UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish

Vainio, Martti; Suni, Antti; Järveläinen, Hanna; Järvikivi, Juhani; Mattila, Ville-Veikko

2005

Acoustical Society of America

http://hdl.handle.net/10138/24705

Vainio, M, Suni, A, Järveläinen, H, Järvikivi, J & Mattila, V-V 2005, 'Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish', Journal of the Acoustical Society of America, vol. 118, pp. 1742-1750. https://doi.org/10.1121/1.1993129

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish

Martti Vainio
Department of General Linguistics and Department of Speech Sciences, University of Helsinki, Helsinki, Finland

Antti Suni
Department of General Linguistics, University of Helsinki, Helsinki, Finland

Hanna Järveläinen
Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, Finland

Juhani Järvikivi
Department of Psychology, University of Turku, Turku, Finland

Ville-Veikko Mattila
Multimedia Technologies Laboratory, Nokia Research Center, Tampere, Finland

(Received 11 May 2004; revised 13 June 2005; accepted 13 June 2005)

A subjective test was developed suitable for evaluating the effect of mobile communications devices on sentence intelligibility in background noise. Originally a total of 25 lists, each list including 16 sentences, were developed in British English and Finnish to serve as the test stimuli representative of adult language today. The sentences, produced by two male and two female speakers, were normalized for naturalness, length, and intelligibility in each language. The sentence sets were balanced with regard to the expected lexical and phonetic distributions in the given language. The sentence lists are intended for adaptive measurement of speech reception thresholds (SRTs) in noise. In the verification of the test stimuli, SRTs were measured for ten subjects in Finnish and nine subjects in English. Mean SRTs were −2.47 dB in Finnish and −1.12 dB in English, with standard deviations of 1.61 and 2.36 dB, respectively. The mean thresholds did not vary significantly between the lists or the talkers after two lists were removed from the Finnish set and one from the English set. Thus the numbers of lists were reduced from 25 to 23 and 24, respectively. The statistical power of the test increased when thresholds were averaged over several sentence lists. With three lists per condition, the test is able to detect a 1.5-dB difference in SRTs with the probability of about 90%.
© *2005 Acoustical Society of America.* [DOI: 10.1121/1.1993129]

PACS number(s): 43.71.Gv, 43.72.Kb [DOS] Pages: 1742–1750

I. INTRODUCTION

In mobile communications, the intelligibility of speech transmitted over a voice communication channel may vary considerably due to a number of interfering factors. Typical examples of such factors are transmission channel errors, electrical channel noise, and environmental noise at the talker. Many subjective test methods making use of speech as test materials have been developed to assess the intelligibility of speech. In telecommunications, subjective audio quality tests include human listeners who give their opinion of the performance of the telephone transmission system under evaluation used either for conversation or for listening to spoken material. In utilitarian assessment methods, a single quality scale is typically used to give the opinion of the overall quality of the system, whereas in analytic methods several quality scales, describing different aspects of quality, can be used to give a multidimensional view of the quality. However, none of the presently available tests are sufficiently applicable to evaluate speech processing algorithms, such as speech coders, channel coders, and noise suppressors, which are the essential components of any modern communication system. As these algorithms have been designed to process continuous speech, taking into account, e.g., temporal masking, their performance cannot be measured with short speech stimuli. Therefore the quality of the coders cannot be assessed by processing only a single word at a time or by measuring the intelligibility only for short segments of speech.

Different languages also impose different needs with respect to the normalization and balancing of the test materials. It can be argued that to be truly universal the tests should be multilingual and should also be built with common criteria with regard to the acoustic as well as the linguistic methodology. This, in turn, calls for additional care with regard to choosing the linguistic material for the test sentences: the material should be normalized and balanced so that it reflects the current adult language usage across the given languages and/or dialects. The current research introduces a reliable and efficient method—based on the measurement of speech reception thresholds in noise—to evaluate the intelligibility of speech in mobile communications for two languages, namely English and Finnish.

The earliest quantitative subjective speech assessment methods have typically focused on speech intelligibility. In addition to this, most early methods were designed to evaluate the intelligibility of single segments or phonemes. Moreover, the evolution of the testing methodology has proceeded from the use of purely nonsensical units such as monosyllables[1] towards a controlled use of real words, as in e.g., the family of rhyme tests[2–5] and whole sentences corresponding more closely to typical language use. In addition to this, a noise source is often used to simulate the interference in a real world situation beginning with the work of Egan.[6]

Kalikow, Stevens, and Elliot[7] introduced the first speech perception in noise test. This test used sentence-level materials to measure word-level intelligibility of the input at fixed speech and noise levels. The listeners were asked to repeat the final monosyllabic noun of a sentence. The length of the sentences was manipulated and both the key-word familiarity and predictability (high versus low predictability) were balanced within lists of 50 sentences. As a percent intelligibility measure the test is, however, limited by floor and ceiling effects.[8]

An alternative to percent intelligibility measures was developed by Plomp and Mimpen.[9] The speech reception threshold (SRT) is free of the above-mentioned limitations. The SRT is defined as the presentation level of test speech necessary for a listener to understand the speech correctly a specified percent of the time, usually 50%. In the SRT, the speech material consists of sentences that are presented either in silence or in the presence of a reference noise signal. In practice the SRT measurement requires an implementation of an adaptive listening test procedure where the intensity level of speech can be varied depending on the listener's responses, i.e., whether sentences are understood correctly or not. If a sentence is not understood correctly, the level of the next sentence is increased, whereas in the opposite case, the level is decreased. In this way, over a sequence of sentences, the level of speech should gradually be reaching a stable value, i.e., a listener's SRT, where speech is just understandable to the listener.

