JASA ARTICLE

Development of the Everyday Conversational Sentences in Noise test

Kelly M. Miles,[1,a)] Gitte Keidser,[2] Katrina Freeston,[3] Timothy Beechey,[4,b)] Virginia Best,[5,c)] and Jörg M. Buchholz[3]

[1]*National Acoustic Laboratories, Macquarie University, Sydney, Australia*

[2]*Eriksolm Research Centre, Snekkersten, Denmark*

[3]*Department of Linguistics, Macquarie University, Sydney, Australia*

[4]*Department of Speech-Language-Hearing Sciences, University of Minnesota, Twin Cities, Minneapolis, Minnesota 55455, USA*

[5]*Department of Speech, Language and Hearing Sciences, Boston University, Boston, Massachusetts 02215, USA*

ABSTRACT:
To capture the demands of real-world listening, laboratory-based speech-in-noise tasks must better reflect the types of speech and environments listeners encounter in everyday life. This article reports the development of original sentence materials that were produced spontaneously with varying vocal efforts. These sentences were extracted from conversations between a talker pair (female/male) communicating in different realistic acoustic environments to elicit normal, raised and loud vocal efforts. In total, 384 sentences were extracted to provide four equivalent lists of 16 sentences at the three efforts for the two talkers. The sentences were presented to 32 young, normally hearing participants in stationary noise at five signal-to-noise ratios from -8 to 0 dB in 2 dB steps. Psychometric functions were fitted for each sentence, revealing an average 50% speech reception threshold (SRT_{50}) of -5.2 dB, and an average slope of 17.2%/dB. Sentences were then level-normalised to adjust their individual SRT_{50} to the mean (-5.2 dB). The sentences may be combined with realistic background noise to provide an assessment method that better captures the perceptual demands of everyday communication. *© 2020 Acoustical Society of America.*

https://doi.org/10.1121/10.0000780

(Received 9 September 2019; revised 3 February 2020; accepted 4 February 2020; published online 6 March 2020)

[Editor: Karen S. Helfer] Pages: 1562–1576

I. INTRODUCTION

Understanding speech in background noise is a major problem for many people with hearing impairment (HI). Clinical and laboratory-based evaluations of this difficulty are typically based on speech perception tests such as the Bamford-Kowal-Bench (BKB) (Bench *et al.*, 1979), the hearing in noise test (HINT) (Nilsson *et al.*, 1994), or Hagerman (1982) sentences presented in various background maskers. While speech-in-noise difficulties can be observed using these tests, it has been noted that such measures are likely poor predictors of an individual's speech understanding in real-world conditions (CHABA, 1988; Cord *et al.*, 2007; Kiessling *et al.*, 2003). This mismatch between predicted and actual performance poses a problem for an accurate assessment of difficulties, provision of optimal aural rehabilitation, and the development and fine-tuning of new technologies within the hearing-devices industry. Without the ability to accurately predict real-world speech understanding using clinical measures, it is very difficult to efficiently select the most appropriate intervention to improve real-world outcomes, or to determine the effectiveness of interventions. The lack of a clear target means that rehabilitation may be prematurely ceased in response to overly optimistic results from clinical assessments. For example, perception of the type of clear speech and simple sentences which characterise clinical speech tests may produce ceiling effects. Equally, the rehabilitation process may be unnecessarily prolonged due to poor results which do not reflect an individual's real-world performance. For example, a person who benefits from meaningful prosodic cues, which are absent from standard sentence materials, may receive an artificially poor score.

The apparent disparity between clinic- and laboratory-based results and real-world outcomes may be attributed, in large part, to a lack of ecological validity of existing speech-in-noise tests (Cord *et al.*, 2007; Jerger, 2009). This is likely due to a combination of factors including, but not limited to, the use of idealised speech materials and the acoustic characteristics of both speech and background noise. The advantage of using such materials is that they provide experimental control. However, when the aim is to predict performance outside the clinic or laboratory, experimental control is not the sole concern. Rather, it is crucial that measures of hearing involve tasks that are as representative of the task of hearing in everyday life as possible.

Idealised speech materials comprise short, simple sentences that are read aloud by a trained speaker in a quiet

[a)]Electronic mail: kelly.miles@mq.edu.au, ORCID: 0000-0002-4104-980X

[b)]Also at: Linköping University, Linköping, Sweden. ORCID: 0000-0001-8858-946X

[c)]ORCID: 0000-0002-5535-5736

sound-treated booth using clear, carefully articulated speech. Such material does not reflect the demands of real-world speech perception for a number of reasons. For example, read speech, which is typically scripted and rehearsed and thus fluent, differs from spontaneous speech in terms of spectral characteristics, segmental duration, and prosody (Laan, 1997). Read speech also differs from spontaneous speech in terms of phonetic reductions and deletions (Ernestus and Warner, 2011). Further, the simple sentences found in standard speech tests lack the complexity and variation that exist in natural conversational speech. For example, complex, embedded clauses and dependencies between sentential elements are typical of natural utterances (Gibson, 1998). Typical conversations in both quiet and adverse listening conditions comprise dynamic and interactive properties in which the speakers use various strategies to facilitate successful communication. For example, it is commonplace for a talker to increase their vocal effort when talking in background noise (i.e., Lombard speech; Lombard, 1911) in order to help the listener differentiate the speech signal from the noise (Cipriano *et al.*, 2017). Vocal modifications produced by talkers in adverse conditions include increases in speech amplitude, pitch, and duration (Davis *et al.*, 2006; Summers *et al.*, 1988), and changes to vowel formant frequencies (Castellanos *et al.*, 1996; Summers *et al.*, 1988). These changes make Lombard speech more intelligible (e.g., Cooke *et al.*, 2014). Even in quiet, the characteristics of speech and language in dialogues contribute to greater listener understanding in comparison to monologistic speech (Branigan *et al.*, 2011). The beneficial listener-oriented characteristics of speech which people encounter in everyday communication scenarios are generally absent from sentences used in standard speech-in-noise tests. These differences each represent ways in which standard speech tests do not reflect the perceptual and cognitive demands of everyday hearing for speech. To the extent that the demands of a standard speech intelligibility test differ from the demands of everyday listening, clinical and laboratory speech measures may fail to accurately predict performance, satisfaction, and preferences outside the clinic or laboratory.

A few attempts have been made to create more realistic speech materials for use in laboratory speech testing. For example, Best *et al.* (2016) recorded scripted conversations using voice actors who were instructed to "act out" the scripts rather than read them aloud verbatim. This resulted in ongoing speech that included natural speech traits such as pauses and disfluencies. There has also been a shift towards improving the realism of the background noise used in speech testing, so that it is more indicative of that encountered in the real world. For instance, Gifford and colleagues (2011) evaluated speech reception thresholds (SRTs) in individuals with cochlear implants using semi-diffuse restaurant noise, while others considered the effect of room reverberation (Geissler and Arweiler, 2014). In recent work, speech intelligibility was compared for BKB sentences that were presented in anechoic diffuse multi-talker babble vs a simulated reverberant cafeteria containing competing conversations (Best *et al.*, 2015). That study found that the more realistic listening background impacted the psychometric properties of the test; for example, SRTs increased more dramatically as a function of hearing loss in the more realistic background noise.

Although this implementation added more realism to a standard speech-in-noise task, it retained a degree of artificiality because the speech stimuli, in terms of speech production, were not matched to the background noise in which they were presented. That is, the difficulty of the task was varied by adjusting the relative levels of speech stimuli and noise, but the vocal effort of the speech stimuli was kept constant across signal-to-noise ratios (SNRs). Therefore, while realistic background noise was used to better align with the types of listening scenarios found in the real-world, the vocal effort of speech stimuli did not match the background noise in which it was presented.

When devising a test to assess the performance of a new directional microphone, Killion *et al.* (1998) recorded an individual producing standard sentences in real background noise scenarios (a quiet restaurant, a noisy restaurant, a street corner with music playing, a party at a museum) and recorded the speech signals at a listener's ear in the same environment. In doing so, the speaker's vocal effort was matched to the SNRs in the realistic listening environments, potentially resulting in stimuli that better reflect real-world speech production. However, since subjects adapt their vocal effort less to background noise when producing sentences in isolation than when actually communicating with a person (Lane *et al.*, 1970), here we have taken the approach of extracting sentences directly from natural conversations between a pair of talkers communicating in different simulated real-world environments. Within this context, the talkers produced natural conversational speech with appropriate levels of vocal effort that matched the background noise of the given listening environment in which they were communicating. Extracting test materials from natural conversations ensured that the speech stimuli reflected the content and style of real-world speech, including speech rate, phonetic reductions and deletions, natural word choices and intonation contours. Materials were recorded in a variety of environments that individuals are likely to encounter in the real world and therefore had a high degree of ecological validity.

