A Danish nonsense word corpus for phoneme recognition measurements

Bo Nielsen, Jens; Dau, Torsten

Published in:
Acta Acustica United With Acustica

Link to article, DOI:
10.3813/AAA.919299

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Acta Acustica united with Acustica
Vol. 105 (2019) 183 – 194

DOI 10.3813/AAA.919299

A Danish Nonsense Word Corpus for Phoneme Recognition Measurements

Jens Bo Nielsen, Torsten Dau
Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. [jbn, tdau]@elektro.dtu.dk

Summary
Signal processing algorithms intended to improve speech intelligibility are common, e.g. in hearing aids. Due to the spectral and temporal variations between speech segments, the effect of such processing is likely to vary across the signal. To facilitate the analysis of these varying effects, a Danish speech corpus for measuring the perception of individual phonemes was developed. More than 1150 nonsense words were created according to a common template and audio-visually recorded with two male and two female talkers. Carrier sentences were also recorded. The audio recordings of all words were presented to ten normal-hearing listeners to ensure that phoneme recognition scores under optimal conditions were close to 100%. Accepted words were compiled into test lists with different characteristics. These lists were evaluated with seven normal-hearing listeners to determine norm data and to investigate memory effects. The speech recognition thresholds (SRTs) of the test lists varied between −7 to −0.9 dB signal-to-noise ratio (SNR) depending on list type and talker. The recognition score sensitivity to the SNR was 5.2 to 8.9%/dB. Memory effects were small and not significant. The presented speech materials seem well suited for measuring phoneme recognition scores and thus for assessing signal processing effects that vary across speech segments.

PACS no. 43.71.Es, 43.71.Gv, 43.72.Dv

1. Introduction

Speech intelligibility in challenging acoustical environments is the main concern for most hearing aid users. With advanced signal processing algorithms, the potential for improving intelligibility has become larger and signal manipulations that represent much more than mere amplification are common. For example, frequency bands that cannot be sufficiently amplified can be transposed to other frequency bands where the hearing-impaired listener's hearing threshold is lower [1], and interfering noise components can be selectively attenuated by noise reduction algorithms [2]. However, despite such strategies, the occurrence of distortions may lead to only minor or no improvement of the overall speech intelligibility. The effects of signal processing may also vary across different segments of the speech signal as a function of the varying spectral and temporal characteristics of these segments. The intelligibility of some speech segments may increase, while other segments may be distorted. An intelligibility test with only a single outcome measure, such as the speech recognition threshold (SRT), is not able to segregate such segmental effects but will produce an average result that leaves potentially conflicting effects unresolved.

A more detailed assessment of processing effects on speech intelligibility may be achieved by analysing smaller units of speech, such as phonemes, cf. [3]. Phonemes represent the shortest speech units that can change the meaning of words in a language. A listener may misperceive a word if a single phoneme is incorrectly identified; the human auditory system is thus naturally tuned to the identification of these. Rather than the 'macroscopic' view, as for example reflected in sentence intelligibility tests, analysing phoneme perception can be considered a 'microscopic' view of intelligibility [4]. A speech-processing algorithm aimed at improving the perception of high-frequency phonemes should ideally achieve this without affecting the perception of other phonemes. A test with recognition scores for individual phonemes may be well suited for assessing whether this goal has been achieved. Such a phoneme-based assessment may generalise well to how speech intelligibility is affected in real-life situations since the spectral and temporal characteristics of the test tokens are representative of the phonetic content in every day speech.

In real-life speech communication, listeners are typically able to correctly identify words and, thus, phonemes even when acoustical information related to the individ-

Received 4 April 2016,
accepted 26 November 2018.

ual phonemes is distorted or missing [5]. The cause of this degradation may be related to the speaker, the listener, or external factors, such as background noise or electroacoustic processing. When a communication is successful, the listener is able to mentally recreate missing acoustical information from the signal redundancy that stems from the phonetic, syntactic, semantic, and grammatical properties of a language [6]. Correctly identified segments of the acoustical signal combined with context information enable the listener to identify distorted speech segments. Pichora-Fuller *et al.* [7] observed that older listeners benefited even more from context than younger listeners, presumably because they have more practice in using context to recover lost speech information. In a phoneme-based test involving meaningful speech stimuli, such as everyday words, a listener's recognition scores for individual phonemes may thus reflect the available contextual information rather than the acoustical information [8]. A closer correlation between the acoustical information available to the listener and the recognition score can be achieved by reducing the contextual information within the speech signal. A straightforward method for doing this is to use nonsense words [8]. This approach was taken in the present study.

Semantic context effects between the phonemes in spoken words stem from the listener's assumption that the words are meaningful and can be found in lexical memory [9]. In everyday speech communication, word recognition is greatly improved by this lexical discrimination process [5] because listeners are routinely able to find a unique word that matches the perceived signal. Typically, this match corresponds to the intended word of the talker, also for words that are severely distorted. The discrimination process is likely to operate even when the listener is exposed to a nonsense word, which thus may be perceived as a meaningful, resembling word. However, in a nonsense-based test, the listeners should report the phonemes they perceive without being biased towards meaningful words. Thus, all words in the present corpus were provided with a fixed 'nonsense ending' that reduces their resemblance to meaningful words and minimizes the likelihood that the listener will perceive them as such. This approach has previously been applied in the Danish speech corpus 'PiTu' [10], where all words were given the ending 'tu', which does otherwise not occur in Danish.

In the initial phase of the present study, an assessment was conducted of whether any existing speech materials that could fulfil the following four requirements were available: (1) The pronunciation should be Danish. Even when the identification task is limited to an isolated phoneme, native pronunciation will facilitate phoneme identification [11]; (2) Co-articulation should be tested both in a CV and a VC combination. Although consonant pronunciation is affected more by the vowel in the CV-combination than in the VC-combination, the material should allow for investigating both effects; (3) The most common Danish consonants should be included to allow thorough investigation of, e.g., signal-processing effects; and (4) The material should be of "nonsense character" to minimize the effects of semantic redundancy on phoneme recognition. No existing speech materials fulfilled all four requirements, which motivated the development of the new corpus.

In addition to the speech corpus, a test procedure based on the scoring of individual target phonemes was developed. A combined evaluation of the corpus and the test procedure was conducted. The purpose was to investigate whether the test could be considered a tool for reliable assessment of phoneme perception. The speech materials themselves can be used in many types of experiments, also experiments that do not rely on the test procedure presented here. The speech corpus was named DANOK ('DAnsk NonsensOrdsKorpus').

2. The speech corpus

2.1. Requirements and design

The nonsense words of the present speech materials comply with Danish phonology. The phonemic content is representative of naturally spoken Danish, although a phonemic distribution similar to spoken Danish was not attempted. Instead, all target phonemes are represented an equal number of times across the corpus. The words were constructed from the standard phonetic inventory of Danish. The included target phonemes represent all classes of production for phonemes in Danish. Since phonetic variations of phonemes, e.g. dialectal, are often considered of minor importance for speech perception [12], phoneme variations have been omitted and only one common form of each phoneme included.

The corpus was primarily designed for the assessment of consonant perception. The phonetic inventory was confined by the following: (1) The selection of target phonemes should allow scoring without any knowledge of phonetic notation, i.e. the scoring task should be based on the standard alphabetical letters of written Danish; (2) Due to assimilation and co-articulation, consonants should be tested in two positions, prior to a vowel and after a vowel; (3) Consonant clusters, e.g. /bl/ and /tr/, should be included since they are not necessarily perceived as two separate consonants but as one entity. In addition, the consonants in an initial cluster are more rapidly pronounced than single consonants [13] and thus likely to be more difficult to perceive; (4) The Danish 'stød' (a short glottal stop) was not included. This characteristic feature of Danish is difficult to detect, difficult to score, and not part of the language in all parts of Denmark [14]; and finally (5) semi-vowels and diphthongs were not included as scoring tokens since lay people cannot identify them without training.

