The French Digit Triplet test: A hearing screening tool for speech intelligibility in noise


Sofie JANSEN¹, Heleen LUTS¹, Kirsten Carola WAGENER², Bruno FRACHET³,⁴ and Jan WOUTERS¹

² Hörzentrum Oldenburg, Germany
³ Hôpital Avicenne, Assistance Publique Hôpitaux de Paris - Université Paris XIII, France
⁴ France Presbyacousie, France

Key words:
Speech intelligibility, self-screening, speech-in-noise, optimization, home telephone use

Corresponding author:
Sofie Jansen
Herestraat 49/721
B-3000 Leuven
Belgium
Tel: +32 16 33 04 77
Sofie.Jansen@med.kuleuven.be

Abstract
A French speech intelligibility screening test in noise that applies digit triplets as stimuli has been developed and evaluated for both telephone and broadband headphone use. After optimizing the speech material based on the intelligibility of the individual digits, norms for normal-hearing subjects were established. Speech reception thresholds (SRTs) of -6.4 ±0.4 and -10.5 ±0.3 dB SNR and slopes of 17.1 and 27.1 %/dB were obtained for telephone and broadband headphone presentation,
respective. The French digit triplet test by telephone was then implemented as an automatic self-screening test by home telephone, and further evaluated in normal-hearing and hearing-impaired listeners. A test-retest variability of 0.7 dB was found and the correlation between SRT and pure-tone average (PTA0.5,1,2,4) was 0.77. One month after launching the test, 20 000 calls were registered. It can be concluded that both versions of the newly developed test have steep slopes and small SRT differences across normal-hearing listeners. The screening test by telephone is highly reliable and proves to fulfill the need for an easily accessible and objective hearing screening.

Introduction

The percentage of hearing-impaired listeners wearing hearing aids is found to be very low. The study of Popelka et al (1998) showed a prevalence of only 31.7 % of hearing aid use among a large group of listeners with pure-tone averages at 500, 1000, 2000 and 4000 Hz (PTA0.5,1,2,4) above 40 dB HL. One of the reasons for this under-treatment is the high prevalence of undiagnosed hearing losses (Trumble & Piterman, 1992). Therefore, a reliable and accurate screening test – with the characteristic to be performed quickly, to be inexpensive, and to have a low (psychological) barrier – is expected to offer a solution for the under-diagnosis in the population of adults who would benefit from amplification.

Subjective screening tests using questionnaires to measure self-reported hearing disability or handicap have been shown in literature to have a quite low accuracy to detect hearing losses. Nondahl et al (1998) found a sensitivity of only 0.34 for the screening version of the Hearing Handicap Inventory for the Elderly when hearing loss was defined as a PTA0.5,1,2,4 higher than 25 dB HL. The correlation between ‘perceived disability’ measured by the Swedish translation of the Hearing Disabilities and Handicaps Scale and PTA4,6 was only 0.26 (Hallberg, 1998). Using the five questions of the self-reported hearing disability questionnaire of the Longitudinal Aging Study Amsterdam, only 69 % of the participants were classified correctly compared to the outcome of a diotic speech-in-noise test (Smits et al, 2006a). Furthermore, an elderly population tends to rate their perceived hearing disability lower than younger subjects with the same hearing loss (Gordon-Salant et al, 1994). Since subjective screening measures seem to result in an underestimation of hearing impairment, the development and validation of objective screening tests started to gain more interest.

Difficulties with understanding speech in the presence of competing sounds are experienced by most people as the greatest handicap related to their hearing impairment (Kramer et al, 1998; Gatehouse & Noble, 2004). Therefore, measuring the ability to understand meaningful everyday sentences in a background noise has become a commonly used method to quantify communication performance. The speech reception threshold (SRT), i.e. the signal-to-noise ratio at which an intelligibility score of 50 % is obtained, is generally used to present this ability and can give additional information about one’s hearing status compared to pure-tone thresholds. Speech-in-noise tests are considered to be suitable for screening purposes because of the following qualities: (1) the
presented stimuli are all supra-threshold, so the test is less demanding with regard to testing environment. (2) Since the SRT is not influenced by the absolute presentation level in a quite large intensity range (Plomp & Mimpen, 1979b), no absolute calibration is required. (3) In case of a closed set of speech materials, the test can easily be implemented as a fully automatic test. (4) Using an adaptive procedure, the SRT can be determined in only a few minutes. Due to these four factors, speech-in-noise tests can easily be implemented as an objective and easily accessible self-screening test which can be performed in a home setting.

