International Journal of Audiology

ISSN: 1499-2027 (Print) 1708-8186 (Online) Journal homepage: http://www.tandfonline.com/loi/iija20

Development of Telscreen: a telephone-based speech-in-noise hearing screening test with a novel masking noise and scoring procedure

Harvey Dillon, Elizabeth Francis Beach, John Seymour, Lyndal Carter & Maryanne Golding

To cite this article: Harvey Dillon, Elizabeth Francis Beach, John Seymour, Lyndal Carter & Maryanne Golding (2016): Development of Telscreen: a telephone-based speech-in-noise hearing screening test with a novel masking noise and scoring procedure, International Journal of Audiology, DOI: 10.3109/14992027.2016.1172268

To link to this article: http://dx.doi.org/10.3109/14992027.2016.1172268

Published online: 03 May 2016.

Taylor & Francis
Taylor & Francis Group

International Journal of Audiology 2016; Early Online: 1–8

International Journal of Audiology

Original Article

Development of Telscreen: a telephone-based speech-in-noise hearing screening test with a novel masking noise and scoring procedure

Harvey Dillon, Elizabeth Francis Beach, John Seymour, Lyndal Carter & Maryanne Golding

National Acoustic Laboratories, Macquarie University, Sydney, NSW, Australia

Abstract

Objective: In 2006 the National Acoustic Laboratories was commissioned to create a telephone-based hearing screening test. *Design:* NAL developed 'Telscreen', a speech-in-noise test modelled on the Dutch and UK telephone tests. The first version, Telscreen I, had several novel features: individual scoring of digits; individual equalization of digit intelligibility; and accuracy-determined test termination. Evaluation of Telscreen I revealed that it did not discriminate satisfactorily between those with and without hearing impairment. Subsequently Telscreen II, which included a novel sensitized masking noise, was developed. *Study sample:* Telscreen I was evaluated by 105 participants (22–86 years), 37% with normal hearing (all thresholds <20 dB HL in the test ear), 63% with hearing impairment (all thresholds >20 dB HL in the test ear). Telscreen II was evaluated by 75 participants (25–86 years), 33% with normal hearing, 67% with hearing impairment. *Results:* Correlations between Telscreen I results and hearing thresholds, $r = 0.57$, and hearing disability scores, $r = 0.51$ were highly significant, but lower than expected. Correlations for Telscreen II were higher: $r = 0.77$ and 0.65, respectively. Telscreen II was found to have high sensitivity: 90%; and specificity: 90.2%. *Conclusions:* Telscreen II is an efficient, reliable, and innovative hearing screening test that provides a solid foundation for future tests delivered via mobile and internet technologies.

Key Words: Telephone, hearing, screening

The primary aim of a hearing screening test is to identify people with hearing impairment or disability, and ultimately for those identified to 'follow up' on their test result and seek further advice in managing their hearing impairment (Grandori et al, 2009). Standard, pure-tone audiometry (PTA), performed by a clinician, is the main method by which hearing impairment is assessed, but it is not suitable as a mass hearing screening method because it is resource- and time-intensive. In recent years, self-administered telephone-based hearing screening tests have been developed, in which the caller is presented with speech masked by a background noise (Smits et al, 2006). Such 'speech-in-noise' tests are preferred as hearing screening tests because they are suited to automated delivery by telephone or the internet and can be completed in considerably less time than standard PTA. Furthermore, results from speech-in-noise tests directly measure hearing disability, i.e. deficits in the ability to understand speech (Kramer et al, 1998; Lutman et al, 1987). Hearing disability, rather than hearing impairment (i.e. hearing threshold shift) per se, is a more likely determinant of whether an individual will seek rehabilitation (Kramer et al, 1998; Lutman et al, 1987), and thus a method which measures disability is likely to be more useful as a screening test. Further, measures of speech understanding in noise seem likely to have a higher face validity than pure-tone tests in quiet, from the perspective of the people being tested.

In 2006, an Australian telephone-based speech-in-noise test was requested by the Australian government which the National Acoustic Laboratories (NAL) was asked to develop, for national implementation by Australian Hearing, the government-funded provider of hearing loss services in Australia. At that time, hearing telephone screening tests had been developed and shown to be reliable in The Netherlands (Smits & Houtgast, 2007; Smits et al, 2004, 2006), and had just been launched in the UK. These tests use digit triplets presented in speech-shaped masking noise with a variable signal-to-noise ratio (SNR). The test sounds are presented via telephone, and the listener responds by entering the digits heard on the telephone keypad. These tests use an adaptive procedure whereby a predetermined number of triplets are presented, and during the test, the noise level is fixed while the triplet level varies. After an incorrect

Correspondence: Elizabeth Francis Beach, National Acoustic Laboratories, Australian Hearing Hub, Macquarie University, Sydney, NSW 2109, Australia. E-mail: Elizabeth.beach@nal.gov.au

(Received 14 October 2015; revised 16 March 2016; accepted 25 March 2016)

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.3109/14992027.2016.1172268

Abbreviations

4FAHL	4-frequency average hearing loss
HDQ	Hearing disability questionnaire
ILTASS	International long-term average speech spectrum
IVR	Interactive voice recognition
NAL	National Acoustic Laboratories
PTA	Pure-tone audiometry
RMS	Root mean square
SNR	Signal-to-noise ratio
SRT	Speech reception threshold

response (i.e. any digit incorrect) the next triplet is presented at a 2-dB higher level, making the task easier; after a correct response the triplet level is lowered by 2 dB. The speech reception threshold (SRT_n), defined as the SNR that corresponds to 50% of triplets correct, is estimated by taking the average of the SNRs of the last 20 presentations, following four practice trials, during which the SNR adapts to close to the eventual SRT_n. This individual SRT_n is then compared with a pre-determined pass/fail criteria and the caller receives an appropriate post-test message telling them their result on the test.

