informa
healthcare

International Journal of Audiology 2014; Early Online: 1–4

International Journal of Audiology

Original Article

Development of a Dutch matrix sentence test to assess speech intelligibility in noise

Rolph Houben*, Jan Koopman[†,‡], Heleen Luts[#], Kirsten C. Wagener[$], Astrid van Wieringen[#], Hans Verschuure[†] & Wouter A. Dreschler*

*Clinical and Experimental Audiology, Academic Medical Center Amsterdam, The Netherlands [†]Erasmus Medical Center, Rotterdam, The Netherlands [‡]Royal Visio, Institute for Visually Impaired and Blind People, Amsterdam, The Netherlands [#]ExpORL, Department of Neurosciences, Katholieke Universiteit Leuven, Belgium, and [$]Hörzentrum Oldenburg GmbH, Oldenburg, Germany

Abstract

Objective: A Dutch matrix sentence test was developed and evaluated. A matrix test is a speech-in-noise test based on a closed speech corpus of sentences derived from words from fixed categories. An example is 'Mark gives five large flowers.' *Design:* This report consists of the development of the speech test and a multi-center evaluation. *Study sample:* Forty-five normal-hearing participants. *Results:* The developed matrix test has a speech reception threshold in stationary noise of − 8.4 dB with an inter-list standard deviation of 0.2 dB. The slope of the intelligibility function is 10.2 %/dB and this is slightly lower than that of similar tests in other languages (12.6 to 17.1 %/dB). *Conclusions:* The matrix test is now also available in Dutch and can be used in both Flanders and the Netherlands.

Key Words: Speech-in-noise; speech test; speech intelligibility; normative data

A matrix test is a sentence-in-noise test that uses sentences of identical grammatical structure. The words, taken from a closed set of alternatives, are combined to form a complete sentence. Each sentence is grammatically and semantically correct and with no redundancy. Because the sentences are constructed from a fixed matrix of words, we refer to this type of test as 'matrix test.' The matrix test was originally developed by Hagerman (1982) for Swedish and is now available in German (Wagener et al, 1999a), Danish (Wagener et al, 2003), British English (Hewitt, 2008), Polish (Ozimek et al, 2010), French (Jansen et al, 2012), Spanish (Hochmuth et al, 2012), American English (Zokoll et al, 2012), Turkish (Zokoll et al, 2013), and Russian (Zokoll et al, 2013). Up to now, there has been no matrix test available in Dutch.

The primary goal of the present study was therefore to develop a Dutch matrix test and to obtain normative data for normal-hearing listeners. The first section of this paper describes the development of the test materials. The second section deals with a multi-center evaluation of these new materials.

Test development

The development of the Dutch matrix test consisted of the following three steps: (1) composition of a base matrix, (2) recording of the speech materials, and (3) homogenization of the materials by equalizing word intelligibility with level adjustments.

The base matrix

The design of the sentence matrix was based on the Swedish matrix test (Hagerman, 1982). Each sentence has the same fixed grammatical structure: 'name, verb, numeral, adjective, object.' In the European Hearcom project (Vlaming et al, 2011) matrix tests have been

Correspondence: Rolph Houben, Clinical and Experimental Audiology, Academic Medical Center, Meibergdreef 9, 1105 AZ, Amsterdam, The Netherlands. E-mail: A.C.Houben@amc.uva.nl

(Received 16 January 2013; accepted 26 April 2014)

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.3109/14992027.2014.920111

Abbreviations

ND	Northern Dutch
SNR	Signal-to-noise ratio
SD	Southern Dutch
SRT	Speech reception threshold

Table 1. The matrix of the Dutch matrix test. Bold words indicate the sentence 'Mark gives five large flowers'.