2.2. Written speech materials

The nonsense words were created according to a common scheme:

$$C(C)VC + /i/,$$

where *C* represents a consonant and *V* a vowel. The *C* in parenthesis indicates that words can have a single consonant or a cluster of two consonants in the initial position. The fixed ending lends a nonsense character to the words since /i/ is seldom as the final phoneme in two-syllable Danish words. In the following, words with a single consonant in the initial position are referred to as 'C words' and words with a consonant cluster as 'CC words'. Words intended for vowel recognition measurements (with a larger selection of middle vowels) were also created according to the common scheme. These are referred to as 'V words'.

For the C words, 15 consonants (/p t k b d g m n l f v s r h j/) were included as the initial consonant. For the CC words, 15 clusters (/bl gl br dr gr sp sj st sk tj pr tr kr kl pl/) were included in this position. For both C and CC words, twelve consonants (/p t k b d g m n l f v s/) were selected as the post-vowel consonant and three vowels (/i ɑ u/) as the middle vowel. /ɑ/ was pronounced as the first phoneme in the Danish word 'arbejde'. The three included vowels represent the boundaries of Danish vowel space. All combinations of the included phonemes were compiled to create $15 \times 12 \times 3 = 540$ words of both the C and the CC type. For the V words, three consonants (/b v n/) were chosen for both the initial position and the post-vowel position. These consonants are characterised by different places of articulation (bilabial, labiodental, and oral) and are thus easily distinguishable by lip reading. Nine vowels (/i e a ɑ y ø u o å/) were selected for the V words, where /a/ was pronounced as in the Danish word 'abe'. The number of V words is $3 \times 9 \times 3 = 81$.

In a large word corpus as the present, the individual words will inevitably have lexical neighbourhoods of different sizes and structures. Lexical neighbours are commonly defined as meaningful words that deviate from the target word by only one phoneme [15]. Due to the discrimination process, the lexical neighbourhood has a large impact on recognition of meaningful words. Words with few lexical neighbours are easier to identify than words with many lexical neighbours due to a lower number of possible confusions [5]. Despite the fact that the words in the present corpus are themselves not meaningful, the size of their neighbourhoods of meaningful words will vary. However, the addition of the fixed 'nonsense' ending /i/ that was introduced to lower the listeners engagement in the lexical discrimination process is also assumed to reduce the impact of the lexical neighbourhood.

Four carrier sentences of different durations were created for the corpus, e.g. for activating the signal processing of a hearing aid before the target word is presented. The duration of the longest carrier (in front of the target word) was required to be at least 2 sec. A call sign was included at the beginning of each carrier to facilitate speech-on-speech masking experiments. The call signs were four girl's names (*Dagmar, Asta, Berit*, and *Gunhild*) and four boy's names (*Bjarke, Kresten, Malthe*, and *Eskild*). These names are relatively unusual Danish first names and of similar duration. The carrier sentences are listed in Appendix A2.

2.3. Recordings

The recordings were conducted in a film studio using professional equipment. The audio was recorded with a TASCAM DR-680 digital recorder (16 bit, 48 kHz) and a DPA 4011 cardioid microphone with a linear frequency response up to 20 kHz. The microphone was mounted on a stand in front and slightly to the right of the talker. The recordings were without salient reflections, but not anechoic. Four professional native Danish talkers, two male (M1 and M2) and two female (F1 and F2), uttered the speech materials. Three of the talkers were selected from the candidates of two speech actor agencies; one female talker was the talker of the DANTALE II speech materials [16]. The voices of the talkers were without special or striking characteristics. They spoke in a neutral voice and clearly pronounced all phonemes in accordance with standard Danish pronunciation. The primary stress was on the first syllable with a long middle vowel. The talkers were instructed to keep a constant sound pressure level (SPL) for the two syllables; to avoid a drifting fundamental frequency (F0) during the recordings; and to maintain a uniform speech rate, slightly slower than everyday speech. All words were recorded individually in a continuous process with a pause of about three seconds between words. Unclearly or incorrectly pronounced words were re-recorded right away or noted for re-recording at the end of the session.

As call signs in the carrier sentences, the two female talkers were assigned the four girl's names and the male talkers were assigned the four boy's names. Each talker recorded the four carrier sentences in combination with each call sign, in total 16 sentences for each talker. The carrier sentences were recorded with the default nonsense words 'marki og marbi' in order to ensure a natural rhythm of the sentences. When using the carrier sentences in experiments, the default words can be replaced by one or two of the individually recorded words.

The audio recordings were supplemented with simultaneous video recordings. A teleprompter enabled the talkers to read the nonsense words while keeping eye contact with the camera. The screening background was a uniform green that can be replaced with any background during post-processing. The talkers wore a black blouse or similar without a collar. For all talkers, the recordings were completed in approximately three hours.

2.4. Post-processing

In the present study, the post-processing primarily targeted the audio recordings. Unfortunately, an audible, low-frequency dominated background noise was discovered in these recordings, most likely stemming from the ventilation system of the film studio. For the most strongly affected recordings, the SNR was down to about 20 dB. All tracks were thus high-pass filtered with a cut-off frequency of 70 Hz, leading to an SNR improvement of about 5 dB. The SNR was further increased by a log-minimum mean square error (MMSE) noise reduction algorithm [17] com-

Table I. Mean phoneme recognition scores [%] across listeners in the identification test. Ten NH listeners participated for each talker. C1 refers to the initial consonant (or consonant cluster), C2 to the post-vowel consonant, and V to the middle vowel.

	Sub-test 1 V words		Sub-test 2 V words	Sub-test 3 C words		Sub-test 4 CC words	
Talker	C1	C2	V	C1	C2	C1	C2
M1	98.3	99.3	97.7	99.1	97.7	97.7	96.6
M2	99.4	99.9	96.2	99.4	98.2	97.1	96.8
F1	97.4	99.5	97.0	98.5	97.2	97.1	95.8
F2	99.9	99.8	98.1	99.4	96.6	97.2	96.0

bined with a MMSE-based noise power estimation algorithm [18], both implemented in MATLAB by Brookes [19]. The overall increase in SNR was about 18 dB, leading to a final, minimum SNR of 38 dB. More typical SNRs for the recordings were about 30 dB before and 50 dB after the processing. The processed word recordings were manually assessed to ensure that they did not contain salient background noises, e.g. coughs, and that the words were pronounced correctly. Words available in more than one recording were compared and the best version was chosen. All carriers and words were stored as separate wav-files and equalized to a common root mean square (RMS) peak level (−18 dB, re: max. digital output) determined in a sliding window of 200 ms. The video recordings were updated with the final, processed audio track, strictly maintaining the alignment between video and audio.

2.5. Phoneme identification test

A phoneme identification test of all nonsense words in the corpus was conducted to determine whether the target phonemes could be correctly identified under optimal listening conditions. Although correctly pronounced, some words were expected to be prone to incorrect phoneme identifications, e.g. due to confusion of similar syllables (e.g. /pibi/). Such words should be identified to give users of the corpus the option to omit them when compiling test stimuli. Four groups of ten normal-hearing (NH) listeners participated in the identification test. Each group listened to all the words uttered by one of the four talkers. The requirements for participation were: (1) Age of 18–45 years; (2) listeners reported to have a normal hearing and no history of hearing problems; (3) Danish as native language; (4) no indication of dyslexia; and (5) linguistically naïve. Participation was approved by The National Committee on Health Research Ethics.

