Original Article

Trends in Hearing
2015, Vol. 19: 1–10

DOI: 10.1177/2331216515583138
tia.sagepub.com
SAGE

Optimization of the Dutch Matrix Test by Random Selection of Sentences From a Preselected Subset

Rolph Houben[1] and Wouter A. Dreschler[1]

Abstract

Matrix tests are available for speech recognition testing in many languages. For an accurate measurement, a steep psychometric function of the speech materials is required. For existing tests, it would be beneficial if it were possible to further optimize the available materials by increasing the function's steepness. The objective is to show if the steepness of the psychometric function of an existing matrix test can be increased by selecting a homogeneous subset of recordings with the steepest sentence-based psychometric functions. We took data from a previous multicenter evaluation of the Dutch matrix test (45 normal-hearing listeners). Based on half of the data set, first the sentences (140 out of 311) with a similar speech reception threshold and with the steepest psychometric function ($\geqslant$9.7%/dB) were selected. Subsequently, the steepness of the psychometric function for this selection was calculated from the remaining (unused) second half of the data set. The calculation showed that the slope increased from 10.2%/dB to 13.7%/dB. The resulting subset did not allow the construction of enough balanced test lists. Therefore, the measurement procedure was changed to randomly select the sentences during testing. Random selection may interfere with a representative occurrence of phonemes. However, in our material, the median phonemic occurrence remained close to that of the original test. This finding indicates that phonemic occurrence is not a critical factor. The work highlights the possibility that existing speech tests might be improved by selecting sentences with a steep psychometric function.

Keywords

speech-in-noise, speech test, speech intelligibility, matrix test, psychometric function

Introduction

Speech-in-noise testing is a powerful tool for both clinical audiology and audiological research. The results allow the determination of a patient's speech perception ability and can help determine the potential benefits of hearing aids or cochlear implants. To assess such benefits with speech tests, one needs tests that are able to detect relevant differences in perception. Improving the characteristics of speech-in-noise tests could lead to better (e.g., more precise or quicker) evaluation of the benefit of hearing aids when used in a noisy environment. Because the development of a new speech-in-noise test takes a large effort, it would be beneficial if it were possible to improve the characteristics of an existing test.

Here, the focus is on closed-set speech materials as used in matrix-type speech tests. Matrix-type speech tests are sentence-in-noise tests that use sentences of identical grammatical structure in which all available words are taken from a closed set of alternatives. An example sentence is "Mark gives five large flowers." The sentences of the matrix test are syntactically fixed (name + verb + number + adjective + objective) but semantically unpredictable. To obtain a reliable measurement, several sentences (e.g., 20) need to be used, and therefore, sentences are grouped to form test lists. Such a test list is used to perform a single measurement of speech recognition.

The matrix test was originally developed by Hagerman (1982) for Swedish and is now available in many languages (e.g., German, Danish, British English, Polish, French, Russian, Spanish, American English, and

[1]Clinical and Experimental Audiology, Academic Medical Center, Amsterdam, The Netherlands

Corresponding author:
Rolph Houben, Clinical and Experimental Audiology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands.
Email: A.C.Houben@amc.uva.nl

Turkish), including Dutch (Hochmuth et al., 2012; Houben et al., 2014; Jansen et al. 2012; Ozimek, Warzybok, & Kutzner, 2010; Vlaming et al., 2011; Wagener, Brand, & Kollmeier, 1999; Wagener, Josvassen, & Ardenkjaer, 2003; Zokoll et al., 2013).

A psychometric function describes the relationship between a physical stimulus (e.g., the signal-to-noise ratio [SNR]) and the performance of a participant (e.g., speech recognition). For a speech-in-noise test, the slope of the psychometric function quantifies the quotient between a change in speech recognition and a change in SNR. A steeper slope of the psychometric function is desirable because steeper slopes allow more accurate estimates of speech recognition (Kollmeier & Wesselkamp, 1997). For the Dutch matrix test, the slope of the psychometric function is 10.2%/dB with a standard deviation over lists of 0.9%/dB (Houben et al., 2014). This slope seems acceptable if it is compared with the slope of speech tests in general: MacPherson and Akeroyd (2014) found a mean slope of 8.5%/dB (with a range of 1–44%/dB) for tests with a single stationary noise masker. However, the slope of the Dutch matrix test is slightly shallower than the slope of matrix tests in other languages that have a range of about 13%/dB to 17%/dB (Houben et al., 2014).

The reason that the slope of the Dutch matrix test is shallower than the slope of other matrix tests with a similar noise masker is likely due to either language properties (Wagener et al., 2003) or to the way the sentences were pronounced and recorded. First, salient speech cues differ between languages and may result in a different robustness against noise. These salient speech cues can occur differently in different sentences and words and this, in turn, can lead to differences in slopes across languages. Second, differences in slope between speech tests can occur, at least in part, due to differences in speaking characteristics during recording (e.g., speed, prosody, timing, and articulation). Hood and Poole (1980) have shown that *easy* words can become *difficult* when pronounced by a different speaker. Even though the matrix test was spoken by a single speaker, it is likely that this pronunciation effect also holds for the stimuli of the matrix test: Some recorded sections will be more difficult than others.

To illustrate that sentences differ, Figure 1 shows the distribution of the sentence-based slopes of the psychometric function for all sentences of the Dutch matrix test. The distribution is skewed to the right as was also found by MacPherson and Akeroyd (2014).