In this paper we describe how the Everyday Conversational Sentences in Noise (ECO-SiN) test was developed, provide a preliminary characterisation of its psychometric properties, and discuss how these sentences can be applied to achieve a more realistic assessment of speech intelligibility that improves the ecological validity of outcomes for people with hearing impairment.

II. DEVELOPMENT OF SENTENCE MATERIAL

A. Talker selection

Two Australian-English speakers (female, 31 yr; male, 32 yr) were recruited to engage in unscripted conversations.

The difference between male and female talkers is an important feature which we aimed to capture in the development of the ECO-SiN test. While differences in intelligibility between male and female talkers are an anecdotal observation familiar to many clinical audiologists, this is a topic which has not been widely studied. Both audio and visual materials were recorded, however only the audio material was used to develop the ECO-SiN test. Professional actors were recruited through a talent agency to ensure they would be comfortable in a studio environment with cameras and audio equipment. Actors were selected as opposed to voice-over artists or those with experience in radio in order to limit speech sounding trained or unnatural. The audio and visual materials will be recombined in the future. At this stage, understanding the psychometrics of the audio-only materials is crucial before introducing the combined audio-visual materials, which would confound the performance intensity functions of the sentences.

B. Recordings

1. Setup and equipment

The talkers were seated approximately 1 m apart from each other inside an acoustically treated film studio. Microphones (Schoeps MK 41 microphone capsule with CMC 6 amplifier) and headphones (Bose SoundSport®) were positioned on each talker and secured in place. These in-ear headphones were chosen to minimise the occlusion effect as well as the acoustic attenuation of the communication partner's voice. Audio was delivered through the headphones from a laptop computer connected to an RME Fireface UC USB soundcard and running Adobe Audition software. The headphones were calibrated and equalised using a minimum-phase finite impulse response (FIR) filter based on measurements made in the ears of three subjects using calibrated custom-built low-noise probe microphones (similar to the Etymotic ER-7 probe microphone). Application of the equalisation filter resulted in a flat frequency response of the headphones when averaged across the ears of the three subjects. The speech recording microphones had super-cardioid directivity and were mounted solidly on booms pointing from above to the mouth of each talker at an approximate distance of 0.3 m. The microphones were connected to a Sound Devices 633 Recorder and recordings were sampled at a frequency of 48 kHz.

2. Recording procedures

Nineteen conversations were recorded, each with a duration of around 6 min. The first conversation was recorded in quiet so the talkers could become familiar with each other and the studio environment. Each of the subsequent conversations corresponded to one of three different background noise conditions to elicit normal, raised and loud vocal efforts. Background noise was played over headphones for the length of the conversation (described in more detail below). The background noises were pseudorandomised with the condition that the loud background noise was not presented first. Prior to each conversation, the talkers were given a topic to choose from to facilitate ease of conversation (e.g., what did you do on your last birthday? how was your commute this morning?). The talkers were then instructed to speak naturally as they typically would in a casual social setting, and not to "over act" or "play a character." This permitted natural variations in speed, duration, and conversational overlaps such as those encountered in the real-world. The talkers were given short breaks after each conversation.

3. Background noise

The background noise conditions were selected from the Ambisonics Recordings of Typical Environments (ARTE) database (Weisser *et al.*, 2019) to elicit normal, raised, and loud vocal efforts. These recordings were made using a 62-channel hard-sphere microphone array, decoded into the higher-order Ambisonics format, and transformed into binaural signals by simulating the playback via a 3 D loudspeaker array to the in-ear microphones of a Brüel & Kjær type 4128 C Head and Torso Simulator. Details of the verified methods can be found in Weisser *et al.* (2019), Oreinos and Buchholz (2015), and Oreinos and Buchholz (2016). The three environments (see Table I) were selected based on the results from Weisser and Buchholz (2019) who measured conversational speech levels and SNRs in all the 13 environments that are currently provided in the ARTE database. Applying these environments, the expected unweighted speech levels can be predicted from Weisser and Buchholz [2019, Eq. (9)] and are given in Table I separately for male and female talkers. These predicted speech levels are well in line with the levels specified by ANSI (1997) for normal [62.4 dB sound pressure level (SPL)], raised (68.3 dB SPL), and loud (74.9 dB SPL) speech.

TABLE I. Summary of realistic acoustic environments used in the conversation recordings and predicted speech levels. RT: reverberation time T30; noise levels in free-field.

		Noise level		Speech levels (SPL)		
ID	Environment	dBA	SPL	Male	Female	RT (s)
Normal	Small church, people arriving and waiting for the sermon	54.7	60.5	65.3	63.7	1.2
Raised	Indoor café at medium occupancy, dominant fridge, kitchen sounds	67.3	71.0	70.0	68.4	1.1
Loud	Very noisy food court in shopping mall, "babble speech"	76.7	79.6	73.8	72.1	1.0

C. Segmenting the sentence materials

In order for the final test to be useful for speech assessment purposes, it was decided that a balanced set of materials was required that included two different talker sexes (male and female) and three different vocal efforts (normal, raised, and loud). As such, the target was to create 384 sentences in total, organised in four equivalent lists of 16 sentences for each of the two talkers at the three vocal efforts.

The recorded conversations were imported into Praat (Boersma and Weenink, 2018) where non-overlapping speech was extracted. Sentences with between two and 14 words were manually coded, and individual sentences were extracted for further analysis. In total, 956 sentences across vocal efforts and talkers were retained.

The first and fourth authors, both native Australian-English speakers with normal hearing, independently listened to the segmented sound files over Sennheiser HD 215 headphones and classified the following parameters: presence of background noise (no/yes), extensive pauses (no/yes), talking speed (fast/slow), length (short/long), predictability (high/low), topical (no/yes), appropriateness (no/yes), filler words (no/yes), slang (no/yes), naturalness (no/yes), and grammaticality (grammatical/ungrammatical). These criteria aided in providing an overall ranking (discard/reasonable/acceptable). All sentences with an "acceptable" overall ranking were retained for the next phase. Sentences ranked "reasonable" were put forward for discussion, and depending on the reasoning, were either discarded or corrected if possible (e.g., the sentence "2002 I was in Switzerland" was ranked as reasonable by both listeners, but after discussion, this sentence was further segmented to "I was in Switzerland"). We also aimed to include sentences that may not necessarily be well formulated with high semantic content, but do reflect the types of speech signals individuals encounter in the real-world (e.g., the utterances "not so much" and "but very expensive"). In total 495 sentences were retained for the next phase.

D. Post-processing sentences

Sentences were RMS-level normalised across all vocal efforts and talkers. As level normalising may result in some sentences sounding unnatural (i.e., overly loud or soft) when compared to the other sentences of the same vocal effort and talker, the first author listened to each sound file before it was included in the next phase. Two sentences were removed as they sounded unnatural, leaving 493 sentences.

E. Sentence scoring

Sentence tests are typically scored using a morphemic, keyword or a word-by-word scoring method. As the current approach extracted sentences from two people having unscripted conversations, unique challenges were present when determining the appropriate scoring method. Morphemic scoring aims to maximise the number of scoring units in a given sentence to improve reliability over a short period of time. While this method has known drawbacks relating to statistical violations (Dawson *et al.*, 2013) it becomes particularly laborious to score unscripted speech which often contains morphologically complex words (e.g., "unaccommodating" and "incomprehensible"). Keyword scoring, on the other hand, uses preselected words that typically give the gist of a sentence, including nouns and verbs (e.g., "A dog chased the ball"). The selection of keywords is particularly challenging for sentences extracted from natural conversations. For example, in the sentence "That's interesting isn't it," the selection of keywords is not straightforward. For these reasons, word-by-word scoring was deemed the most appropriate scoring method. Of note, if a participant correctly repeated a root word, but not the suffix (e.g., said "sail" instead of "sailed") or incorrectly repeated a root and suffix (e.g., "sailed" instead of "sail"), the word was scored as incorrect. In addition to this, the word "like" was removed from scoring ($n = 12$) along with repetitions ($n = 1$; e.g., the second "with a" in "It must be hard to travel with a, with a, baby though"), and ums/ahhs ($n = 1$). The full list of materials are available in the Appendix.