The identification test was self-administered by the listeners and conducted during one visit of about three hours, including instruction. The listeners could set their own pace and were allowed to pause whenever they wanted. The nonsense words were presented at a comfortable listening level without any background noise or other degradations. Each session consisted of four sub-tests that all listeners conducted in the same order. The tasks in the four sub-tests were: (1) Scoring the two consonants of the 81 V words; (2) scoring the middle vowel of the 81 V words; (3) scoring the two single consonants of the 540 C words; and (4) scoring the consonant cluster and the single consonant of the 540 CC words. Although the consonants in the vowel words were not intended as target phonemes, sub-test 1 was conducted to check whether these consonants were perceived correctly. Prior to each sub-test, the listeners received oral instructions and ran a short training session. Alphabetical letters on separate response buttons represented the phonemes. A '?' button was not included and the words could not be repeated.

The mean phoneme recognition scores of the identification test (across listeners and phonemes) are shown in Table I. 'C1' refers to the initial consonant or consonant cluster and 'C2' refers to the post-vowel consonant. A more detailed analysis of the results for the C words revealed a large number of confusions between /k/ and /g/ and between /b/ and /p/. These confusions occurred in both consonant positions, but were particularly frequent for position C2. The CC words displayed similar confusions for C2, while confusions between /kr/ and /gr/, between /br/ and /pr/, and between /bl/ and /pl/ were frequent for C1. Table I reveals a variation in the recognition scores across talkers, but this should be interpreted with caution, since a different group of listeners participated for each talker.

The distributions of the recognition scores in Table I were analysed. One incorrect scoring (from one of the ten listeners) occurred for a relatively large number of target phonemes, up to 96 for position C2 in the CC words (18% of 540). These incorrect scorings were considered coincidental and not as an indication of any phoneme identification problem related to the word itself. Only words with two or more incorrect scorings of a target phoneme were assumed to have characteristics that could cause difficulties with the phoneme identification. These words were maintained in the corpus but listed in Appendix A1 as a hint to users of the corpus. In the appendix, the number of incorrect scorings is detailed for each target phoneme and each talker. In a few of the V words, confusions of the non-target phonemes /b/ and /v/ occurred more than once; these words are also listed in Appendix A1. Two words ('prubi' by talker F2 and 'vybi' by talker M1) achieved a score of 10 due to clearly incorrect pronunciations; these two words were omitted from the final corpus.

3. Evaluation of the corpus

3.1. Rationale

An evaluation test of the speech corpus and an associated test procedure was conducted. The evaluation had four specific goals: (1) To establish the SNR where 50% of the phonemes were scored correctly when presenting the nonsense words in noise to NH listeners; (2) to assess the sensitivity, i.e. the dependency of the recognition score on the SNR; (3) to assess the test-retest variability, i.e. the repeatability of the phoneme recognition scores under similar listening conditions; and (4) to investigate memory effects, which are considered to be involved when the listener's performance improves (beyond the effect of practice) during repeated use of the same speech material. Memory effects are undesirable since they prevent reuse of the same word list as a method to reduce test-retest variability.

The evaluation test included four separate test lists compiled from words spoken by each of the four talkers. Additionally, four test lists with other characteristics were compiled for talker M1. The purpose of these additional lists was to measure recognition scores for subsets of phonemes, e.g. phonemes with a high frequency spectrum. Observed differences between the lists were assumed to be representative for all talkers; only one set of additional lists was thus compiled.

3.2. Methods

3.2.1. Listeners

Seven NH listeners (four male, three female) participated in the evaluation test; they fulfilled the same requirements as in the phoneme identification test. The listeners were between 20 and 22 years of age (mean 21.1 years) and paid on an hourly basis for their participation. Five of the listeners had also participated in the identification test. Participation was approved by The National Committee on Health Research Ethics.

3.2.2. Stimuli

The presented nonsense words were organised in lists of 54 to 90 words. The lists were compiled from words that had passed the identification test, i.e. the words were not among the ones listed in Appendix A1. A few additional words were excluded, either because they were real words and very common, e.g. 'mini', or because they had resemblance to offensive words.

For each of the four talkers, a 'general consonant type' list (GC type) with 90 C words was compiled. These lists included all the C1 single consonants and all the C2 consonants. Four additional word lists were compiled for talker M1: (1) A 'high-frequency consonant type' (HF type) containing 30 C words and 60 CC words with an obstruent in position C1; (2) a 'short-duration consonant type' (ShD type) containing 30 C words and 30 CC words with common Danish consonant clusters; the clusters have a shorter duration than the two phonemes pronounced separately; (3) a 'low-intensity consonant type' (LI type) containing 60 C words without high-frequency noise; and (4) a 'vowel type' (V type) containing 54 words that included all nine middle vowels.

The phoneme selection in the additional consonant lists deviated from the GC type in the following positions: HF list, C1: /f k s t p sp sj st sk tj pr tr kr kl pl/; ShD list, C1: /b l r d g bl gl br dr gr/; LI list, C1: /b d g m n l v j r h/; and LI list, C2: /b d g m n l v/. Note that the GC, HF, and ShD lists only differ with respect to the C1 phonemes. The nonsense words of the eight word lists are listed in Appendix A2.

Phonemes and phoneme combinations were repeated an equal number of times (as closely as possible) within each word list. The numbers of repetitions for the GC lists were: 6 for the C1 consonants; 7 or 8 for the C2 consonants; 30 for the middle vowel; 2 for each CV-combination; and 2 or 3 for each VC-combination. The two consonants were not allowed to be identical within a word. Due to this systematic phoneme repetition pattern, each word differed by only one phoneme from at least two other words in the list, making it difficult to keep them apart and thus potentially reducing memory effects. For the remaining list types, the criteria were similar (with varying numbers of repetitions), although for the V list, the consonants were allowed to be identical.

The target words were presented without a carrier sentence in a stationary, speech-shaped background noise, spectrally matched to the individual talker. For each list presentation, the order of the test words was randomized. The noise was windowed around the presented words, starting 1 sec. before the word and ending 0.4 sec. after. The SPL of the noise was fixed at 65 dB; different SNRs were achieved by adjusting the level of the speech signal. The applied SNRs depended on the list type and the talker; they were maintained constant during each list presentation.

3.2.3. Apparatus and procedure

The test procedure for the evaluation test was intended as an example of how phoneme perception and recognition scores can be measured using the present speech corpus. The procedure was designed to be executable in less than ten minutes per word list, and without the presence of an experimenter, i.e. it was self-administered. A self-administered test saves the listener-to-experimenter report time for each presented stimulus, and it is not prone to errors due to misinterpretations by the experimenter. Several listeners can also run the test in parallel if the facilities allow this. The test procedure was implemented in MATLAB.

The test was divided into two sessions (two visits separated by several days) for each listener. In session 1, the listeners scored the four GC type lists, one for each talker. The order of the talkers was counterbalanced across listeners using Latin squares. In session 2, the listeners scored the four additional list types for talker M1 in the order HF – ShD – LI – V for all listeners. Finally, the listeners

scored the target phonemes in five repetitions of the GC list (talker M1) to investigate test-retest variability and memory effects. This repetition test was conducted as the last of all tests to avoid potential improvements due to practice. Improved recognition scores during the repetition test could thus be ascribed to memory of the words in the list. Note that the repetition test was separated from a previous presentation of the list (the M1 GC list) by several days and at least 12 other list presentations (the HF, ShD, LI, and V lists at three SNRs). The listeners were thus unlikely to have any recollection of the list before starting the repetition test.