NAL developed the Australian telephone-based speech-in-noise test, 'Telscreen', modelled on the Dutch and UK telephone screening services (Smits & Houtgast, 2006) with a number of amendments, which are the focus of this article.

Individual scoring of digits

Telscreen scores each digit in the triplet individually, that is, the consequence of making an incorrect response differs according to the number of digits entered incorrectly. This was motivated by the desire to have the response to every digit contribute to the adaptation of level, irrespective of the correctness of the other digits in the triplet, with the aim of improving test efficiency. In contrast, in the Dutch and UK tests a response in which one, two, or three numbers are entered incorrectly all result in the next digit triplet being presented at a higher SNR.

Individual equalization of digit intelligibility

Because each digit is individually scored in Telscreen, preliminary experiments were used to equalize the intelligibility of all digits in the test, rather than equalize the intelligibility of digit triplets. Both theoretical principles (Dillon, 1983) and the experimental results of Smits and Houtgast (2006) indicate that such equalization increases test precision by maximizing the slope of the psychometric function when the test items are combined.

Accuracy-determined test termination

Unlike other tests, Telscreen continually estimates the accuracy of the measurement. Rather than continue for a pre-determined number of trials, the test length is variable, stopping when a target accuracy is achieved.

Sensitized masking noise

Because people with sensorineural hearing impairment are less able to take advantage of temporal gaps in a fluctuating masker, and less able to take advantage of spectral gaps than those with normal hearing (Larsby & Arlinger, 1999), it was decided to use a modulated masking sound that also contained a series of spectral gaps, to improve the correspondence between SRT_n and hearing thresholds. A further aim of this modification was that the SRT_n would be more correlated with perceived disability in real life, in which masking sounds often have marked temporal and/or spectral variation.

Telscreen was launched in 2007 (Telscreen I), with the first three of the above four amendments and an evaluation period commenced. The evaluation revealed that the test did not discriminate as well as required between those with and without a pure-tone hearing impairment (i.e. hearing threshold shift). To rectify this, it was decided to modify the test by introducing the novel masking noise (Telscreen II).

This paper describes the two-stage development of Telscreen and the evaluation processes which were undertaken to improve the test and ensure that results obtained from Telscreen were reliable and well-matched to both perceived hearing disability and measured pure-tone hearing thresholds.

Experiment 1: Telscreen I development and evaluation

Method

Speech recordings

Speech files were recorded by a mature adult female speaker of standard Australian English, who was also an experienced audiologist and speech researcher. Multiple examples of nine triplets which comprised: 1, 2, 3, 4, 5, 6, 8, 9, and 0, were recorded, in which each digit occurred in each position in the triplet once. The digit '0' was pronounced as 'oh', which is commonly used and recognized by Australian (and British) speakers. The digit '7' was excluded because it has two syllables, and would therefore be easier to identify compared to the single-syllable numbers. The carrier phrase: 'The numbers are' was also recorded, and this preceded the presentation of each digit triplet throughout the Telscreen test.

Recordings took place in an anechoic chamber using a 4155 microphone connected to a B&K sound level meter type 2230 and then to a Lynx Two sound card. The sampling rate for recording was 44.1 kHz, but the files were re-sampled to 8 kHz, with a 4 kHz bandwidth to better match the typical bandwidth of an analogue telephone (3.4 kHz) and to keep the file sizes manageable for storage and playback.

The triplets were segmented into individual digits and one example of each of nine digits in each of the three positions (first, second, or third) was selected. Care was taken to ensure that the digits selected were clearly spoken, free from artefacts, and between 400 and 500 ms in length. The selected 27 individual digits were recombined to produce 81 pseudo-random triplets in which there were no repeated digits, and each digit appeared equally often in each position (i.e. nine times). There was a 600 msec gap between digits, and each digit occurred in its original position within the triplet to ensure natural intonation was maintained. The average root mean square (RMS) level of each triplet was adjusted to ensure consistency within and across triplets.

The digit triplets in noise were provided to a third-party communications company who were engaged to build Telscreen on an interactive voice recognition (IVR) platform, commonly used in automated telephone systems, using the algorithm described later in this section.