Name	*Verb*	*Numeral*	*Adjective*	*Object*
Anneke	**geeft**	twee	dure	**bloemen**
Christien	had	drie	goede	boeken
Heleen	kiest	vier	groene	boten
Jan	koopt	**vijf**	**grote**	dozen
Mark	maakte	zes	kleine	fietsen
Monique	tekent	acht	mooie	messen
Pieter	telde	negen	nieuwe	munten
Sarah	vond	tien	oranje	ringen
Tom	vroeg	twaalf	vuile	schoenen
Willem	wint	achttien	zware	stenen

used in several languages. The respective base matrices, including that for the Dutch version, were described in the Hearcom report by Dreschler et al (2006). For each of the five word categories we selected ten alternative words, leading to a total of 50 unique words (Table 1). To obtain relevant results from our sentence-in-noise test, we ensured that the occurrence of phonemes in our matrix mirrored that of standard Dutch. Figure 1 shows the phoneme occurrence in our matrix as well as the phoneme occurrence in the reference corpus by Luyckx et al (2007). The occurrence of the phonemes in our matrix was close to that of the reference data. The average absolute difference in occurrence between the base matrix and the average of Northern Dutch (ND) and Southern Dutch (SD) from Luyckx et al was 1.1 percentage points.

Recordings

The sentences were spoken by a Dutch female (24 years) who originated from the border region between the area where ND and SD is spoken. To obtain a naturally sounding sentence test we took into account the co-articulation between words (Wagener et al, 1999b). The speech materials were recorded in the recording studio of the University of Oldenburg (Wagener et al, 2003) with a near-field microphone (AKG C-1000S) on digital tape (AIWA HHB1 Pro DAT recorder). The sampling frequency was 44.1 kHz and the resolution was 16 bit. The 100 recorded sentences were cut into sections. The sections were concatenated to form 360 unique sentences with the correct co-articulation between the words.

These sentences were subsequently checked to verify if the words were spoken correctly and with a clear articulation, and if there were any artifacts from the recording and the cutting and recombination of the sentences. Thirteen sentences were discarded and a total of 347 naturally sounding sentences remained. A masking noise was created by superimposing sentences on top of each other (Wagener et al, 2003). First, 100 sentences were concatenated. Then, 30 versions of this sound file were superimposed onto each other while each superimposition was delayed with a random time between 5 ms and 2.5 s. This resulted in a stationary noise with an average power spectrum equal to that of the sentences.

Homogenization of the materials

The outcome measure of our sentence-in-noise test was the speech reception threshold (SRT): the signal-to-noise ratio (SNR) where 50% of the words were correctly understood. For a reliable measurement of the SRT a steep intelligibility function is required (Kollmeier & Wesselkamp, 1997). The intelligibility function is steepest if the intelligibility of the individual words of the sentence is the same (Kollmeier, 1990). To equalize the intelligibility of each word we first measured the intelligibility for the words at SNRs of -12, -9, -6, -3, and 0 dB. Ten normal-hearing participants (native ND, all having hearing thresholds ≤ 20 dB HL for the octave frequencies between 0.5 and 4 kHz, median age 24 years with a range of 19 years to 26 years) listened to stimuli presented monaurally via

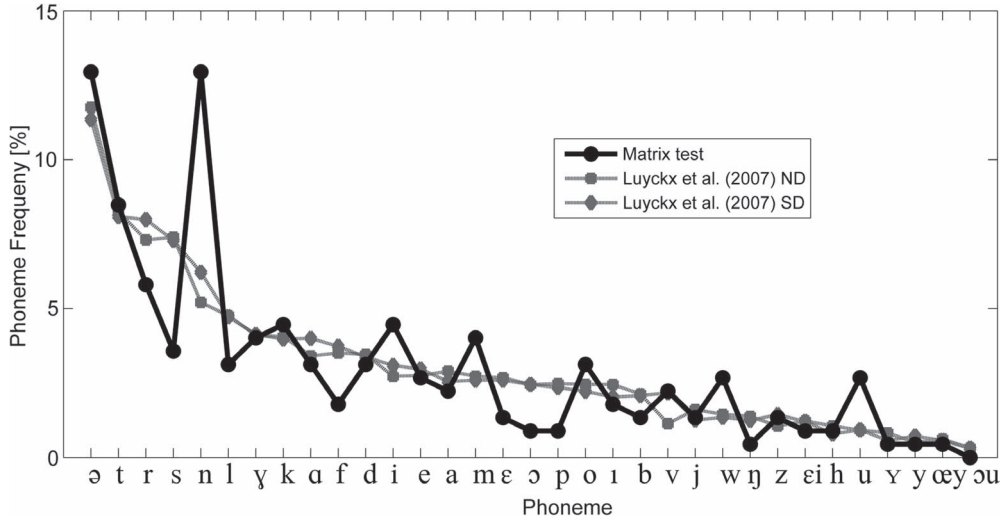

Figure 1. Phoneme distribution for the base matrix and for the reference corpus for Dutch from Luyckx et al (2007). Phonemes are ordered according to the average phoneme occurrence from the reference distribution.