The evaluation test was conducted simultaneously in up to four adjacent soundproof booths. The stimuli were presented over Sennheiser HD 650 headphones with a calibrated SPL. Each word list was presented at three SNRs in a randomized order before continuing with the next list. Based on the results of a pilot test conducted by the first author, the SNRs were set to obtain average phoneme recognition scores of approximately 25% to 75%. The same SNRs were maintained across listeners. For the GC lists, the SNRs were [dB SNR]: −7, −3, 1 for talker M1; −7, −3, 1 for talker M2; −6, −2, 2 for talker F1; and −5, −1, 3 for talker F2. For talker M1, the SNRs for the four additional list types and the repetition test were [dB SNR]: −9, −4, 1 for the HF list; −9, −4, 1 for the ShD list; −4, 2, 8 for the LI list; −11, −7.5, −4 for the V list; and −4 dB for all five repetitions in the repetition test.

The scoring task did not require knowledge of phonetic notation but was based on Danish alphabetical letters in the scoring panels. When the listener scored a target phoneme, the selected phoneme/letter was displayed on the computer screen in a text field where the non-target phonemes were pre-filled. When the listener had completed the scoring of a word, the text field would spell out the listener's perception of the nonsense word in an easily readable format. No indication of correctness was given. The listener could change the phoneme scoring (and the corresponding letter would change in the text field) until he/she advanced to the next stimulus by pressing 'Next'. The listener did not have the option to replay the stimulus or to go back to the previous word.

The response options available to the listener changed according to the type of list under test. The response buttons were grouped in a separate panel for each target phoneme, and the buttons corresponded 1:1 to the phonemes included in the word list. A '?' response button was also displayed. The intention was to prevent guessing from confounding the outcome data when a phoneme was inaudible. For each presented word, the listener's phoneme responses were logged in a table, and a confusion matrix was compiled (stimulus phoneme vs. response phoneme).

The listeners received oral instructions before running the tests. They were encouraged to guess if they were in doubt about the target phonemes and instructed to press '?' only when the phoneme was inaudible, not as an expression of uncertainty about the correct response. A short training task was conducted before each change of talker or list type. The training was conducted at the middle SNR of the three pre-defined levels. The training lists contained 20 words of the same type as in the test list that followed. Additionally, five training words were included at the beginning of each test list; to the listener, these appeared to be part of the list. The purpose was to eliminate possible incorrect scorings due to the sudden SNR changes between list repetitions, especially when the SNR dropped. The listeners were given a short pause after each set of three list presentations.

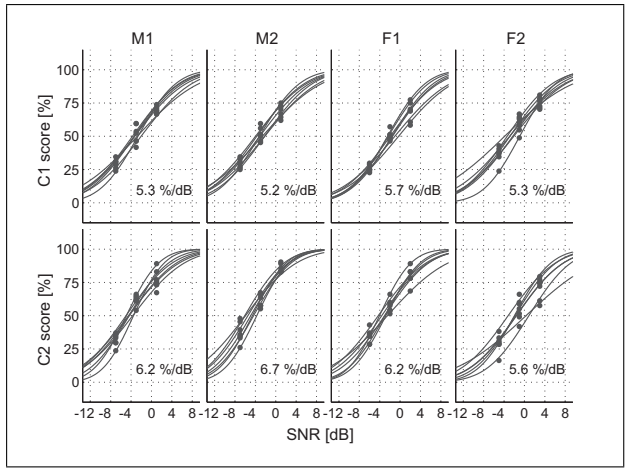

Figure 1. Mean phoneme recognition scores (across phonemes) for the GC lists. Dots indicate the score for each of the seven listeners at the three SNRs. A cumulative normal distribution curve was fitted to the data for each listener. The steepest slope (mean across listeners) is indicated in each panel.

3.3. Results

Phoneme recognition scores were defined as the number of correctly scored phonemes divided by the total number of presented phonemes. Since the test procedure was self-administered with a limited number of scoring options, the recognition scores were affected by the chance probability of a correct scoring. All recognition scores and SRTs reported here were adjusted to compensate for this. For example, with ten scoring options, a recognition score of 10% equals the chance probability of a correct scoring and therefore represents a true recognition score of only 0%.

3.3.1. The general consonant (GC) lists

The general consonant list type (GC) was tested with all four talkers. The listeners scored the two target phonemes, C1 and C2. For each listener, the mean recognition scores (across phonemes) at the three SNRs are shown as dots in Figure 1. Individual cumulative normal distribution functions were fitted to the results. These performance-intensity (PI) functions were quite shallow; the steepest slopes were 5.2-6.7%/dB, as indicated in the panels. A separate repeated measures analysis of variance (ANOVA) of the slopes for C1 and for C2 showed no significant difference between the talkers. Between C1 and C2, the only significant slope difference was observed for talker M2

Table II. Mean SRTs and SDs across listeners for the GC word lists [dB SNR]. The SRTs for C1 include the scores for phonemes /j h r/ that were presented in this position only.

	C1		C2		mean
Talker	SRT	SD	SRT	SD	SRT
M1	-3.0	0.6	-4.3	0.6	-3.7
M2	-3.0	0.7	-5.0	0.9	-4.0
F1	-1.6	0.8	-3.3	0.7	-2.5
F2	-2.6	1.0	-1.3	1.4	-2.0
mean	-2.6		-3.5		-3.0

Table III. Mean consonant recognition scores [%] for the four talkers in the GC test as function of the middle vowel. Scores are averaged across target phonemes, SNRs, and listeners.

	C1				C2			
Vowel	M1	M2	F1	F2	M1	M2	F1	F2
/i/	42.3	45.6	48.0	53.6	53.0	59.3	50.3	50.2
/ɑ/	73.0	69.9	58.7	68.7	66.2	60.0	63.9	43.7
/u/	34.9	34.5	37.8	49.0	46.4	60.9	55.3	59.8

Table IV. Mean SRTs and SDs across listeners for the additional test lists [dB SNR]. The GC results are the same as the M1 results in Table II.

	C1		C2		mean
List type	SRT	SD	SRT	SD	SRT
GC	-3.0	0.6	-4.3	0.6	-3.7
HF	-7.0	0.5	-3.3	0.8	-5.2
ShD	-1.3	0.9	-5.0	0.9	-3.2
LI	-0.9	0.5	-1.9	1.2	-1.4
mean	-3.1		-3.6		-3.4

[F(1,13) = 17.92, p = 0.006]. Based on the fitted functions, an SRT was estimated for each listener. The mean SRTs and their standard deviations (SDs) across listeners are listed in Table II. A repeated measures ANOVA of the SRTs with the talkers and the position (C1 and C2) as factors, showed a significant difference between talkers [F(3,55) = 41.67, p < 0.0001], a significant difference between C1 and C2 [F(1,55) = 38.07, p < 0.0001], and a significant interaction between position and talker [F(3,55) = 25.73, p < 0.0001]. When averaged across C1 and C2, the highest SRT was observed for talker F2 (−2.0 dB SNR), the lowest for talker M2 (−4.0 dB SNR). The mean SRT was 0.9 dB lower for C2 than for C1, indicating that the consonants were slightly easier to identify in the post-vowel position than in the initial position. Talker F2 deviated from this pattern by displaying a lower SRT for C1 than for C2.