Recently Nilsson *et al.*[8] introduced a hearing in noise test (HINT) for sSRT (sentence speech reception threshold) measurements. Their starting point was the Bamford-Kowal-Bench (BKB)[10] sentences originally designed for use with British children and representative of children's speech. The sentences were revised by removing British idioms and controlling for sentence length and the number of present and past tense verbs, after which the sentences were normalized for naturalness, difficulty, and reliability.

While the methodology presented in Nilsson *et al.*[8] is very good for developing SRT tests in general, the sentences they used suffer from a number of shortcomings: (1) they do not represent current adult language usage and (2) they are not translatable to other languages without losing such critical features as phonetic and structural balance. The latter is naturally true for any set of sentences. Developing an SRT test for a new language, therefore, requires the development of a completely new set of sentences which is representative of that language alone. Furthermore, developing a more universal test requires the development of more than one language simultaneously. The situation clearly calls for methodology that uses common criteria for the languages in question. That is, the languages should be treated with criteria which are determined by those languages together.

The existence of large text corpora and high-quality linguistic tools for analyzing them offers new possibilities meeting the requirements listed above. The linguistic material can be balanced both phonetically and structurally in such a way that the end result is representative of current adult usage.

Our present study is largely based upon the HINT methodology, but intends to refine it by developing new test sentences for British English and Finnish, which would be representative of current adult speech. This, we hope, will reflect the real-world situations of the speech coding devices at work in a more realistic manner.

Following the procedure introduced in Nilsson *et al.*[8] we first created the new sentence lists, recorded and edited them, created speech-shaped masker noise spectra from the recordings, matched the recorded sentences for difficulty, tested the interlist reliability, as well as estimated the statistical power of the test itself.

In conjunction with this study, a fully computer-based test system which allows for a reliable sound reproduction, a fast test administration, and easy data collection, as well as an automated way to conduct statistical data analysis, was developed. With the system, the duration of a threshold measurement with a single list usually takes less than 2 min. The computerized system is not discussed further in this paper.

The following section outlines in more detail the development process of the test sentences in British English and Finnish. The rest of the article presents a study on the suitability of the sentences for SRT measurements followed by an experimental evaluation of the performance of the proposed intelligibility test.

II. DEVELOPMENT OF SENTENCE MATERIALS

Our goal was to create sentence sets that would be representative samples of the target language as used by adults in a relatively formal context with a minimal amount of phonological and phonetic reduction. Our aim was to produce a certain number of sentences (here 400, divided into 25 groups of 16 sentences each) rather than finding an unforeseen number of sentences meeting certain criteria from corpora. Nilsson *et al.*[8] report that list sizes of 10 and 12 are adequate for the adaptation in the SRT. As we could not know *a priori* what the list size for Finnish would be, we used 16 sentences per list to start with.

A. Source material

The source material for both languages consisted of different text corpora available for our use. For Finnish we used a selection of regional newspapers from the years 1998–2000 (Turun Sanomat, Hämeen Sanomat, Keskisuomalainen, and Alasatakunta) from SKTP-B archives.[11] The material contained approximately 20 million word tokens.

For English we used the Brown Corpus, Lancester-Oslo/Bergen Corpus, and Grolier Electronic Encyclopedia, containing about three million words. We also used a number of books from the Project Gutenberg (http://www.gutenberg.net) collection with approximately six million additional words to achieve the necessary number of suitable sentences for the sets. The Gutenberg collection includes a number of works dating from before the 20 century which have outdated and archaic word forms (see Sec. II B 2 below). Therefore these materials were excluded from the phone and word frequency calculations.

B. Selection procedure

1. Extraction of phonetic and lexical frequencies

The Finnish corpora were transcribed to phonetic form following the standards of carefully articulated spoken language. In Finnish this can be achieved with a fairly small set of rules as the grapheme to phoneme correspondence is very high. We did not take into account assimilation and coarticulation phenomena and indicated glottalization only in the case of a word starting with a vowel and the preceding word ending with one within a sentence.

The phonetic transcription of the English corpora was performed using the front end of an English version of the Festival speech synthesis system[12] (http://www.festvox.org/). The system uses decision trees trained with the CMU phonetic dictionary of approximately 100 000 words and a statistical parser to remove possible ambiguities.[13] The biphone and monophone frequencies were calculated from the transcriptions. Word frequencies and corresponding base forms for both languages were extracted from the corpora using a *functional dependency grammar* syntactic parser[14] (as implemented by Connexor Inc.) At this stage, we concluded that the possible errors introduced by the automatic syntactic analyses and letter-to-sound conversion would be so small as not to have any effect on the final results. That is, the phonetic balance is centered around the most common units and the grammatical well-formedness of the sentences in the final sets was verified by humans.