Smits et al (2004) were the first to develop such a screening tool for the Dutch language, based on the intelligibility of digit triplets in noise, and which could be conducted at home by telephone. The SRTs determined by this adaptive self-screening test show a small measurement error (0.8 dB), and a steep slope of 20 %/dB at the SRT is found for normal-hearing listeners. Strong correlations with a standard Dutch sentence-in-noise test (0.87) and with the PTA_{0.5, 1, 2, 4} (0.77) were obtained. The test also has a high sensitivity and specificity of 0.91 and 0.93, respectively (Smits et al, 2004). Four months after making the test publically available, more than 65 000 people dialed the test, showing the large interest and need for a hearing screening test (Smits & Houtgast, 2005). Through an additional questionnaire, Smits et al (2006b) found that about 50% of the Dutch callers, who received an ‘insufficient’ or ‘poor’ result, did follow the recommendation to visit a hearing specialist. As a result, an increase in identification and treatment of hearing impairment can be expected.

Because of the large success of the Dutch automatic speech-in-noise screening test, the test was also realized in several other languages within the framework of the European HearCom project. The present study, being part of HearCom, involves the development of a digit triplet test for Francophone listeners. The first section of this paper deals with the optimization and evaluation in normal-hearing listeners of the new French speech material, for both telephone and broadband headphone presentation. The reference intelligibility functions were determined and the equivalence of test lists was verified. In the second section, implementation of the screening test by telephone, performance of both normal-hearing and hearing-impaired listeners was evaluated using the final setup of the screening test. Test-retest variability, training effect and the correlation between SRT and PTA_{0.5, 1, 2, 4} were determined and cutoff values between a ‘good’, ‘insufficient’ and ‘poor’ SRT outcome were established. Section III treats the application of the screening test by telephone in France. Calls in the first month after public release were examined with regard to the gender and age of the callers, and the distribution of the SRTs.

**Section I: Optimization and evaluation in normal-hearing listeners**

**Introduction**

In the development of this new speech intelligibility test, the recorded speech materials were first optimized to get a steep slope of the reference intelligibility function. Since the shape of the intelligibility function of a speech-in-noise test depends on the mean word-specific slope and the distribution of the word-specific SRTs, an optimization procedure in two steps was applied on the
level of the individual digits. The first step was a selection of digits with steep slopes and with SRTs near the average SRT, in the second step the level of the selected digits was adjusted to further reduce the spread of the SRTs. By selecting digits that needed only small level adjustments, the second step was expected not to result in unnaturally sounding intensity changes in the triplets. This final speech material was than evaluated in normal-hearing listeners to determine the reference intelligibility function of the French digit triplet test.

To enable this screening test to be available through several channels, both a telephone version (with limited bandwidth) and a broadband version (which can be presented through headphones or speakers, for use in clinics, through internet, etc.) were developed. Since the optimization procedures for one version are not expected to be optimal for the other, these procedures were executed for both versions separately. Norms for normal-hearing listeners were also obtained for both presentation modes through separate evaluation measurements.

**Methods**

**Speech and noise material**

The monosyllabic digits 1 to 9 combined into triplets were used as speech material. The triplets, uttered by a French female speaker, started each time with the announcement words ‘les numéros ...’ (meaning ‘the numbers ...’). They were pronounced with a natural intonation, but without coarticulation from one digit to the next. Recordings were made through an AKG C100S microphone with a Tascam DA-P1 DAT recorder. A first level equalization of the speech material was done based on the average root mean square (RMS) level of the announcement words, as RMS levels of very short single digits are too diverse for equalization. The recorded triplets were then segmented into single digits, omitting the silences in between. For each digit at each position in the triplet two recordings were available, resulting in 54 different stimuli. During the speech intelligibility measurements, the measurement software itself generated new triplets by concatenating the announcement words and the three digits with silence intervals of 200ms in between. Only one recording of the announcement words was used throughout the measurements.

Five test lists were created for the optimization measurements in order to determine the intelligibility functions for each single digit at each position in the triplet. These lists were presented in two variants, each containing one of the two recordings of a digit. For the evaluation of the French digit triplet test, ten final lists were generated. All lists consisted of 27 triplets, always including each of the nine digits at each position exactly three times. Within a list, triplets were presented in random order.

A quasi stationary noise with the same spectrum as the long-term spectrum of the speech material was generated by repeatedly superimposing sequences of digit repetitions (Wagener et al, 2003). The RMS level of the noise was chosen to match the average RMS level of the single digits.
Test setup

The French digit triplet test was administered with a computer using the Oldenburg Measurements Applications software (www.hoertech.de). It was implemented as a closed-set test; therefore subjects had to enter their answer on the computer screen by selecting three digits in a response box with ten digit buttons representing a telephone pad. Subjects could also enter the digit ‘0’ in their answer, since they were instructed that all digits were possible.

Participants were seated in a quiet room in front of a computer screen. When conducting the telephone version of the test, a ‘Telephone Handset Audio Tap’ (JK Audio) was used to play sounds from the PC over a fixed phone line (ISDN), so that the subjects heard the stimuli through the telephone receiver at the other side of the connection. During the broadband headphone version, the triplets were sent via an RME sound card and presented monaurally through Sennheiser HDA200 headphones, with free-field equalization. Both test versions were calibrated to dB SPL with a B&K Sound level meter 2260 and a B&K Artificial ear 4153.