F. Screening for intelligibility in quiet

A screening task was conducted to ensure that all of the sentences were highly intelligible when presented in quiet. As sentences were extracted from natural conversations, the speed of some utterances could potentially render them incomprehensible, or the length of a sentence could be problematic for verbatim recall. Three naive Australian-English participants performed the sentence screening task (mean age = 31.65 yr, SD = 8.08 yr). All had hearing thresholds better than 20 dB hearing level (HL) from 250 Hz to 4 kHz in both ears.

A graphical user interface was created in MATLAB to administer the screening task. The participants were seated in a sound attenuated test booth and listened to the sentence materials via Beyerdynamic DT990 Pro open headphones connected to a desktop computer via an RME Fireface UC USB sound card. The sentences were presented in quiet at 65 dB SPL. The sentence materials were blocked per vocal effort and talker, with the order of sentences randomised within a block. Each sentence was presented once to each participant. After a sentence was presented, participants were required to type the sentence, or any of the words they heard, in a dialogue box. In the event there was an issue with the sound file, the participants could click a radio button which would activate a dialogue box where they could document the issue (e.g., sounded unnatural, too quick, slurred, first/last word sounded cut-off). Participants were then required to type a comment before moving on to the next sentence, which occurred 23 times.

Participant responses were compared to the sentence transcripts. A sentence was included in the next phase if 2/3 participants correctly identified all of the words in the sentence. Thirty-eight sentences were removed that did not meet this criterion. One sentence was removed as 2/3 participants noted the final word sounded cut-off. There were 32 sentences in disagreement, however most of the errors included word insertions (n = 18; e.g., "but he decided to run all the way around" had the response "but he just decided to run all the way around") word reversals (n = 2; "kind of falling over all the time" had the response "falling over kind of all the time") and missed sentences due to pressing the next button too quickly (n = 3). Of the remaining 9 disagreements, these included tense switching (e.g., "it is nice when they all have a bit of flair" had the response "it was nice when they all had a bit of flair") and morpheme exclusions which may also be spelling errors (e.g., "they're comfortable there" had the response "they comfortable there").

After sentence exclusions, there was a surplus of sentences over the original goal to create a balanced set of 64 sentences per talker and vocal effort (384 sentences in total), and so the first and fourth authors made several more exclusions. Reasons for exclusion included word repetition (e.g., there were multiple sentences with the word "Santa") and sentences containing words subjectively judged to be uncommon (e.g., "pilates"). Each sentence in the final set had a minimum length of three words, and a maximum of 13 words. The distribution of the number of words (or scoring units) for all 384 sentences is shown in Fig. 1. The sentences ranged in duration from 0.7 to 4.1 s (mean = 1.7 s; SD = 0.6 s).

III. INTELLIGIBILITY NORMALISATION AND LIST EQUIVALENCE

A. Method

1. Participants

In total, 32 young normal-hearing adults with Australian-English as their native language participated in the intelligibility normalisation and list equivalence phase (mean age = 27.5, SD = 5.0 yr). All participants had hearing thresholds better than 20 dB HL from 250 Hz to 4 kHz in both ears.

2. Test environment and setup

Testing took place with the same equipment and in the same room as the screening task. Sentences were presented in stationary background noise, which was created from white noise that was filtered to match the average speech spectrum of each talker at each effort level (i.e., there were six different background noises). As a result, the average SNR was constant across frequency, removing at least some of the likely effects of talker and vocal efforts on speech intelligibility. Realistic background noise was not used at this point as this would have created additional variance in sentence intelligibility.

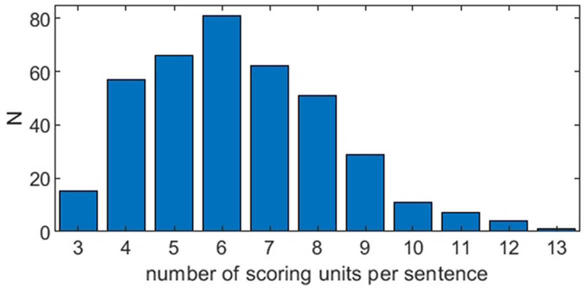

FIG. 1. (Color online) Distribution of the number of scoring units (words) per sentence across all talkers and effort levels.

3. Choice of test SNRs

Prior to intelligibility normalisation, it was necessary to first estimate the SNRs that covered the psychometric function. Five normal-hearing participants (mean age = 29.4, SD = 6.9 yr) were recruited for this task. Participants were asked to rate their estimated intelligibility (in percentage of words understood) of five sentences presented at each of 11 SNRs (−10 to 10 dB in 2 dB steps). This was done separately for the two talkers at the three efforts, using the spectrally matched noise described above. Psychometric functions were fitted separately for each participant, vocal effort, and talker using the logistic function given by

$$\Psi(SNR) = \frac{100}{1 + \exp\left(-4\frac{k}{100}(SNR - SRT_{50})\right)}, \qquad (1)$$

with k the estimated slope of the psychometric function at the inflection, and SRT_{50} the corresponding SNR-shift in dB. Final psychometric functions were then estimated for each vocal effort and talker by averaging the estimated individual slopes and shifts across participants. From this, it was determined to use five SNRs in 2 dB steps between −8 and 0 dB to capture the psychometric functions during the intelligibility normalisation phase.

4. Procedure

Background noise was the same stationary noise used to select the test SNRs. To ensure the overall playback level was not too loud, the noise was fixed to 65 dB SPL, and the speech presentation level was adjusted. Minimum-phase equalization filters were applied to all stimuli to ensure a flat frequency response of the Beyerdynamic DT990 Pro headphones when measured with an artificial ear provided by a Brüel&Kjær type 4128-C Head and Torso Simulator. A specialised graphical user interface was created in MATLAB to administer and score the speech-in-noise test at fixed SNRs.

Before the experiment commenced, the 32 participants were presented 16 practice sentences in 0 dB SNR background noise to familiarise them with the experiment format. Participants were informed they would hear sentences in noise and were instructed to repeat back the sentence, or any words they heard. The experimenter recorded the responses using the user interface provided. After the practice block, 16 sentences were presented for each of the three vocal efforts, two talkers and five SNRs. The combination of sentences and the four lowest SNRs was counterbalanced across participants. Sentences presented at −8 dB SNR (the lowest SNR) were repeated at 0 dB SNR (the highest SNR) for each participant, and pseudorandomised so that the lowest SNR was always presented first, and the sentences at the highest SNR were always presented at least five blocks (80 sentences) after the lowest SNR to avoid any learning effects. This was done to increase the number of test SNRs, and thus to increase the robustness of the fitting of psychometric functions, without further increasing the number of test participants. Testing was conducted in one session of approximately 2 h. In total, across subjects, each sentence was presented 8 times at each SNR.

B. Data manipulation

For each sentence, the word-correct scores were averaged across the eight participants separately for each of the five SNRs. The resulting data were then fitted with a logistic function given by Eq. (1) and the reliability of the fit (and the data) were evaluated. The psychometric function for a given sentence was considered unreliable if (i) the range of (average) scores measured across all five SNRs was less than 45%, (ii) the RMS error between the fitted psychometric function and the measured scores was larger than 11%, or (iii) the predicted SNR-shift was outside the range from $-10\,\mathrm{dB} \leq \mathrm{SRT}_{50} \leq 2\,\mathrm{dB}$. These rather arbitrary exclusion criteria were developed iteratively by visually evaluating the measured data for each sentence and the corresponding fits.

Applying these exclusion criteria, fits for 20 out of the 384 sentences were identified as unreliable and further analysed. For 16 out of the 20 sentences, the problem appeared to be related to particularly large variations in the vocal level across the sentence. This resulted in large variations in the local SNR across words and a psychometric function with a very shallow slope. This issue was resolved by readjusting the level of individual words or by removing overall trends in the envelope, such as steady level increases or decreases over the duration of the sentence. Either way, the goal was to flatten the overall sentence envelope while maintaining the naturalness of the spoken sentence. In the case of individual words, gain transitions were realized by half-sided Hanning windows with a duration of at least 100 ms, and no gain adjustment exceeded 8 dB. Following this manipulation, the naturalness of each sentence was verified by two independent listeners with significant experience in evaluating the sound quality of speech recordings, and who were instructed to attend specifically to the naturalness of the spoken sentences. Four other unreliable sentences were replaced by one of the surplus sentences that were originally excluded in the screening phase.

