Comparison of three types of French speech-in-noise tests: A multi-center study

Cite as: Jansen S., Luts H., Wagener K.C., Kollmeier B., Del Rio M., et al 2012. Comparison of three types of French speech-in-noise tests: a multi-center study. *Int J Audiol*, 51 (3), 164–73.

Sofie JANSEN¹, Heleen LUTS¹, Kirsten Carola WAGENER², Birger KOLLMEIER^{2,3}, Matthieu DEL RIO⁴, René DAUMAN⁴, Chris JAMES⁵, Bernard FRAYSSE⁵, Emilie VORMES⁶, Bruno FRACHET⁶, Jan WOUTERS¹, & Astrid VAN WIERINGEN¹

¹ ExpORL, Dept. Neurosciences, K.U.Leuven, Belgium

² Hörzentrum Oldenburg, Germany

³ Carl von Ossietzky-Universität Oldenburg, Germany

⁴ Hôpital Pellegrin, Université Bordeaux Segalen, France

⁵ Service d'ORL, Hôpital Purpan, CHU Toulouse, France

⁶ Hôpital Avicenne, Université Paris XIII, France

Key words:

Speech recognition in noise, French speech tests, everyday sentences, matrix sentences, digit triplets, discriminative power

Corresponding author:

Sofie Jansen

ExpORL, Dept. Neurosciences, K.U.Leuven

Herestraat 49/721

B-3000 Leuven

Belgium

Tel: +32 16 33 04 77

Sofie.Jansen@med.kuleuven.be

Abbreviations:

ANOVA: Analyses of variance

FIST: French Intelligibility Sentence Test

FrDigit3: French Digit Triplet test

FrMatrix: French Matrix test

HI: Hearing impaired

LTASS: Long-term average speech spectrum

NH: Normal hearing

NH-: Sub-normal hearing

PTA: Pure-tone average

SNR: Signal-to-noise ratio

SRT: Speech reception threshold

Abstract

Objective: To compare results on the everyday sentence test 'FIST', the new closed-set sentence test 'FrMatrix', and the digit triplet screening test 'FrDigit3'. *Design:* First, the FrMatrix was developed and normative values were obtained. Subsequently, speech reception thresholds (SRTs) for the three types of tests were gathered at four study centers representing different geographic regions in Belgium and France. *Study Sample:* Fifty-seven normal hearing listeners took part in the normative study of the FrMatrix and 118 subjects, with a wide range of hearing thresholds, participated in the comparative study. *Results:* Homogenizing the individual words of the FrMatrix with regard to their intelligibility resulted in a reference SRT of -6.0 (± 0.6) dB SNR and slope at the SRT of 14.0 %/dB. The within-subject variability was only 0.4 dB. Comparison of the three tests showed high correlations between the SRTs mutually (>0.81). The FrMatrix had the highest discriminative power, both in stationary and in fluctuating noise. For all three tests, differences across the participating study centers were small and not significant. *Conclusions:* The FIST, the FrMatrix, and the FrDigit3 provide similar results and reliably evaluate speech recognition performance in noise both in normal hearing and hearing impaired listeners.

Introduction

For many years it has been known that testing a listener's speech recognition in background noise yields important information in addition to the pure-tone audiogram. Since the introduction of the first sentence-in-noise tests by Kalikow et al (1977) and Plomp & Mimpen (1979), many similar sentence tests have been developed in different languages, but with lots of variations in the style and the content of the sentences, speaker selection, type of noise, presentation method, response scoring, et cetera. A very extensive and recent overview of these variables is given by Theunissen et al (2009). One important similarity is that most of these tests contain everyday sentences, presented in an open-set paradigm. In 1982, Hagerman introduced a new type of speech-in-noise test in

Swedish with syntactically fixed sentences (name-verb-numeral-adjective-object), based on a closed-set of 5 x 10 words. These sentences are semantically unpredictable, thus being less redundant than the everyday sentences. This type of test was adopted and further improved by Wagener et al (1999c, 1999a, 1999b) for the development of the German OLSA, and later for the Danish Dantale II (Wagener et al, 2003). Within the European HearCom project, this so-called *matrix test* has been developed for many other languages (e.g. the Polish Sentence Matrix Test by Ozimek et al, 2010). The major advantage of a matrix test over an everyday sentence test is primarily that it can be used for multiple testing in the same subject, since the sentences are less liable to be memorized. This is especially valuable in research settings (e.g. for evaluation of new hearing instrument algorithms) or in rehabilitation (e.g. comparison of different hearing aid settings), where subjects are typically tested in a series of different conditions.

However, one already known disadvantage of matrix tests is the training effect in inexperienced and even experienced listeners (Hagerman, 1982; Wagener et al, 1999b, 2003; Hernvig & Olsen, 2005). This effect can be reduced by starting each test session with a training phase. Furthermore, when analyzing results on the level of a whole group of subjects, the training effect can be outbalanced by testing different conditions (e.g. different hearing aid algorithms) in a different order for each subject. Matrix tests are thus less suitable for routine clinical testing, where testing time restrictions prohibit sufficient training.

A frequently used measure for speech-in-noise tests is the speech reception threshold (SRT), which is the stimulus presentation level (relative to the noise level) yielding a 50 % recognition score. For everyday sentence tests, the simple up-down adaptive procedure of (Plomp & Mimpen, 1979) is generally used to estimate the SRT on the basis of only one list of 10 to 20 sentences. In this procedure, sentence scoring is applied (all words or all key words within the sentence need to be correctly repeated), and a fixed 2-dB step size is employed. The SRT is simply calculated as the average signal-to-noise ratio (SNR) of the last N sentences in the list (with N depending on the used speech material). As matrix tests contain sentences with equal numbers of words, Hagerman & Kinnefors (1995) proposed a new adaptive procedure where the step size depends on the number of correctly understood words. When repeating correctly 0, 1, 2, 3, 4, or 5 words out of 5, the SNR of the next sentence is changed by +2, +1, 0, -1, -2, or -3 dB respectively, thus converging to a 40% correct score. This procedure was further improved by Brand & Kollmeier (2002), with a convergence to a 50% correct score and a reduction of the step size per reversal within a list. Furthermore, a maximum likelihood fit was applied to estimate the SRT. An important measure to evaluate a used measurement procedure and a speech material itself is the test-retest reliability (or within-subject variability). Reliability is typically calculated as the root mean square of the within-subject standard deviations of repeatedly measured adaptive SRTs. In this way, a reliability of 1 dB means that in 95 % of cases the adaptively measured SRT differs less than 2 dB (2 standard deviations) from a listener's 'true' SRT (Plomp & Mimpen, 1979). With their new word scoring procedure and with lists of 20 sentences of the German OLSA, Brand & Kollmeier (2002) showed that

the SRT estimate proved to be highly reliable, with an average within-subject standard deviation of less than 1 dB.