Subjects

Twenty Belgian and French Francophone listeners took part in the optimization measurements. For the evaluation of the French digit triplet test, 30 Belgian Francophone subjects participated. All participants were screened for normal hearing: thresholds of all octave frequencies from 250 to 8000 Hz were below or equal to 20 dB HL. Subjects had a mean age of 29 years, ranging between 16 and 49 years.

Procedures

Optimization

To determine the intelligibility function for each single digit, subjects performed the test at 7 fixed signal-to-noise ratios (SNRs), chosen around the expected SRT given by pilot experiments. Measurements by telephone were performed at -13.7, -11.7, -9.7, -7.7, -5.7, -3.7, and -1.7 dB SNR, broadband headphone measurements at -17.9, -15.9, -13.9, -11.9, -9.9, -7.9, and -5.9 dB SNR (decimals result from a posterior rescaling of the speech versus the noise). Ten subjects did the telephone version, the other ten the broadband headphone version. The noise presentation level was fixed at 65 dB SPL. The noise started 500 ms before and ended 500 ms after each triplet presentation, and was thus interrupted in between the test stimuli. The onset and the offset of the noise stimulus were ramped with a 50 ms Hanning window. All participants performed 14 test lists (each SNR twice), so that each digit was presented 210 times in total (30 times per SNR). Since only 5 different lists were used in this phase, lists were reused during one session. Subjects performed the test in one session of about two hours. Although a triplet scoring method will be used in the final test, responses during the optimization measurements were scored per digit. For each of the 54 digits and for both versions (telephone and broadband headphone) separately, intelligibility scores per SNR were obtained by pooling data of all subjects and test lists. Then, a non-linear regression fit was performed, using the logistic function
The optimization was first done by selecting the best of the two recordings of each digit according to the following three criteria: (1) steepest slope of the intelligibility function, (2) individual SRT closest to the mean SRT across digits (resulting in the smallest level adjustment in the next step), and (3) RMS after level adjustment closest to the mean RMS across digits. The second and third criterion both aim to prevent the final triplets from sounding unnatural due to large intensity changes between successive digits. The digit meeting at least 2 of the 3 criteria was selected. In a second step, level adjustments were applied to the individual digits towards the mean SRT. The largest adjustments applied were 5.4 and 3.1 dB for the telephone and broadband headphone version, respectively. Naturalness in sound melody of the triplets, evaluated by a native listener, was not affected by these adjustments. It is important to note that selection and level adjustments were done separately for telephone and broadband headphone version, resulting in two different speech materials.

**Evaluation**

A fixed procedure was used to determine the reference intelligibility function of the French digit triplet test. Twenty-four Belgian Francophone normal-hearing listeners, divided over the two versions, participated. Lists were presented at -9.2, -7.7, -6.2, and -4.7 dB SNR for the telephone version and at -12.9, -11.4, -9.9, and -8.4 dB SNR when presenting through headphones (decimals result from a posterior rescaling of the speech versus the noise). The noise presentation level was fixed at 65 dB SPL, and the noise was interrupted between triplet presentations like in the optimization measurements. Each subject performed 20 lists (5 at each SNR), so all ten test lists were presented twice to each subject. Participants did the measurements in one session of about two hours. Since the simple up-down adaptive procedure of the final test (described hereafter) requires triplet scoring, norms for normal-hearing listeners were established with this scoring method as well. A triplet was thus judged to be correct only when all three digits were identified correctly and chance level was now around zero. SRTs and slopes of each subject individually, with data averaged across lists, were determined by means of non-linear regression fits to the logistic function given by equation 1 (with \( \gamma = 0 \)). Intelligibility functions of each test list separately were determined as well, by pooling all data of the different subjects. To examine the effect of the optimization steps, SRTs of the individual digits were also analyzed after recalculating the scores on a digit level.

The French digit triplet test was also administered using the simple up-down adaptive procedure (Plomp & Mimpren, 1979a) of the final setup. Now the level of each triplet varied depending on the response to the previous triplet, whereas the level of the noise was held constant. When a complete triplet was identified correctly, the level of the next triplet was decreased by 2 dB. If not, the level was increased by this amount. The SRT was calculated by averaging the SNRs of triplet 7 to 27, plus

\[
SI(SNR) = \gamma + (1 - \gamma) \frac{1}{1 + e^{s(SRT - SNR)}}
\]  

**(equation 1)**

where \( SI = \) speech intelligibility, \( \gamma = \) guess level (in this case \( \gamma = 0.1 \)), and \( s = \) slope at the SRT.
the SNR of the imaginary 28th triplet determined from the response on the 27th. In this phase, six normal-hearing subjects performed each 10 test lists via telephone and 10 via headphones (broadband). The level of the first triplet was set at -4 and -8 dB SNR for the telephone and broadband headphone version, respectively. The noise presentation level was again fixed at 65 dB SPL, and the noise was interrupted between triplet presentations as in the optimization measurement.