To provide an estimate of the psychometric function (i.e., slope and SNR-shift) for the 20 revised sentences, each sentence was mixed with the stationary noise (see above) at different SNRs. Percent word correct scores were then rated by ten new normal-hearing Australian-English speaking participants who listened to the noisy sentences in randomised order. For each sentence and participant, the resulting scores were fitted with the function given in Eq. (1). The fitted slopes and SNR-shifts were then averaged across subjects.

C. Results

1. Intelligibility normalisation

The average RMS error between the measured scores and the fitted psychometric function for the 364 reliable sentences was 4.7% (SD of 2.7%). The estimated SRT_{50} average across all 384 sentences was -5.2 dB (SD of 1.9 dB) and varied from -5.7 to -4.3 dB when averaged individually for the different vocal efforts and talkers. This rather small variation between vocal efforts and talkers was due to the RMS-level normalisation that was applied to all sentences as well as to the use of spectrally matched noises. The distribution of the estimated slopes was significantly skewed towards lower values, with a median value of 17.4%/dB and 0.25 and 0.75 quantiles of 12.8% and 23.5%/dB, respectively. The histograms for the estimated slopes and SNR-shifts of the psychometric functions for all 384 sentences are shown in Fig. 2. Note that the slope estimate was limited to 50%/dB, which explains the small increase in frequency at this maximum slope value.

Based on the psychometric functions derived in Sec. III C, the intelligibility of each sentence was normalised by applying a linear gain that compensated the deviation from the mean SRT_{50} of -5.2 dB. In this way, the resulting psychometric functions for the 384 sentences were all shifted such that their inflection points all coincided at an SNR of -5.2 dB. This step is illustrated in Fig. 3. In the left panel, the average scores measured for three example sentences at five different SNRs are shown (points) together with their fitted psychometric functions (curves). After intelligibility normalisation (right panel), the three psychometric functions and their corresponding scores (data points) are shifted along the SNR axis such that they all have their inflection point at -5.2 dB as indicated by the dashed lines.

IV. LIST EQUIVALENCE

The intelligibility normalised sentences were organised into equivalent lists of 16 sentences each. To achieve four equivalent lists for each of the vocal efforts and talkers, the following optimisation (i.e., error) criteria were applied:

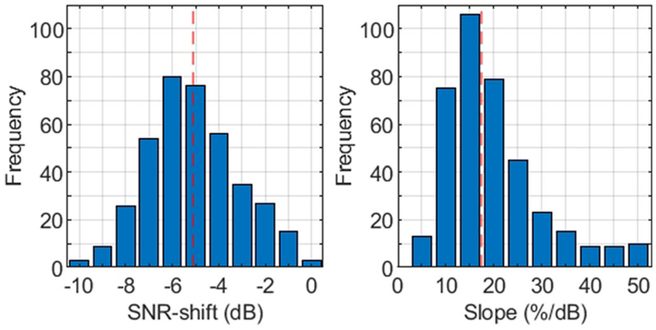

FIG. 2. (Color online) Histogram of the SNR-shift (SRT_{50}) and slope at the inflection point of the psychometric functions fitted to each of the 384 sentences. Note that the slope estimate was limited to 50%/dB.

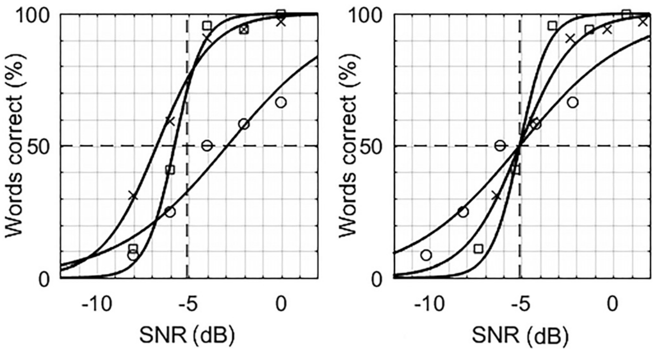

FIG. 3. Psychometric functions using three sentences as an example. In the left panel, the measured data and the corresponding fitted psychometric functions are shown. The right panel demonstrates the applied intelligibility normalisation resulting in a shift in SNR to pin the 50% inflection point (SRT_{50}) at −5.2 dB SNR.

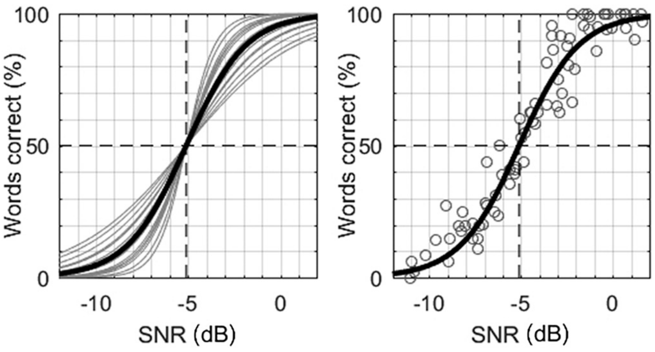

FIG. 4. The left panel illustrates the individual psychometric functions (grey lines) for one list (16 sentences) of intelligibility-normalised sentences together with the average psychometric function of the list (solid black line). The right panel shows the average psychometric function (solid black line) plotted with the average word-correct scores (grey circles) after normalisation for the same sentence list.

- Each list should produce a similar average psychometric function. Since all sentences were already intelligibility normalised with respect to their SRT_{50}, this was derived by calculating the average slope across all 16 sentences of a list. The error was then defined by the standard deviation of the average slopes across lists.
- Each list should have a similar number of total scoring units (i.e., words). This was derived by simply counting all scoring units across the 16 sentences of a list. The error was then defined by the variance of the total number of scoring units across lists.
- Each list should have a similar distribution of short and long sentences. This was derived from histograms showing the frequency of the number of words per sentence calculated across all 16 sentences of a list. The error was then defined by the variance of the histograms across lists derived separately for each number of words per sentence averaged across the number of words per sentence.

The three errors were then weighted and summed into a single error. The weights were derived iteratively by manually inspecting the similarity between the resulting psychometric functions, total scoring units, and histograms. To find equivalent lists that best fit the above error criteria, 10^7 sentence combinations were generated randomly for each of the vocal efforts and talkers, and the version with the smallest overall error was selected.

The left panel of Fig. 4 shows individual psychometric functions (grey curves) for 16 intelligibility-normalised sentences that were grouped into an example list (female talker with raised effort, list 4) together with the average psychometric function for that list (black curve). A significant spread in the slopes across psychometric functions can be seen. In the right panel, the average psychometric function is replotted together with the average word-correct scores (grey points) for the 16 sentences of the example list measured at five different SNRs after intelligibility normalisation. For this sentence list the RMS error between the average scores (in total $16 \times 5 = 80$ data points) and the average psychometric function was 8%. The RMS error averaged across all lists was 8.8% with a standard deviation of 0.8%.

Table II summarises the mean value and standard deviation across equivalent lists for all parameters of the average psychometric functions (i.e., slopes and SNR-shifts), the total number of scoring units, and the RMS error between the raw data and the average fit for each of the vocal efforts and talkers.

Unlike some speech materials (e.g., BKB and HINT sentences), phonemic balance within sentence lists was not attempted. First, it is uncommon for conversations in the real-world to be phonemically balanced and therefore it was not a dimension we wished to control. Second, recall that the ECO-SiN materials were extracted from natural conversations between a pair of talkers. Due to the nature of how the sentence materials were therefore acquired, phonemic balance was not possible to achieve.

V. DISCUSSION

In this paper we presented the development of the ECO-SiN test, a new sentence test that uses naturally spoken sentences extracted from real conversations. We have provided a description of its psychometric properties from an adult population with normal hearing. We aimed to create sentence materials that captured potentially important characteristics of real-life speech signals, such as variations in vocal effort and speed, and a spontaneous conversational vocabulary. These factors are absent from standard speech tests that use well formed, well-articulated sentences read in quiet from a script. We argue that these new materials represent a critical step towards the broader goal of adding realism to laboratory and clinical speech-in-noise tests, while maintaining desirable elements of control.