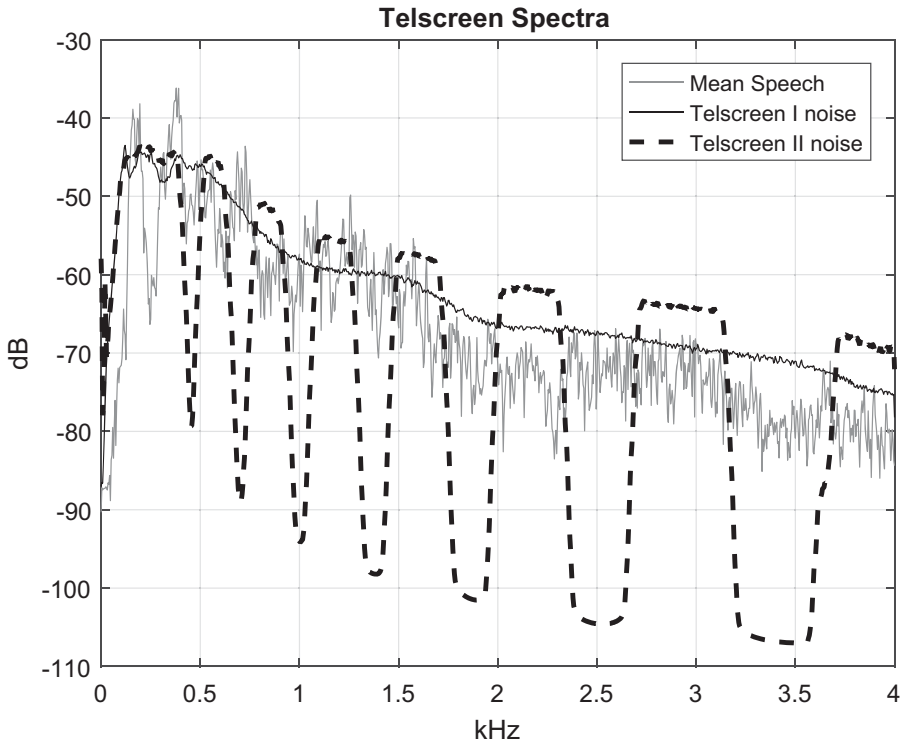

Figure 1. Long-term average spectra of the speech digits (measured after concatenation) and the masking noise used in Telscreen I and in Telscreen II, all plotted with the same long-term RMS level.

Masking noise

The masking noise was created by generating white noise, which was filtered to approximate the average spectrum of the triplets. The 81 digit triplets were mixed with the filtered white noise at 11 SNR levels ranging from −12 to +8 dB SNR, and the final set thus comprised 891 files. It can be seen from Figure 1 that the long-term average spectrum of the masking noise is similar to, but has slightly more high-frequency emphasis than, that of the concatenated speech digits.

Intelligibility matching

Forty-one normal-hearing participants, with pure-tone hearing thresholds ≤20 dB HL, were recruited from within NAL to complete Telscreen installed on a local computer, with a telephone handset connected to the soundcard. To emulate how the test would be delivered 'in the real world', the volume level at NAL was fixed to a level subjectively adjusted by the experimenters to be typical of telephone output levels. The response to every triplet was recorded and used to fit individual psychometric functions (the logistic function, with slope and offset as free parameters) for each of the 27 digits. The SRT_n (defined as the SNR at which each digit was correct 67% of the time) varied from −20.6 dB (for 'six' in the third position) to −5.8 dB (for 'two' in the second position). For most of the remaining digits, SRT_n was in the range −8 to −12 dB. The level of each digit was adjusted by an amount equal to the difference between its individual SRT_n and the average of all the digit SRT_n values, which was −10.3 dB.

Adaptive algorithm and Monte Carlo simulation

The adaptive test targeted the SNR at which each digit was correctly perceived 66.6% of the time. This value was chosen as it is close to the mid-point, and hence steepest point, of the psychometric function for each digit (bounded by 100% and the chance level of 10%, or 11% if subjects deduce that the digit 'seven' is never used). Conveniently, when three digits are presented, the current SNR can be estimated as being below, at, or above the 67% SRT_n target depending on whether the number of digits correct is less than two, equal to two, or three, respectively. The appropriate adaptation for each of these outcomes is to increase, leave unchanged, or decrease the SNR by one step, respectively. The step size was 4 dB during the familiarization phase, and the first digit triplet was presented at a SNR of 4 dB. The familiarization phase continued until the number of digits correct was two or less on two different trials. The test phase then commenced.

During the test phase, the step size was reduced to 2 dB. This phase continued for a maximum of 24 digit-triplets within the test phase or until the variability in test performance, measured by estimating the standard error of measurement of the SNRs presented during the test phase, was <1.0 dB. This accuracy-related termination condition means that the test is as short as possible for every caller, and those that are more consistent experience fewer trials. At the end of the test phase, the system calculated the SRT_n as the average of the SNRs presented during the test phase.

The decision to adopt an algorithm that targeted the 67% point on the digit psychometric function instead of the 50% point on the word psychometric function was based on theoretical principles and Monte Carlo modelling, as was the method used to estimate the standard error of measurement. Figure 2 shows a psychometric function for digits with a 67% SRT_n at −10 dB and a mid-point slope of 10% per dB, typical of the psychometric functions for the digits in Telscreen I. The probability of getting the complete triplet correct at any SNR is equal to the probability cubed of

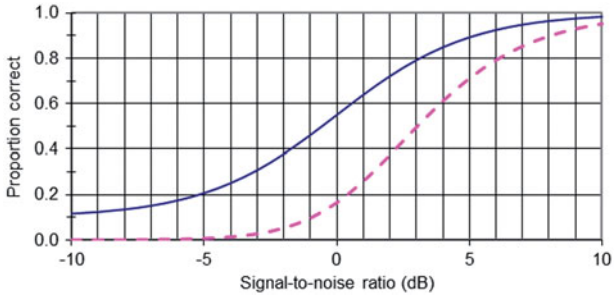

Figure 2. Psychometric functions for individual digits (solid line), with a mid-point slope of 0.1 dB^{-1}, and digit triplets (dashed line) as used in the Monte Carlo modelling.

getting a single digit correct. This is also shown in Figure 2. Inevitably, the triplet psychometric function is steeper than the digit function. However, each trial is based on a single right/wrong outcome in the case of the triplet function, but on the average of three right/wrong outcomes in the case of the digits. These psychometric functions were used as the basis of Monte Carlo modelling of an adaptive algorithm with 2-dB up/down step sizes, and 20 trials per measurement, commencing at the anticipated SRT_n. These runs were completed 500 times, and the mean and standard deviation of the SRT_n of all the runs were computed. When the conventional algorithm of increasing and decreasing SRT based on the entire triplet being correct was modelled, the standard deviation of SRT_n values was 0.93 dB. For the algorithm described earlier that targeted the 67% point on the digit function, the standard deviation was 0.79 dB, a small increase in precision which was expected.