The influence of the middle vowel is considered in Table III, where the recognition scores are shown as a function of the vowel. Scores were averaged across target phonemes, SNRs, and listeners. An ANOVA of the recognition scores with the vowel, the talker, and the position (C1 and C2) as factors and listeners as repeated measures was conducted. A highly significant effect of the vowel was found [F(2,167) = 177.01, p < 0.0001]. The scores in the table show that the effect is larger in position C1 than in position C2. Typically, the highest scores were observed in combinations with the vowel /ɑ/. Highly significant interaction effects were found between vowel and talker [F(6,167) = 14.42, p < 0.0001] and between vowel and position [F(2,167) = 76.87, p < 0.0001]. The interaction between middle vowel and talker can be observed, for example, in the C1 recognition scores for talkers M1 and F2. While the consonant score varied with 37% between /ɑ/ and /u/ for M1, the variation was only 19.2% for F2. In Table III, no effect of talker was found for the recognition scores [F(3,167) = 2.25, p = 0.0861], despite the significant effect of talker on the SRTs in Table I. This apparent inconsistency is explained by the individually SNRs that were selected in order to achieve recognition scores in the interval 25–75 % for all four talkers. The non-significant effect of talker on the recognition scores indicate that these SNRs were selected adequately.

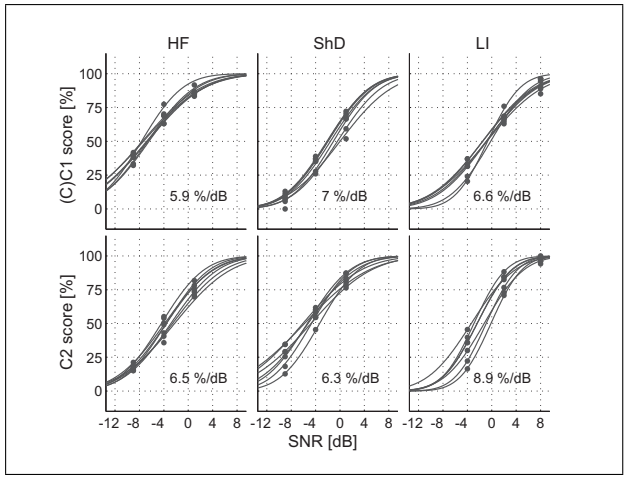

Figure 2. Mean phoneme recognition scores (across phonemes) for the additional test lists. Dots show the score for each of the seven listeners at the three SNRs. A cumulative normal distribution curve was fitted to the data for each listener. The steepest slope (mean across listeners) is indicated for the curves in each panel.

3.3.2. Additional consonant lists

For the HF, ShD, and LI lists, the listeners scored the consonant or consonant cluster in position C1 and the single consonant in position C2. These lists were only tested with talker M1. The mean recognition scores (across phonemes) are shown as dots in Figure 2 as a function of the three tested SNRs. The slopes of the additional PI functions are slightly steeper than those observed for the M1 GC list. Two separate repeated measures ANOVAs of the slopes of all M1 functions, i.e. including the two M1 functions in Figure 1, showed a significant difference in slope between list types for position C1 [F(3,27) = 4.00, p = 0.024] and for C2 [F(3,27) = 11.08, p = 0.0002]. Sepa-

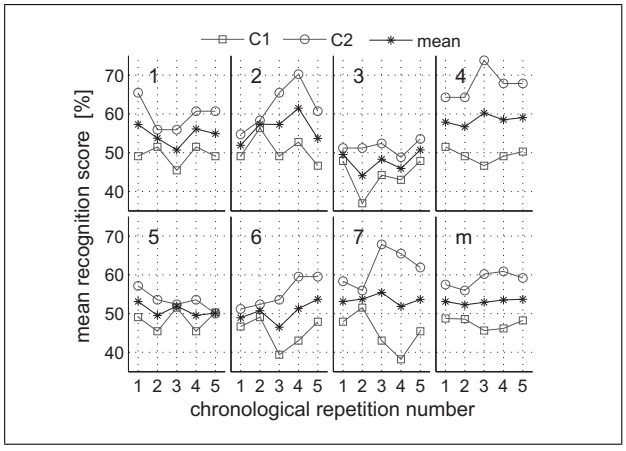

Figure 3. The mean recognition scores (across phonemes) for C1 and C2 in five repetitions of the GC word list with talker M1. The lower, rightmost panel shows the mean across the seven listeners. The SNR was fixed at −4 dB. The repetition of the lists did not have a significant influence on the listeners' performance.

rate ANOVAs of the slope differences between C1 and C2 only showed a significant difference for the LI list [F(1,13) = 12.4, p = 0.013].

The mean SRTs and SDs across listeners are shown in Table IV; the results for the GC list of talker M1 (from Table II) are included. The SRTs of all four lists were included in two repeated measures ANOVAs with list type as factor; the position was not included as a factor since it was confounded with a difference in phonemes. The effect of list type was highly significant in position C1 [F(3,27) = 157.84, p < 0.0001] and in position C2 [F(3,27) = 24.2, p < 0.0001]. While the SRT variations between list types were expected for C1, the rather large SRT differences between the GC, HF, and ShD lists for C2 are worth noting, since the phonemes in this position were identical across lists. A separate ANOVA of the SRTs for C2 for these three lists showed a highly significant effect of list [F(2,20) = 21.85, p = 0.0001]. The listeners' scoring of the C2 phoneme is thus affected by the C1 phoneme, since only C1 phonemes differed between the GC, HF, and ShD lists.

3.3.3. Repetition test

In the repetition test, the GC list uttered by talker M1 was repeated five times. For each of the seven listeners, Figure 3 shows the recognition score of C1 and C2 (averaged across phonemes) and the mean of the two scores. In the lower, rightmost panel, the overall mean across the seven NH listeners is displayed. The within-subject SD across the five repetitions was 3.6% for C1, 4.1% for C2, and 2.4% for the mean. A repeated measures ANOVA of the recognition scores showed no significant effect of repetition for the mean of the C1 and C2 scores [F(4, 34) = 0.35, p = 0.84]. Neither did separate ANOVAs of the C1 scores [F(4, 34) = 1.15, p = 0.35] and C2 scores [F(4, 34) = 1.92, p = 0.14] show a significant effect of repetition. These results indicate that memory effects did not improve the listeners' performance during the course of the test; nor did the performance deteriorate due to fatigue.

Assuming the within-subject SDs found in the repetition test, the required number of NH listeners for a statistically significant detection of an improved recognition score of 1%, 2%, or 5%, respectively, at a significance level of 5% (one-tailed) and a statistical power of 80%, would be 81, 21, or 4 for the C1 phonemes. For the C2 phonemes, the required number of listeners would be 104, 26, or 5. For the mean recognition score of C1 and C2, the numbers would be 36, 9, and 2. Improved recognition scores of 1%, 2%, and 5% would correspond to an increase in the SNR of 0.17 dB, 0.35 dB, and 0.87 dB, respectively, assuming a PI slope of 5.8%/dB (the mean of C1 and C2 for list M1 GC). In experiments involving HI listeners, the SDs are likely to become larger, thus requiring more listeners to make statistically significant detections.