Considering the high reliability and the complementary applicability a matrix test can have for testing speech recognition in noise, the current study involves the development of the French Matrix test ('FrMatrix'). To this day, three standardized speech-in-noise tests have been available for Francophone listeners: (1) the Canadian French Hearing In Noise Test (Vaillancourt et al, 2005, 2008) and its adapted version for children (Laroche et al, 2006), (2) the French Intelligibility Sentence Test ('FIST') for European Francophone listeners (Luts et al, 2008), and (3) the French Digit Triplet test ('FrDigit3', Jansen et al, 2010). The first two tests are everyday sentence tests, with the largest difference being the vocabulary and pronunciation of the sentences. The FrDigit3 is an automatic self-screening test, with lists of 27 triplets of the digits 1 to 9 presented in a stationary speech-weighted noise. The recordings of both the FIST and the FrDigit3 were made with speakers from Paris.

In the first part of this paper, the *development of the FrMatrix* is described. In a first phase, the 50 words of the matrix were selected. Recordings were optimized by improving the homogeneity of the words with regard to their intelligibility by normal hearing (NH) listeners. The recordings and optimization were done in exactly the same way as for the development of the Danish test (Wagener et al, 2003), and is highly similar to the German (Wagener et al, 1999c, 1999a, 1999b) and the Polish test (Ozimek et al, 2010). The optimized FrMatrix test was then finally evaluated in a large group of NH listeners: the reference curve of the psychometric function was established, and the list equivalency, the extent of the training effect, and the reliability of two different adaptive procedures were studied.

The second part of this study concerns the *comparison of the FIST, the FrMatrix, and the FrDigit3* in a large group of NH and hearing impaired (HI) listeners across Belgium and France. The main aim of this second study is to determine to what extent these three different types of speech-in-noise tests yield similar SRTs in similar noises. There might be substantial differences with regard to linguistic and general cognitive demand, influencing the outcome of the speech-in-noise tests in different ways. For example, subjects might experience more difficulty repeating the words of the matrix sentences, which do not have any context, than repeating the more redundant sentences of the FIST. Table 1 summarizes the main differences between the three tests with regard to the characteristics of the speech material and of the test procedure generally used with that test. The comparison of the three tests in this study involves: (1) the effect of test site on the SRT, (2) the test-retest reliability, and (3) correlation and linear regression analyses for the three different tests and their discriminative power. In addition, results of the FrMatrix in both stationary and fluctuating background noise will be compared as well, since the latter is expected to differentiate better between listeners (Festen & Plomp, 1990; Wagener et al, 2006).

Table 1. Characteristics of the speech material and test procedure of the FIST, the FrMatrix, and the FrDigit3.

	FIST	FrMatrix	FrDigit3
Speech material:			
semantical context	✓		
representative vocabulary/ phonetic distribution	✓	✓	
unlimited answer possibilities	✓		
cognitive demanding	✓	✓	
speech rate (syllables/sec)	3.6	4.2	1.9
speaker	male	female	female
Test procedure:			
# trials in one list	10	20	27
scoring level	complete sentence	word	complete triplet
adaptive procedure	Plomp & Mimpen (1979)	Brand & Kollmeier (2002)	Plomp & Mimpen (1979)
SRT calculation	mean of last 6 SNRs	max. likelihood fit	mean of last 22 SNRs

Part 1: Development of the FrMatrix

Methods

Speech and noise material

Ten names, 10 verbs, 10 numerals, 10 objects, and 10 colors were chosen to build up the FrMatrix (Table 2). Within a column, the 10 words all have the same number of syllables. These 50 words were selected based on their phonetic content, to represent the mean phonetic distribution of the French spoken language (Wioland, 1972, Figure 1). The recording of the speech material was consistent with the procedure described by Wagener et al (2003). In short: 100 different sentences were recorded so that for each word 10 different versions were available, always with the coarticulation to another adjacent word. The sentences were uttered by a French female speaker from Paris. After equalizing the level of the 100 sentences (based on the average root-mean-square of a complete sentence, with silence parts included), they were cut into single words, preserving the coarticulation to the next word. Fifty lists of 10 sentences were now generated, by recombining the 500 single words into new sentences. Each of these lists contained all 50 words of the matrix, and was thus phonetically balanced. A final set of 28 lists was obtained after the exclusion of lists containing one or more unnaturally sounding sentences. The stationary long-term average speech spectrum (LTASS) noise was generated by superimposing the 280 sentences several times (Wagener et al, 2003).

Table 2. The closed set of 50 words of the FrMatrix.

	Name	Verb	Numeral	Object	Color
1	Agnès	achète	deux	anneaux	blancs
2	Charlotte	attrape	trois	ballons	bleus
3	Emile	demande	cinq	classeurs	bruns
4	Etienne	déplace	six	crayons	gris
5	Eugène	dessine	sept	jetons	jaunes
6	Félix	propose	huit	livres	mauves
7	Jean-Luc	ramasse	neuf	pions	noirs
8	Julien	ramène	onze	piquets	roses
9	Michel	reprend	douze	rubans	rouges
10	Sophie	voudrait	quinze	vélos	verts

Subjects

For the development of the FrMatrix, 57 NH listeners (17 males, 40 females) were tested. Each participant had pure-tone thresholds below or equal to 20 dB HL for the octave frequencies between 250 and 8000 Hz. For the optimization, 27 listeners participated. They had a mean age of 26 years, ranging between 20 and 54 years. The 30 participants of the evaluation measurements had a mean age of 22, with a range from 20 to 29 years old. They were all from Belgium and were native French-speaking.