Trial-by-trial SNR values in the same model were used to calculate the standard error of measurement, obtained by first calculating the standard deviation of all the SNR values presented during the test phase, and then dividing by the square root of the number of trials. If the SNR values for each trial are truly independent of each other, this value would be an unbiased estimator of the standard error of the SRT_n measurement. However, the SNRs of each trial in an adaptive paradigm are not independent, because the SNR presented in any trial cannot be more than one step size away from the SNR presented in the preceding trial. Because of this lack of independence, the Monte Carlo modelling showed that the calculated standard error of the means, on average, equalled 2.02 times the standard deviation of the actual mean SRT_n values. Consequently, when the algorithm was implemented, the equation for standard error of measurement was estimated at twice the value obtained from the conventional calculation of standard error of the mean.

Participants for evaluation of Telscreen I

One hundred and five participants, aged from 22 to 86 years (mean age: 53 years) participated in the evaluation of Telscreen I. There were 60 males and 45 females. Hearing thresholds were less than 20 dB HL in the test ear from 0.5 to 8 kHz inclusive for 39 of the adults, and the remaining 66 had sensorineural loss in the test ear. For the group as a whole, four-frequency average hearing loss (4FAHL; average of 0.5, 1, 2, and 4 kHz) in the test ear ranged from 0 to 62.5 dB, with a mean of 21.2 dB. The 10th percentile 4FAHL was 2.5 dB and the 90th percentile was 47.5 dB.

Procedure

Participants completed PTA at 0.5, 1, 2, 4, and 8 kHz, administered using a standard clinical audiometer in a soundproof booth by a qualified audiologist. Over a four-week period, participants completed Telscreen at the lab using their office phone to dial into the IVR platform, and they also completed Telscreen on two further occasions using their home telephone. Each participant completed the test unilaterally. To simulate real-world conditions, participants were simply instructed to 'hold the telephone to their preferred ear'—i.e. the ear they usually use when talking on the telephone. It was assumed that the volume of the telephone handset, if adjustable, would be set at the user's usual level. All participants completed a hearing disability questionnaire (HDQ) which comprised nine items relating to ease of listening in various situations and the impact of hearing problems on their life (Lutman et al, 1987).

Data analysis

Participants' responses on the HDQ were coded according to the procedure outlined in Lutman et al (1987) and a final score out of 54 was calculated, with larger numbers indicating greater disability. The participants' Telscreen results (SRT_n) were accessed through the IVR provider and then correlated with the HDQ scores and the 4FAHL of the test ear.

Results

A home trial test and retest was available for 98 of the participants. The difference in SRT_n values were within ±2 dB for 85 cases, within ±3 dB for 92 cases, and within ±4 dB for 96 cases. The remaining two participants had differences of 5 dB and 9 dB, the latter of which was regarded as a spurious outlier and was discarded from further analysis. For the remaining 97 participants, the scores were highly correlated, $r = 0.76$, and the mean test-retest difference was 0.34 dB, with 95% confidence interval from −0.01 dB to +0.68 dB. This difference, representing a small improvement in performance from test to retest, bordered on significance in a paired t-test ($p = 0.053$). The standard deviation of test-retest differences was 1.71 dB. As this represents the difference of two tests with independent random measurement error, the inferred standard deviation of individual scores was $1.71/\sqrt{2} = 1.21$ dB.

Table 1 shows the correlation between the Telscreen I SRT_n averaged across the two home administrations, the SRT_n measured in the lab, the 4FAHL in the test ear and the disability questionnaire scores, as well as the means and standard deviations for each of these measures. While all correlations were highly significant ($p < 0.001$), the correlations between the SRT_n values (either home trial or lab trial) and each of the questionnaire results and hearing thresholds were disappointingly low.

Figure 3 shows the relationship between 4FAHL in the test ear and SRT_n, averaged across the two home administrations. Although the increase in SRT_n with increasing hearing loss is clear, there is a large overlap in SRT_n values between those with hearing thresholds better than 10 dB HL and those with hearing loss, even when hearing thresholds exceeded 40 dB HL.

Discussion

The standard deviation of test-retest differences of 1.21 dB slightly exceeded the target value of 1.0 dB on which the stopping rules of the adaptive algorithm were based. The difference presumably

Table 1. Correlation matrix for Telscreen I home trial, Telscreen I lab trial, HDQ, and 4FAHL.