An important step in the development of matrix tests is to equalize the intelligibility of each word of a sentence. This equalization is required because the psychometric function of the sentence is steepest if the individual words of the sentence have about the same

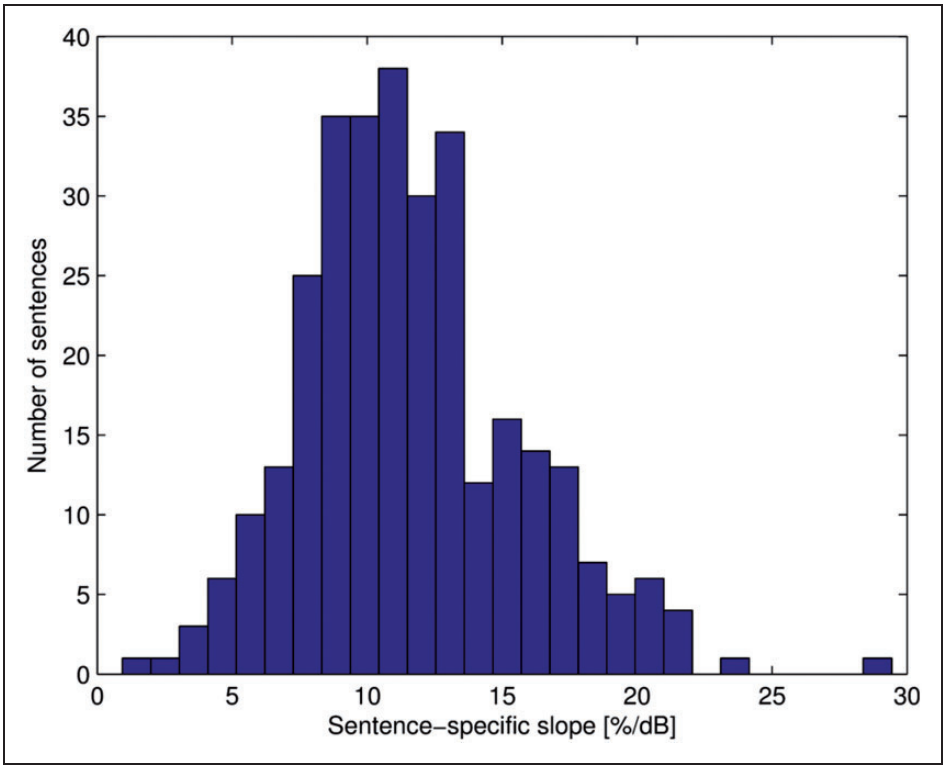

Figure 1. Distribution of the sentence-specific slopes for all 311 available Dutch matrix sentences.

intelligibility. However, the level corrections used for equalization are limited to prevent an unnatural prosody of a sentence. For the Dutch matrix test, the maximum correction was limited to 3 dB. For sentences that required a correction larger than 3 dB, this limitation might have contributed to a shallow psychometric function.

Given a set of speech material recorded by a specific speaker in a specific language, the inherent differences between languages as mentioned earlier cannot be optimized easily. However, there might be room for improvement with respect to differences in speaking characteristics that occurred during the recording session. In this article, an attempt is made to optimize the Dutch matrix test by selecting the recordings that had the highest sentence-specific psychometric slopes. To achieve this optimization, data from the previously published Dutch Matrix evaluation study (Houben et al., 2014) were used. From these data, the recordings with the steepest sentence-specific psychometric functions were selected. Then, the resulting list-specific psychometric functions were calculated, and it was determined if the slope had increased by the selection process.

Matrix tests are developed such that the occurrence of phonemes in each list mirror reasonably well that of the language the test is derived from. To ensure that our selection process did not lead to a subset of sentences with a strongly deviating phonemic occurrence relative to the original base matrix, the phonemic occurrence was compared with the occurrence of both the original matrix test and to that of Dutch reference corpora.

Methods

The Data Set

For the Dutch matrix, one single validated recording is available for each of the 311 unique sentences. The test additionally contains a stationary speech-shaped noise that has the same long-term average spectrum as the sentences (Houben et al., 2014). All materials were evaluated through listening tests with normal-hearing listeners in a multicenter study (Houben et al., 2014). Three centers participated in that study, and each center recruited 15 local normal-hearing adults. The participants were recruited from outside the departments (students and coworkers who responded to a call for participation). The participants reported no otological problems, and their hearing thresholds did not exceed 20 dB HL at each octave frequency between 250 Hz and 8 kHz. In the study, speech recognition was measured at SNRs of −5 dB, −7 dB, and −9 dB, with the noise level at 70 dB SPL. The sentences were balanced across the SNRs so that each subject listened to every sentence only once. For each sentence, 45 measurements are available (15 per SNR). The order of presentation of the SNRs was also balanced across the subjects and sentences to minimize order effects on the slope. The outcome measure of a matrix test is the speech reception threshold (SRT): the SNR where 50% of the words are correctly repeated. In the evaluation study, the SRT and the slope of the psychometric function were determined by fitting a logistic model, in a similar way as was done by MacPherson and Akeroyd (2014). 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 listener correctly repeated the sentence. The resulting SRT in noise was −8.4 dB SNR with an interlist standard deviation of 0.2 dB. The list-specific slope of the psychometric function was 10.2%/dB with an interlist standard deviation of 0.9%/dB (Houben et al., 2014).

Improving the Slope of the Psychometric Function

In an attempt to increase the steepness of the list-specific psychometric function, a relatively homogeneous set of the sentences was selected that had the steepest sentence-specific slope.

To obtain reliable results, this selection was made by applying twofold cross-validation (Steyerberg, 2008, chap. 17). In twofold cross-validation, the data are partitioned into two complementary subsets: one subset was used to select the sentences (training subset) and the other (nonoverlapping) subset was used to test if the sentence selection increased the list-specific slope (validation subset). This partitioning of the data prevents an overestimation of the improvement (Steyerberg, 2008). The training subset consisted of the measurement data of subjects 1 through 7 of each center (in total 21 subjects). The validation subset was filled with the data of subjects 8 through 15 of each center (in total 24 subjects).