TDH-39P headphones and were asked to indicate which sentence they had heard by using a response box displaying the matrix of 50 words (see Table 1).

To equalize word intelligibility, the individual words of the sentences were amplified or attenuated individually based on their SRT. The maximum correction was limited to ± 3 dB to avoid unnatural intensity jumps between words that may influence the prosody of the sentences. Even with this limitation the amplification/attenuation led to some unnaturally sounding sentences and these were discarded. Of the 311 remaining sentences we constructed 14 lists, each containing 20 sentences. Each of these 14 lists contained all words from the base matrix exactly twice. Together, the 14 lists contained 198 unique sentences and some of these unique sentences were present in more than one list.

Evaluation of the speech materials in normal-hearing listeners

Methods

Experimental setup

The materials were evaluated through listening tests in three centers. One center was located in Flanders in Belgium (SD): 'LEU' (ExpORL, Department of Neurosciences, KU Leuven, Belgium). Two centers were located in the Netherlands (ND): 'ROT' (Erasmus Medical Center, KNO-Audiologie, Rotterdam) and 'AMS' (Academic Medical Center, University of Amsterdam, Amsterdam). All testing was done in double-walled soundproof booths. LEU used an RME Multiface sound card with Sennheiser HDA200 headphones; ROT used an Echo Gina 24 sound card with Telephonics TDH39P headphones; and AMS used an RME Fireface 800 sound card with Sennheiser HDA200 headphones. Testing was done with the Oldenburg measurement applications software package (OMA) developed by Hörtech, Oldenburg.

Subjects and measurement procedure

Each center recruited 15 local normal-hearing adults. The participants reported no otological problems and their hearing thresholds did not exceed 20 dB HL at each octave frequency between 0.25 and 8 kHz. The median age of participants was 26 years (range: 20 to 42 years), 22 years (range 19 to 25 years), and 24 years (range: 19 to 44 years) for LEU (SD), ROT (ND), and AMS (ND), respectively. All stimuli were presented to the participant's best ear. The participants used a response box on the computer screen to select the words that they heard. Participants had to choose one of the alternatives and this represented a chance level of 10%. All participants started with two practice lists and these data were discarded in further analyses. The subsequent measurements were done at fixed SNRs of − 5 dB, − 7 dB, and − 9 dB with the noise level at 70 dB SPL. For each test center, the test lists were balanced over the SNRs so that while each subject listened to every list, a specific list was tested at one SNR only. In short, each subject did not hear a specific sentence more than once.

Results

Normative data

For each sentence the average percentage correct score was calculated for each center (see Figure 2).

To check whether there were differences in SRT and slope between the centers, we applied a logistic regression model to the data of each center. This logistic model describes the intelligibility function because it models the intelligibility as a function of SNR. From this logistic model we calculated the SRT and the slope. The logistic regression model[1] took into account the chance level of 10% that originates from the use of a closed test corpus with ten alternatives for each word. The results are shown in Table 2 and plotted in Figure 2. Both the SRT and the slope did not differ significantly between the centers (ANOVA, for the SRT $F(2,42) = 0.04$, $p = 1$; and for the slope $F(2,42) = 0.9$, $p = 0.4$).