2. Filtering the sentences

A set of good candidates for the final set selection process was selected, in effect, by filtering the big corpora incrementally from millions of words into a set of approximately 1000 sentences per language. The first step in extracting suitable candidate sentences was to exclude the sentences of unwanted length from the large corpora. We measured the length of the sentences in syllables. Due to structural differences between English and Finnish, we could not use the same number of syllables for both languages. Due to suffixation, Finnish words tend to be much longer than English and a greater number of syllables is necessary to prevent ungrammatical and clipped sentences to be introduced into the sets. Therefore, we used 9 to 12 syllable sentences for Finnish, whereas for English 7 to 9 syllables were considered sufficient. For similar reasons, we omitted the most frequent words (such as articles and prepositions) from the base-form frequency counts in English. The Finnish material was syllabified with an algorithmic method whereas the English material was syllabified by using the CMU phonetic dictionary.[12] After the selection procedure approximately 30 000 Finnish and 20 000 English candidate sentences remained in the source material. At this stage the sentences were checked for grammatical correctness; the remaining sentences were analyzed with a syntactic parser[14] and incomplete sentences with respect to main syntactic roles were removed. This ensured that all of the sentences were syntactically complete and sentences lacking, for instance, a predicate verb were removed.

Cumulative biphone frequencies of candidates were then calculated based on the frequencies observed in the source corpora. Sentences which deviated the most from the average value were removed. Additionally, sentences with rare diphones (less than 1000 occurrences in Finnish corpora and less than 200 in English) were left out. The numbers reflect the size of the corpora as well as the size of the diphone inventories in each language, Finnish having a larger corpus and fewer diphones—ca. 700 as opposed to ca. 1500 in English.

A similar procedure was applied on the lexical level, removing sentences with extreme values with regard to cumulative frequencies of the base forms of words.

At this point we also used certain language-specific filtering. Finnish sentences containing words with nonpredictable pronunciation were removed i.e., words of foreign origin, as well as certain loan words, which could simply be identified by foreign letters (c, q, w, z, and x). The English material from the Gutenberg project included books from the late 19th and early 20th centuries and contained, therefore, archaic material such as, for instance, the words "thou," "hither," and "ye," which had to be accounted for. Although most of the problematic forms had already been removed on the grounds of low base-form frequencies of the archaic words, nevertheless, we still had a native English talker check the final 1000 candidates for naturalness.

3. Balancing of the sets

After reducing the number of candidates to approximately 1000 sentences per language, the final 25 groups of 16 sentences were created. Intergroup variation was minimized according to average sentence length in phonemes, average word base-form frequency and distribution of phones. Simultaneously the monophone frequencies of the groups were matched with the source corpora using χ^2 test for goodness-of-fit. Minimization was performed by first creating randomized sets containing 400 sentences, normalizing the observed variances for each variable (average sentence length, average word base-form frequency) and then using a brute-force algorithm to find balanced sets minimizing the total variance using the whole set of 1000 sentences. All pairwise changes of sentences were tested between the groups and between the sentences left out of the groups. The changes which reduced variance to a significant degree were kept. This process was allowed to continue until no improvement was observed.

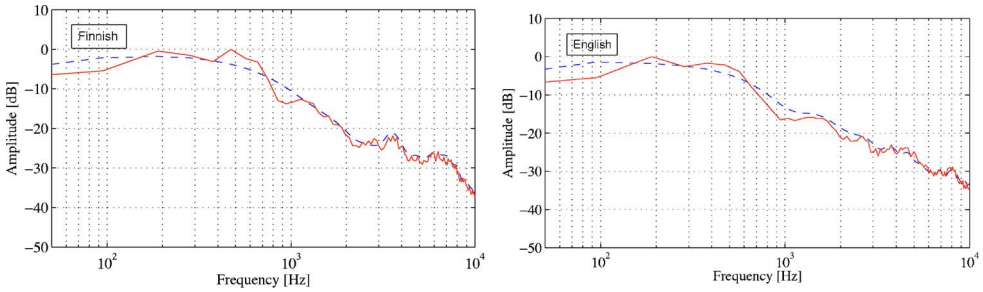


FIG. 1. (Color online) Magnitude spectra of the SRT test sentences in Finnish and English (solid lines), and the matching respective noise spectra (dashed lines).

C. Resulting materials

As a result 25 sets of 16 sentences for both languages were obtained. The χ^2 measure for the English sets varied from 9.45 to 17.55 (mean 13.27) ensuring that each set did not significantly differ from the whole corpora at chance level of 95% (i.e., α=0.005, df=50); the respective values for Finnish were 10.39 to 16.74 (mean 13.11) for 41 degrees of freedom. The average length of the sentences in the sets varied between 23.3 and 24.0 phones for English and 31.9 and 32.9 phones for Finnish. Therefore, we could expect that the Finnish sentences would also be somewhat longer in duration than the English ones. The cumulative word base–form frequencies in the Finnish sets varied ±1.6% from the grand average and the English sets varied between ±1.0%. The average lengths of the sentences (in syllables) were Finnish, 12.7 syllables (standard deviation 1.1), and English, 7.49 syllables (standard deviation 0.76). The resulting average utterance durations were 2.1 s (standard deviation 0.27 s) for Finnish and 1.7 s (standard deviation 0.32 s) for English.

III. RECORDING AND EDITING OF SENTENCE MATERIALS

The sentence materials in each language were recorded by four talkers. Two talkers in each language were male and two female. All were native speakers of British English or Finnish. Sentence lists 1–12 and 13–25 were spoken by one male and one female in each language. The Finnish talkers were voice professionals: two actors, a phonetician, and a speech therapist. All of the English talkers were teachers accustomed to public speaking but not specially trained to use their voice. None of the talkers represented an extreme voice type of any kind.

