indicating that the stimuli and the procedures used, except the alternative descending procedure, allowed for a quite precise estimation of the threshold.

If we accept that the transition interval (i.e. the interval of intensities between complete non-recognition and complete recognition) covers several standard deviations of the recognition curve, we can assess the appropriateness of the chosen step sizes. In the last column of Table 2 the predicted transition intervals are presented for different procedures. Predicted transition intervals cover the interval of 3.29 median standard deviations, therefore representing the middle 90% of the recognition curve (i.e., the interval between the intensity at which the stimuli are recognized correctly in 5% of cases and the intensity at which the stimuli are recognized correctly in 95% of cases). We can see that in case of the staircase method, the descending procedure and the ascending procedure there would be a 4–7 dB SPL difference between the point of 5% recognition and the point of 95% recognition. It therefore seems that for these procedures, the chosen step size of 2 dB SPL was a minimal step size for measuring an SRT in normal-hearing subjects––with three to five presentation levels we could cover the whole transition interval. This is at the lower limit of the number of stimuli required for the method of constant stimuli (see Gescheider, 1997). Therefore, a smaller step size (a step size of 1 dB SPL) would perhaps be better.

To assess the convergent validity of the applied procedures, we calculated correlation coefficients between the SRTs assessed by different procedures. The results are shown in bold in Table 3. Overall, the correlations between the results of different procedures were positive and moderate, both in the first and in the second measurement. We may conclude that the adaptive methods showed a satisfactory convergent validity. The correlations could hardly be higher, due to low inter-individual variation of thresholds (see Figure 1). Another indicator of the criterion validity can be the correlation of the threshold assessed by a certain procedure with the average threshold of the other three procedures. This coefficient was largest for the staircase method both for the first and the second measurement (see the last row in Table 3).

Table 3. *Pearson correlation coefficients for thresholds assessed by different procedures, and corrected discrimination coefficient (r) for each procedure*

Procedure	SM1	DP1	ADP1	AP1	SM2	DP2	ADP2	AP2	AP-N
DP1	**.56****								
ADP1	**.48****	**.12**							
AP1	**.61****	**.52****	**.42***						
SM2	.43*	.41*	.21	.28					
DP2	.09	.29	–.02	.06	**.47***				
ADP2	.63**	.55**	.48*	.68**	**.68****	**.41***			
AP2	.46*	.29	.13	.24	**.46***	**.35**	**.47***		
AP-N	.33	.30	.27	.19	**.47***	**.31**	**.38**	**.46***	
r	.73	.45	.39	.66	.72	.50	.66	.56	.51

Note. SM – the staircase method, DP – descending procedure, ADP – alternative descending procedure, AP – ascending procedure, AP-N – ascending procedure with a new pool of words. The numbers in the names of the procedures represent the measurement order: 1 – the first measurement (N = 36), 2 – the second measurement (retest; N = 24; 24 data were also included in the correlations of the first and the second measurement). Corrected discrimination coefficient (r) is the correlation of the threshold, assessed by a certain procedure, with the average threshold of the other three procedures. Convergent validity coefficients are written in bold and test-retest reliability coefficients are underlined.
$^{*}p < .05$. $^{**}p < .01$.

We also examined the test-retest reliability of different procedures. The correlations of the SRTs obtained with a certain procedure in the first and the second measurement (see the underlined values in Table 3) were positive, but reached statistical significance only for the staircase method and the alternative descending procedure. Thus, the correlation of the two measurements was not as high as desired. Moreover, some of the correlations between the results of different procedures in different measurements were close to zero. This may indicate that the methods are unreliable or do not measure the same thing. However, if we take into account that we examined a normal-hearing sample and used a 2 dB step size, these results are not surprising, because the individual thresholds as assessed by different procedures were often within the size of a single step. In the adaptive methods, momentary uncontrolled factors may have affected the final estimate of an SRT slightly. For example, with a 2 dB step size, incorrect recognition of a single stimulus may result in the SRT increase in the order of magnitude of 0.2 dB in the staircase method and in the order of magnitude of as much as 1 dB in both descending procedures. With the normal-hearing listeners this may easily affect the distribution of individual SRTs and consequently lower the correlations between SRTs of different procedures. Low correlations can therefore be attributed to relatively small inter-individual variability of the thresholds. It can be expected that the correlations reflecting convergent validity and reliability of different procedures would be higher on a clinical population with more diverse SRTs (cf. Bauman, 1984). To measure an SRT in a normal-hearing

population the use smaller step size might be better, but would probably also result in much longer measurements.

Conclusions

The convergent results of the three adaptive methods and the ascending procedure using the pool of selected words show that the adaptive methods could effectively replace the time-consuming ascending procedure. The moderate positive correlations among procedures support the conclusion about the validity of the adaptive procedures for measuring an SRT. The staircase method showed a slight advantage over the other two adaptive procedures: in this method we found the highest corrected discrimination coefficient, moderate test-retest reliability and a low dispersion of intra-individual thresholds. The advantage of this method is that the accuracy of measurements and their duration can be adjusted with the size of the step and the required number of reversals. Its disadvantage, on the other hand, is that stimulus intensity varies closely around the threshold level most of the time, which may be stressful for the participants because the near-threshold stimuli constantly require their full attention and make them feel uncertain when responding.

Future studies of the used procedures should include samples from clinical populations. Because higher age is characteristic of such a population, the problems of attention might appear more relevant, and limitations of procedures like the staircase method might become more prominent. Studies should also be extended to the population of children, for whom fast and efficient measurement of communication function is even more essential in order to provide proper rehabilitation as soon as possible. Regional accents should be taken into account (see Feldman, 2004). Furthermore, methods should be extended to utterances of different length and linguistic complexity, such as rhymes (e.g., Sukowski, Brand, Wagener, & Kollmeier, 2007) and other bi- or multisyllabic words and sentences, to grasp speech reception in every day communication, which is far more complex than usually considered in artificial situations in the laboratory (Kollmeier, 2007).