Figure 1. Phonetic distribution of the FrMatrix (open triangles), compared to the mean phonetic distribution of spoken French (Wioland, 1972; filled diamonds).

Test setup

All measurements across all clinics were carried out using a PC running the Oldenburg Measurement Applications software (www.hörtech.de), a high quality 24-bits sound card (with a dynamic range of 144 dB), and Sennheiser HDA200 headphones. The speech and noise were always presented monaurally to the subject's best ear. The setup was calibrated with a B&K sound level meter 2260 and a B&K artificial ear 4153. The noise started 500 ms before and ended 500 ms after each

sentence, and was a randomly selected segment from the noise file. The subjects used a response box to answer (as in Table 2).

Optimization procedure

In order to reach the steepest possible slope at the SRT of the final reference psychometric curve, the speech material of the FrMatrix was further optimized according to the procedure described by Wagener et al (2003). First, all 280 sentences were presented to the NH listeners at a large range of fixed SNRs (-18 to +6 dB SNR, steps of 2 dB) at a noise level of 65 dB SPL. The SRT of each of the 500 single words was then determined by applying a logistic regression fit to the data of all listeners together. To improve the homogeneity of the words with regard to their intelligibility, the level of each word was now adjusted towards the mean SRT (e.g. a word with an SRT that is 2.3 dB better than the mean SRT was attenuated by 2.3 dB). To avoid unnatural loudness fluctuations within a sentence, the level of the individual words was adjusted by maximally 4 dB. The 280 sentences were then regenerated and recombined into the original 28 lists of 10 sentences.

Evaluation procedure

Two groups of subjects were tested to establish norms for NH listeners for the final speech material of the FrMatrix. In group 1, 20 participants were tested using the standard *word scoring* procedure. First, six double lists (i.e. combination of two original lists of 10 sentences into one list of 20 sentences) were presented using the adaptive procedure of Brand & Kollmeier (2002): This allowed the evaluation of training effects and test-retest reliability. Thereafter, the participants listened to all 28 lists at one of four fixed SNRs (7 lists per SNR). The presented SNRs were -8.0, -6.5, -5.0, and -3.5 dB, since pilot measurements showed that these would yield intelligibility scores below and above 50 % for normal hearing listeners. By applying logistic regression fits to the data of each subject, the reference psychometric curve for NH listeners was determined.

In the second group, 10 other subjects listened to six double lists with the simple up-down 2dB-step adaptive procedure (Plomp & Mimpen, 1979), in which *sentence scoring* was applied. The SRT was calculated by averaging the SNRs of the last 16 sentences, including the SNR of the imaginary 21st sentence. Again, training effects and test-retest reliability could be estimated this way. During all evaluation measurements, the level of the noise was fixed at 65 dB SPL.

Results

Based on the evaluation measurements with fixed SNRs (word scoring procedure only), the psychometric curve for each subject could be precisely determined for the final speech material. This yielded an average SRT of -6.0 dB SNR (with a standard deviation of 0.6 dB SNR across listeners), and an average slope at the SRT of 14.0 %/dB. Results per subject and the reference psychometric curve are shown in Figure 2 and Table 3. Before the optimization, the average subject-specific slope was only 7.8 %/dB, indicating that the homogenization of the individual words resulted in a significant improvement in the steepness of the psychometric function. List-specific SRTs were determined as well by pooling data of all subjects and applying a logistic regression fit per test list. The standard

deviation of the SRTs across all 28 lists was only 0.1 dB, with a maximal deviation of an individual list from the overall mean SRT of only 0.4 dB.

Figure 2. Reference psychometric function for NH listeners for the FrMatrix using word scoring. Squares represent subject scores per SNR.

The training effect of the FrMatrix was studied based on the adaptive evaluation measurements. Figure 3 shows the improvement in SRT with increasing number of presented lists, for both word scoring and sentence scoring procedure. As expected, the largest improvement was found from the first to the second list (1.2 dB for word scoring, and 1.0 dB for sentence scoring). Repeated-measures analyses of variance (ANOVA) proved that there was a significant effect of training for both word scoring [$F(5,95)=42.0$, $p<0.001$] and sentence scoring [$F(5,45)=4.9$, $p=0.002$]. Pairwise comparisons (with Bonferroni correction) for word scoring results demonstrated that the first list differed significantly from all the following lists ($p<0.001$). The second list also differed significantly from the 5th and 6th list ($p<0.05$). For the sentence scoring results, pairwise comparisons only showed a significant difference between the first and the sixth list ($p=0.016$).

The test-retest reliability, defined as the root mean square of the within-subject standard deviations of repeatedly measured adaptive SRTs, was determined for both procedures as well. When taking into account only the SRTs from the third to the sixth list to exclude training effects, a within-subject variability of 0.4 dB was found with the word scoring procedure. For the sentence scoring procedure, the variability increased to 1.1 dB.

Discussion

The evaluation of the FrMatrix in a large group of NH listeners shows that the newly developed test is accurate and reliable, especially when using the standard word scoring procedure. The reference psychometric function has a slope at SRT of 14.0 %/dB. This is somewhat shallower than that found for the OLSA (Wagener et al, 1999b) and the Polish Sentence Matrix Test (Ozimek et al, 2010), with both tests having a slope of 17.1 %/dB. However, it is steeper than the slope of Dantale II (12.6 %/dB, Wagener et al, 2003). Redundancy is lower in matrix tests than in everyday sentence tests, therefore a shallower slope is expected compared to the open-set FIST (20.2 %/dB, Luts et al, 2008).

Since the slope of the FrMatrix lies in between the above reported slopes for other matrix tests, it can thus be considered adequate.