	Mean	*Std. dev.*	*Home trial SRT_n*	*Lab trial SRT_n*	*HDQ*	*4FAHL*
Home trial SRT_n (dB)	−8.0	2.3	1.00	0.77	0.51	0.57
Lab trial SRT_n (dB)	−9.5	2.8		1.00	0.60	0.73
HDQ	11.4	12.2			1.00	0.81
4FAHL (dB)	20.8	17.6				1.00

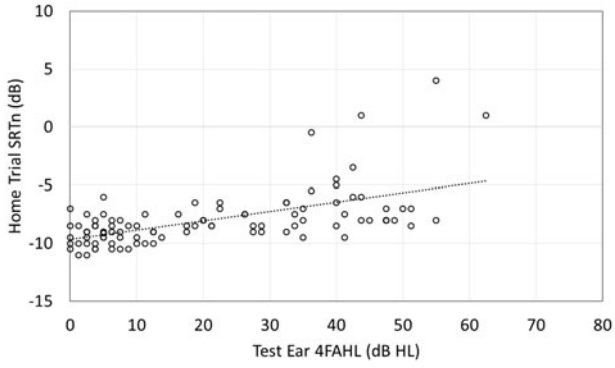

Figure 3. The relationship of Telscreen I SRT_n, averaged across two home-trial measurements, and 4FAHL.

arises because there are sources of measurement error other than those implicit in the adaptive measurement itself, and because some adaptive tracks terminated because the maximum allowable number of trials were presented rather than because the standard error of measurement criterion was met. The other sources of measurement error include changes in attention of the participants and the influence of background noise, if sufficient to rise above the level of masking noise provided by the test.

Although the correlation between SRT_n and both HDQ scores and hearing thresholds were highly significant, they were not as high as expected, nor were they as high as reported previously for a digit triplet test. Smits et al (2004) reported a correlation of 0.77 between the SRT_n measured with their triplet test and 4FAHL, compared to only 0.57 (with 95% confidence interval from 0.42 to 0.69) in this experiment. This difference may partly arise from the more restricted range of hearing loss in this experiment (10% to 90% range of 2.5 to 47.5 dB 4FAHL) compared to Smits et al (10% to 90% range from −1 to 59 dB 4FAHL).

The low correlation in this experiment in part arises from the spread of results in those with normal hearing. This spread is caused by both random measurement error and genuine differences amongst people with normal thresholds. For the 42 participants with 4FAHL ≤10 dB HL, the standard deviation of test-retest differences was 1.55 dB. This means the intra-participant standard deviation for a single test was 1.1 dB, and the standard deviation for the average of two tests would be 0.77 dB. By contrast, the inter-participant standard deviation for these same participants was 1.18 dB. Consequently, only 42% of the variance in their SRT_n values can be accounted for by random measurement error. The remaining 58% reflects true differences in understanding speech in noise, and it is quite possible that some people with a mild hearing loss could hear better than those with normal hearing thresholds. Such a result would not be surprising if recent findings in animal studies about damage to high-level nerve fibres, without affecting hearing thresholds, also applied to human hearing (Kujawa & Liberman, 2009).

Nonetheless, it was considered desirable to improve the ability of the hearing test, if possible, to differentiate those with mild pure-tone hearing loss from those with normal hearing. The approach taken was to change the masking noise in an attempt to make the degree of masking more dependent on the presence of hearing loss, as outlined in Experiment 2.

Experiment 2: Telscreen II development and evaluation

Telscreen II was created from Telscreen I by changing the masking noise, and as a consequence, changing the method of adapting the SNR.

Method

Speech recordings

The same 81 digit-triplet speech files recorded for Telscreen I were used in Telscreen II.

Masking noise

White noise, which was filtered to match the international long-term average speech spectrum (ILTASS; Byrne et al, 1994) with an extra 2 dB per octave slope added from 0.1 to 4 kHz was generated. The noise was then filtered so that every second critical band from 0.4 to 4 kHz was attenuated by 40 dB, and a 0.02 kHz sinusoidal modulation was applied with a modulation depth of 20 dB. The additional slope was applied to give additional low-frequency emphasis to the noise. The aim of this modification was to increase the masking of the low frequency parts of the stimuli, and hence make detection of the stimuli more dependent on high-frequency hearing ability. This was considered important because high-frequency hearing loss is by far more common, and because it is possible to identify most of the stimuli from the vowels alone, for which low frequency hearing ability is often sufficient. The spectral gaps in the masker can be seen in the spectrogram shown in Figure 4, and in the long-term average spectrum shown in Figure 1. Portions of the three target digits are most evident within these gaps visually in this representation, just as they are by audition. The temporal modulations of the masker can just be discerned in the spectrogram, but are more evident in the waveform shown in Figure 5.

In Telscreen I, the variation in SNR was created by combining digit triplets of fixed intensity with masking noise of variable intensity. Because of the spectral and temporal gaps in Telscreen II, preliminary measurements indicated that people with normal hearing could obtain 66% of digits correct at SNRs around −20 dB. When the target stimuli in quiet were at a comfortable level, the combined stimulus often produced an uncomfortably loud sound. Consequently, the algorithm was changed so that SNR was varied by adjusting the target level rather than the noise level.