References

American Speech-Language-Hearing Association (ASHA). (1979). Guidelines for determining threshold level for speech. *ASHA, 21*, 353–356.

American Speech-Language-Hearing Association (ASHA). (1988). Guidelines for determining threshold level for speech. *ASHA, 30*, 85–89.

Bangert, H. (1980). Probleme bei der Ermittlung des Diskriminationsverlustes nach dem Freiburger Sprachtest [Problems in investigating the laws in disrimination with the Freiburg Speech Test]. *Audiologische Akustik, 19*, 166–170.

Bauman, N. (1984). *Suprathreshold levels for pure tones and speech in subjects with nor-*

mal hearing (Educat. D. diss., Columbia University Teachers College). Retrieved on 30 June 2008 from Dissertations & Theses: A&I (database on-line, publication number AAT 8411256): http://www.proquest.com/

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America, 111*, 2801–2810.

Buss, E., Hall, J. W., Grose, J. H., & Dev, M. B. (2001). A comparison of threshold estimation methods in children 6–11 years of age. *Journal of the Acoustical Society of America, 109*, 727–731.

FidaPlus. (2007). *FidaPlus – korpus slovenskega jezika [FidaPlus – the corpus of Slovenian language]*. Retreived on 20 May 2007 from: http://www.fidaplus.net/

Gescheider, G. A. (1997). *Psychophysics: The Fundamentals* (3rd Ed.). Mahwah, NJ: Lawrence-Erlbaum Associates.

Hahlbrock, K. H. (1953). Über Sprachaudiometrie und neue Wörterteste [On speech audiometry and new word test]. *Arch Ohren Nasen Kehlkopfheilkd, 162*, 394–431.

Hahlbrock, K. H. (1960). Kritische Betrachtungen und vergleichende Untersuchungen der Schubertschen und Freiburger Sprachteste [Critical reflection and comparative examination of the Schuberts and the Freiburg test]. *Zeitschrift für Laryngologie, Rhinologie, Otologie und Ihre Grenzgebiete (Stuttg)*, *39*, 100.

International Electrotechnical Commission. (1996). *ISO 8253-3: 1996. Acoustics. Audiometric test methods - Part 3: Speech audiometry*. Geneva: Author.

Kollmeier, B. (2007). *Speech recognition*. Paper presented at the 8th EFAS Congress / 10th Congress of the German Society of Audiology, Heidelberg, Germany, June 6–9, 2007. Retrieved on 22 July 2008 from: www.unizh.ch/orl/dga2007/program/Kollmeier__B..pdf

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*, 467–477.

Martin, F. N., & Stauffer, M. L. (1975). A modification of the Tillman-Olsen method for obtaining the speech reception threshold. *Journal of Speech and Hearing Disorders*, *40*, 25–28.

Podlesek, A., Komidar, L., Sočan, G., Bajec, B., Bucik, V., Brenk, K. M., Vatovec, J., & Žargi, M. (2007). *Razvoj preizkusov procesiranja govornih dražljajev: kognitivnopsihološki in avdiološki vidiki [Development of speech audiometry tests: Cognitive psychological and audiological perspective]*. Research report L5-6240. Ljubljana: Slovenian Research Agency.

Pompe, J. (1968). *Razvoj avdiometrije na ORL kliniki v Ljubljani [Development of audiometry at ORL Clinic in Ljubljana]*. Unpublished manuscript, University Medical Center Ljubljana, Ljubljana, Slovenia.

Robinson, D. O., & Koenigs, M. J. (1979). A comparison of procedures and materials for speech reception thresholds. *Journal of the American Audiology Society*, *4*(6), 227–230.

SLOG. (2002). *Slovenian Speech Audiometry Tests* [CD]. Ljubljana: University Medical Center.

Smoski, W. J. (2007, August). *Speech audiometry*. Retrieved on 20 October 2007 from eMedicine Clinical Reference database: http://www.emedicine.com/ent/topic371.htm

Sukowski, H., Brand, T., Wagener, K. C., & Kollmeier, B. (2007). *The relationship between tone- and speech-audiometry based assessments of hearing loss*. Presentation at the 8th EFAS Congress / 10th Congress of the German Society of Audiology, Heidelberg, June 2007. Retrieved on 22 July 2008 from: www.unizh.ch/orl/dga2007/program/scientificprogram/Sukowski__H._et_al.pdf

Sukowski, H., Brand, T., Wagener, K. C., & Kollmeier, B. (2008). *Vergleich des Freiburger Sprachtests mit moderneren Sprachtestverfahren im Rahmen der Begutachtung beruflicher Lärmschwerhörigkeit [Comparison of the Freiburg speech test with modern speech test procedures within the framework of assessment of professional noise-induced hardness of hearing].* Presentation at the 39. DGMP Tagung, September 2008, Oldenburg, Germany. Retreived on 22 July 2008 from: https://www.hoertech.hausdeshoerens-oldenburg.de/dgmp2008/abstract/Sukowski.doc

Wall, L. G., Davis, L. A., & Myers, D. K. (1984). Four spondee threshold procedures: A comparison. *Ear and Hearing, 5*(3), 171–174.

Wilson, R. H., Morgan, D. E., & Dirks, D. D. (1973). A proposed SRT procedure and its statistical precedent. *Journal of Speech and Hearing Disorders*, *38*, 184–191.

Prispelo/Received: 31.07.2008
Sprejeto/Accepted: 08.12.2008