Table 3. Speech recognition scores per SNR and estimated SRT and slope for 20 NH listeners for the FrMatrix using word scoring.

subject	Speech recognition score (%)				SRT (dB SNR)	Slope (%/dB)
	-8.0 dB	-6.5 dB	-5.0 dB	-3.5 dB		
1	28	45	66	83	-6.2	14.2
2	20	36	61	76	-5.6	14.5
3	26	58	66	82	-6.5	13.4
4	29	54	66	79	-6.5	12.1
5	22	50	70	85	-6.3	16.5
6	31	58	73	84	-6.8	14.0
7	20	33	60	79	-5.6	15.6
8	25	44	66	82	-6.1	14.5
9	24	44	74	78	-6.2	15.2
10	18	34	50	68	-5.0	12.2
11	34	53	73	86	-6.7	14.1
12	21	42	58	78	-5.7	13.7
13	37	58	75	79	-7.0	11.3
14	12	31	63	76	-5.5	18.2
15	16	28	51	71	-5.0	14.7
16	22	37	56	74	-5.5	13.0
17	25	45	65	78	-6.0	13.2
18	19	40	51	70	-5.2	11.7
19	26	51	74	77	-6.4	14.2
20	19	46	58	79	-5.8	14.3
min	12	28	50	68	-7.0	11.3
max	37	58	75	86	-5.0	18.2
average	24	44	64	78	-6.0	14.0
stdev	6	9	8	5	0.6	1.6

List-specific SRTs were highly homogeneous as well. The fitted SRTs of the 28 lists of 10 sentences were situated closely to the mean SRT, with a maximal discrepancy of only 0.4 dB. For everyday sentence tests, this maximal deviation is typically much higher: for the FIST (Luts et al, 2008), the LIST (Van Wieringen & Wouters, 2008), and the HINT (Nilsson et al, 1994) this was around 1dB. The high homogeneity in intelligibility of the test lists of the FrMatrix can be explained by the word content which is exactly the same in each list. Lists of OLSA (Wagener et al, 1999b) and the Polish Sentence Matrix Test (Ozimek et al, 2010) also showed high homogeneity, with maximal deviations of only 0.4 and 0.3 dB, respectively.

Results of the adaptive measurements show that the training effect for the FrMatrix, with an average improvement of 1.8 dB over six lists of 20 sentences, is comparable to other matrix tests. Wagener et al (1999b) found a training effect of about 2 dB SNR when presenting six lists of 20 sentences of OLSA. For Dantale II (Wagener et al, 2003) this effect was 2.2 dB for eight lists, and about 1.9 dB for the first six. For both tests and for the FrMatrix, the highest improvement was

always found from the first to the second list (± 1 dB), and from the second to the third (± 0.4 dB). We therefore recommend for future speech recognition measurements with the FrMatrix to start with a training phase of two double lists.

Figure 3. Training effect of the FrMatrix for two different procedures. For each measurement, one double list (20 sentences) was used.

When comparing the within-subject variability of the FrMatrix for two different scoring procedures, the word scoring procedure of Brand & Kollmeier (2002) proves to be the most reliable to determine the SRT. The within-subject variability was only 0.4 dB, whereas the simple up-down sentence-scoring procedure of Plomp & Mimpen (1979) resulted in a higher variability of 1.1 dB. This increase in variability can be explained by the decrease in sample size of the scored items ($n=100$ for word scoring versus $n=20$ for sentence scoring). Compared to the test-retest reliability found for the FIST (1.1 dB for lists of 10 sentences; Luts et al, 2008), the Plomp & Mimpen sentences (0.9 dB for lists of 13 sentences; Plomp & Mimpen, 1979), the VU-sentences (1.1 dB for lists of 13 sentences; Versfeld et al, 2000), the LIST sentences (1.2 dB for lists of 10 sentences; Van Wieringen & Wouters, 2008) and the FrDigit3 (0.7 dB for lists of 27 triplets; Jansen et al, 2010), the FrMatrix thus yields the most accurate SRTs of these speech recognition tests, when the word scoring procedure is applied. This is probably also a result of the very high list equivalency of the FrMatrix as described above.

Part 2: Comparison of the FIST, the FrMatrix, and the FrDigit3

Methods

Subjects

Two groups of subjects, 118 in total, participated in this study. The first group consisted of 49 NH listeners. Forty-four met the audiometric criterion as described above. The remaining five showed an elevated threshold (up to 35 dB HL) at one or two tested frequencies, and will further be referred to as sub-normal hearing (NH-) listeners. The second group comprised 69 mildly-to-moderately HI participants, both hearing aid users and non-users. They had a maximum pure-tone average at 500, 1000, 2000, and 4000 Hz ($PTA_{0.5,1,2,4}$) of 76 dB HL, and had no known middle-ear problems. The age

range of all 118 listeners varied widely from 18 up to 87 years, with an average of 51. The male/female ratio was 0.87. Participants were recruited in four different regions in Belgium and France: Brussels, Paris, Bordeaux, and Toulouse (see Table 4).

Table 4. Number of tested subjects and speech-noise combinations per test site. ('NH-' designates listeners with sub-normal hearing with elevated thresholds up to 35 dB HL at only one or two frequencies).

		Brussels	Paris	Bordeaux	Toulouse
N	All	31	20	29	38
	NH (NH-)	13 (0)	7 (1)	8 (2)	16 (2)
	HI	18	12	19	20
FIST	Stationary LTASS	✓	✓	✓	
FrMatrix	Stationary LTASS	✓	✓	✓	✓
FrDigit3	Stationary LTASS	✓	✓	✓	
FrMatrix	ICRA4-250	✓			✓

Test setup and procedure

Exactly the same test setup was used as described for the development of the FrMatrix (see above). All setups were controlled and calibrated by the first author, on-site, to guarantee comparability.