From the training subset, a homogeneous set of sentences was selected whose SRT differed less than 1.0 dB from the mean SRT ($-9.4\,\text{dB} < \text{SRT} < -7.4\,\text{dB}$). Figure 2 shows the distribution of the slopes of the sentences from the training set whose SRT differed less than 1.0 dB from the mean. Note that the data underlying Figures 1 and 2 differ in that for Figure 1 all data were used, whereas for Figure 2 only the data from the validation subset were used.

To avoid memorization of the sentences by the subjects and to be able to generate different lists, the goal was to obtain as many sentences as possible with a minimum of about 100 sentences. However, the number of sentences that could be included was limited, because low-slope sentences needed to be excluded. Based on these two constraints, a slope of 9.7%/dB was found to split the set of sentences successfully. A set of 140 sentences was obtained in which each sentence had a slope

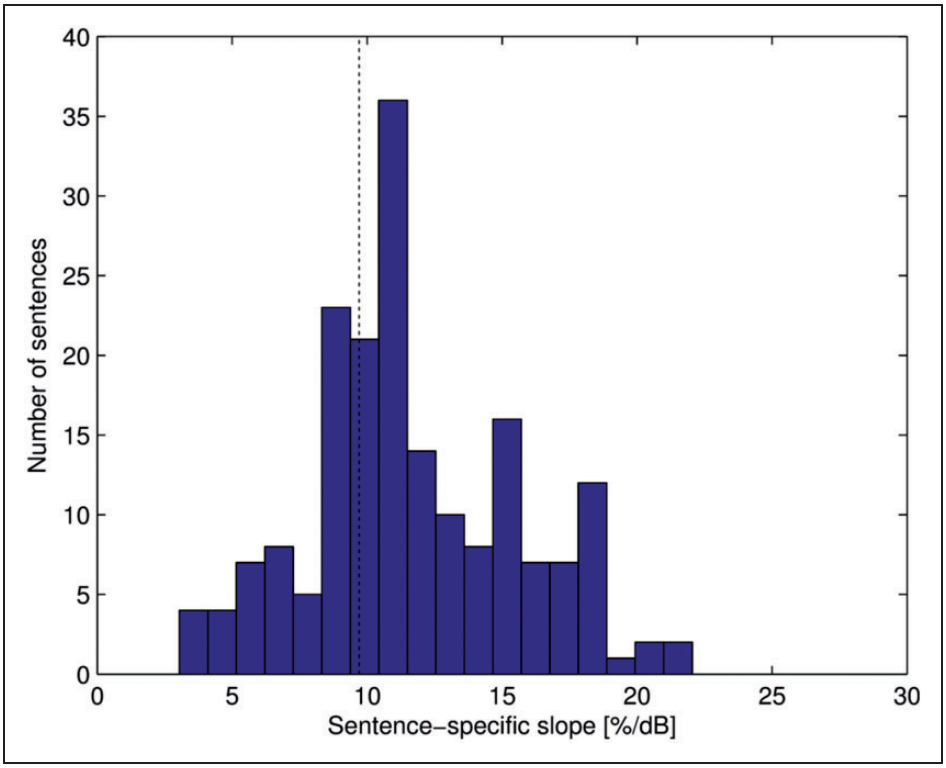

Figure 2. Distribution of the sentence-specific slopes for the sentences whose SRT was within 1 dB of the mean. Data are based on the training set only. The striped line indicates the cut-off value of 9.7%/dB.
Note. SRT: speech reception threshold.

of 9.7%/dB or steeper. The selected sentences are provided as Supplementary Data.

It is important that specific words do not occur too frequently in the subset, as recurring words would make the matrix test too predictable when used repeatedly with the same participant. For our set of 140 sentences, a single word occurred at most twice the times that it occurred in the original set of 311 sentences.

A comparison of the discarded sentences to the selected sentences showed no clear word pattern that distinguishes the discarded from the selected sentences. Both sets contain all 50 words of the base matrix.

As noted earlier, the psychometric function of a sentence is steepest if the individual words of the sentence have about the same intelligibility. Thus, for the steepness of the sentence-specific slope, the differences in word scores within a sentence are relevant. To investigate the differences in word intelligibility within a sentence, the standard deviation of the measured word recognition scores was calculated. This was done on the validation subset, and the results are therefore independent from the training subset that was used to select the sentences. The median standard deviation of the word recognition score was statistically significantly higher for the discarded set (24.0%) than for the selected set (17.5%; Wilcoxon rank-sum test: $Z = 5.4$, $p < .001$, two sided).

Test Procedure

In regular use of the matrix test, all sentences are arranged into fixed sentence lists. These phonemically balanced lists of 20 sentences contain each word exactly twice. In a balanced list, it is not possible to replace a sentence by a different sentence without reducing the degree of phonemic balance because the replacement leads to some words to occur more than twice and some less. From the selected 140 sentences, it was not possible to generate enough (>5) new balanced test lists. An alternative way to use matrix materials, besides using them to form balanced test lists, would be to draw sentences at random during testing. To do this, the testing procedure was changed from using fixed-test lists to random sampling of the sentences from the subset. This procedure will be called *subset-based random sampling*.