A. Comparison with other tests

The sensitivity of a speech-in-noise test is captured in the slope of its psychometric function; a steeper slope indicates that a small change in SNR in either direction results in large performance differences. In the newly developed test reported here, the average slope of the psychometric function for the female talker was 16.3%/dB, and for the male talker was 18.0%/dB. The slopes are comparable to those reported for other tests such as the Leuven Intelligibility Sentence Test (17.5%/dB for the female talker, van Wieringen and Wouters, 2008; 18.7%/dB for the male talker, Jansen *et al.*, 2014) and slightly shallower than the Australian-English matrix sentence test that uses a closed set of words spoken by five female and five male talkers (19.3%/dB, Kelly *et al.*, 2017), and the QuickSIN with a reported 15%/dB slope in speech-shaped noise and 14.3%/dB in four-talker babble noise (Brungart *et al.*, 2014).

In contrast, the slope of the psychometric functions are considerably steeper than those reported for the commonly used HINT and BKB tests. The average psychometric slope of the American English HINT is 10.6%/dB, and across languages is 10.3%/dB (Soli *et al.*, 2009). Similarly, Keidser *et al.* (2002) reported a slope of 10%/dB for BKB sentences presented in speech babble.

Differences in slope may stem from a range of factors including the variety of background maskers employed across the different sentence tests (MacPherson and Akeroyd, 2014; Hochmuth *et al.*, 2015) and the linguistic characteristics of the sentences including syntactic complexity, the range of sentence lengths, and the amount of contextual and semantic information (Bronkhorst *et al.*, 2002; van Rooij and Plomp, 2005; and Holmes *et al.*, 2018). While steep psychometric functions are desirable for maximising test sensitivity, the most important characteristic of the psychometric function for a "realistic" speech test is that it reflects how real-world performance would change given a similar change in conditions. That is, test material that returns a much larger performance change than would be observed in everyday communication given the same change in background noise conditions would sacrifice predictive power for sensitivity. The relationship between the psychometric functions of any particular speech tests and real-world communication performance is not well understood. However, the increased realism of the speech materials we have employed provide some basis for believing that these functions may reflect real-world performance more closely than many of the more traditional tests.

B. Differences between vocal efforts

Conversing in background noise requires interlocutors to modify their vocal levels if successful communication is to take place. In the current test, the recordings were extracted from a pair of talkers conversing whilst listening to background noise played over headphones. This resulted in natural variations of vocal effort given the playback of the acoustic environment. This is illustrated in Fig. 5, which shows third-octave levels averaged across all sentences for the female (left panel) and male (right panel) talker separately for the three vocal efforts. Since there was no opportunity to calibrate the recording equipment provided by the hired film studio, the speech levels were normalised such that the RMS level of the speech with loud vocal effort was equal to 0 dB for each talker. In this way the differences across vocal efforts were preserved for each talker.

There is a clear demarcation between the vocal efforts (normal, raised, loud) corresponding to the acoustic environments, which is pronounced in the mid-to-high frequencies. Relative to the normal effort levels, the overall RMS levels for the female and male talker increase by 4.4 and 3.8 dB for the raised effort and by 12.4 and 9.9 dB for the loud effort. These changes in overall RMS level are within range of the SII standard (ANSI, 1997), the continuous speech levels reported by the International Collegium of Rehabilitative Audiology (ICRA) (Dreschler *et al.*, 2001), and the levels reported in Weisser and Buchholz (2019) and (Beechey *et al.*, 2018) who both used the same background noise environments to elicit different levels of vocal effort between pairs of talkers. The increase in mid-to-high frequency emphasis with increasing vocal effort is also in agreement with these publications. Additionally, there appears to be more energy in the F0 range for the loud vocal effort, with "F0 raising" a known phenomenon when increasing vocal intensity (Alku *et al.*, 2001).

C. Spectral differences between female and male speech

As described above, the female talker adjusted her vocal effort slightly more than the male talker in response to the increasing background noise level, in particular in the loudest environment (i.e., with an increase in RMS level of 12.4 dB versus 9.9 dB relative to their normal effort level). Given that individual talkers adjust their vocal effort very differently to increasing background level (e.g., Weisser and Buchholz, 2019), this difference is expected and in agreement with the relevant literature. Besides this difference in overall level, the long-term spectra of the female and male speech shown in Fig. 5 differ mainly at frequencies below

TABLE II. Average psychometric functions, total number of scoring units, and RMS error for the final (equivalent) lists. Shown are mean values and ±1 standard deviation across the four equivalent lists per vocal effort and talker.

	Male talker			Female talker		
Analysis	Normal	Raised	Loud	Normal	Raised	Loud
Slope (%/dB)	17.2 ± 0.2	17.5 ± 0.2	19.3 ± 0.1	16.1 ± 0.2	15.6 ± 0.1	17.1 ± 0.3
SNR-shift (dB)	−5.2	−5.2	−5.2	−5.2	−5.2	−5.2
Total scoring units	103.3 ± 2.1	97 ± 1.4	95 ± 1.4	104.7 ± 1.2	102 ± 1.4	107.8 ± 0.9
RMS error (%)	9.8 ± 0.2	8.4 ± 1.0	8.7 ± 0.3	9.2 ± 0.4	8.0 ± 0.4	8.3 ± 0.3

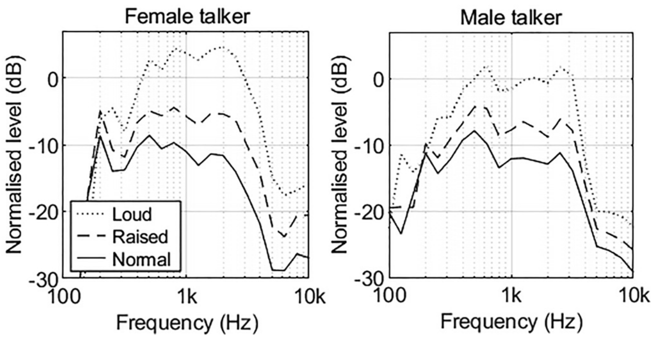

FIG. 5. Average speech spectra in 3rd-octave bands for the female (left panel) and male (right panel) talker at the three different vocal efforts. The spectra were normalised so that the RMS level for the loud vocal effort was equal to 0 dB for either talker.

200 Hz, which is due to the differences in fundamental frequency. Otherwise, the spectra at each vocal effort do not show any other systematic differences between talkers, with absolute differences at any third-octave frequency below 4 dB. This is in agreement with Byrne *et al.* (1994), who measured long-term average spectra for read speech across a large number of talkers and languages, and reported no considerable effect of gender for frequencies above 200 Hz. This observation is also in line with Weisser and Buchholz (2019), who measured long-term speech spectra for conversational speech that was produced under similar simulated noisy conditions as used in this study.

D. Implications when using the materials

In particular, when combined with realistic background noise, the realistic speech materials in the ECO-SiN test may provide a more valid measure of an individual's listening ability in the context of real-world conversations. As such, outcomes of the ECO-SiN test may allow the individual, researcher, and/or clinician to make better informed decisions about device settings, device selections, and intervention/rehabilitation strategies that may ease listening for the individual in everyday life.

The availability of equivalent male and female speech materials may provide the basis for investigating whether female speech is, as is anecdotally reported in clinical settings, more difficult for hearing impaired listeners to accurately perceive. Should such a finding be confirmed, this would have important implications for our understanding of how pure-tone audiometry results relate to speech perception. Since the difference between male and female speech, in terms of long-term average spectra, is limited to the low frequencies, poorer perception of female speech is unlikely to be primarily attributed to elevated high-frequency thresholds relative to high-frequency consonant spectra (see Diehl *et al.*, 1996).

When applied in a speech-in-noise test, it is important that the speech materials are combined with appropriate background noise levels. Recall that the speech materials were elicited when the interlocutors were conversing in different levels of background noise which resulted in normal, raised, and loud vocal efforts. The speech materials therefore have inherent properties that require playback at specific SNRs in order to sound "natural." Simply adjusting the speech signal (for example, attenuating the speech), may result in unusual sounding playback. Care must therefore be taken if the speech materials are to be used for an adaptive track due to the potential mismatch of vocal efforts and SNRs (Weisser and Buchholz, 2019).