For the FIST (Luts et al, 2008), lists of 10 everyday sentences were presented with a simple up-down adaptive procedure. Sentence scoring and a 2-dB step size were applied, and the first sentence was repeated with +2dB-steps until this sentence was understood correctly. The SRT was calculated by averaging the SNRs of the last 6 sentences, including the SNR of the imaginary 11th sentence. The FrMatrix on the other hand, was presented in lists of 20 sentences. Since each word was scored separately, the adaptive procedure and SRT estimation of Brand & Kollmeier (2002) was used. For the FrDigit3 (Jansen et al, 2010), lists of 27 triplets were used. Again the simple up-down 2dB-step adaptive procedure was used, but without repetition of the first triplet. If one or more digits in a triplet were misunderstood, the complete triplet was scored as incorrect. The SRT was calculated by averaging the SNRs of the last 22 triplets, including the SNR of the imaginary 28th triplet.

The three tests were each presented in their own stationary LTASS noise. For the FrMatrix also the fluctuating ICRA4-250 noise was used. This noise has the temporal envelope of a single talker, but is unintelligible. Dreschler et al (2001) created the ICRA4 by sending an English speech sample of a single talker through a three-channel filter bank to remove the temporal fine structure, and filtering it with a gender filter to resemble the overall spectral shape of female speech. Later on, silence gaps within this signal were limited to 250 ms. The spectrum of the noise is thus not identical to the LTASS of the FrMatrix.

At the start of a test session, pure-tone audiometry was conducted. After that, two lists of the FrMatrix were presented for training. Thereafter a test and retest of the four speech-noise combinations took place. Not all four combinations were carried out at each of the four test sites. The division of the speech-noise combinations and the number of subjects per site are given in Table 4. The order in which the tests were presented was randomized (separately for test and retest). The

lists were also distributed at random over the subjects, taking into account that a specific list was only presented once to a subject.

During these adaptive tests, the level of the speech was kept at a constant subject-dependent level. For the NH listeners, the speech level was fixed at 60 dB SPL. This level was raised for each HI participant individually according to his/her pure-tone thresholds, to (partially) compensate for audibility. Since the highest energy of the three speech materials lies within the 250-1000 Hz range, the overall gain was based on their $PTA_{0.25,0.5,1}$. Only half of this $PTA_{0.25,0.5,1}$ was added to the 60 dB SPL, in order to avoid an uncomfortable loudness due to recruitment. With a maximal $PTA_{0.25,0.5,1}$ of 52 dB HL within the participant group, the highest speech presentation level was thus 86 dB SPL.

For all three types of tests, subjects just had to repeat what they had heard. However, to maintain the closed-set character of the FrMatrix and the FrDigit3, the subjects received a sheet with the possible words.

Results

For all 118 tested subjects of the multi-center study, the following data were available: group (NH, NH-, or HI), test site, audiometric thresholds at octave frequencies from 250 to 8000 Hz, and test and retest SRTs for the four speech-noise combinations. The results of the two training lists of the FrMatrix were not taken into consideration. Since not all four speech-noise combinations were carried out at all sites (see Table 4), only results from Brussels, Paris, and Bordeaux (N=80) are available for the analyses in which the three tests in stationary noise are compared. Similarly, for the comparison between the FrMatrix in stationary noise versus fluctuating ICRA4-250 noise, only results from Brussels and Toulouse are considered (N=69).

Effect of test site

In a first analysis, results from the NH subjects of the four different test sites were compared to examine a possible region effect. Five subjects, tested in Toulouse, were excluded since they originally came from another region in France. Figure 4 shows the average SRTs per test site and speech-noise combination. In general, differences between test sites were small (< 1 dB). For each of the four speech-noise combinations separately, a univariate ANOVA did not reveal a significant difference across test sites. P-values were 0.085, 0.117, 0.575, and 0.852 for FIST, FrMatrix in stationary LTASS noise, FrDigit3, and FrMatrix in ICRA4-250 noise respectively.

The same analyses were also executed on the complete group of subjects. Since the four test sites were not matched for $PTA_{0.5,1,2,4}$ and audiometric slope, these two factors were added as covariates. The test site effect was not significant in this large population as well ($p=0.443, 0.589, 0.221$, and 0.864 for FIST, FrMatrix in stationary LTASS noise, FrDigit3, and FrMatrix in ICRA4-250 noise respectively).

Figure 4. Average SRT per test site and per speech-noise combination for the NH subjects. Error bars represent the standard deviation on the average.

Test-retest reliability

The test-retest reliability of the different speech-noise combinations is given in Table 5, for all listeners, for NH/NH- listeners, and for HI listeners. In general, reliability was worse (higher values) for the HI group, especially for the FrMatrix and the FrDigit3 (0.4 versus 1.1 dB for NH versus HI listeners respectively). The FIST gave also less reliable SRTs than the FrMatrix and the FrDigit3 (1.3 versus 0.9 dB). Nevertheless, if we account for the number of trials within a list (which was done by dividing the reliability by $\sqrt{20/10}$ for the FIST and by $\sqrt{20/27}$ for the FrDigit3), reliabilities are much more similar for the three tests in stationary LTASS noise. This implies that similar reliability can be obtained if the FIST uses double lists of 20 sentences. When using a fluctuating noise, the reliability of the FrMatrix deteriorated from 0.5 to 1.4 dB and from 1.1 to 1.9 dB, for NH/NH- and HI listeners, respectively.

Table 5. Test-retest reliability (dB) of the FIST, the FrMatrix, and the FrDigit3 in stationary LTASS noise (results from Belgium, Paris, and Bordeaux) and of the FrMatrix in stationary LTASS and ICRA4-250 noise (results from Belgium and Toulouse).

		Stationary LTASS				FrMatrix		
		N	FIST	FrMatrix	FrDigit3	N	Stationary	ICRA4-250
standard	All	80	1.3	0.9	0.9	69	0.9	1.7
	NH/NH-	31	1.2	0.4	0.4	31	0.5	1.4
	HI	49	1.3	1.1	1.1	38	1.1	1.9
corrected to 20 trials	All	80	0.9	0.9	1.0			
	NH/NH-	31	0.9	0.4	0.5			
	HI	49	0.9	1.1	1.2			

Correlation and regression

In order to obtain a better insight on how the four speech-noise combinations correspond to each other, correlation coefficients were calculated between the SRTs of the different tests and the

PTA_{0.5,1,2,4}. Analyses were executed both on the whole group of subjects and for HI listeners only (see Table 6).