Calculation of the Slope of the Psychometric Function

To verify if the selection of sentences increased the slope of the psychometric function, the previously unused validation data were used to calculate the list-specific SRTs and slopes for the selected 140 sentences. This was done by using subset-based random sampling instead of by using fixed-test lists, as described earlier. Random lists

of 20 sentences were repeatedly (1,000 times) drawn from the 140 sentences. The method used was sampling with replacement, that is, after each draw of 20 sentences, and the sentences were returned to the pool after which the next 20 sentences were drawn. For each random list, the list-specific slope and SRT were calculated as if that list was directly measured (see Houben et al., 2014). These calculations were done with the logistic model described earlier. Data were available for 24 subjects (validation set) who listened to a specific sentence only once (at one of the three SNRs).

Calculation of the Phonemic Occurrence

While balanced tests lists are specifically designed to have a phonemic distribution close to that of the reference corpora, random selection from a subset does not. Because it is important that the phonemic occurrence of the test remains close to that of the reference corpora, the phonemic distribution was determined for subset-based random sampling and compared that with the available reference corpora.

Results

Psychometric Function

For subset-based random sampling, the calculated mean list-specific slope was 13.7%/dB with a standard deviation over lists of 0.9%/dB. The median slope was 13.0%/dB. Figure 3 shows the distribution of the list-specific slope of the 1,000 drawn lists. The mean slope was 3.5%/dB steeper than the list-specific slope of the original set (10.2%/dB). The SRT for subset-based random sampling was −8.4 dB with a standard deviation of 0.1 dB. This SRT is the same as for the original test.

Phonemic Balancing

A possible drawback of subset-based random sampling is that the phonemic occurrence in the drawn set can deviate from that of the original base matrix and also from the natural language.

To obtain valid results, the occurrence of phonemes in each list of the original matrix test mirrored that of standard Dutch. Figure 4 shows the phoneme occurrence in the original full matrix test (Houben et al., 2014). For comparison, the phoneme occurrence in six Dutch reference corpora is also shown. The most recent Dutch reference data stem from a large survey by Luyckx, Kloots, Coussé, and Gillis (2007). They used nearly 900,000 spoken words to measure the distribution of phonemes. Luyckx et al. (2007) make the distinction between northern and southern Dutch. This is the only available reference corpus that makes this distinction. Since the distribution of phonemes between northern Dutch and southern Dutch is very similar, and since there was no difference in slope found for

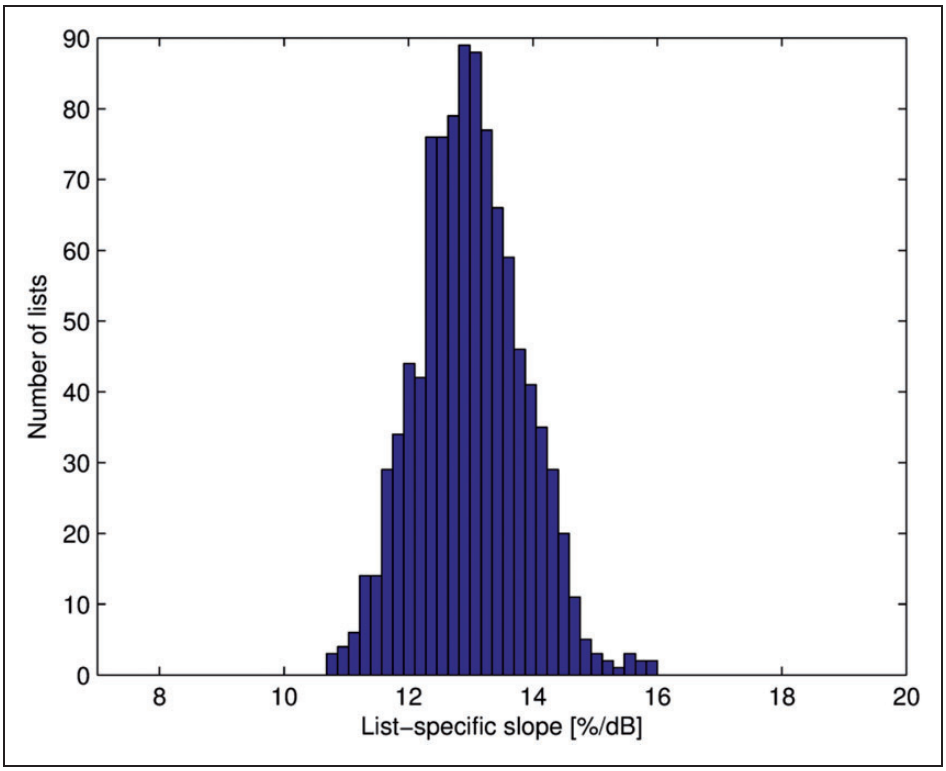

Figure 3. Distribution of the list-specific slopes for the 1,000 randomly drawn lists. Data are based on the validation subset only.