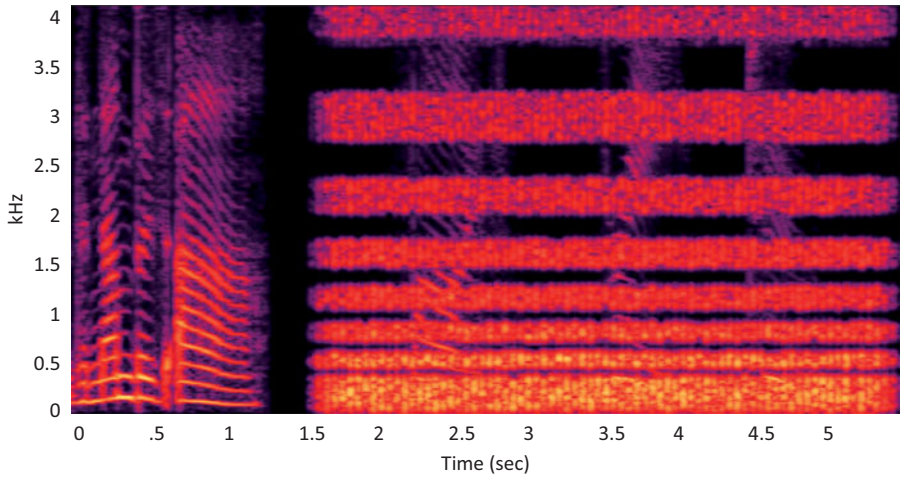

Figure 4. Spectrogram of the target material 'The numbers are five, three, two' embedded in the temporally modulated and spectrally serrated masking noise with a SNR of −14 dB.

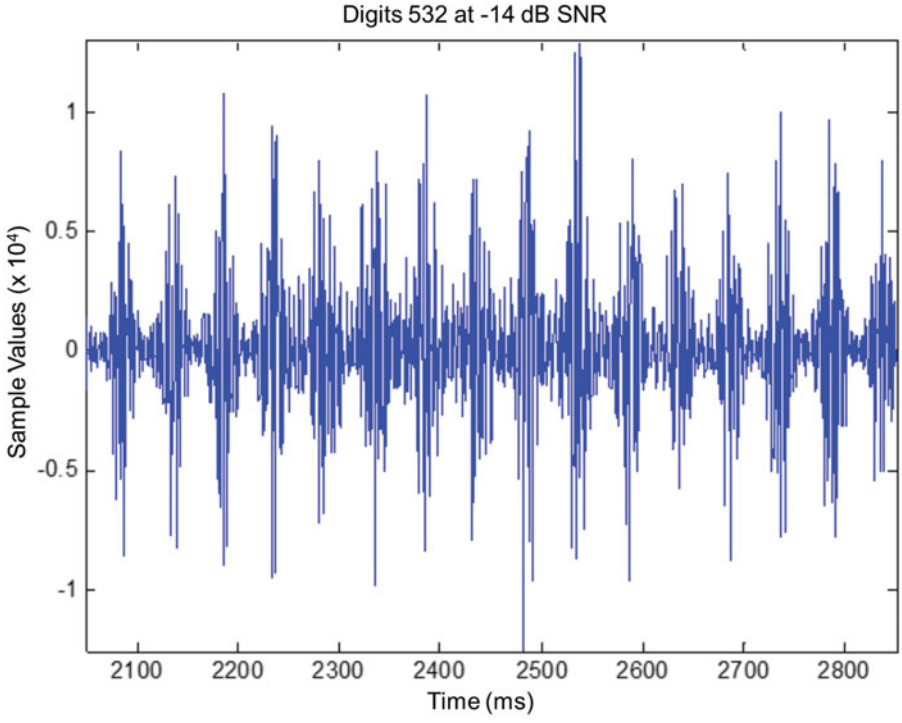

Figure 5. Waveform of the target and masker material from 2.1 to 2.8 seconds for the stimulus item shown in Figure 4.

Each of the 81 digit triplets were scaled to varying RMS levels and then mixed with the new masking noise to create files at 17 SNR levels ranging from −28 to +4 SNR. The final set thus comprised 1377 files. These were provided to the communications company to replace the previous sound files on the IVR platform.

Algorithm

The same test-flow algorithm used in Telscreen I was used in Telscreen II.

Participants

Seventy-five participants participated in the evaluation of Telscreen II. Sixty of these (22 with 4FAHL in the test ear <20 dB HL and 38 with 4FAHL >20 dB HL) had completed the Telscreen I evaluation six weeks earlier. The group was aged from 25 to 86 years (mean age: 61 years) and there were 29 males and 46 females. Hearing thresholds were <20 dB HL in the test ear from 0.5 to 8 kHz inclusive for 25 of the participants. The 4FAHL in the test ear ranged from 0 to 76.3 dB, with a mean of 31.7 dB. The 10th

Table 2. Correlation matrix for Telscreen II trials, HDQ, and 4FAHL.

	Mean	*Std. dev.*	*Home trial SRT_n*	*HDQ*	*4FAHL*
Home trial SRT_n (dB)	−14.3	4.6	1.00	0.65	0.77
HDQ	17.9	13.1		1.00	0.65
4FAHL (dB)	31.7	20.1			1.00

percentile 4FAHL was 3.8 dB and the 90th percentile was 53.5 dB. Results from an additional two adults were collected but were not recoverable from the Telscreen database, and were therefore excluded from further analysis.

Procedure

Participants were asked to complete two home-based trials of Telscreen II using their home telephone. Those who had not already participated in Experiment 1 completed the HDQ and underwent PTA at 0.5, 1, 2, 4, and 8 kHz, which was performed in a soundproof booth by a qualified audiologist to confirm hearing status.

Data analysis

The participants' Telscreen results (SRT_n) were accessed through the IVR provider. All but four of the participants had completed the test twice. A paired t-test showed no significant difference in the two test results for the 71 participants who had completed the test twice, $t(70) = 0.47$, $p > 0.6$, so for each person, an average result was calculated and this was correlated with the 4FAHL of the test ear as well as their score on the disability questionnaire (out of 54). Criteria for passing and failing the test were devised from the correlation graph.