First, test and retest scores were correlated for each test, with the lowest correlation found for FIST (0.78) and the highest correlation for FrMatrix both in stationary and in fluctuating noise (>0.90). For the next analyses, test and retest scores were averaged.

Table 6. Pearson correlations for the FIST, the FrMatrix, and the FrDigit3 in stationary LTASS noise (results from Belgium, Paris, and Bordeaux) and for the FrMatrix in stationary LTASS and ICRA4-250 noise (results from Belgium and Toulouse). All correlations are significantly different from 0 ($p < 0.01$).

	Stationary LTASS			FrMatrix	
	N=80			N=69	
All listeners:	FIST	FrMatrix	FrDigit3	Stationary	ICRA4-250
Test-retest	0.78	0.96	0.94	0.96	0.96
FIST		0.83	0.81		
FrMatrix stationary			0.89		0.94
PTA _{0.5,1,2,4}	0.69	0.82	0.84	0.84	0.93
HI only:	N=49			N=38	
	FIST	FrMatrix	FrDigit3	Stationary	ICRA4-250
Test-retest	0.78	0.93	0.88	0.92	0.90
FIST		0.79	0.74		
FrMatrix stationary			0.81		0.89
PTA _{0.5,1,2,4}	0.53	0.69	0.56	0.61	0.76

The open-set everyday sentence test *FIST*, which can be considered as the standard type of speech-in-noise test, showed a rather low correlation with the PTA_{0.5,1,2,4} (0.69 for all listeners and 0.53 for the HI group). Figure 5 also shows this low correspondence: listeners with similar PTAs (e.g. around 40 dB HL) had widely different SRTs, going from -8 to +3 dB SNR. This low correlation was not unexpected, since pure-tone audiometry and speech-in-noise tests measure a different aspect of hearing (Glasberg & Moore, 1989). Therefore, further analyses only include comparisons across SRTs in noise.

The *FrMatrix*, developed as an alternative speech-in-noise test for multiple testing within the same listener, correlated well with the FIST (0.83 for all listeners and 0.79 for the HI group). This correlation between the FIST and the FrMatrix was significantly higher than the correlation between the FIST and the PTA_{0.5,1,2,4} (Steiger's $Z=3.49$, $p < 0.001$ for all listeners and Steiger's $Z=3.19$, $p=0.001$ for the HI group). The scatter plot of the FrMatrix versus the FIST is given in Figure 6. Besides the high correlation, this figure also shows a steep slope of the linear regression (1.4, $SE=0.1$). Small differences in SRT for the FIST thus correspond with larger differences for the FrMatrix. Given the equal measurement precision (cf. test-retest reliability) for the two tests when using the same number of trials, a slope above 1 indicates that the FrMatrix had a higher discriminative power than the FIST. This discriminative power was also determined for the FrMatrix in ICRA4-250 noise. Since a direct comparison to the FIST is only possible in a small group of listeners (13 NH and 18 HI listeners), it was only compared to results of the FrMatrix in stationary noise. As expected, the correlation

between the two types of noise was very high (0.94). And, as shown in Figure 7, the slope of the linear regression line was very steep (2.2, SE=0.1). Even taking into account the poorer reliability of the FrMatrix in ICRA4-250 (1.7 dB), which was a factor 1.9 greater than the reliability in stationary noise (0.9 dB), the discriminative power in ICRA4-250 remained higher than in stationary noise.

Figure 5. Scatter plot of the FIST results versus the PTA_{0.5,1,2,4} for all subjects from Brussels, Paris, and Bordeaux (N=80). Open and filled circles represent NH/NH- and HI listeners respectively.

The third type of speech-in-noise test, the *FrDigit3*, deviates from the other two types mainly in its limited and unrepresentative vocabulary and phonetic distribution. Still, the results of this screening test correlated highly with the two sentence tests. The correlation between the *FrDigit3* and the FIST (0.81 for all listeners and 0.74 for the HI group) was high and significantly stronger than the correlation between the FIST and the PTA_{0.5,1,2,4} (Steiger's $Z=3.04$, $p=0.002$ for all listeners and Steiger's $Z=2.15$, $p=0.031$ for the HI group). The correlation between the *FrDigit3* and the FrMatrix (0.89 for all listeners and 0.81 for the HI group) tended to be even higher (Steiger's $Z=2.82$, $p=0.005$ for all listeners and Steiger's $Z=1.30$, $p=0.193$ for the HI group). The slope of the linear regression of the *FrDigit3* versus the FIST was 1.1 (SE=0.1) and the *FrDigit3* versus the FrMatrix was 0.8 (SE<0.1). With a reliability of 0.9 dB when using 27 triplets in a test, the discriminative power of the *FrDigit3* lies in between that of the FIST and the FrMatrix.

Discussion

The results of 49 normal hearing participants and 69 mildly-to-moderately hearing impaired listeners, tested at four test sites across Belgium and France, made it possible to thoroughly compare three types of speech-in-noise tests. First, the effect of test site was studied. Since the recordings of all three speech-in-noise tests were made with speakers from Paris, it was presumed that the listeners from Belgium (Brussels) or from the south of France (Bordeaux and Toulouse) might have slightly higher (worse) SRTs than listeners from the region of Paris, due to the somewhat

less familiar accent (Van Wijngaarden et al, 2002). Furthermore, the everyday sentences of the FIST possibly contain a more typically French vocabulary, which might deteriorate further the recognition performance of the Belgian subjects. Our results show indeed a tendency for this region effect, but only for the FIST scores. However, the differences are small (< 1 dB) and not significant. All three tests are thus employable both in Belgium and in all regions in France, and the same normative values can be used as a reference.

Figure 6. Scatter plot of the FrMatrix versus the FIST for all subjects from Brussels, Paris, and Bordeaux (N=80). Open and filled circles represent NH/NH- and HI listeners respectively. The linear regression line, its slope and the correlation coefficient is given for all listeners.