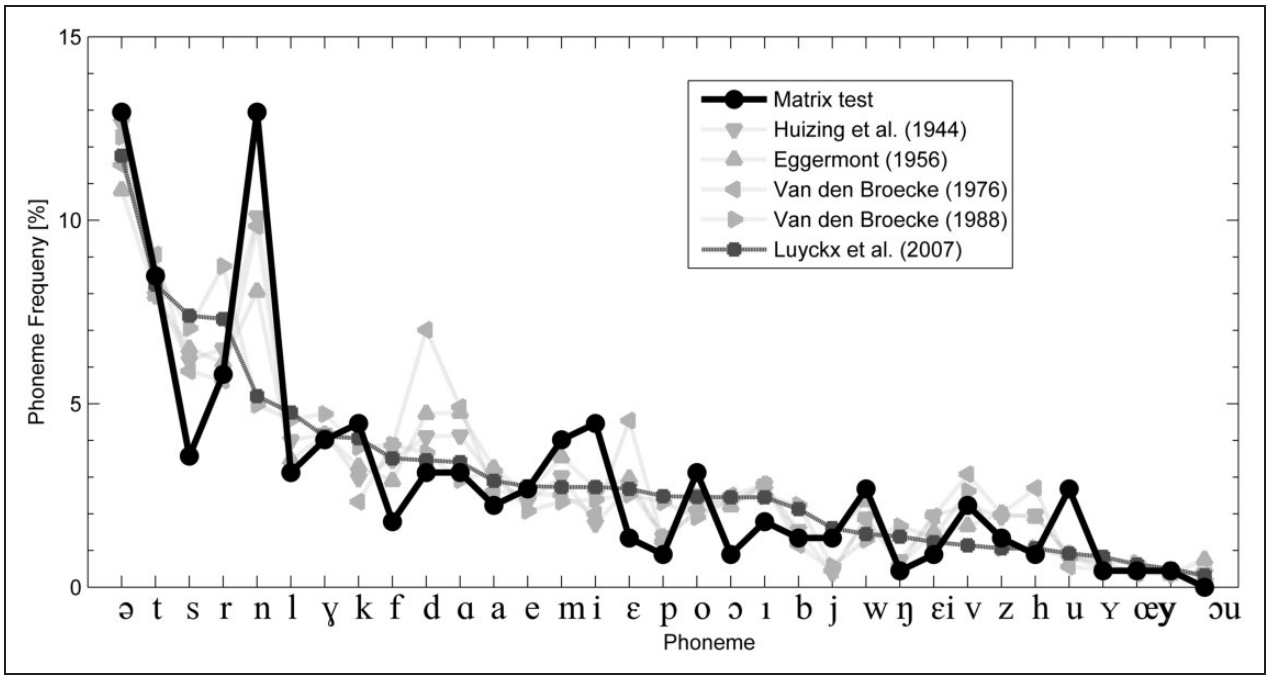

Figure 4. Phoneme distribution for the original matrix test and for five reference corpora for Dutch. Phonemes are ordered according to the phoneme occurrence reported by Luyckx et al. (2007).

the matrix test measured in a northern and southern region (Houben et al., 2014), we do not make that distinction here and use the data for northern Dutch only. The matrix test was recorded with a neutral speaker who originated from a Northern area that lies close to the border between the Netherlands and Belgium. The reference data shown in Figure 4 come from Huizing and Moolenaar-Bijl (1944, based on 10,000 written words), Eggermont (1956, based on 10,000 spoken words), Van den Broecke (1976, based on the 1,000 most frequent words taken from 50,000 written words), and Van den Broecke (1988, based on the 12,000 most frequent words taken from 2.3 million written words).

The occurrence of the phonemes in the original Dutch matrix test was close to that of the reference data. The average absolute difference in occurrence between the matrix test and the data of Luyckx et al. was 1.1 percentage points.

To estimate the size of the deviations for subset-based random sampling, the phonemic occurrence was calculated for the entire subset of 140 selected sentences as well as for each of the 1,000 randomly generated test lists of 20 sentences. Figure 5 shows the results, including the median phonemic occurrence for 1,000 drawn sets and the phoneme distribution that occurred in 95% of the drawn lists. The phonemic occurrence of the simulations closely resembled that of the original full matrix test. The maximum absolute difference between the median of the simulations and the original matrix was 1.7 percentage points for /t/, see Figure 5.

Word Occurrence

Another possible drawback of subset-based random sampling is that the word occurrence in the drawn set can happen to be high for specific words. A high word occurrence might alter the guess rate or might alter the behavior of the subject. Figure 6 shows the word occurrence for the 1,000 randomly generated lists.

The average median word occurrence was 2.0, indicating that averaged over all 1,000 lists, each word occurred about twice in a test list. This value is the same as that for the base matrix.

Discussion

An appropriate selection of sentences can increase the slope of the psychometric function. For the test lists investigated in this study, the list-specific slope increased from 10.2%/dB to 13.7%/dB.

The new slope is about 5%/dB larger than the mean slope for a single static noise masker as was used here, reported by MacPherson and Akeroyd (2014) in their comprehensive survey of the slope of the psychometric function. The new slope might seem high; however, MacPherson and Akeroyd (2014) also found a large variation in slope (ranging from 1%/dB to 44%/dB) between different speech materials. They showed that for a single stationary masker, the choice of target corpus was important: The median slope of the IEEE materials