The recordings were made in an acoustically anechoic chamber. A Macintosh G3 computer with a 16-bit sound card (dynamic range 90 dB and frequency response ±0.5 dB between 30 Hz and 18 kHz) was connected to a Bruel & Kjaer (B&K) 2238 Mediator sound level meter, which includes a B&K 4188 condensator microphone and a BZ 7126 preamplifier. Headphone monitoring was provided for the talkers, and they were instructed to maintain a neutral speech style and volume. The microphone was always placed at about 0.50 m distance from the talker's mouth. The sentences were recorded directly to individual sound files by using the QuickSig measurement software,[15] which also allows filtering and amplitude monitoring. The sentences were checked for clipping and they were linear phase high-pass filtered at 70 Hz to remove induction noise. The sampling rate was 44.1 kHz. The resulting signal-to-noise ratio of the recordings was approximately 50 dB.

The sentences were then edited by removing any extra silence at the beginning and end. Other unwanted sounds, such as inspiration and lip smacks, were also removed. The signals were then up-sampled to 48 kHz and their mean-squared (MS) amplitudes were equated to 60 dB (relative to one sample unit in a 16-bit digital representation).

At this point we created the speech-shaped masker noise spectra by computing a long-term spectrum of both Finnish and English speech material. We summed up all the sentences in both languages separately, computed short-term spectra from these samples in 512-point slices, and averaged to obtain an estimate of the long-term spectrum for each language. Infinite impulse response (IIR) filters were then designed with frequency responses matching the long-term spectra in 128 frequency points up to 10 kHz. The criteria produced a 64-zero, 2-pole filter for Finnish and a 64-zero, 16-pole filter for English. White noise was then generated and filtered with the designed filters to produce language-dependent, spectrally matched, masking noise for the speech samples. The thick lines in Fig. 1 show the long-term spectra of Finnish and English, and the dashed lines show the corresponding filter responses.

IV. MATCHING SENTENCE DIFFICULTY

It was expected that the individual intelligibility of the sentences would not be equal in spite of the equal MS amplitude and overall phonetic balance. Therefore, initial intelligibility of each sentence was tested and the sentences were then rescaled according to the intelligibility scores with the aim that reverse variations in the signal level would compensate for the observed variations in initial intelligibility.

A. Experimental setup and procedure

(a) *Participants*. Two groups of eight listeners participated in both the Finnish and English tests. The participants were native talkers of English and Finnish. The English group consisted of people of British, American, and Australian origin. Prior to the test, the participants were screened for normal hearing in the range of 125 Hz to 8 kHz. The participants were paid for their participation.

(b) *Procedure*. The sentences were presented through headphones in an acoustically controlled reference listening room which satisfied the ITU-R BS.1116 Recommendation.[16] The speech and noise samples were mixed by the computer and played at −4 dB S/N ratio. The participants were instructed to type in what they had heard after each sentence. No feedback was given as to the correctness of the responses.

All the sentences were presented to each listener in four 75-min sessions. Prior to the actual SRT test material, a set of 20 practice sentences was presented. The sentences were similar to the actual test materials, but lacked the phonetic balance. The playback order of the lists as well as the sentences within lists was randomized. The materials from different talkers were presented to the participants in different order.

After the first test group, intelligibility scores were computed for each sentence of each individual talker. The score was based on the percentage of correct words in the responses to a given sentence averaged over all listeners. Each sentence that was less intelligible than all sentences on average was scaled up relative to 1 dB in MS level for each 10% difference. Sentences with higher intelligibility scores than average were scaled down in a similar manner. Thus, for instance, a 15% increase in the intelligibility score led to −1.5-dB decrease in MS level. The test was repeated for the next group of listeners with the rescaled sentences, and new level adjustments were made based on the new listeners' responses.

B. Results

After the first group of listeners, the overall mean intelligibility was 73% (standard deviation, std=32%) for the Finnish and 61% (std=34% correct) for the English sentences. After the second round, the mean scores were 70% and 71%, respectively, and the standard deviations had decreased to 29% correct and 28% correct, respectively. The distribution of the final MS adjustments are depicted in Fig. 2 for both the Finnish and English sentences. About 50% of the adjustments for Finnish and about 60% for English fell between ± 1 dB. However, the adjustments were not distributed evenly between the talkers. Sentences spoken by Finnish males were generally scaled down (−1.8 and −0.7 dB for male 1 and male 2 on average), whereas the sentences spoken by females were scaled up (+1.8 and +0.6 dB on average). For the English sentences, one of the male voices was generally scaled down (−1 dB in average), while the other male voice and one female voice were scaled up (+0.5 dB in average). The other female voice had a 0-dB mean scaling. Although the differences in intelligibility between talkers were mostly compensated by the sentence difficulty equalization, some talker effect could still be observed in the final adaptive SRT measurements. This will be discussed further below.