The test-retest reliability for the group of NH/NH- listeners corresponds perfectly to the values found during the development of each test. Luts et al (2008) reported a reliability of 1.0 and 1.1 dB for the FIST for French and Belgian NH listeners respectively, where we found a reliability of 1.2 dB. The FrMatrix reached a reliability of 0.4 dB, which is exactly the same as what has been found during development (described above). Also for the FrDigit3, similar reliabilities were found: 0.4 dB for the current study and 0.6 dB reported by Jansen et al (2010). It should be noted that the reliability of the FrMatrix was determined for measurements which were always preceded by two training lists. Since the extent or lack of training might affect reliability, the reported value is thus only valid when this condition has been met. The poorer reliability of the FIST compared to the FrMatrix and the FrDigit3 disappears completely when correcting for the number of trials within a test list. The use of double

lists would thus be worthwhile for future application of the FIST. All other differences in the characteristics of the speech material and the used test procedure for the three tests, as summarized in Table 1, thus do not seem to affect the test-retest reliability. In spite of the general increase in within-subject variability in the group of HI listeners compared to the NH listeners, a reliability of up to 1.2 dB (FrDigit3) can still be considered to be sufficient.

Figure 7. Scatter plot of the FrMatrix in ICRA4-250 versus the FrMatrix in stationary noise for all subjects from Brussels and Toulouse (N=69). Open and filled circles represent NH/NH- and HI listeners respectively. The linear regression line, its slope and the correlation coefficient is given for all listeners.

The deterioration in test-retest reliability when presenting the FrMatrix in the fluctuating ICRA4-250 noise instead of the stationary LTASS noise (from 0.9 to 1.7 dB), was not unexpected. (Wagener et al, 2006) presented the OLSA to 10 HI listeners using a stationary speech-weighted noise (ICRA1) and the highly fluctuating ICRA5-250 noise, and found intra-individual standard deviations of 0.8 and 1.5 dB respectively. Also for three different everyday sentence tests, Francart et al, (2011) showed an increase of around 0.8 dB in mean absolute test-retest difference for ICRA5-250 compared to the stationary LTASS noise. Nevertheless, the increase in test-retest variability for fluctuating noise was more than compensated by the increase in between-subjects variability (the slope of the linear

regression was 2.2), resulting in an increased discriminative power compared to that for stationary noise.

SRTs in noise are known to be related more to supra-threshold discrimination abilities rather than absolute detection thresholds (Glasberg & Moore, 1989; Noordhoek et al, 2001). Glasberg & Moore (1989) reported correlation coefficients of 0.62 (both normal and impaired ears) and 0.56 (impaired ears only) between the SRT in 75 dB SPL noise and the $PTA_{0.5,1,2}$. In the study of Smoorenburg (1992) a correlation of 0.67 was found between the SRT in noise and the $PTA_{0.5,1,2,4}$ for noise-exposed workers both with and without hearing loss. The used speech materials in these studies were everyday sentences, presented in an open-set paradigm. A correlation of 0.69 (all listeners) and 0.53 (HI only) between the FIST and the $PTA_{0.5,1,2,4}$ is thus in line with the previous studies, endorsing that a speech-in-noise test measures a partly different aspect of hearing than pure-tone audiometry does. The correlation coefficients of the FrMatrix (0.82) and the FrDigit3 (0.84) with the $PTA_{0.5,1,2,4}$ were rather high. However, these results are probably biased by the contrast between the NH and HI subjects. Analyses based on the HI listeners only, yielded lower correlations between the SRT and the $PTA_{0.5,1,2,4}$ for FrMatrix (0.69) and FrDigit3 (0.56). These were more in line with the above results.

More important than the relation between the SRT and the PTA, was – in this study – the correlation between the different types of speech-in-noise tests. Since all three tests aim to measure the same performance of a listener, these correlations should be very high. Taking into account the measurement error in psychophysical tests, both visible in the within-subject variability values (Table 5) and in the test-retest correlation coefficients (Table 6), correlations of 0.83 (FrMatrix vs. FIST), 0.81 (FrDigit3 vs. FIST), and 0.89 (FrDigit3 vs. FrMatrix) were considered adequate.

The relatively high discriminative power found for the FrMatrix, both in stationary and even more in ICRA4-250 noise, is advantageous for identifying subtle performance differences. When studying the potential benefit of e.g. a new speech enhancement algorithm, using a test with a higher discriminative power might imply that a smaller number of repeated measurements per condition, or even a smaller number of subjects, would be sufficient to find significant improvements.

Since currently there is no specific framework for the application of the FrDigit3 as a screening test, determination of sensitivity and specificity was beyond the scope of this study: both the choice for the reference test and for the cutoff point between a ‘positive’ and ‘negative’ outcome would have been arbitrary. However, notwithstanding its limited vocabulary (only the digits 1 to 9), results for the FrDigit3 were in general not inferior to the results for the two sentence tests: the differences between the test sites were smallest for the FrDigit3 (only 0.2 dB), the test-retest reliability was as high as for the FrMatrix (0.4 dB and 1.1 dB for NH and HI listeners respectively, when using 27 triplets), and the correlation with the FIST (0.81) and the FrMatrix (0.89) was strong. Therefore, this easy and accessible test is expected to have a high sensitivity and specificity when it will be used for screening of large populations.

Conclusions

The results of the current study show that the FIST, the FrMatrix, and the FrDigit3 provide similar results and reliably evaluate speech recognition performance in noise both in normal hearing and hearing impaired listeners.

The newly developed FrMatrix showed a steep slope (14.0 %/dB) of the reference psychometric curve. The list equivalency of the 28 lists of 10 sentences was also very high. When measuring the SRT adaptively with the procedure of Brand & Kollmeier (2002), using lists of 20 sentences and applying a word scoring, a very high reliability (0.4 dB) was obtained for NH subjects. Since SRTs of listeners typically improve significantly during the first lists, it is required to start each test session with a short training phase of two lists.