Results

Optimization

For each single digit, SRTs and slopes were obtained as described above. Averaged across all 54 digits, telephone and broadband headphone SRTs were -8.1 (SD=2.9) and -11.8 (SD=1.9) dB SNR, and slopes were 16.2 (SD=4.9) and 20.7 (SD=6.2) %/dB, respectively. To obtain a steep slope of the final reference intelligibility function for normal-hearing listeners, it is necessary to have basic items with steep slopes and to decrease the variation in intelligibility of these items. For the French digit triplet test, this was done on the level of the single digits, using the selection and level adjustment procedures described above. The achieved reduction in SRT spread is shown in Table 1. Both the range between the highest and the lowest measured SRT and the standard deviation of SRTs across all digits is given for telephone and broadband headphone version. A first, rather small reduction in spread could be seen after removing the most deviant digits with regard to slope, SRT, and RMS. After adopting the level adjustments – which theoretically reduces the range and standard deviation to zero – and presenting the test to 24 new subjects (evaluation measurements), SRTs of the single digits showed indeed to be very concentrated. In general, the spread in SRT was higher for the telephone version than for broadband headphone measurements. A steeper slope of the intelligibility function for the broadband headphone version was thus expected.

Table 1. Spread of the fitted SRTs of the individual digits, represented as the range between the highest and lowest SRT and as the standard deviation (SD) of SRTs across the digits. Three phases are shown: before any optimization, after the selection of the most homogenous digits, and after level adjustments towards the mean SRT (final speech material).

<table>
<thead>
<tr>
<th></th>
<th># digits</th>
<th>Telephone</th>
<th>Broadband</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Range (dB)</td>
<td>SD (dB)</td>
</tr>
<tr>
<td>Before optimization</td>
<td>54</td>
<td>11.2</td>
<td>2.9</td>
</tr>
<tr>
<td>After selection</td>
<td>27</td>
<td>10.1</td>
<td>2.6</td>
</tr>
<tr>
<td>After adjustment (evaluation measurements)</td>
<td>27</td>
<td>4.6</td>
<td>0.9</td>
</tr>
</tbody>
</table>

The spreading of the SRTs before optimization is also given in Figure 1. For all 54 digits, the telephone SRT is plotted against the broadband headphone SRT. The larger dispersion of telephone SRTs compared to broadband headphone SRTs is visible in this scatter plot as well. At first sight, a linear relation could be seen between the two versions. Nevertheless, six data points diverged strongly, since relatively, a very high telephone SRT was found against a rather low broadband headphone SRT. These deviating data points represented all recordings of the digit ‘six’ (3 positions, 2 different recordings). This divergence could be explained by the spectral content of the phonemes
of the word ‘six’ (pronounced as /sis/). Most energy of the consonant /s/ is located in the higher frequencies above 3000 Hz. By sending the word ‘six’ through a telephone line, which is acting like a band-pass filter from 300 to 3400 Hz, speech intelligibility deteriorated seriously, compared to other speech sounds. The rather low correlation (Pearson, R=0.59) between telephone and broadband headphone SRT, was a strong argument to perform a separate optimization for both versions.

![Figure 1. Scatter plot of the SRTs of the single digits for telephone (vertical axis) versus broadband headphone presentation (horizontal axis) before optimization. The six encircled data points represent the recordings of the digit ‘six’.

Evaluation
SRTs and slopes at 50 % scores were obtained for each individual subject based on a non-linear regression fit of the intelligibility function of the fixed measurements. Table 2 gives the arithmetic mean of the SRTs and slopes, the standard deviation across the normal-hearing subjects, and the precision on both parameters, deduced from the quadratically averaged standard errors of the individual fits. Also given in Table 2 is the mean SRT and standard deviation across subjects measured with the adaptive measurements. Since triplet scoring was used, the SRTs were shifted, as expected, to higher values compared to the digit SRTs of the optimization phase.

A first analysis showed that the SRTs measured in an adaptive way did not differ significantly from the fitted SRTs of the fixed procedure, neither for telephone version \([t (16) = 1.3, p = 0.23]\), nor for broadband headphone version \([t (16) = -1.5, p = 0.16]\). Furthermore, small standard deviations (≤ 0.5 dB) across normal-hearing subjects and high precisions (≤ 0.2 dB) of the fitted SRTs were found. Intelligibility functions of the French digit triplet test through telephone and through headphones, using the fixed method are given in Figure 2.
Table 2. SRTs and slopes of the intelligibility function, averaged across subjects, for fixed and adaptive measurements. Standard deviations across subjects and precision values deduced from the quadratically averaged standard errors of the individual fits are given.