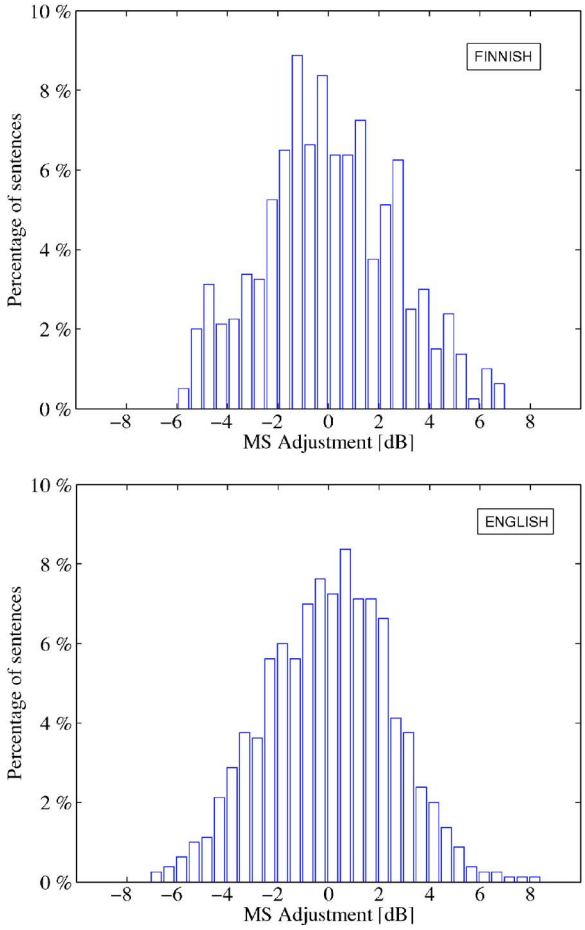


FIG. 2. (Color online) Distribution of MS amplitude adjustments for Finnish and English sentences (bar width=0.5 dB).

V. TESTING INTERLIST RELIABILITY

In order to determine whether each of the 25 sentence lists would produce a reliable and replicable measure of speech intelligibility, the SRTs were measured using an adaptive procedure and the results were analyzed statistically with analysis of variance (ANOVAs). The power of the test was also studied in order to determine the threshold differences.

A. Method

(a) *Participants*. A total of ten native talkers participated in the SRT measurement test for Finnish and nine in English. Prior to the test, the participants were screened for normal hearing in the frequency range of 125 Hz to 8 kHz.

(b) *Method*. The sentences were presented through headphones using the GuineaPig3,[17] computer-based, subjective test system in the same listening room as in the previous experiment.

The noise and the speech samples were mixed by the system at a S/N level that was dependent on subjects' performance in the previous trial. After a correct response, the S/N ratio of the following sentence was decreased by 2 dB, while an incorrect response increased it by the same amount. The S/N ratio in the first sentence of the test block was −5

dB relative to the overall rms level before scaling. The first sentence was repeated and its S/N ratio was increased in 2-dB steps until the subject had given a correct answer. All the other sentences were presented only once, regardless of whether the response was correct or not. No feedback was given to the subjects about their responses.

(c) *Procedure*. The SRT measurements were divided into two blocks of 12 and 13 lists each. Each subject was presented with the first 12 lists spoken by one talker and lists 13–25 spoken by another talker. The distribution of different talkers was balanced, so that the sentences from each talker were presented a total of five times to the subjects. The order of test blocks was varied between subjects and the list order was randomized by the test system.

The task was to type in the sentence the listener had heard using the GuineaPig3 test interface. This time minor variations were allowed in articles ("a/the") and verb tense ("is/was," "are/were," "has/had") for the English sentences, and in verb tense [as in "on/oli" ("is/was"), or "puhuu/puhui" ("speaks/spoke")] and the nearly identical singular versus plural forms [as in "syytä/syitä" ("reason/reasons"), or "maata/maita" ("land/lands")] in Finnish. No capital letters nor punctuation were required, nor was the system sensitive to added or missing spaces in compound words. Other than the exceptions mentioned above, a correct response required getting all words correct. The sentence presentation levels, the responses, and their correctness were recorded by the computer system. Since the sentence scoring was automatic, there was no way to treat spelling errors.

B. Results

The SRT was computed for each list as the average of the fifth and all the subsequent presentation levels within that list, including the determined level of the 17th trial that would be presented next. The mean S/N ratio at threshold over all lists was −2.47 dB in Finnish with a standard deviation of 1.61 dB, and in English −1.12 dB with a standard deviation of 2.36 dB. The current mean threshold for English was higher and the variability was larger than those measured by Nilsson *et al.*,[8] who obtained a mean threshold of −2.92 dB, with a standard deviation of 0.78 dB. The current mean threshold for Finnish was closer to the previous study. The use of four talkers instead of one as in Nilsson *et al.*[8] is a source of extra variability in both languages.

There are a number of possible explanations for the observed differences between the Finnish and English SRTs. The differences could be due to either the talkers, the listeners, or the sentence materials *per se* and, consequently, the languages themselves. While the Finnish talkers were either actors or phoneticians, the English talkers had had no professional voice training. That is, the differences could be explained simply by the quality of the speech and articulation. This can, however, only be a partial explanation. Another fact about the tests concerns the English speaking listeners; although the talkers were British, not all the listeners were. The results were therefore compared between British and American subjects, and a significant difference was at first observed ($p=0.007$). However, we discovered that two out of the nine subjects scored poorly compared to all the others, and both of them happened to be British whereas the best performing subjects were not American. In fact, if those two listeners are removed from the results, the difference between British and American subjects becomes nonsignificant with $p=0.41$. Therefore, we can conclude that the language differences, again, cannot be explained by the listeners.