The comparison of the FIST, the FrMatrix, and the FrDigit3 showed in general a high correspondence in test outcome. For all three tests, the differences across the participating study centers were very small and not significant. The measured reliability values were all below 1 dB for the NH listeners and around 1 dB for the HI listeners, when using double lists for the FIST and applying a short training phase for the FrMatrix. Strong correlations were found between the SRTs of the different tests: 0.83 (FrMatrix vs. FIST), 0.81 (FrDigit3 vs. FIST), and 0.89 (FrDigit3 vs. FrMatrix). The FrMatrix had the highest discriminative power, both in stationary LTASS and even more in fluctuating ICRA4-250 noise.

Acknowledgements

The authors would like to thank Marian Saels, Sarah Vinck, and Kyra Visser for their help in gathering the optimization and evaluation results of the FrMatrix. Caroline Diercxsens and Alice Lamy are gratefully acknowledged for administering the measurements in Brussels and Toulouse, and we thank Patrick Verheyden, Evelien Bienstman, Dominique Dehaze, and Olivier Vales for recruiting the subjects. Finally, we wish to thank Daniel Berg and Melanie Zokoll for their technical and methodological support. Part of this work was supported by a grant from the European Union FP6, Project No. 004171 HEARCOM.

References

- Brand T. & Kollmeier B. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J Acoust Soc Am*, 111, 2801–10.
- Dreschler W.A., Verschuure H., Ludvigsen C. & Westermann S. 2001. ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment. *Audiology*, 40, 148–157.
- Festen J.M. & Plomp R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J Acoust Soc Am*, 88, 1725–36.
- Francart T., Van Wieringen A. & Wouters J. 2011. Comparison of fluctuating maskers for speech recognition tests. *Int J Audiol*, 50, 2–13.
- Glasberg B.R. & Moore B.C.J. 1989. Psychoacoustic abilities of subjects with unilateral and bilateral cochlear hearing impairments and their relationship to the ability to understand speech. *Scand Audiol*, 32, 1–25.
- Hagerman B. 1982. Sentences for testing speech intelligibility in noise. *Scand Audiol*, 11, 79–87.
- Hagerman B. & Kinnefors C. 1995. Efficient adaptive methods for measuring speech reception threshold in quiet and in noise. *Scand Audiol*, 24, 71–77.
- Hernvig L.H. & Olsen S.O. 2005. Learning effect when using the Danish Hagerman sentences (Dantale II) to determine speech reception threshold. *Int J Audiol*, 44, 509–512.
- Jansen S., Luts H., Wagener K.C., Frachet B. & Wouters J. 2010. The French digit triplet test: a hearing screening tool for speech intelligibility in noise. *Int J Audiol*, 49, 378–87.
- Kalikow D.N., Stevens K.N. & Elliott L.L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J Acoust Soc Am*, 61, 1337–51.
- Laroche C., Vaillancourt V., Melanson C., Renault M.-E., Thériault C., et al 2006. Adaptation du HINT (Hearing in Noise Test) pour les enfants francophones canadiens et données préliminaires sur l'effet d'âge. *Revue d'orthophonie et d'audiologie*, 30, 95–109.
- Luts H., Boon E., Wable J. & Wouters J. 2008. FIST: A French sentence test for speech intelligibility in noise. *Int J Audiol*, 47, 373–374.
- Nilsson M., Soli S.D. & Sullivan J.A. 1994. Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am*, 95, 1085–1099.
- Noordhoek I.M., Houtgast T. & Festen J.M. 2001. Relations between intelligibility of narrow-band speech and auditory functions, both in the 1-kHz frequency region. *J Acoust Soc Am*, 109, 1197–212.
- Ozimek E., Warzybok A. & Kutzner D. 2010. Polish sentence matrix test for speech intelligibility measurement in noise. *Int J Audiol*, 49, 444–454.

- Plomp R. & Mimpen A.M. 1979. Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18, 43–52.
- Smootenburg G.F. 1992. Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *J Acoust Soc Am*, 91, 421–37.
- Theunissen M., Swanepoel D.W. & Hanekom J. 2009. Sentence recognition in noise: Variables in compilation and interpretation of tests. *Int J Audiol*, 48, 743–757.
- Vaillancourt V., Laroche C., Mayer C., Basque C., Nali M., et al 2005. Adaptation of the HINT (Hearing in Noise Test) for adult Canadian Francophone populations. *Int J Audiol*, 44, 358–369.
- Vaillancourt V., Laroche C., Mayer C., Basque C., Nali M., et al 2008. The Canadian French Hearing in Noise Test. *Int J Audiol*, 47, 383–385.
- Van Wieringen A. & Wouters J. 2008. LIST and LINT: sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. *Int J Audiol*, 47, 348–55.
- Van Wijngaarden S.J., Steeneken H.J.M. & Houtgast T. 2002. Quantifying the intelligibility of speech in noise for non-native talkers. *J Acoust Soc Am*, 112, 3004–13.
- Versfeld N.J., Daalder L., Festen J.M. & Houtgast T. 2000. Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J Acoust Soc Am*, 107, 1671–84.
- Wagener K.C., Brand T. & Kollmeier B. 1999a. Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil II: Optimierung des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38, 44–56.
- Wagener K.C., Brand T. & Kollmeier B. 1999b. Entwicklung und Evaluation eines Satztests für die deutsche Sprache Teil III: Evaluation des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38, 86–95.
- Wagener K.C., Brand T. & Kollmeier B. 2006. The role of silent intervals for sentence intelligibility in fluctuating noise in hearing-impaired listeners. *Int J Audiol*, 45, 26–33.
- Wagener K.C., Jøsvassen J.L. & Ardenkjar R. 2003. Design, optimization and evaluation of a Danish sentence test in noise. *Int J Audiol*, 42, 10–17.
- Wagener K.C., Kühnel V. & Kollmeier B. 1999c. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38, 4–15.
- Wioland F. 1972. Estimation de la fréquence des phonèmes en français parlé. *Travaux de l'Institut de Phonétique de l'Université de Strasbourg*, 4, 177–204.