3.3.4. Individual phoneme SRTs

The average recognition scores (across phonemes) reported so far do not reveal the differences in recognition score between phonemes. Large variations in the recognisability of different consonants have previously been demonstrated, e.g. [4, 20, 21]. In the present study, recognition scores for the individual target phonemes in the GC word lists of the four talkers (averaged across listeners) were determined and an SRT was estimated for each phoneme. The results are shown in Figure 4. Due to floor and ceiling effects, a few estimates led to extreme values of the SRT; these were replaced by a minimum value, −12 dB SNR, or a maximum value, 8 dB SNR. A large spread of the SRTs can be observed, with some clustering in the interval −2 to 2 dB SNR. Similarly, for the HF, ShD, and LI lists (talker M1), individual SRTs were determined for the consonants and consonant clusters. The results are shown in Figure 5. As in Figure 4, large SRT differences between phonemes can be observed.

3.3.5. Vowel list

For the vowel list (talker M1) where the middle vowel is the target phoneme, a mean SRT of −8.2 dB SNR was calculated across all nine vowels and the seven listeners, i.e. a lower SRT than for consonants in the other list types. The SD of the SRT between listeners was relatively low, 0.59 dB. As for the consonants, large differences in the recognition scores were observed for the individual vowels. At the middle SNR (−7.5 dB), the following mean scores (across listeners) were measured: /a/: 97%, /ɑ/: 100%, /e/: 46%, /i/: 0%, /y/: 44%, /ø/: 63%, /o/: 44%, /å/: 54%, /u/: 33%. The mean steepest slope of the PI functions across listeners was 8.1%/dB.

4. Discussion

4.1. Observed SRTs for the word lists

In the present investigation, the overall SRTs were determined for lists where the scoring tokens were single consonants, a combination of single consonants and consonant clusters, or vowels. The lowest SRT (−8.2 dB SNR) was

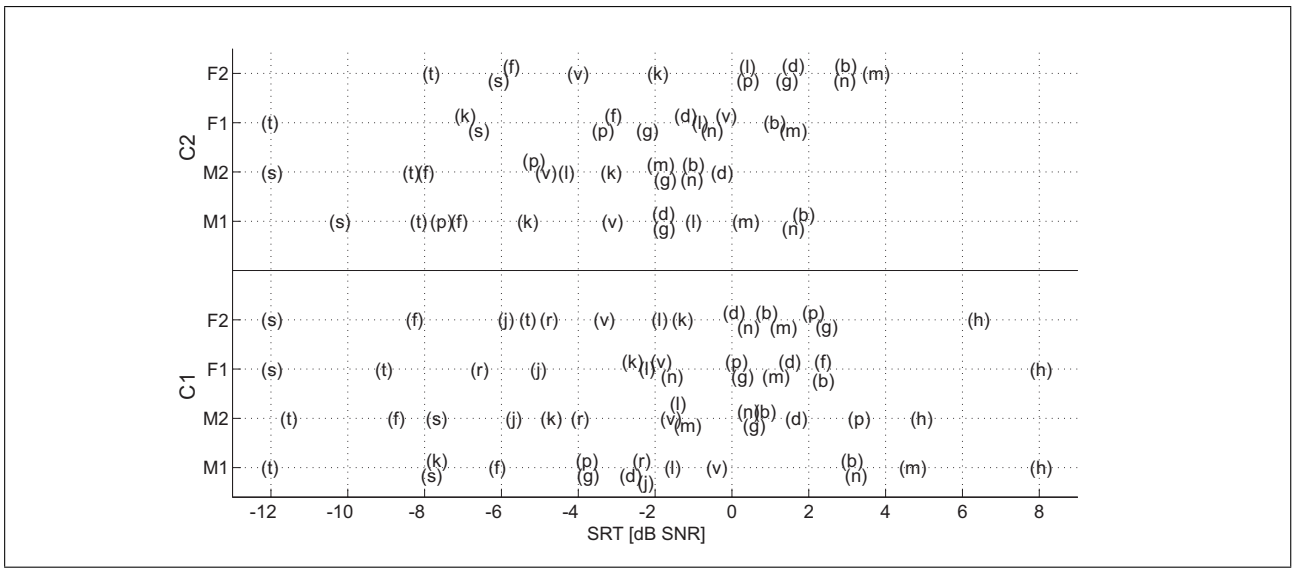

Figure 4. SRTs for the individual phonemes in the GC word list for each of the four talkers. The SRTs are estimated from PI functions that were fitted to the recognition scores at the three SNRs in the test. The minor vertical shifts of the phonemes are for readability.

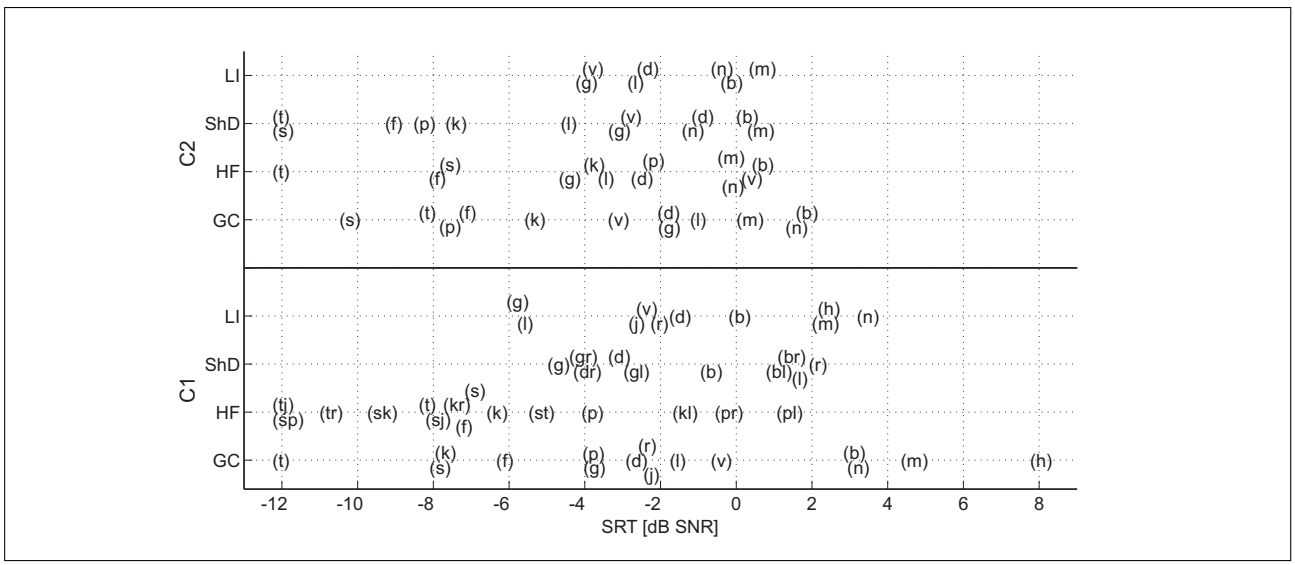

Figure 5. SRTs for the individual phonemes (including consonant clusters) in the tests of the GC, HF, ShD, and LI lists uttered by talker M1. The SRTs are estimated from PI functions that were fitted to the recognition scores at the three different SNRs for each list type. The minor vertical shifts are for readability. The GC list corresponds to the M1 list in Figure 4.

observed for the vowels, in correspondence with the findings in previous investigations, e.g. [22], that vowels generally achieve lower SRTs than the majority of consonants. For the different consonant list types of the present study (tested with talker M1), the relative SRTs for the C1 tokens were as expected. The order was: (1) the 'high-frequency' HF list (−7.0 dB SNR), (2) the 'general consonant' GC list (−3.0 dB SNR), (3) the 'short duration' ShD list (−1.3 dB SNR), and (4) the 'low intensity' LI list (−0.9 dB SNR), see Table IV. While high-frequency consonants are challenging for hearing-impaired listeners, they are typically easy to identify for NH listeners due to their spectral deviation from the speech-shaped background noise. This corresponds well with Figure 5, where the SRTs of the individual consonants and consonant clusters are plotted. The majority of the HF C1 tokens are below −6 dB SNR, while the C1 tokens of the other three lists mainly fall above this limit. Figure 5 also corresponds well with the higher SRTs for C1 in the ShD and LI lists. The ShD C1 tokens are clustered at a relatively high SRT, and the LI C1 tokens are shifted slightly towards even higher SNRs.