<table>
<thead>
<tr>
<th></th>
<th>Telephone</th>
<th>Broadband</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fixed (fitted)</td>
<td>Adaptive</td>
</tr>
<tr>
<td>N</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>SRT (dB SNR)</td>
<td>Mean -6.4</td>
<td>-6.7</td>
</tr>
<tr>
<td></td>
<td>SD 0.4</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Precision 0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Slope (%/dB)</td>
<td>Mean 17.1</td>
<td>27.1</td>
</tr>
<tr>
<td></td>
<td>SD 2.5</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>Precision 2.1</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Test lists were constructed to contain all exactly the same digit material, improving the homogeneity of different lists. However, the combination of the digits into different triplets might bring along some variation between test lists. Therefore, the intelligibility of each test list was determined. By averaging the adaptively measured SRTs of different subjects, the SRT per list was obtained. Figure 3 shows the deviation of each test list from the overall mean SRT for telephone and broadband headphone version. The largest deviation was only 0.6 dB for both versions. A repeated measures ANOVA on the measured SRTs of the adaptive method (2 versions, 10 lists, 6 listeners) also showed no significant main effect of test list \([F(1,5) = 2.3, p = 0.19]\), nor an interaction effect of list and version \([F(1,5) = 0.8, p = 0.42]\). Fitted SRTs per list based on a non-linear regression fit of the fixed method results, with subject scores pooled together, deviated from the overall mean by at most 0.3 dB for both versions.

![Figure 2. Speech intelligibility scores per subject at the 4 measured SNRs for telephone (▲) and broadband headphone (○) version. Also the reference intelligibility function (with SRT and slope calculated as the arithmetic mean of the fitted SRTs and slopes of each subject individually) is given for telephone version (solid line) and broadband headphone (dashed line).](image-url)
The root mean square of the within-subject standard deviation of repeated measurements (Plomp & Mimpen, 1979a) was considered to determine the measurement error of the adaptively measured SRTs. This proved to be 0.7 dB for the telephone and 0.6 dB for broadband headphone presentation.

Figure 3. Deviation of the SRT of an individual test list from the overall mean. Data are taken from the adaptive measurements. Error bars show standard deviations across subjects. Upper panel: Telephone presentation; Lower panel: Broadband headphone presentation.

Discussion

Both in optimization and evaluation results, a difference in SRT of approximately 4 dB can be seen between telephone and broadband headphone version. This corresponds well to the difference found in the digit triplet test in Dutch (Smits et al, 2004) and in German (Wagener et al, 2006): Telephone SRTs of -7.1 and -6.4 dB SNR, and headphone SRTs of -11.2 and -9.3 dB SNR were obtained for the Dutch and German test, respectively. In Smits & Houtgast (2006) it was shown that a telephone network simulation with only signal filtering and signal compression and decompression, presented under headphones, resulted in comparable SRT values as for measurements through a real telephone and telephone network. The decrease in speech intelligibility for the telephone version can thus only be attributed to the signal transmission through the telephone network (limited bandwidth, distortions), and not to the inferior quality of the telephone transducer.

The broadband headphone SRT of the French digit triplet test is also much lower compared to an SRT of -7.1 dB SNR for the French Intelligibility Sentence test (FIST, Luts et al, 2008), which is a open-set sentence-in-noise test for Francophone listeners. This difference in SRT between a digit- and a sentence-in-noise test (both administered monaurally through headphones, with a stationary speech-weighted noise), could also be seen in other languages. The Dutch digit triplet SRT equals
-11.2 dB SNR (Smits et al, 2004), whereas a sentence SRT of -4.1 dB SNR is reported by Versfeld et al (2000). In Germany too, a much lower SRT (-9.3 dB SNR) is found for the digit triplet test of Wagener et al (2006), compared to -6.2 dB SNR of the sentence test of Kollmeier & Wesselkamp (1997). Although triplets are judged to be correct only when all three digits are identified correctly - decreasing chance level from 10 % to 0.1 % - there is still an advantage of the closed response format of a digit-in-noise test together with a strongly limited test word material, leading to a lowered SRT.

The fitted slope of the French digit triplet test by telephone (17.1 %/dB) is steep and corresponds well to the slopes of the German (16.1 %/dB) and Dutch version (20 %/dB). Presenting the French digit triplet test through headphones (broadband) resulted in an even steeper slope of 27.1 %/dB. This was expected since the broadband headphone optimization resulted in a very small spread of SRTs of the single digits. The Dutch digit triplet test through headphones reached a much lower slope of 16 %/dB, which might be explained by the fact that the optimization steps were based only on telephone measurements. For the French version, a low correlation between telephone and headphones SRTs of the optimization measurements was shown. The application of a separate selection and level adjustment procedure finally resulted in the steepest possible slope for both versions.

Taking into account the number of triplets/sentences used to calculate the SRT, a measurement error of 0.7 dB (telephone) and 0.6 dB (broadband headphone) for 22 triplets is comparable to 0.7 dB for 20 triplets of the Dutch telephone digit triplet test (Smits et al, 2004), and to 1.1 dB for 6 sentences of the FIST (Luts et al, 2008). Compared to the variability of 1.4 dB for 10 digits of the Leuven Intelligibility Number Test (LINT, Van Wieringen & Wouters, 2008), 1.1 dB for 17 sentences of the Canadian French HINT (Vaillancourt et al, 2005), and 0.9 dB for 20 triplets of the Dutch headphone digit triplet test (Smits et al, 2004), the test-retest variability of the French digit triplet test is relatively high.