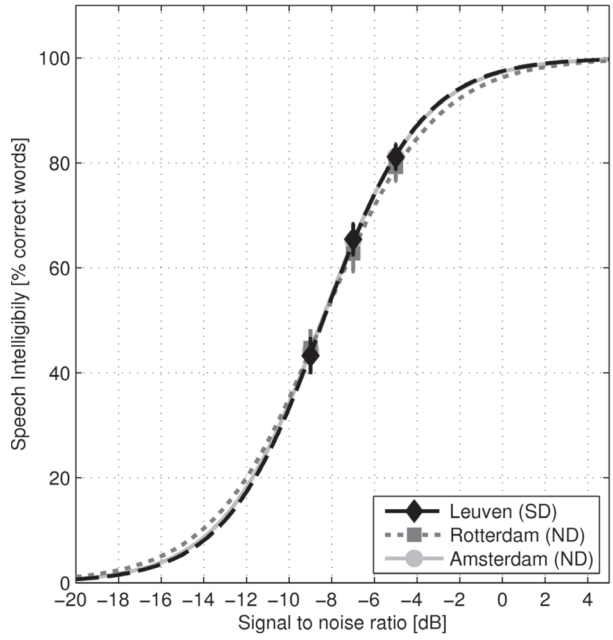

Figure 2. Speech intelligibility functions per center. For each center, the markers show the average data across all subjects (n = 15) and test lists (n = 14). Chance level was 10%. The data is not corrected for the average performance level of individual subjects. Error bars represent standard error of the mean.

We calculated the list-specific SRTs and the slope of the intelligibility function. To do this we first corrected the data for the relative performance level of each subject. Then the adjusted data of all subjects was pooled and one logistic model was fitted for each of the 14 lists. The average list-specific SRT across the 14 lists was − 8.4 dB SNR with a standard deviation of 0.2 dB SNR, and the maximum deviation of a list from the overall average was 0.3 dB. The average list-specific slope was 10.2 %/dB with a standard deviation of 0.9 %/dB and this was slightly lower than the average subject-specific slope (10.5 %/dB, see Table 2).

Table 2. The estimated SRT and slope for each center.

	SRT (dB)	*Standard deviation across listeners (dB)*	*Slope of intelligibility function (%/dB)*	*Standard deviation across listeners (%/dB)*
Leuven	− 8.4	0.9	10.8	1.6
Rotterdam	− 8.4	0.4	10.0	1.1
Amsterdam	− 8.5	0.8	10.6	1.5
Overall	− 8.4	0.7	10.5	1.4

Discussion

Our developed matrix test had a speech reception threshold in noise of −8.4 dB with an inter-list standard deviation of 0.2 dB. The list-specific steepness of the intelligibility function was 10.2 %/dB. The SRTs and slopes of the intelligibility function did not differ significantly for listeners in Leuven (SD), Rotterdam (ND), and Amsterdam (ND).

Differences in accent of the listener (ND versus SD) did not result in differences in speech scores. Thus, we can conclude that there is no significant difference in SRT and slope between ND listeners in ND (tested in Rotterdam and Amsterdam) and SD (tested in Leuven). The speech materials can therefore be used in both the Netherlands and in Flanders.

The inter-list standard deviation (0.2 dB SNR) was smaller than the inter-subject standard deviation (0.7 dB SNR, Table 2), indicating that differences between lists are smaller than differences between subjects. This is comparable to that of matrix tests in other languages. For instance the inter-list standard deviation for the Danish matrix test is also 0.2 dB SNR and its inter-subject standard deviation is 1.0 dB SNR (Wagener et al, 2003).

For other languages the slopes of the intelligibility functions are reported either as list-specific slopes or subject-specific slopes. First, several authors reported list-specific slopes that were measured *without* visual response matrices (Danish, 12.6 %/dB; Swedish, 16.0 %/dB; German, 17.1 %/dB; Polish, 17.1 %/dB). Second, some authors did use a visual response matrix; those authors reported inter-subject slopes of 11.5 %/dB (English) and 13.8 %/dB (French). The influence of whether or not the response matrix is shown seems limited. Hewitt (2008) found a 1.3 %/dB *steeper* subject-specific slope if the participants did not have access to a visual matrix, while Hochmuth et al (2012) found a 0.9 %.dB *shallower* list-specific slope (0.9 %/dB shallower) for the measurements without visual matrix.

The list-specific slope of the Dutch version (10.2 %/dB, no visible response matrix) is 2.4 to 6.9 %/dB shallower than the slopes of the intelligibility functions for the other languages with visible response matrix. To compare our data to that of the French and the English tests, we need to look at the subject-specific slope. For the Dutch materials this slope (10.5 %/dB, Table 2) is 1.0 %/dB shallower than that for English and 2.3 %/dB shallower than that for French. To summarize, the intelligibility function of the Dutch matrix test is less steep (the slope is 1.0 to 6.9 %/dB shallower) than that of matrix tests in other languages.