We next considered the sentence materials themselves as a source for the differences in SRT scores. The most conspicuous difference between the sentence materials is that the Finnish sentences are made up of fewer, although longer, words than the English ones. The Finnish sentences are also longer with respect to the number of syllables and, therefore, duration. This provides them with more redundancy which could make them easier to perceive in noise. However, the sentence length had no significant effect on the SRTs in either language. This was checked by comparing the intelligibility scores of the longest and shortest 10% of the sentences in both languages. No significant correlations were found.

The final, and perhaps the most crucial, difference has to do with the languages themselves and how certain sounds are distributed within them. Finnish has a much greater relative number of voiceless stops than English, 20.53% as opposed to 12.1% in the current materials, respectively. Thus, there are significantly more voiceless gaps in Finnish speech signals than in the English ones. The average power values that the Finnish signals were normalized with are, therefore, more affected by the voiceless gaps in Finnish than in English. Consequently, the Finnish signals have slightly better S/N ratios, which could explain the differences in SRTs. There are, thus, a number of factors which work additively in favor of the Finnish materials in the tests. Fortunately, these factors have no bearing on the final applicability of the sets themselves.

A reliable SRT measure should reveal true differences in intelligibility. The criteria for this are that the SRTs obtained from different lists should not differ statistically, and that the probability of observing a difference in SRT by chance should be reasonably low. The test results were therefore analyzed in terms of variance and statistical power.

(a) *List equivalence*. Differences were scored between the mean SRT for each list and the mean SRT over all lists and subjects. They are presented in Fig. 3. The error bar shows one standard deviation below and above the average. The mean differences in the Finnish sentences were less than 1.0 dB for all lists, except for lists 12 and 16. After the removal of these lists, the one-way ANOVA did not reveal significant differences with respect to the list means ($p=0.14$). Having only list 16 removed, the ANOVA remained nonsignificant at the 0.05 level with $p=0.07$. None of the English lists differed significantly from the overall mean, the ANOVA being nonsignificant with $p=0.64$, although for lists 10, 16, and 22 the mean deviation exceeded 1.0 dB.

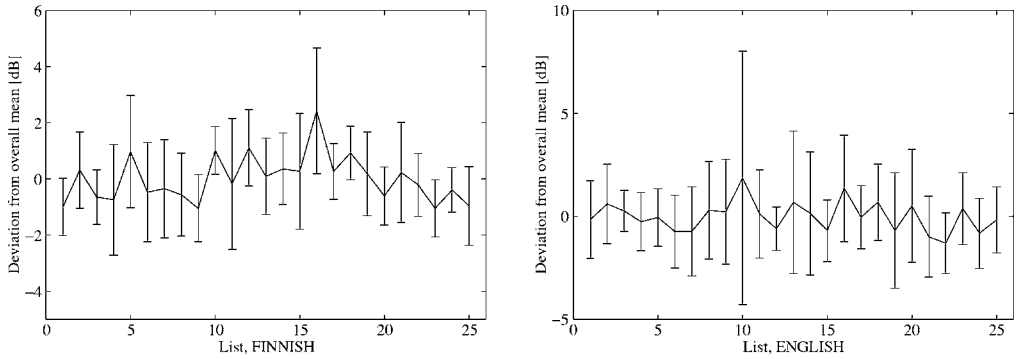


FIG. 3. Differences between mean SRTs of each list against the overall mean SRT in Finnish and English. The error bar shows one standard deviation above and below the average.

(b) *Statistical power*. The probability of correctly detecting a threshold difference can be estimated by studying the power of the test. Intuitively, it is clear that using a larger number of lists per condition reveals a true change in SRT more reliably than using only one list.

Let us first study the difference between two SRTs obtained from repeated measures within subjects, i.e., SRTs from lists spoken by the same talker and heard by the same listener. The standard deviation of the difference scores from all listeners was 1.61 dB for the Finnish material and 1.15 dB for the English material, resulting in 95% confidence intervals (±1.96 standard deviations) of ±3.15 and ±2.27 dB, respectively. Averaging two difference scores decreased the standard deviation to 1.40 and 0.97 dB, and by averaging three difference scores it was decreased to 1.01 and 0.81 dB. Thus, the confidence interval is also reduced when more lists are used per condition, and the threshold can be estimated more reliably.

Statistical power means the probability of correctly rejecting the null hypothesis that the true means of SRTs measured from different conditions are equal. The power measure is obtained by first finding the acceptable region for the observed sample mean of differences between the repeated measures, assuming the true mean difference to be 0 dB. This was done by a one-tailed Student's t test on a 5% risk level. Using this critical mean difference as decision criterion, it was computed how probably an even greater difference between two conditions would be observed, assuming now that the true mean difference is greater than zero. This was done by using the cumulative t distribution. Figure 4 shows the statistical power as a function of the mean difference in SRTs between conditions. It is seen that a 0.5-dB difference can be detected with no higher than 25% probability, while a difference of 1.5 dB can be detected with about 80% probability in the Finnish test and with about 95% probability in the English test. It is also evident from Fig. 4 that calculating the SRT as an average of two or three lists per condition increases the power of the test. It is thus recommended to use three lists per condition.

C. Required list length

The SRTs were recalculated for the reduced list lengths of 14, 12, and 10 sentences in addition to the original length of 16 sentences per list. The mean SRTs between the different list lengths varied within 0.25 dB in Finnish and 0.07 dB in English. For the Finnish sentences, the standard deviation of SRTs increased from the original value of 1.61 dB to 1.69, 1.78, and 1.91 dB, respectively. For the English sentences, it increased from the original value of 2.36 dB to 2.45, 2.55, and 2.69 dB. As the increase in standard deviation was relatively small, it was concluded that the list length could well be reduced to 10 sentences per list without a significant difference in the thresholds.