Although the optimization of the French digit triplet test was based on the intelligibility of single digits, the final test assesses the intelligibility of complete triplets. Theoretically, to reach the steepest possible slope at the 50%-point at triplet level, one needs to align the 79%-points of the single digits (Smits & Houtgast, 2006). Since the slopes of the single digits of the French digit triplet test were very similar to each other, aligning the SRTs of the digits was expected to be as optimal as adjusting the levels of the digits based on their 79%-point.

**Section II: Implementation of the French screening test by telephone:**

- **cutoff values, test-retest variability, and SRT-PTA correlation**

**Introduction**

Since it has been shown that the French digit triplet test is an accurate speech-in-noise test with a steep slope and a small measurement error, the telephone version could now be implemented on an
interactive voice response (IVR) system as an automatic self-screening test by home telephone. As opposed to the typical binary screening outcome ‘pass/refer’, it was chosen to divide the SRT results in three categories of ‘good’, ‘insufficient’, and ‘poor’ hearing. As no gold standard for speech intelligibility in noise is available in French (the recently developed FIST (Luts et al, 2008) was only evaluated yet in normal-hearing listeners), a proper validation of the newly implemented French digit triplet screening test by telephone was not possible. Therefore, the cutoff values between the three outcome categories were determined based on measurements with a group of normal-hearing listeners. The final setup was also presented to hearing-impaired subjects to verify the correlation between SRTs obtained with the screening test and PTAs. The test-retest variability and the training effect were examined for both groups as well.

Methods

The screening test by telephone was implemented by a French telecom company. The same adaptive method as described in the evaluation measurements of section I was adopted. Sound files with pre-mixed triplets and noise were used. SNRs ranged from -12 to +8 dB SNR, with a step size of 2 dB. In case subjects gave a correct response on -12 dB SNR or an incorrect response on +8 dB SNR, the next triplet was presented at the same SNR. The first triplet was presented at an SNR of +2 dB. When connecting to the test via telephone, a short introduction and instruction message was presented. People were also asked to give in their age and gender. Then a test list was randomly selected and the triplets within the list were presented in randomized order. The listener answered by pressing the keys on the telephone pad. At the end, the test result (‘good’, ‘insufficient’ or ‘poor’) and an appropriate recommendation were given. The complete phone call took about 5 minutes, including the measurement of one ear. Since calls from mobile phones give worse and less reliable results (Smits & Houtgast, 2005) a method to block calls from mobile phones was applied.

Twenty-four French subjects participated. Most of them were outpatients of the ENT department of Hôpital Avicenne, Bobigny. The other participants were employees at the same department. Pure-tone thresholds were measured at both ears for all octave frequencies between 250 and 8000 Hz.

Sixteen participants did the screening test by telephone with both ears successively. For the remaining eight only one ear was tested. Subjects performed the test twice per ear. However, no retest data were available for eight out of the 40 ears due to time restrictions. Participants who did the complete test (audiogram, test and retest for both ears) finished the measurements in one session of about 30 minutes. The test was performed in a quiet room at the clinic and all participants used the same telephone.

Results

The audiograms of 19 out of the 40 tested ears showed normal hearing with all measured thresholds below or equal to 20 dB HL. The remaining 21 ears showed an elevated threshold for at least one of the frequencies, and were thus considered as hearing-impaired. Some hearing-impaired participants
got an invalid SRT by giving an incorrect response at +8 dB SNR. Although the outcome of the test ('poor hearing') was still appropriate, these data were omitted when determining the test-retest variability, the training effect, and correlations between SRT and PTA.

The test-retest variability was determined based on data of test-retest scores of 28 ears, by considering the root mean square of the within-subject standard deviation of repeated measurements (Plomp & Mimpfen, 1979a). For both groups of normal-hearing and hearing-impaired ears, the variability was 0.9 dB. When taking the average of the difference scores, a significant training effect of 0.8 and 0.9 dB was obtained for normal-hearing and hearing-impaired ears, respectively. Since a training effect influences the measured variability, the latter was re-determined using equation 2 which balances out a possible training effect (Smits & Houtgast, 2005). A pure test-retest variability (τ) of 0.7 dB was now found.

$$\tau = \sqrt{\frac{\sum (\text{difference}\text{-}\text{difference})^2}{n}}$$  

(equation 2)

In the following analyses only test results were taken into account, since retest results showed a decreased SRT due to training effect, and are thus not representative for people who will do the test only once.

To define the SRT cutoff value between 'good' and 'insufficient' hearing, the upper 95 % confidence limit for SRT scores for normal-hearing listeners was determined. With an average SRT of -5.4 dB SNR and a standard deviation of 0.7 dB for 19 normal-hearing ears, this limit was set to -4.0 dB SNR. A second border - between 'insufficient' and 'poor' hearing - was determined by adding up once more two standard deviations, resulting in a limit of -2.7 dB SNR.