The differences in slope between the different matrix tests possibly reflect differences in speaking characteristics (e.g. speed, prosody, timing, and articulation) on the recordings.

Conclusions

We developed a matrix type sentence-in-noise test for the Dutch language that can be used in both Flanders and the Netherlands.

The developed matrix test has a speech reception threshold in noise (−8.4 dB SNR) and inter-list standard deviation (0.2 dB) comparable to that of other languages. The list-specific slope of the intelligibility function (10.2 %/dB) was lower than that of similar speech tests in other languages (12.6 to 17.1 %/dB).

Acknowledgments

We would like to thank E. Boon and M. Krone, E. Visser, M. Nelissen, and R. Maas for their contributions and assistance in the evaluation measurements. We also thank our colleagues from Hörtech and the Carl von Ossietzky University of Oldenburg for the cooperation in the production of the test material, and B. Kollmeier for his valuable suggestions. Authors Rolph Houben and Jan Koopman contributed equally to this work and are considered joint first authors.

Note

1. We used a generalized linear model with the following link function: log((p-a)/(1-p)). In this equation p represents the probability that the sentence is correctly repeated by the listener.

Declaration of interest:The authors report no declarations of interest.

References

Dreschler W.A., van Esch T.E.M., Lijzenga J., Wagener K., Larsby B. et al. 2006. *Procedures for the Tests Included in the Auditory Profile in Four Languages*, HearCom Report D2-2. Available at: www.HearCom.eu.

Hagerman B. 1982. Sentences for testing speech intelligibility in noise. *Scand Audiol*, 11, 79–87.

Hewitt D.R. 2008. *Evaluation of an English Speech-in-Noise Audiometry Test*. Southampton, UK: MSc. Thesis, Faculty of Engineering, Science and Mathematics, University of Southampton.

Hochmuth S., Brand T., Zokoll M.A., Castro F.Z., Wardenga N. et al. 2012. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int J Audiol*, 51, 536–544.

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 multicenter study. *Int J Audiol*, 51, 164–173.

Kollmeier B. 1990. *Messmetodik, Modellierung, und Verbeserung der Verstandlichkeir von Sprache. (Methodology, modeling, and improvement of speech intelligibility measurements). Habilitation thesis*. Göttingen: University of Göttingen.

Kollmeier B. & Wesselkamp M. 1997. Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *J Acoust Soc Am*, 102, 2412–2421.

Luyckx K., Kloots H., Coussé E. & Gillis S. 2007. Klankfrequenties in het Nederlands. In: D. Sandra (ed). *Tussen taal, spelling en onderwijs*.

Ozimek E., Warzybok A. & Kutzner D. 2010. Polish sentence matrix test for speech intelligibility measurement in noise. *Int J Audiol*, 49, 444–454.

Vlaming M.S.M., Kollmeier B., Dreschler W.A., Martin R., Wouters J. et al, 2011. HearCom: Hearing in the Communication Society. *Acta Acust United Ac*, 97, 175–192.

Wagener K., Brand T. & Kollmeier B. 1999a. Evaluation des Oldenburger Satztests. In: *2. Jahrestagung der Deutschen Gesellschaft für Audiologie*. München, Germany. *Z Audiol*, Suppl. 52–54.

Wagener K., Brand T. & Kollmeier B., 1999b. Enwicklung und Evaluation eines Satztests für die Deutsche Sprache I: Design des Oldenburger Satztests. *Z Audiol*, 38, 4–15.

Wagener K., Josvassen J.L. & Ardenkjaer R. 2003. Design, optimization, and evaluation of a Danish sentence test in noise. *Int J Audiol*, 42, p. 10–17.

Zokoll M., Wagener K.C., Warzybok A., Hochmuth S. & Kollmeier B. 2012. International vergleichbare und multilingual einsetzbare Sprachtests im Störgeräusch: Erweiterung um einen Amerikanisch Englischen Test. In: 43. Jahrestagung der Deutschen Gesellschaft für Medizinische Physik. Jena,Germany, pp. 153–156.

Zokoll M., Hochmuth S., Warzybok A., Wagener K.C., Buschermöhle M. et al. 2013 Speech-in-noise tests for multilingual hearing screening and diagnostics. *Am J Audiol*, 22, 175–178.