Results

A significant relationship was evident between the Telscreen II results and the 4FAHL ($r = 0.77$, $p < 0.001$) as shown in Table 2 and Figure 6. The correlation was also significant between the Telscreen II results and the HDQ score ($r = 0.65$, $p < 0.001$). Importantly, the correlations obtained were stronger than those found for Telscreen I, demonstrating that the novel stimuli used in Telscreen II produced superior performance over Telscreen I in distinguishing between individuals with normal hearing and those with hearing disability. The *increase* in correlation coefficient was significant in the case of correlation with 4FAHL ($p = 0.02$), but not in the case of correlation with the HDQ score ($p = 0.18$). The pass/fail criteria were set at: Pass: $SRT_n \leq -18$ dB; Fail (or 'Refer'): $SRT_n > -16$ dB; Borderline: SRT_n between −16 and −17 dB. Using 4FAHL >20 dB HL as the measure of hearing impairment, the sensitivity was 90% and the specificity was 90.2%.

Discussion

Telscreen I and II, the Australian telephone-based hearing screening tests, were developed using comprehensive evaluation procedures at every stage of the process. Telscreen I was made publicly available from 4 September 2007, but from January 2008 it was replaced by Telscreen II. Both versions of Telscreen included several novel features which set it apart from other similar tests developed in The Netherlands and the UK, including equal intelligibility of digits, individual scoring of digits, and real-time assessment of accuracy. The distinguishing feature of Telscreen II is the novel masking noise that takes advantage of differences in the ability of those with and without a hearing loss to process frequency and temporal information. The inclusion of this noise in Telscreen II resulted in a better correlation between the test results and hearing disability (self-reported) and hearing impairment (4FAHL), with correlation results similar to those found for other similar hearing screening tests. For example, the 0.77 correlation found here between the test results and the 4FAHL is comparable to the 0.73 and 0.77 correlations reported for the Dutch test results with 3FPTA and 4FPTA respectively (Smits et al, 2004), and higher than the .64 and .66 correlations with left and right ear 4FPTA respectively that were reported for the American English-language telephone hearing test (Williams-Sanchez et al, 2014).

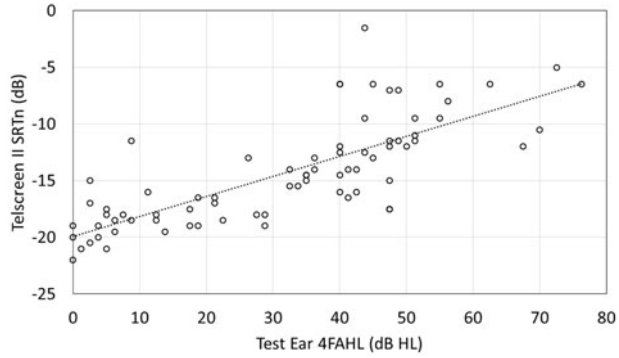

Figure 6. The relationship between the SRTn for Telscreen II and the test ear 4FAHL.

The fact that Telscreen I was only moderately correlated with hearing disability and impairment (4FAHL) was unexpected, given the similarity in masking noise used in Telscreen I and the other similar screening tests. A possible reason relates to the use of correlation coefficients to indicate the degree of agreement between the two data sets. Although widely used, correlation measures depend to a large extent on the range and distribution of the data involved. In the Telscreen I dataset, the 10%–90% range of 4FAHLs was 2.5–47.5 dB HL, while in the Dutch study, the 10%–90% range of 4FAHLs was −1 to 59 dB HL (Smits et al, 2004).

A caveat to the comparison of our Telscreen I and Telscreen II results is that the two samples, although very similar, were not identical. The 10th to 90th percentile range of hearing losses in the Telscreen II study was 5 dB greater than in the Telscreen I study. Nevertheless, both samples were considered to be representative of the population most likely to use the Telscreen service, i.e. people with hearing thresholds from normal through to a moderate loss. It was assumed that those with more severe losses would be aware of their loss and have no need for a screening test. Our participants were deliberately recruited to cover the expected range from normal to moderate loss, and in fact, the loss profile of our samples appears to be remarkably similar to that of actual

callers to Telscreen. In the first nine months of operation, 36% of Telscreen callers passed, and the remainder failed or returned borderline results. Of the participants reported here, 35% had thresholds <20 dB HL, and the remainder had thresholds between 20 and 76.25 dB HL.

Differences in intelligibility of the individual digits could arise from differences in clarity of articulation, differences in the extent to which each digit is acoustically similar to any other digit, and differences in the extent to which the spectrum of each digit is masked by the spectrum of the noise. The combined effects of these reasons for differences were removed by level adjustments within the development of Telscreen I. However, these same level adjustments were re-applied in Telscreen II, the noise of which had a different spectrum, so it is likely that the stimuli were not exactly equally intelligible in Telscreen II. This offers the potential to further improve the precision of Telscreen II. Another option for improving Telscreen II relates to the masking noise. Although it was our intention to increase low-frequency masking, as Figure 1 shows, this was not entirely successful likely because the spectrum used to shape the Telscreen I masking noise differed somewhat from the ILTASS that we used to shape the Telscreen II masking noise. If the low-frequency component of the masking noise was increased further, we believe Telscreen II could be an even more effective hearing screening test, particularly for those with high-frequency hearing loss.