Previous investigations of consonant perception in a speech-shaped noise have shown large variations of the observed SRTs. For 16 American English consonants in the initial position of CV tokens, Phatak and Allen [22] measured recognition scores that corresponded to a mean SRT of about −20 dB SNR. For 20 American English consonants in the initial position of CVC tokens, Woods *et al.* [23] determined the SNR at the 65% recognition point to be 6.5 dB, corresponding to an SRT of about +3.5 dB SNR

(assuming a PI function slope of 5%/dB). The SRTs for the initial consonant in the GC lists of the present corpus were between −3.0 to −1.6 dB SNR, i.e., intermaediate to the two American investigations. Possible causes for the large SRT variations between these investigations are the precise selection of consonants, the format of the tokens (CV, CVC, CVC+i), the talkers of the materials, the test procedure, and the procedure for determining the signal SPLs and hence the SNRs. Although primarily intended for comparative A/B experiments, the absolute SRT of the present speech corpus is not irrelevant. In order to ensure ecologically valid measurements, the SRT of a speech test should be as close as possible to the values observed in challenging real-life situations [24]. The typical SNR at a 'cocktail party' was estimated at 0 dB SNR by Plomp [25], while Smeds *et al.* [26] estimated the SNR in challenging situations to be approximately 5 dB higher. The latter result is likely to be biased towards higher SNRs because the informants were hearing-aid users, who may have become accustomed to avoiding more challenging situations in their daily life. The SRTs observed in the evaluation test of the present corpus are slightly lower than these real-life values. If considered too low to ensure ecological validity, steps can be taken to raise the SNR when conducting investigations. Such measures could be to apply a different background noise, to determine the SRT at a higher recognition score than 50%, or to consider the choice of talker. In Table II, SRT differences of up to 3.7 dB can be observed between talkers.

The noise in the present study was speech shaped and individually matched, resulting in similar spectral deviations between the speech and the background noise for all four talkers. If the same background noise, e.g. white noise, were applied for all talkers, different spectral deviations would emerge and presumably lead to larger differences in recognition scores, c.f. [22, 23]. Thus, when not relying on a speech-shaped noise in an investigation, the influence of the talker may be larger than expected from the slopes in Figure 1 and the SRTs in Table II.

4.2. Memory effect and test-retest variability

In the repetition test, recognition scores were determined for the same GC test list (talker M1) in five repetitions within a short time frame. As reflected in the mean score across listeners (Figure 3, lower rightmost panel), no significant effect of repetition was found. The five measurements can thus be regarded as independent with no systematic variation due to memory or fatigue. This result implies that a list can be compiled from the present speech corpus and used repeatedly, e.g. for comparisons of recognition scores under different experimental conditions. Although the corpus is large enough for compiling several lists of the same type, reusing the same list is preferable in order to reduce the test-retest variability. However, the issue of a memory effect should be reconsidered if investigations are conducted with test lists that are substantially shorter than in the repetition test (90 words). Obviously, a shorter list would be easier to remember and memory effects would thus be more likely to occur.

4.3. Sensitivity of the speech materials

MacPherson [28] compared the PI functions of a large number of tests and observed that the recognition scores of very short tokens, e.g. phonemes, were relatively insensitive with respect to changes in the SNR compared to longer tokens, e.g. sentences. MacPherson observed an average steepest slope of 5%/dB SNR for short target tokens. Thus, the 5–9%/dB slopes observed for the present corpus (see the PI functions in Figure 1 and Figure 2) are in the upper range of what could be expected, although low compared to the slope of a sentence-based test, e.g. the Danish HINT (16.8%/dB for NH listeners [29]). The relatively low sensitivity of the recognition score to the SNR is partly caused by the between-phoneme variations of the SRT. These variations are much larger than between the tokens of a sentence-based test. In the GC lists, SRT differences of up to ±8 dB SNR were common between the target phonemes (Figure 4), while the SRT differences of the equalised sentences in, for instance, a Hearing in Noise Test (HINT), are assumed to be below ±1.5 dB SNR [30]. In a phoneme-based test, the recognition scores of the individual target tokens can be partly equalized by including only phonemes with similar SRTs, thus presumably increasing the sensitivity of the test. In the present study, this seems confirmed by the relatively steep PI functions for C2 in the LI type list (Figure 2, panel C2-LI). These functions were based on a subset of phonemes (/b d g m n l v/) with little variation in the SRT (see Figure 5, panel C2- LI). Using subsets of phonemes will also reduce the likelihood that floor and ceiling effects occur, i.e. recognition scores reaching 0% or 100%. A detailed analysis of the test data showed many occurrences of such effects, e.g. for C1 in the M1 GC list. Although the mean score at 1 dB SNR was 67.6% (Figure 1, upper leftmost panel), the score for /t/ and /s/ was 100% for all listeners. At the low SNR, −7 dB, the opposite effect could be observed for /r/ with a score of 0% for all listeners. When attempting to design experiments that avoid floor and ceiling effects, the strong influence of the middle vowel should also be considered. Although these effects may be absent when regarding mean scores across the three middle vowels, a ceiling effect may still occur in combination with /ɑ/ or a floor effect in combination with /u/, cf. Table III.

4.4. The nonsense corpus as a tool for speech intelligibility measurements

The statistical power of the repetition test showed that an SNR improvement of 0.35 dB would be detectable in the mean of the C1 and C2 scores with nine list presentations in each condition. To detect a similar SRT improvement between two conditions, the Danish HINT (within-subject SD of 0.86 dB [29]) would require 38 list presentations in each condition, i.e. four times as many. This implies that the present test is a quite effective measure of intelligibility.

However, the efficiency can partly be explained by the long GC lists (90 words, test duration about 9 min.) compared to the HINT lists (20 sentences, test duration about 3 min.) Furthermore, the calculated efficiency of the nonsense test requires that it should be run at the SNR where it is most sensitive, i.e. at the 50% point of the PI function. In practice, this may be difficult to achieve and additional pilot testing will be required. Identification of the 50% point is not necessary for the HINT that is based on an adaptive procedure. Finally, the SRTs measured with the sentence-based HINT are probably more closely correlated with real-life speech intelligibility than the recognition scores of the phoneme-based test.

Despite these reservations, the nonsense test seems to be effective as a test of speech intelligibility, especially when outcomes at a microscopic level are desired. Whether the present nonsense corpus is applicable as a tool for general measurements of speech intelligibility needs to be investigated in following experiments.

4.5. Linguistic considerations

In the identification test, a relatively large part of the words received incorrect scorings across talkers (see Appendix A1). For these words, the word characteristics and not the pronunciation by the individual talker are likely to have caused the incorrect scorings. A large part of these words had a /g/ in position C2. Closer inspection of the scorings revealed that the /g/ was usually scored as a /k/. The reason is presumably that a written <k> in this position is typically pronounced as [g] in Danish. The listeners are so accustomed to this shift that a stimulus [g] evokes the mental image of the letter <k>. A longer training session may reduce the number of these confusions, especially if the issue was brought to the listeners' attention. Nevertheless, the more reliable approach might be to avoid the words in Appendix A1.