E. Limitations and outlook

In developing a speech perception test that better captures the real-world challenges of listening in noise, there are multiple variables to consider. Here, we attempted to greatly increase the realism of the speech materials—eliciting speech that captured naturalistic variation in speech rate, phonetic reductions and deletions, natural word choices, intonation contours, and vocal effort. In further developing the ECO-SiN test, the materials will be combined with the visual recordings and realistic acoustic environments from the ARTE database (Weisser *et al.*, 2019) which will result in a task that combines both realistic audio-visual speech and realistic background noise. One limitation in the development of this task, however, is that the sentence intelligibility task itself does not target higher-level speech processing, such as comprehension and response formulation, and fails to capture the dynamic interplay of listening and conversing with a conversational partner in the real-world (Pichora-Fuller, 2007; Schneider *et al.*, 2010). In developing more realistic assessments of listening ability, future studies may wish to consider employing a more naturalistic task combined with more realistic speech and background noise (e.g., see Beechey *et al.*, 2018). Finally, future studies will need to further validate the speech material with normal-hearing and hearing-impaired listeners and provide a systematic assessment of its test-retest reliability when combined with different background noises.

ACKNOWLEDGMENTS

Thank you to Elizabeth Beach for valuable feedback on the manuscript. Funding from: Oticon Foundation, the Commonwealth Department of Health and Ageing, and the HEARing CRC, established under the Australian Government's Cooperative Research Centres (CRC) Program. The CRC Program supports industry-led collaborations between industry, researchers, and the community. V.B. was supported by a grant from NIH/NIDCD (R01DC015760).

APPENDIX

Full list of ECO-SiN sentence 1 materials. Only capitalised words contribute to scoring units. Lower case words excluded from scoring include "like" (n = 12), the repetition "with a" (n = 1), and "uhh" (n = 1).

Normal vocal effort			
Female		Male	
Sentence	Scoring units	Sentence	Scoring units
EVEN THOUGH THEY WERE DOING THAT	6	IT'S NOT AS GOOD AS IT WAS WHEN I WAS A CHILD	13
COMPARED TO FOUR WHEEL DRIVES	5	IF WE'RE GOING ON A ROAD TRIP	8
THE LIST OF MOVIES	4	WHAT IS NORMAL	3
ONE OF MY AUNTIES	4	THERE'S A LOT OF MYSTERY AROUND WHY HE CAME TO AUSTRALIA	12
NOT SO MUCH	3	I WASN'T IN IT VERY MUCH	7
GOING ON THIS BIG JOURNEY	5	GET THIS NEIGHBOUR SOME NEW SANDALS	6
BUT THEY'RE HUGE TRUCKS DRIVING REALLY FAST	8	MY MOTHERS BROTHER	3
BUT I DIDN'T REALLY KNOW THE FIRST ONE	9	WHEN YOU REWATCH IT AS AN ADULT	7
YOU HAVE TO GO IN THE LEFT LANE TO TURN RIGHT	11	THERE'S NO SPEED LIMIT	5
WE RECENTLY DID SOME DRIVING IN ITALY	7	HE WOULD GO AROUND AND VISIT FARMS	7
I THINK PEOPLE SHOULD GET like A TRUCK LICENSE	8	IT'S EARLY IN THE FILM	6
I DON'T REALLY EVEN SEE THAT KINDA POSTER IMAGE	10	THEY NEVER LOSE THAT CHILDHOOD	5
THEY JUST KEEP RAISING THE TOLL	6	NOT LIKE SOME FAMILIES	4
I USED TO LOVE SOFT SERVE	6	WERE THE FAMILIES QUITE SEPARATE	5
THEY DON'T INDICATE	4	ADHERING TO REGULATIONS	3
TO MAKE SURE THAT THEY KNOW WHAT THEY'RE DOING	10	I'VE NEVER SEEN ANYONE GET FINED FOR IT	9
THROUGH THE CROSS CITY TUNNEL	5	BUT I DON'T REALLY REMEMBER THAT STUFF	8
WHEN YOU'RE ON A ROADTRIP	6	THE SHOW MUST GO ON	5
IN QUITE A DRAMATIC SCENE	5	I KIND OF RESENT TOLL ROADS	6
THERE'S A LOT OF OPPORTUNITIES	6	THERE WAS A CINEMA NEAR WHERE I GREW UP	9
AND YOU KIND OF HAVE TO GO THROUGH THEM	9	ONE OF THE WORST FILMS EVER MADE	7
YOU KNOW WHAT I MISS FROM WHEN I WAS YOUNG	10	I DON'T KNOW IF I HAVE ANY OF THOSE STEROTYPICAL MEMORIES	12
PARTICULAR TO THE HISTORY	4	BUT YOU CERTAINLY DON'T SEE THEM IN THE STREET	10
WHAT DO YOU DO WHEN YOU DRIVE ON THE ROAD?	10	I THINK I DID	4
I HAD TO GIVE HER A SCROLL	7	BUT THAT WOULDN'T WORK HERE	6
IF YOU'RE AT THE AIRPORT	6	I REMEMBER BEING VERY UPSET	5
I GUESS CLOSER TO AUSTRALIA	5	BEING QUITE TRADITIONAL	3
PEOPLE JUST LIKE DRIVING IN THE MIDDLE LANE	8	BUT I LOVE STORIES LIKE THAT	6
THEY WEREN'T VERY GOOD LIARS	6	I DO REMEMBER LEAVING like MILK AND COOKIES	7
SO WE HAD A TURTLE	5	THESE RANDOM SHOTS OF THIS SPOON	6
THREE FAMILIES OR SOMETHING	4	THE CITY GENERALLY IS A BIT AGGRESSIVE	7
I THINK THEY'RE ACTUALLY CLOSER THAN I THOUGHT	9	I THINK IN THE CINEMA	5
YOU KNOW THEY TELL YOU WHERE THEY GO	8	THAT'S INTERESTING ISN'T IT	6
TEXTING AS HE WAS DRIVING	5	SANTA BRINGS ME THIS THING IF I'M GOOD	9
I DON'T REMEMBER SEEING THE FILM SO MUCH	9	THE WORST MOVIE EVER MADE	5
WHEN YOU'RE NOT OVERTAKING	5	I WOULDN'T EAT IT AS A MEAL	8
WE LIVED ON THE FOURTH FLOOR	6	SEND HIM ON THAT PATH	5
I SAW IT AT THE THEATRE ROYAL I THINK	9	I TRY NOT TO DRIVE IN SYDNEY	7
THEY KIND OF JUST SWING OUT WHEN THEY WANT TO	10	BRANCHES OF THE TREE	4
WHEN YOU WERE YOUNG	4	THE SONG IS ABOUT HIM BEING POOR	7

Continued

Normal vocal effort			
Female		Male	
Sentence	Scoring units	Sentence	Scoring units
DID YOU BELIEVE IN SANTA CLAUS	6	I DON'T LIKE THEM	5
AND THEN THERE'S OUR GRANDMA	6	A WAY TO DEAL WITH SANTA	6
SOMEHOW THAT SCROLL GOT LOST IN THE COSTUME	8	SO MY MOTHERS MOTHERS FATHER	5
BUT IT WAS ACTUALLY MORE INTEGRATED	6	NO ONE INDICATES IN SYDNEY	5
IT WAS JUST KIND OF A NICKNAME	7	SHE HAS A MUCH BETTER MEMORY OF HER CHILDHOOD	9
THEY'RE COMFORTABLE THERE	4	I MEAN THEY STILL HAVE THAT	6
I DON'T THINK THEY MAKE MOVIE POSTERS	8	THIS WHOLE CULT FOLLOWING SPRUNG UP AROUND IT	8
THAT'S ALWAYS FUN	4	WHICH I JUST FIND ABSOLUTELY CRAZY	6
MY PARENTS DIDN'T REALLY TALK ABOUT IT	8	SANTA BROUGHT YOU THIS	4
THEY KNEW WHAT THEY WERE DOING	6	WHEN I WAS A KID	5
THE OVERTAKING RULES KIND OF DON'T WORK	8	I WAS CHECKING MY PHONE AS I WAS GETTING READY	10
HOW IT'S ADVERTISED	4	THE GIRL THAT HE'S KIND OF IN LOVE WITH	10
THIS IS A BUSY AREA NOW	6	THEY STILL SHOW MOVIE POSTERS	5
THIS AREA'S GROWING	4	HE WAS THE ONE DELIVERING THE PRESENTS	7
TAKE LOTS OF PHOTOS	4	AS I WAS WALKING ON STAGE	6
KIDS THAT I HARDLY KNOW	5	THERE WERE A FEW THINGS PARTICULARLY FOOD THINGS	8
WALKING INTO THE CINEMAS AND SEEING MOVIE POSTERS	8	I WAS TALKING ABOUT THIS THE OTHER DAY	8
I LOVE IT WHEN PEOPLE SWING OUT TO TURN	9	THE PEOPLE WALKING DOWN THE AISLES	6
I MEAN I WOULDN'T AGREE WITH IT NOW	9	I HEAR THE MUSIC ON STAGE	6
BUT IT WAS like SO NORMAL	5	IT WOULD'VE BEEN THE CAPITOL THEATRE	7
I DO LOVE WHEN THINGS GO WRONG	7	IN THE APARTMENT	3
HOW WAS YOUR DRIVE THIS MORNING	6	DO YOU HAVE ANY FAVOURITE DISNEY FILMS	7
OH IT'S TEN DOLLARS NOW	6	SO THEY'RE VERY NOSTALGIC	5
I GUESS IT'S LIKE WATCHING A CARTOON	8	IT'S CALLED THE SAME THING	6