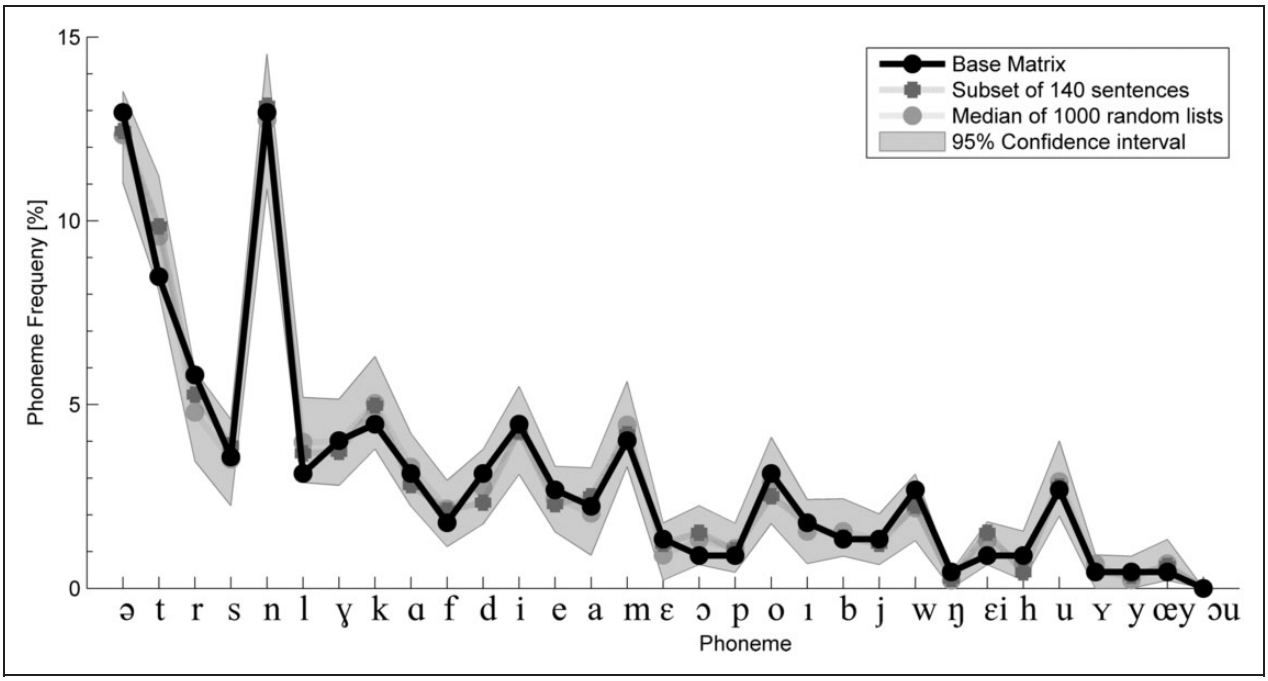

Figure 5. Phonemic occurrence for a random draw of 20 sentences out of the set of 140 sentences. Shown are the median and 95% confidence interval of 1,000 random drawings. The phonemic occurrence for both the original matrix test and the selected subset is also shown.

was 4.8%/dB and the median slope of the SSI materials was 17.1%/dB.

Compared with the slope of matrix tests in other languages (12.6–17.1%/dB, Houben et al. 2014), the new slope falls within the range of slopes of the other tests.

The determination of the slope of the psychometric function was based on previously gathered data. The calculation was the same as for the fixed lists (Houben et al., 2014) except for two differences. First, data were only available for 24 subjects instead of 45, due to the cross-validation procedure. This difference could lead to a somewhat larger variance in the model predictions because the model is fitted with fewer data. Second, the data for each list were not measured as an actual list but as separate sentences that were combined post hoc to form a list. This second difference could lead to a smaller standard deviation in SRT and slope between lists. In the original experiment, early lists might have been influenced by a residual learning effect, while later lists might have been influenced by fatigue. In the calculation for the randomly drawn lists, both these effects are combined in a single list due to the random sampling. Thus, the sentence selection process favors homogenous materials. It is expected that this does not greatly influence the calculated mean SRT and list-specific slope, but their standard deviation might be influenced because the differences between lists could be artificially smaller. The results showed that the interlist standard deviation of the slope remained the same (0.9%/dB for both the original set and subset-based random sampling), suggesting that order effects are limited. The standard deviation of the SRT became slightly smaller (0.2 dB in the original set and 0.1 dB for subset-based random sampling) while, as expected, the mean SRT was not influenced by the selection process (SRT for both sets was −8.4 dB).

The results show that the steepness of the psychometric function can be influenced by selecting materials. Since both the selected and the discarded subsets contained all words that are present in the original matrix test, it seems likely that the increased slope was caused by the removal of stimuli that had a less steep psychometric function due to the way the sentence was spoken and constructed during development of the matrix test. However, because only a single recording was available for each sentence, it is not possible to prove that limited steepness of the slope was caused by these non-language-specific factors. The results suggest that for situations where materials with a steep slope are required, existing speech tests might be further optimized by selecting appropriate sentences. For the development of future tests, one might contemplate recording and evaluating extra materials, or more versions of the same sentences, so that the sentences with the steepest psychometric function can be selected.

A possible limitation of the random selection procedure is that the phonemic distribution of the selected set may differ from that of the original matrix. Even though the validity of phonemic balancing as a design criterion