The included target phonemes in the present study were required to be unambiguously represented by alphabetical letters in the scoring panel. During the phoneme identification test, this requirement was revealed as unfulfilled by two of the consonant clusters, <sk> and <sp>. In Danish, the orthography and the pronunciation do not match for these two clusters. The cluster <sk> is pronounced as [sg] and <sp> is pronounced as [sb]. During testing, only one listener seemed to notice the inconsistency and the problem might thus be negligible. Nevertheless, it does represent a breach of the intended correspondence between the graphic representation of the scoring options and the phoneme pronunciation. A solution could be to replace 'sk' and 'sp' in the scoring panel by 'sg' and 'sb'. Although Danish orthography does not include these two consonant clusters, listeners will presumably consider them as valid representations of the phonemes.

The phonetic inventory of the present corpus is assumed to have sufficient similarities to the inventory of other (European) languages to produce test results that are relevant for these. However, the results of the present study can only be assumed to be valid for native Danish speakers.

5. Conclusion

In the present project, a corpus of nonsense words was developed. The corpus was evaluated using a self-administered test procedure. The evaluation test demonstrated that the speech materials, in combination with this procedure, represent an effective method for measuring phoneme recognition scores. The overall SRT of the general consonant (GC) word lists was about −3 dB SNR, which presumably will allow for ecologically valid measurements. The present corpus includes video recordings of all speech materials. This makes the corpus suitable for testing with hearing-impaired listeners where lip reading is required to achieve reliable results.

Acknowledgement

This study was conducted in collaboration with the Danish hearing aid companies Oticon, GN Resound, and Widex. Thank you to the following colleagues for their contributions and involvement: Lise Bruun Hansen (Oticon), Anja Kofoed Pedersen and Michael Nielsen (Widex), Charlotte T. Jespersen (GN Resound), and Ruben Schachtenhaufen (University of Copenhagen). Finally, thank you to all the listeners who took time to participate. This research was supported by the Centre for Applied Hearing Research (CAHR).

Appendix / Supplementary material

The file
'v105n01_nielsen_dau_supplementary_files.zip',
containing Appendices A1 and A2 can be downloaded via

http://aaua-material.com/t_AX2616

References

[1] J. D. Robinson, T. Baer, B. C. J. Moore: Using transposition to improve consonant discrimination and detection for listeners with severe high-frequency hearing loss. Int. J. Audiol. **46** (2007) 293–308.

[2] J. M. Alexander, R. L. Jenison, K. R. Kluender: Real-Time Contrast Enhancement to Improve Speech Recognition. PLoS One. **6** (2011) e24630.

[3] C. Scheidiger, J. B. Allen, T. Dau: Assessing the efficacy of hearing-aid amplification using a phoneme test. J. Acoust. Soc. Am. **141** (2017) 1739–1748.

[4] J. Zaar, T. Dau: Sources of variability in consonant perception of normal-hearing listeners. J. Acoust. Soc. Am. **138** (2015) 1253–1267.

[5] P. A. Luce, D. B. Pisoni: Recognizing Spoken Words: The Neighborhood Activation Model. Ear Hear. **19** (1998) 1–36.

[6] R. M. Warren: Perceptual Restoration of Missing Speech Sounds. Science **167** (1970) 392–393.

[7] M. K. Pichora-Fuller, B. A. Schneider, M. Daneman: How young and old adults listen to and remember speech in noise. J. Acoust. Soc. Am. **97** (1995) 593–608.

[8] A. Boothroyd, S. Nittrouer: Mathematical treatment of context effects in phoneme and word recognition. J. Acoust. Soc. Am. **84** (1988) 101–114.

[9] D. E. Broadbent: Word-frequency effect and response bias. Psychol. Rev. **74** (1967) 1–15.

[10] T. U. Christiansen, P. J. Henrichsen: Sense Meets Nonsense – a dual-layer Danish speech corpus for perception studies. - In: Proc. 8th Int. Conf. Lang. Resour. Eval., Istanbul, 2012, 3356–3361.

[11] S.-H. Jin, C. Liu: English vowel identification in quiet and noise: effects of listeners' native language background. Front. Neurosci. **8** (2014) 1–8.

[12] P. Lyregaard: Towards a Theory of Speech Audiometry Tests. - In: M. Martin (Ed.), Speech Audiometry, Taylor & Francis, London, 1987, 33–61.

[13] D. O'Shaughnessy: Consonant Durations in Clusters. IEEE Trans. Acoust. ASSP-**22** (1974) 282–295.

[14] E. Fischer-Jørgensen: Phonetic analysis of the stød in standard Danish. Phonetica **46** (1989) 1–59.

[15] V. Taler, G. P. Aaron, L. G. Steinmetz, D. B. Pisoni: Lexical neighborhood density effects on spoken word recognition and production in healthy aging. J. Gerontol. B. Psychol. Sci. Soc. Sci. **65** (2010) 551–560.

[16] K. Wagener, J. L. Josvassen, R. Andenkjær: Design, optimization and evaluation of a Danish sentence test in noise. Int. J. Audiol. **42** (2003) 10–17.

[17] Y. Ephraim, D. Malah: Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. IEEE Trans. Acoust. **33** (1985) 443–445.

[18] T. Gerkmann, R. C. Hendriks: Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay. IEEE Trans. Audio. Speech. Lang. Processing **20** (2012) 1383–1393.

[19] M. Brookes: VOICEBOX: Speech Processing Toolbox for MATLAB. (2015). www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (accessed July 6, 2015).

[20] G. A. Miller, P. E. Nicely: An Analysis of Perceptual Confusions Among Some English Consonants. J. Acoust. Soc. Am. **27** (1955) 338–352.

[21] J. B. Allen: Consonant recognition and the articulation index. J. Acoust. Soc. Am. **117** (2005) 2212–2223.

[22] S. A. Phatak, J. B. Allen: Consonant and vowel confusions in speech-weighted noise. J. Acoust. Soc. Am. **121** (2007) 2312–2326.

[23] D. L. Woods, E. W. Yund, T. J. Herron, M. A. I. Ua Cruadhlaoich: Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise. J. Acoust. Soc. Am. **127** (2010) 1609–1623.

[24] J. B. Nielsen, T. Dau, T. Neher: A Danish open-set speech corpus for competing-speech studies. J. Acoust. Soc. Am. **135** (2014) 407–420.

[25] R. Plomp: Acoustical Aspects of Cocktail Parties. Acta Acust. United with Acust. **38** (1977) 186–191.

[26] K. Smeds, F. Wolters, M. Rung: Estimation of Signal-to-Noise Ratios in Realistic Sound Scenarios. J. Am. Acad. Audiol. **26** (2015) 183–196.

[27] S. A. Phatak, A. Lovitt, J. B. Allen: Consonant confusions in white noise. J. Acoust. Soc. Am. **124** (2008) 1220–1233.

[28] A. MacPherson: The factors affecting the psychometric function for speech intelligibility, University of Strathclyde & MRC Institute of Hearing Research, Glasgow, UK, 2012.

[29] J. B. Nielsen, T. Dau: The Danish hearing in noise test. Int. J. Audiol. **50** (2011) 202–208.

[30] S. D. Soli, L. L. N. Wong: Assessment of speech intelligibility in noise with the Hearing in Noise Test. Int. J. Audiol. **47** (2008) 356–361.