D. The talker effect

The material was recorded using two male and two female talkers in both languages. Based on the differences in

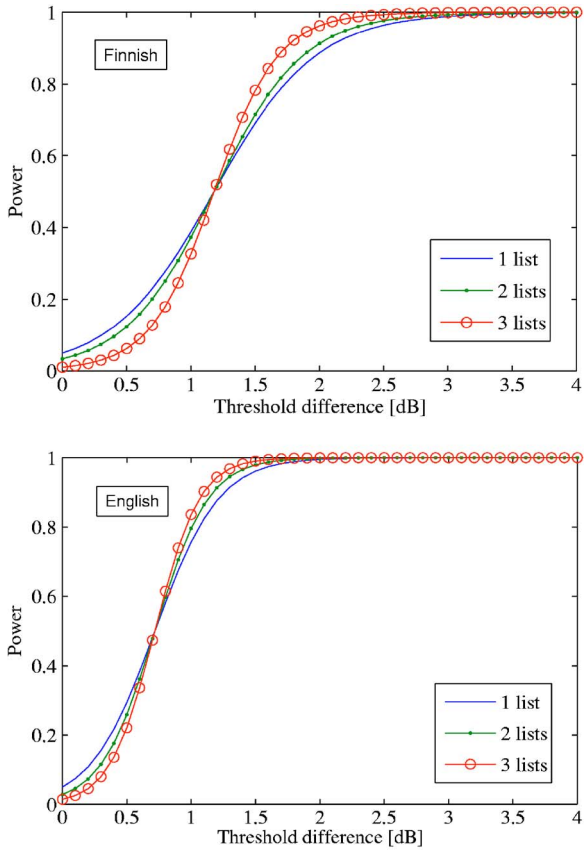


FIG. 4. (Color online) Statistical power of the SRT test in Finnish and English for one, two, and three lists.

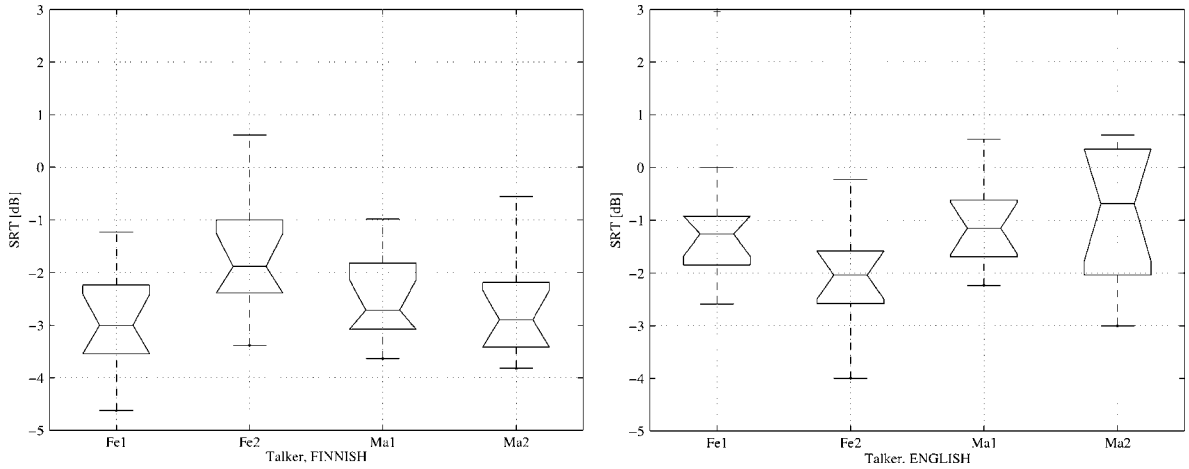


FIG. 5. Box plots of the SRT results for the different talkers in Finnish (left panel) and English (right panel) including all lists. The box plot shows the median, the upper, and the lower quartiles of the data.

the distribution of the sentence rescaling operations between talkers, it was assumed that some talker-induced variation may have been present in the SRT test. Therefore, the results were grouped according to the talker and an ANOVA was carried out between groups, revealing a significant effect in Finnish ($p=0.047$) and a marginally significant effect in English ($p=0.074$) (see Fig. 5). The Tukey HSD (honestly significant difference[18]) *posthoc* test was used to find significant differences in mean SRTs over each talker. The critical difference at the 0.05 uncertainty level was 0.68 dB for Finnish and 0.80 dB for English. In both Finnish and English, the second female talker differed slightly from the other talkers. All other differences in pairs of talker means were nonsignificant.

For female 2, the individual lists, having a mean SRT differing the most from the general mean, were lists 19 and 13 in English and lists 15 and 16 in Finnish. Removing these lists from both female 2 and male 2 diminished the talker effect in both languages. After the removal of the lists, the ANOVA was nonsignificant ($p=0.174$) in Finnish. In English, removing list 19 alone resulted in a nonsignificant ANOVA ($p=0.213$) and the removal of list 13 did not have a significant effect ($p=0.386$).