Raised vocal effort			
Female		Male	
Sentence	Scoring units	Sentence	Scoring units
I LOVE DOING THOSE THINGS	5	THEY PUT OUT A FEW ALBUMS AND THINGS	8
MISSING MY BABY WHEN I'M DOING SOMETHING ELSE	9	DO YOU PLAY ANY SPORTS	5
KIDS PLAYING BASEBALL IN AUSTRALIA	5	IF IT'S ON I'LL WATCH IT	8
THINGS THAT MAKE ME HAPPY NOW	6	I DON'T MIND SNAKES	5
I THINK IT'S ONE SPORT THAT I COULD WATCH	10	JAPAN'S REALLY BIG	4
THERE'S like TWO EXTREMES AFTER HAVING A CHILD	8	PART OF WHERE IT COMES FROM	6
I USED TO PLAY ALL THE LOCAL COMPETITIONS	8	DON'T LOOK IN THE NEXT TANK	7
THINKING ABOUT THE FEEDBACK AND ATTENTION	6	I'VE ALWAYS PLAYED SPORT	5
GET A BIT WOBBLY AT THE KNEES	7	I THINK A LOT OF THEM ARE QUITE VENOMOUS	9
MOMENTS IN ALL SEASONS THAT I LOVE	7	IF IT'S THOSE TWO PLAYING	6
HOW I LISTEN TO MUSIC	5	I STILL HAVE IT	4
PART OF THE COUNTRYSIDE EXPERIENCE	5	THERE'S A LOT TO CHOOSE FROM	7
JUMPING IN PUDDLES	3	THE ONLY REASON THAT YOU DO IT	7
MY MUM HAS SOME FINE CHINA	6	IT'S MORE IN MY HEAD	6
AND JUST SEEING THE JOY IN HIM	7	A LOT OF PHOTOS	4
I HAVE like VHS VIDEOS	4	WHEN SHE SITS AT THE COMPUTER	6
WHAT IT MEANS TO BE AN ACTOR	7	I WENT TO THE AQUARIUM	5
AN INJURY AROUND MY WRIST	5	JUST A BUNCH OF MATES REALLY	6
WHEN WE WERE YOUNG	4	I WOULD LOVE TO GO TO A GAME	8

Continued

Raised vocal effort			
Female		Male	
Sentence	Scoring units	Sentence	Scoring units
SO WHAT OTHER ARTISTS DO YOU LISTEN TO	8	I HAVE A FEW ITEMS	5
GIVING IT THEIR ALL AND PLAYING SO WELL	8	I PLAYED TENNIS WHEN I WAS YOUNGER	7
THE MOMENT HE WAKES UP	5	WHAT ARE SOME THINGS THAT BRING YOU JOY	8
IT’S AN ACTUAL REAL THING	6	IF IT’S A REAL THING IT’S EASIER TO AVOID	11
FUN THINGS I USED TO DO	6	ALL THE LEAVES CHANGING	4
THE FREEDOM TO DO STUFF	5	IT’S like A DANISH WORD	5
SEEING HIM DISCOVER THINGS	4	THINGS FROM MY CHILDHOOD	4
THE END OF A FUN SUMMERY SEASON	7	THEN YOU GO SNORKELING	4
WINTER IN MELBOURNE’S A BIT CHALLENGING	7	MOST OF THE SAME GROUP OF PEOPLE	7
WAKING UP AND SEEING HIM EVERYDAY	6	HAVE YOU KEPT THE TAPES	5
IT IS NICE WHEN THEY ALL HAVE A BIT OF FLAIR	11	EVERYTHING’S JUST EASY AND NICE	6
THEY’RE NOT ANTIQUE OR ANYTHING	6	THE GOOD ONES CREATE THIS FEELING	6
BUT WE USED IT WHEN WE WERE OVER THERE	9	THAT FEELING OF IT UNDER YOUR FOOT	7
WHEN WE WENT AND VISITED SWITZERLAND	6	WHEN I WAS YOUNGER	4
IF YOURS IS ACTUALLY A REAL THING	7	AND I DID SEE ONE	5
I BOUGHT IT SECOND HAND FROM A VIDEO STORE	9	I LIKE THAT TOO	4
GOING TO THE SNOWY MOUNTAINS	5	SHE LIVED IN MELBOURNE A LOT	6
THAT KIND OF MADE IT A BIT HARD	8	I DON’T KNOW IT’S A WEIRD THING	9
THE FIRST MOMENT HE STARTED PULLING UP	7	WHEN YOU USED TO PRINT OUT PHOTOS	7
IT’S KIND OF MY IMAGINATION	6	I LOOK FORWARD TO HAVING KIDS	6
DO YOU DO ANY SPORTS	5	SO I HAVE A FEW OF HER PAINTINGS	8
I USED TO PLAY A LOT OF TENNIS	8	VERY MIDDLE CLASS	3
I DO LIKE THE SNOW	5	IT’S MORE THE FEELING	5
TALENT SCOUTS IN AUSTRALIA	4	I LOVE WATCHING THEM	4
THAT KINDA SUPERFICIAL IMAGE	4	THE IDEA OF JUMPING IN PUDDLES	6
THAT DISCOVERY WAS like REALLY INTERESTING FOR ME	7	WHICH I STILL PLAY NOW	5
IF I COULD EVER USE THEM AGAIN	7	HE RENOVATED A HOUSE WE LIVED IN	7
JUST WATCHING HIM CRAWL NOW IS PRETTY FUN	8	I DON’T THINK YOU’D GUESS IT IF I ASKED YOU	12
HE WENT INTO THE SANDPIT RECENTLY	6	A SHELF OF THEM	4
THE OBSESSION THEY HAVE UNTIL THEY MASTER IT	8	I PLAYED IT SINCE I WAS TEN	7
I’M GONNA PLAY DISNEY SONGS	6	HOW ABOUT THE WEATHER	4
KINDA FALLING OVER ALL THE TIME	6	IT’S QUITE A BIG SPORT HERE ACTUALLY	8
WHAT HAVE YOU DONE TO OVERCOME THEM	7	HARD ON YOUR KNEES, TOO	5
I CAN’T REMEMBER EXACTLY HOW HIGH IT IS	9	IT’S LIKE WHAT YOU’RE TALKING ABOUT	8
WHEN YOU GO TO THE BEACH	6	RANDOM LITTLE THINGS LIKE THAT	5
SOMETHING SILLY AND FUN	4	BUT I DO LIKE TO WATCH IT	7
AND BOTH OF THEM PLAY NOW	6	IT’S A GREAT SPORT TO WATCH	7
THINK OF ALL THE OTHER AUSTRALIAN SPORTS	7	TAP BACK IN TO THAT	5
GOING INTO WINTER IS A BIT IN BETWEEN	8	I LOVE THAT SORT OF STUFF	6
I LOVE WATCHING IT	4	IT’S COLD OUTSIDE BUT INSIDE IT’S WARM	9
WHEN YOU KNOW YOU’RE IN A WARM PLACE	9	THAT MUST BE AMAZING TO WATCH	6
NOT MISSING THE FUN THINGS	5	SO THAT’S like TWENTY TWO YEARS	6
SO THAT WAS KINDA CUTE	5	PULLS ME BACK IN	4
A LITTLE BIT STILL BUT NOT AS MUCH	8	YOU GO OUT FOR DAYTRIPS	5
THE MAIN ALBUM	3	YOU CAN SEE A SEA SNAKE	6