Since 2006–2008, other speech-in-noise tests have been developed, some of which incorporate some of the novel features included in Telscreen. For example, the speech-in-noise tests developed in France matched each digit for intelligibility (Jansen et al, 2010), and in The Netherlands, the developers of the original Dutch screening test have explored the use of a temporally modified masking noise. Although it has not been incorporated into a telephone test, the use of masking noise, 'interrupted' at 16 kHz, has been found to successfully discriminate between listeners with impaired and normal hearing (Smits & Houtgast, 2007). The most recent development in this area is the shift to using the internet and mobile phones as the platforms for the test. The earliest examples of an internet-based test were the Dutch National Hearing Test (Smits et al, 2006), and 'Earcheck' also from The Netherlands, the latter specifically designed to identify noise-induced hearing loss (Leensen et al, 2011). The developers of this test trialled several different types of masking noise types, and found that a temporally modulated low-pass filtered masking noise greatly improved the specificity of the test (Leensen et al, 2011). A UK-based internet test (www.actiononhearingloss.org.uk) is available on both the internet and as a mobile app; and in 2012 Telscreen II was adapted for internet use for the 'Sound Check Australia' project, and is now permanently available online at knowyournoise.nal.gov.au.

Future directions

One of the challenges facing hearing screening in Australia is the need to cater for the large number of people from non-English speaking backgrounds wishing to use the service. The need for multilingual capability also arises in Europe, and this has been addressed by the development of screening tests in multiple languages in which speech intelligibility is optimized between tests (Zokoll et al, 2012). In Australia, a different approach has been taken whereby the digit triplets have been replaced by tone pulses, thus eliminating the need to develop multiple sets of speech stimuli. The caller presses any button on the phone keypad when a test sound is heard in the presence of the masking noise. The multilingual capacity of the tone-based version Telscreen stems from its language-independent sound files that can be coupled with a set of instructions in any language, resulting in a universal hearing screening test that currently covers eleven of the most commonly spoken languages in Australia. Details of the development of this test (Telscreen III) will be published separately.

Conclusions

Over the past decade, the demand for efficient, reliable, self-administered hearing screening has grown as governments and hearing service providers seek inexpensive tests that can quickly identify those with hearing problems. An ageing population, an increase in access to hearing health in developing countries, and the spread of online and mobile technologies will ensure that the demand continues to grow in the foreseeable future. The innovative approach and sound principles underlying Telscreen should ensure its adaptability to new modes of delivery and thus its viability should be maintained even as the traditional telephone approaches obsolescence.

Acknowledgements

NAL was commissioned by Australian Hearing to produce Telscreen. NAL is funded by the Australian Government Department of Health.

Declaration of interests: The authors declare no conflicts of interest.

References

Byrne D., Dillon H., Tran K., Arlinger S., Wibraham K., Cox R. et al. 1994. An international comparison of long-term average speech spectra. *J. Acoust Soc Am*, 96, 2108–2120.

Dillon H. 1983. The effect of test difficulty on the sensitivity of speech discrimination tests. *J Acoust Soc Am*, 73, 336–344.

Grandori F., Parazzini M., Tognola G. & Paglialonga A. 2009. *Hearing screening in older adults is gaining momentum-The European project AHEAD III on adult hearing.* Paper presented at the Proceedings of the 2nd Phonak International Adult Conference: Hearing care for adults.

Jansen S., Luts H., Wagener K.C. & Frachet B. 2010. The French digit triplet test: a hearing screening tool for speech intelligibility in noise. *Int J Audiol*, 49, 378–387.

Kramer S.E., Kapteyn T.S. & Festen J.M. 1998. The self-reported handicapping effect of hearing disability. *Audiology*, 37, 302–312.

Kujawa S.G. & Liberman M.C. 2009. Adding insult to injury: cochlear nerve degeneration after ''temporary'' noise-induced hearing loss. *J Neurosci*, 29, 14077–14085.

Larsby B. & Arlinger S. 1999. Auditory temporal and spectral resolution in normal and impaired hearing. *J Am Acad Audiol*, 10, 198–210.

Leensen M.C.J., de Laat J.A.P.M., Snik A.F.M. & Dreschler W.A. 2011. Speech-in-noise screening tests by internet, part 2: improving test sensitivity for noise-induced hearing loss. *Int J Audiol*, 50, 835–848.

Lutman M., Brown E.J. & Coles R.A. 1987. Self-reported disability and handicap in the population in relation to pure-tone threshold, age, sex and type of hearing loss. *Br J Audiol*, 21, 45–58.

Smits C. & Houtgast T. 2006. Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. *J Acoust Soc Am*, 120, 1608–1621.

Smits C. & Houtgast T. 2007. Recognition of digits in different types of noise by normal-hearing and hearing-impaired listeners. *Int J Audiol*, 46, 134–144.

Smits C., Kapteyn T.S. & Houtgast T. 2004. Development and validation of an automatic speech-in-noise screening test by telephone. *Int J Audiol*, 43, 15–28.

Smits C., Merkus P. & Houtgast T. 2006. How we do it: the Dutch functional hearing–screening tests by telephone and internet. *Clin Otolaryngol*, 31, 436–455.

Williams-Sanchez V., McArdle R., Wilson R., Kidd G., Watson C.S. & Bourne A.L. 2014. Validation of a screening test of auditory function using the telephone. *J Am Acad Audiol*, 25, 937–951.

Zokoll M.A., Wagener K.C., Brand T., Buschermöhle M. & Kollmeier B. 2012. Internationally comparable screening tests for listening in noise in several European languages: the German digit triplet test as an optimization prototype. *Int J Audiol*, 51, 697–707.