VI. SUMMARY AND CONCLUSIONS

A subjective test suitable for evaluating the effect of mobile communications devices on the sentence intelligibility in background noise was developed. A total of 400 sentences were carefully developed and recorded in British English and Finnish to serve as the test stimuli representative of adults' language use. The sentences, produced by two male and two female talkers for each language, were matched for naturalness, length, arid intelligibility. The sentence sets were balanced with regard to the expected lexical as well as phonetic distributions in the given language and the grammaticality of the sentences was checked. This resulted in 25 balanced sets of 16 sentences for both British English and Finnish.

The sentence lists were designed to be used in adaptive measurement of speech reception thresholds (SRT) in noise. The adaptive procedure together with the sentence material will make it possible to compare devices with a high resolution and accuracy.

SRTs were measured for both Finnish and English speaking subjects using the currently developed system and sentence lists. Some differences were found between the language groups which could be explained by linguistic differences, differences between listeners in the experiments, and differences between the talkers who read the sentences. The differences do not, however, impede the use of the test in any way.

The reliability of the SRTs was determined by statistical analysis of the results. The mean thresholds did not vary significantly between lists with the exception of two of the Finnish lists. The list length required to reach a stabilized SRT could be reduced to ten sentences per list without any significant difference in the obtained thresholds. A significant talker effect was found in both languages. A *posthoc* analysis verified that in both Finnish and English, the voice of one female produced slightly different SRTs than with the other three talkers. This problem was overcome by removing two lists from the Finnish set and one list from the English set. Thus the final material consists of 23 Finnish and 24 English lists.

The sensitivity of the test stimuli has been verified through the use of within-subject repeated measurements of SRTs with different lists. The 95% confidence intervals for the difference scores were $\pm$ 3.15 and $\pm$ 2.27 dB for Finnish and English, respectively. The confidence interval describes the sensitivity of the SRT test: a greater threshold difference can be expected to be caused by a true difference in the intelligibility. When the difference scores were averaged over three repeated measurements, the confidence interval reduced to $\pm$ 1.97 dB in English and $\pm$ 1.58 dB in Finnish. Thus, by using multiple lists per condition, smaller differences in SRT can be reliably detected.

The SRT test presented here provides an accurate, reliable, and efficient method for measuring speech intelligibility in noise. Based on the results, the test is directly appli-

cable to testing the intelligibility of various speech communication devices such as those used for mobile communications.

ACKNOWLEDGMENTS

The research presented in this manuscript was partly funded by Nokia Research Center.

[1]H. Fletcher and J. C. Steinberg, "Articulation testing methods," Bell Syst. Tech. J. **8**, pp. 806–854 (1929).
[2]G. Fairbanks, "Test of phonemic differentiation," J. Acoust. Soc. Am. **30**, 596–600 (1958).
[3]A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation testing methods: Consonantal differentiation with a closed response set," J. Acoust. Soc. Am. **37**, 158–166 (1965).
[4]W. D. Voiers, *Speech Intelligibility and Speaker Recognition* (Dowden, Hutchinson, and Ross, Stroudsburg, PA, 1977), Vol. **2**, Chap. 32.
[5]W. D. Voiers, "Evaluating processed speech using the Diagnostic Rhyme Test," Speech Technol. **1**, 30–39 (1983).
[6]J. P. Egan, "Articulation testing methods," Laryngoscope **58**, 955–991 (1948).
[7]D. N. Kalikow, K. N. Stevens, and L. L. Elliot, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**, 1337–1351 (1977).
[8]M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099 (1994).
[9]R. Plomp and A. Mimpen, "Improving the reliability of testing the speech reception threshold of sentence," Audiology **18**, 43–52 (1979).
[10]J. Bench and J. Bamford, *Speech-Hearing Tests and the Spoken Language of Hearing-impaired Children* (Academic, London, 1979).
[11]CSC-Center for Scientific Computing, "Finnish language bank," http://www.csc.fi/kielipankki/ (2003).
[12]A. Black and P. Taylor, *Festival Speech Synthesis System: system documentation (1. 1.1)*, Tech. Rep., Human Communication Research Centre Technical Report HCRC/TR-83 (1997).
[13]S. DeRose, "Grammatical category disambiguation by statistical optimization," Comput. Linguist. **14**, 31–39 (1988).
[14]P. Tapanainen, "Parsing in two frameworks: finite-state and functional dependency grammar," Ph.D. thesis, University of Helsinki, 1999, URL citeseer.nj.nec.com/tapanainen99parsing.html
[15]M. Karjalainen, T. Altosaar, and P. Alku, "QuickSig—An object-oriented signal processing environment," in *Proc. IEEE (ICASSP)* (IEEE, New York, 1988), pp. 1682–l685.
[16]A. Järvinen, L. Savioja, H. Möller, V. Ikonen, and A. Ruusuvuori, "Design of a reference listening room—A case study," in *Proceedings of the Audio Engineering Society; 103th International Convention* (Audio Engineering Society, New York, 1997), preprint number 4559.
[17]J. Hynninen and N. Zacharov, "GuineaPig-A generic subjective test system for multichannel audio," in *Proceedings of the Audio Engineering Society; 106th International Convention* (Audio Engineering Society, New York, 1999), preprint number 4871.
[18]R. Lehman, *Statistics and Research Design in the Behavioral Sciences* (Wadsworth, Pacific Grove, CA, 1991), pp. 369–370.

Copyright of Journal of the Acoustical Society of America is the property of American Institute of Physics and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.