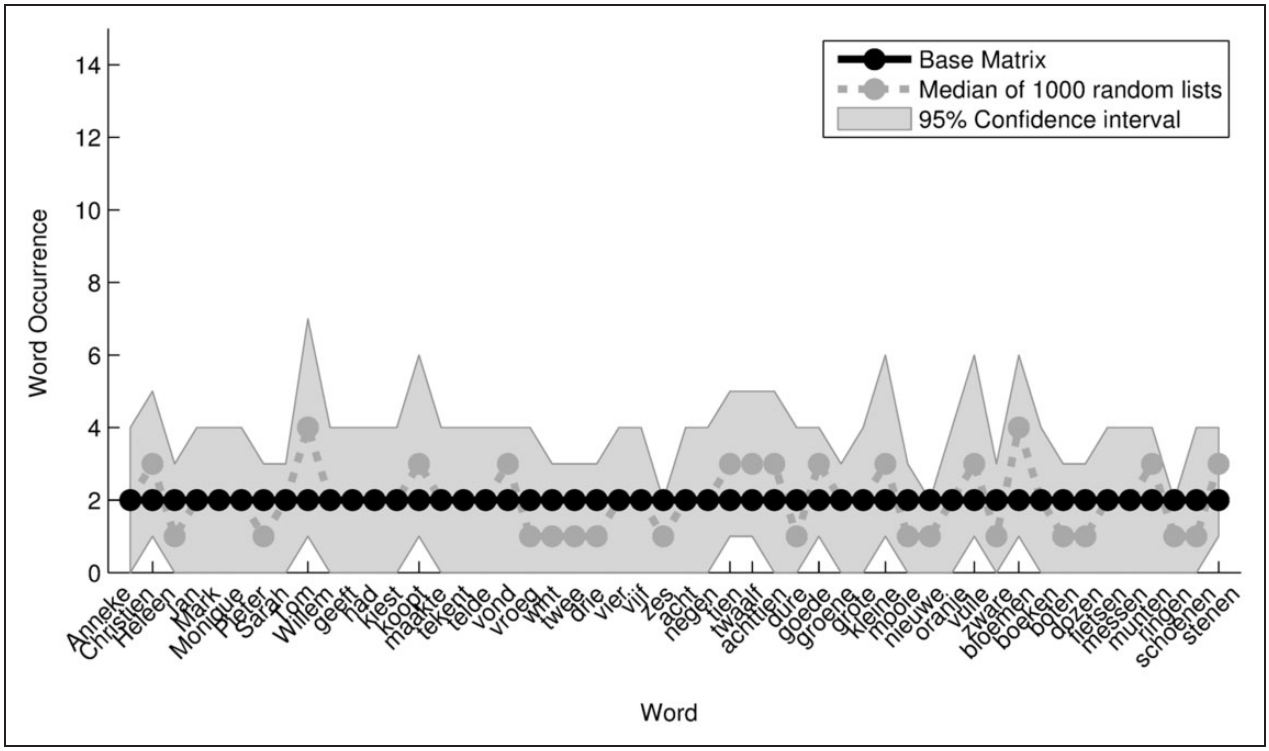

Figure 6. Word occurrence for a random draw of 20 sentences out of the set of 140 sentences. Shown are the median and 95% confidence interval of 1,000 random drawings as well as the word occurrence for the original matrix test.

for speech tests has been debated (Martin, Champlin, & Perez, 2000; Wang, Mannell, Newall, Zhang, & Han, 2007) and some authors have abandoned phonemic balancing in developing speech materials (Harris et al., 2007; Kirk, Pisoni, & Osberger, 1995; Nissen, Harris, Jennings, Eggett, & Buck, 2005), recent matrix tests in other languages were all developed to be phonemically balanced. Regardless of the validity of phonemic balanced speech materials, for the current optimization it is important that the differences between subset-based sampling and the original matrix test with respect to the phonemic distributions remain relatively small. One method to control the phonemic balance is to preconstruct quasi-random lists with an acceptable phoneme occurrence. Another approach, the one taken here, is to investigate the phonemic occurrence for a large amount of randomly drawn lists. The simulations showed that the differences in phonemic occurrence between lists of 20 sentences obtained with subset-based random sampling and those from the original full matrix were rather small. In 95% of the lists, the average absolute difference in phonemic occurrence was at most 1.7% (see Figure 5). This maximum difference seems acceptable in comparison to an average difference of 1.1% between the occurrences in the original matrix test with that of the most recent reference data from Luyckx et al. (2007). Moreover, there are large differences in the occurrences from the different reference studies themselves (see Figure 4).

Another possible limitation of the random selection procedure is that specific words might occur relatively often in a test list. Even though the average median occurrence for the randomly drawn lists was the same as that for the base matrix (Figure 6), this was not the case for every list. If a list with a high word occurrence is drawn repeatedly, the behavior of a subject might be affected. Note that, in a way, the original test could be regarded as similarly biased because the behavior of subjects could depend on whether they expect each word to occur exactly twice. However, if with subset-based random sampling a specific word occurrence is very high, the subject might think that that word is present in all sentences. The use of completely random lists cannot exclude the possibility that word occurrence may become high in incidental cases. Figure 6 illustrates that for some lists (95% confidence interval line in the figure), some words (e.g., Tom) have shown a high word occurrence. The value in the figure stems from the 95% confidence interval of the simulation of a 1,000 random lists, and the highest median was 4. In practice, this effect might be limited because only a limited number of randomly drawn lists have the mentioned high word occurrence. Additionally, even if such a list with a high word occurrence is used, the effect on the subject's perception

might be limited because it is expected that a subject will not be able to correctly repeat every occurrence because the adaptive procedure can make some words unintelligible.

To prevent lists with a too high occurrence of certain words, one could discard lists that do not meet a certain criterion, for example, a maximum word occurrence of, for instance, 3 or 4. As an alternative, one could use preconstructed quasi-random lists that are selected to avoid lists with a high word occurrence.

The current results show that subset-based random sampling might be a viable method to change the characteristics of a speech-in-noise test. However, instead of selecting lists at random during testing, one could also preconstruct quasi-random lists. This has two advantages. First, one can omit lists with deviating phoneme distributions. Second, the use of quasi-random lists might be easier to implement with existing measurement software. For instance, one could use new quasi-random lists with existing software without the need to extend the software with a new procedure for online sentence selection.

Another limitation of the optimization of the materials is that it is not yet known how the obtained results (in SRT) compare against results obtained with the original matrix test when applied to hearing-impaired listeners. For normal-hearing listeners, the SRT was the same for both the original matrix test and subset-based random sampling. But this is not necessarily the case for the hearing impaired. Of course, this limitation holds for most speech materials, since they are developed and refined for normal-hearing listeners and applied in hearing-impaired listeners (such as all matrix tests). For these materials, a perceptual equivalence of the test stimuli for normal-hearing listeners does not guarantee a perceptual equivalence for hearing-impaired listeners, and this requires further investigation. For the Dutch matrix test, there are indications that these effects are of secondary importance, at least for cochlear implant users (Theelen–van den Hoek, Houben, & Dreschler, 2014). However, in their review MacPherson and Akeroyd (2014) noted the trend that for speech tests with a static noise masker, the slope seemed to decrease with increased hearing impairment, including cochlear implant users.