In Figure 4 a scatter plot of the SRT scores versus PTA_{0.5, 1, 2, 4} is shown for all 40 tested ears. The correlation between both variables was 0.77 and proved to be significant (p<0.001). When leaving out all data of the normal-hearing ears, which were possibly subject to a ceiling effect, the correlation between SRT and PTA even increased to 0.83. The borders between the three result categories of the screening test are also shown in Figure 4. As can be seen, several listeners labeled as hearing-impaired still reach the outcome of 'good hearing'. This was not surprising, since most of them still had a PTA_{0.5, 1, 2, 4} below 25 dB HL. A decreased performance on speech intelligibility in noise is thus not expected for those subjects.

**Discussion**

It is well known that no perfect correlation exists between speech-reception thresholds in noise and pure-tone thresholds. In literature, correlations of 0.70 to 0.80 are reported between sentence SRTs in noise and PTAs for normal-hearing and hearing-impaired listeners (Bosman & Smoorenburg, 1995; Smoorenburg, 1992; Smits et al, 2004). The correlation of 0.77 in this study is thus relatively high. It has to be noted that since the age of the subjects was not taken into consideration, the SRT-PTA relation might be influenced in a limited way by a possible cognitive factor. Furthermore, it lies in the
nature of a screening test via telephone that hearing losses at frequencies above the upper telephone band limit could not be identified.

A small test-retest variability was found for the French screening test by telephone. Both for normal-hearing and hearing-impaired listeners the measurement error was within 1 dB, making the test reliable for screening purposes.

Figure 4. Scatter plot of the digit triplet SRTs (only test, no retest) versus the average pure-tone thresholds at 500, 1000, 2000 and 4000 Hz. Crosses represent normal-hearing (NH) ears; Circles represent hearing-impaired (HI) ears. Arrows indicate data points with invalid SRT due to an incorrect response on the maximal SNR (+8 dB). The horizontal dashed lines represent the borders between a ‘good’, ‘insufficient’, and ‘poor’ outcome for the digit triplet screening test by telephone.

The average SRT of -5.4 dB SNR for the normal-hearing subjects was noticeably higher compared to the evaluation results described in the first section. Considering the estimated training effect of 0.8 dB from the first to the second test list, a decrease in SRT of 1.3 dB is not surprising for subjects performing 10 lists during adaptive evaluation measurements. Also an increased standard deviation across normal-hearing subjects was now found (0.7 dB versus 0.5 dB in adaptive evaluation measurements). This can also be explained by the fact that, during evaluation, each subject performed all ten test lists, balancing out the variation between different lists, and resulting in a lower standard deviation.

The broadband version of the French digit triplet test is expected to show similar results in a study with hearing-impaired listeners. However, when this test will be used for screening through internet, an extra normative study will be necessary. The advantage of stereo presentation and disadvantage of mp3-compressed sound files, lower quality sound cards, and lower quality transducers, might influence important qualities of the test. Also, more variation between users is expected since people can perform the test using headphones, in-ear phones, external loudspeakers, or built-in computer speakers.
Section III: Application of the French screening test by telephone

Introduction

In February 2009, the French screening test by telephone was launched nationwide in France. Television spots were transmitted before and after favorite programs for the elderly, and articles were published in newspapers for pensioners. In the first month, about 20 000 people called the test. Distributions of the SRTs, ages, and genders and possible interactions are considered in this section.

Methods

Details on the test setup of the automatic screening test by telephone were reported in the previous section. As opposed to the controlled setting of the measurements described in section II, here the people performed the test at home, used different telephones, and were situated in different environments. Smits et al (2004) and Wagener et al (2006) showed that reference values are not significantly influenced by different (non-mobile) telephones or quiet environments. Therefore, the cutoff values determined above can also be applied to the public test.

Results

In total, 18 123 completed calls were registered in the first month. For further analyses, two exclusion criteria were applied: calls from subjects who gave no answer at three or more triplets, and calls with missing or inappropriate (<18 and >100 years) age information were removed. This resulted in a data set of 15 391 calls.

Overall, the proportion of female callers (62.3 %) was larger than that of males (37.7 %). The median age was 58 and 59 years for female and male callers, respectively. About 60 % of all callers fell within the 50 to 70 years range. A clear peak in age distribution occurred at the age of 60.

After dividing callers in 10-year-wide age groups, box plots of obtained SRTs were generated (Figure 5). Since the age groups 18-20 and 91-100 contained less than 50 observations, these groups are not shown and not further analyzed. Due to a maximal SNR of +8 dB, a ceiling effect can be seen at positive SRTs. The distribution of the SRTs also proved to be positively skewed, with all – except one – outliers (deviation of 1.5 to 3 times the interquartile range from the upper or lower edge of the box) and extreme outliers (deviation of more than 3 times the interquartile range from the upper or lower edge of the box) located at one side of the distribution. Also shown is the border between ‘good’ and ‘insufficient’ hearing (-4.0 dB SNR) and between ‘insufficient’ and ‘poor’ hearing (-2.7 dB SNR). For the age group 61-70, less than 50 % of callers reached a ‘good’ SRT. For the age groups 71-80 and 81-90, this proportion even dropped below 25 %. In a first analysis, male results were compared to female results for each age group separately. Since the SRT data did not have a normal distribution, the Mann-Whitney U test was used. When applying a correction for multiple comparisons (α=0.05/7=0.007), no significant differences were found between both sexes. In a
second analysis, data of two genders were taken together, and paired comparisons (Mann-Whitney U) between one age group and the subsequent were executed. Six comparisons were made, so the significance level was set at $\alpha=0.008$. A small decrease in SRT ranking occurred between the 21-30 and 31-40 group, but this was not significant ($p=0.683$). The increases in SRT between the 31-40, 41-50, 51-60, 61-70, 71-80 and 81-90 age groups, were all significant ($p<0.001$).

In Figure 5, the distribution and cumulative distribution of the SRTs of all 15,391 callers are given. Again, due to the maximal SNR of +8 dB, the right tail of the distribution curve was suppressed. Forty-six percent of the callers obtained a ‘good’ test result, whereas 21 % and 33 % turned out to have ‘insufficient’ and ‘poor’ hearing, respectively.

At last, Figure 7 shows the proportion of test outcomes plotted against age. Ages were grouped per two years. Only test results for the 20-21 to 86-87 years old are shown, to have at least 50 observations per age group. Up to the age of 50, more than 60 % reached a ‘good’ SRT. This proportion gradually decreased to about 10 % for the eighty-year-old listeners.
Figure 6. Distribution and cumulative distribution of obtained SRTs. The vertical dashed lines indicate the borders between ‘good’, ‘insufficient’ and ‘poor’ hearing.

Discussion

The analysis of approximately 15,000 calls in the first month after launching the screening test by telephone in France, revealed that more women were reached with the test than men. This was also found for the Dutch screening test by telephone (Smits et al., 2006b). French callers also showed a similar age distribution to Dutch callers. The French median age was only 3 to 4 years higher than the Dutch (Smits & Houtgast, 2005). Male and female callers reached similar SRT results and, as expected, the SRT increased significantly with age. The percentage of people with a ‘good’ SRT outcome (46%) is noticeably lower than in the Netherlands (67%). This large difference can only partially be explained by the small difference in age and by the fact that all calls with an invalid SRT (incorrect response at +8 dB SNR) were omitted in the Dutch analysis by which the proportion of ‘poor’ outcomes was somewhat reduced. We can conclude that it is more difficult to reach a ‘good’ outcome with the French than with the Dutch screening test by telephone, probably resulting in a somewhat larger proportion of false-positives (>7%), but a smaller proportion of false-negatives (<9%) compared to the Dutch test (Smits et al., 2004).

General conclusions

The French digit triplet test is a speech-in-noise test that can be used as an automatic self-screening test by telephone or internet. After optimizing the test based on the intelligibility of the single digits, this triplet-in-noise test was evaluated in normal-hearing subjects in a fixed and adaptive way. For both the telephone and broadband headphone version, normative data for the SRT and slope were obtained, and small standard deviations across subjects were found. The fitted values for both parameters of the intelligibility function also showed a very high precision. Steep slopes and small measurement errors were established.
Figure 7. Proportion of different SRT outcomes ('good', 'insufficient', 'poor') versus age.

After implementation of the telephone version as an automatic self-screening test by home telephone, measurements with both normal-hearing and hearing-impaired listeners revealed a test-retest variability below 1 dB. We can thus conclude that the French screening test by telephone is able to reliably measure one’s speech reception threshold. The large number of calls proves to fulfill the need for an easily accessible hearing screening test.

Due to the large amount of people conducting this screening test, we can expect to have an increased prevalence of hearing-impaired listeners assessed by a hearing specialist and diagnosed to have a hearing loss. Through the help of follow-up programs where the hearing-impaired person is counseled by a professional, more hearing-impaired listeners will hopefully benefit from good amplification, and an improved quality of life can be expected for this large segment of the population.

Acknowledgements
The authors would like to thank Ellen Boon and Celine Jacobs for administering the optimization measurements. Daniel Berg, Michael Buschermöhle and Birger Kollmeier are gratefully acknowledged for their technical and methodological support. Our thanks go to the association France Presbyacousie for financing the implementation of the digit triplet screening test by telephone, to Jens Kofoed for the technical realization, and to Emilie Vormès for performing the measurements at the ENT department of Hôpital Avicenne. We thank Marcel Vlaming for the first contacts with France Presbyacousie, and for his excellent HearCom management. This research was supported by a grant from the European Union FP6, Project 004171 HEARCOM.
References