The use of variable but not phonetically balanced test lists with a steeper psychometric function might in some situations (e.g., in a research setting) be preferable over the use of fixed and phonetically balanced test lists with shallower psychometric functions. However, if perceptual equivalence is indicated to be a significant factor, it might complicate comparison of results that are obtained with the full matrix test and by subset-based random sampling, especially across specific subject groups and across languages.

Conclusions

Subset-based random sampling can increase the slope of the psychometric function of the Dutch matrix test from 10.2%/dB to 13.7%/dB.

Subset-based random sampling leads to a difference in phoneme occurrence that is limited in view of (a) the fact that the phonemic occurrence of the original matrix test also slightly deviates from that of the most recent literature value and (b) the fact that the phonemic occurrence also differs between the available reference data sets themselves.

Acknowledgments

This work was based on earlier measurements that were carried out in Rotterdam, Amsterdam, and Leuven. We would like to thank Jan Koopman, Hans Verschuure, Heleen Luts, and Astrid van Wieringen for their pleasant and fruitful cooperation. We thank Birger Kollmeier for his valuable suggestions on an earlier version of this work.

Declaration of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Eggermont, J. (1956). De klankfrequentie in het hedendaagse gesproken Nederlands. *Nieuwe Taalg*, *49*, 221–223.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian audiology*, *11*, 79–87.

Harris, R. W., Nissen, S. L., Pola, M. G., McPherson, D. L., Tavartkiladze, G. A., & Eggett, D. L. (2007). Psychometrically equivalent Russian speech audiometry materials by male and female talkers. *International Journal of Audiology*, *46*, 47–66.

Hochmuth, S., Brand, T., Zokoll, M. A., Castro, F. Z., Wardenga, N., & Kollmeier, B. (2012). A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology*, *51*, 536–544.

Hood, J., & Poole, J. (1980). Influence of the speaker and other factors affecting speech intelligibility. *International Journal of Audiology*, *19*, 434–455.

Houben, R., Koopman, J., Luts, H., Wagener, K. C., van Wieringen, A., Verschuure, H.,...,Dreschler, W. A. (2014). Development of a Dutch matrix sentence test to assess speech intelligibility in noise. *International Journal of Audiology*, *53*, 760–763.

Huizing, H. C., & Moolenaar-Bijl, A. (1944). De beteekenis der klankfrequentie in het Nederlandsch voor de oorheelkunde. *Ned Tijd Gen*, *88*, 435–437.

Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R.,...,van Wieringen, A. (2012). Comparison of three types of French speech-in-noise

tests: A multi-center study. *International Journal of Audiology*, *51*, 164–173.

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing*, *16*, 470–481.

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, *102*, 2412–2421.

Luyckx, K., Kloots, H., Coussé, E., & Gillis, S. (2007). Klankfrequenties in het Nederlands. In D. Sandra, et al. (Eds), *Tussen taal, spelling en onderwijs: Essays bij het emeritaat van Frans Daems* (pp. 141–154). Gent: Academia Press.

MacPherson, A., & Akeroyd, M. A. (2014). Variations in the slope of the psychometric functions for speech intelligibility: A systematic survey. *Trends in Hearing*, *18*, 1–26.

Martin, F. N., Champlin, C. A., & Perez, D. D. (2000). The question of phonetic balance in word recognition testing. *The Journal of the Acoustical Society of America*, *11*, 489–493.

Nissen, S. L., Harris, R. W., Jennings, L.-J., Eggett, D. L., & Buck, H. (2005). Psychometrically equivalent mandarin bisyllabic speech discrimination materials spoken by male and female talkers. *International Journal of Audiology*, *44*, 379–390.

Ozimek, E., Warzybok, A., & Kutzner, D. (2010). Polish sentence matrix test for speech intelligibility measurement in noise. *International Journal of Audiology*, *49*, 444–454.

Steyerberg, E. W. (2008). *Clinical prediction models*. New York, NY: Springer.

Theelen–van den Hoek, F. L., Houben, R., & Dreschler, W. A. (2014). Investigation into the applicability and optimization of the Dutch matrix sentence test for use with cochlear implant users. *International Journal of Audiology*, *53*, 817–828.

Van den Broecke, M. P. R. (1976). *Hierarchies and rank orders in distinctive features*. Assen, The Netherlands: Van Gorcum.

Van den Broecke, M. P. R. (1988). Frequenties van letters, lettergrepen, woorden en fonemen in het Nederlands. In V. den B. Van den Broecke (Ed.), *Ter sprake: Spraak als betekenisvol geluid in 36 thematische hoofdstukken* (pp. 400–407). Dordrecht, The Netherlands: Foris Publications.

Vlaming, M. S. M., Kollmeier, B., Dreschler, W. A., Martin, R., Wouters, J., Grover, B.,. . . ,Houtgast, T. (2011). HearCom: Hearing in the communication society. *Acta Acoustica United with Acoustica*, *97*, 175–192.

Wagener, K., Brand, T., & Kollmeier, B. (1999). Enwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie/ Audiological Acoustics*, *38*, 4–15.

Wagener, K., Josvassen, J. L., & Ardenkjaer, R. (2003). Design, optimization and evaluation of a Danish sentence test in noise. *International Journal of Audiology*, *42*, 10–17.

Wang, S., Mannell, R., Newall, P., Zhang, H., & Han, D. (2007). Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *International Journal of Audiology*, *46*, 719–731.

Zokoll, M. A., Hochmuth, S., Warzybok, A., Wagener, K. C., Buschermohle, M., & Kollmeier, B. (2013). Speech-in-noise tests for multilingual hearing screening and diagnostics1. *American Journal of Audiology*, *22*, 175–178.