Loud vocal effort			
Female		Male	
Sentence	Scoring Units	Sentence	Scoring Units
WHAT’S HER OTHER SIDE	5	HAVE YOU BEEN ELSEWHERE IN EUROPE	6
BUT I DON’T GET TO FINISH BOOKS ANYMORE	9	I PLAY A LOT OF VIDEO GAMES	7
WE SAW HIM RUN ALL THE WAY AROUND	8	WHAT WOULD IT BE	4

Continued

Loud vocal effort			
Female		Male	
Sentence	Scoring Units	Sentence	Scoring Units
OUR BABY’S BEEN PRETTY GOOD	6	IN A MARKET WHERE THERE’S STREET FOOD	8
WHAT DO YOU PLAY	4	DO YOU EXPERIENCE THAT HERE IN AUSTRALIA OR MORE OVERSEAS	10
HOW MUCH OF THE RECORDS ARE STILL AVAILABLE	8	I READ THROUGH HIGH SCHOOL AS WELL	7
EVERYONE’S JUMPING UP GETTING THEIR LUGGAGE	7	YOU CAN BE WHO YOU WANT	6
BUT WE THOUGHT WE’D TRY IT	7	YOU DON’T KNOW IF THEY’RE REAL MEMORIES	9
WHO DO YOU THINK YOU ARE	6	WALKING REALLY SLOWLY	3
BUT THERE WAS A LOT OF CONFLICT	7	WE WOULD HAVE THESE AMAZING HOUSE PARTIES	7
PUT EFFORT INTO A DRESS UP PARTY	7	DO YOU EVER PLAY TOGETHER?	5
HOW TO RECONCILE THE DIFFERENCES	5	PEOPLE ALL OVER EUROPE	4
YOU’D LIKE IT TO BE UNDER WATER	8	SAME WITH GRASSHOPPERS	3
I WENT ON ONE OF THOSE AT like THE GREAT BARRIER REEF	11	PLAY FOR FIVE MINUTES	4
THOUGHT ABOUT THE CONCEPT MORE	5	I’VE TRIED like A GERMAN ONE	6
MAKING SOUNDS AT THEM	4	SOME VERY INTERESTING KIND OF REVELATIONS	6
EVERYTHING’S JUST BEAUTIFUL	4	DO YOU HAVE ANY OTHER HOBBIES	6
THERE’S NO KINDA THINKING TIME	6	I READ A LOT WHEN I WAS A KID	9
THERE’S A PLACE WE GO SKIING	7	DO YOU READ NOW	4
STAY WITH IT AND CONTINUE	5	THAT’S WHERE YOUR FAMILY’S FROM	7
EVERYTHING IS SO CLOSE	4	STANDING AND QUEING TO GET ON A PLANE	8
HE PLAYS like MAYBE A COUPLE OF HOURS A WEEK	9	PURPOSE BUILT BUILDINGS	3
YOU CAN GO UP ON THE SWISS SIDE	8	RUDE PEOPLE DRIVING ANNOY ME	5
IN THE SAFETY OF THEIR OWN CAR	7	MILLIONS AND MILLIONS OF THESE INSECTS	6
SO IT’S ALL DESIGNED THAT WAY	7	PARTICULARLY BEING FROM THE NEW ENGLAND AREA	7
BUT WE DO KIND OF GET VIP TREATMENT	8	I THINK THAT MANIFESTS IN LOTS OF DIFFERENT WAYS	9
DRIVING UP AND DOWN A SUBURBAN STREET	7	I WOULD BET THAT MOST OF THEM ARE MEN AS WELL	11
THEY ARE JUST KIND OF CHICKEN SKIN AROUND CHICKEN BONES	10	WHAT WOULD YOUR THEME BE	5
BUT HE DECIDED TO RUN AROUND	6	I SOMETIMES STRUGGLE GOING TO SLEEP	6
WE TRY TO BUILD IN A LITTLE HOLIDAY	8	THAT ANNOYS ME	3
WHAT KIND OF PICTURES THEY’RE LOOKING AT	8	I HAVE OUTGROWN IT	4
YEAH THAT WOULD BE COOL	5	BUT VERY EXPENSIVE	3
WHO ARE YOU SHOWING OFF TO	6	I WAS ONLY THERE FOR A FEW DAYS	8
THERE’S A BIT OF WELSH THERE AS WELL	9	IN ONE OF THE HOTELS THERE	6
HOW FAR THEY REACH	4	I PLAY ALL SORTS OF GAMES	6
I THINK THEY’RE OK	5	WE WENT TO DISNEYLAND	4
SO FAR SO GOOD	4	DRIVING THESE GIANT THINGS	4
PUT HIM IN A REALLY CUTE OUTFIT	7	BUT IT’S ALSO JUST THE CITY ITSELF	8
YOU JUST FEEL THEM ON YOUR BACK	7	VERY LITTLE CARBON FOOTPRINT	4
NOT EATING AND DOING ANYTHING ELSE	6	THEY RAISE THEM IN THESE BUILDINGS	6
I GUESS IT IS YOUR HOBBY	6	I LOVE THAT SHOW	4
I CAN’T TELL THEY’RE IN HELMETS	8	I ALWAYS MAKE TIME FOR BERLIN	6
THERE WAS A TIME WHEN IT WAS	10	THE WHOLE THING IS PROTEIN	5

Continued

Loud vocal effort			
Female		Male	
Sentence	Scoring Units	Sentence	Scoring Units
DIFFICULT TO SLEEP			
MY HUSBAND IS SWISS	4	BEING A BIT CHILDISH IN THAT WAY	7
WE HAVE SO MANY HAND ME DOWNS	7	MAYBE NOT THE WHOLE HOTEL	5
MY FAVOURITE'S STILL ITALY	5	WELL I MEAN I LOVE CHINESE FOOD	7
I'VE HAD THAT EXPERIENCE WITH SHOPPING TROLLEYS	8	WHERE IT'S ALL CRAZY ON NEW YEARS EVE	9
WHEN HE GETS SOMETHING NEW I'LL PLAY A LITTLE BIT	11	MEN WHO JUST WALK DOWN THE STREET	7
IT'S PRETTY TASTY	4	ONE HALF OF MY FAMILY	5
SHE FOUND THESE OLD PHOTOS	5	FOR THE REST OF YOUR LIFE	6
ABOUT TO FLY FOR FIFTEEN HOURS	6	THERE MUST HAVE BEEN SOME INTERACTION	6
WE PROBABLY HAVE like FORTY MINUTES	5	I'VE COME CLOSE A FEW TIMES	7
GOING TO HONG KONG AS A STOP OVER'S PRETTY GOOD	11	I WATCHED A VIDEO	4
IN A DARK ROOM FOR TWENTY HOURS A DAY	9	I'VE ONLY BEEN TO ITALY ONCE	7
I THINK I WOULD FEEL A BIT CLAUSTROPHOBIC	8	WHAT ELSE ANNOYS ME	4
RAISING THEIR FRONT WHEEL	4	TO MAKE TIME FOR IT	5
AND THEN WE SAW HIM SIT DOWN	7	THAT'S WEIRD TOO I THINK	6
I FIND PEOPLE SURPRISINGLY uhh UNACCOMODATING	5	BUT THERE'S CERTAINLY THEMED ASPECTS TO THE HOTEL	9
BECAUSE YOUR SEAT IS ALREADY ASSIGNED	6	DEPENDS ON THE GAME	4
I DID A LOT OF THAT DURING PREGNANCY	8	I'VE BEEN IN ONE THAT WAS DOCKED	8
WHY DO YOU LOVE IT SO MUCH	7	HAVE A REALLY RICH STORY	5
WHICH I QUITE LIKED AS WELL	6	IT MUST BE HARD TO TRAVEL WITH A with a BABY THOUGH	10
I CAN EAT LOTS AND LOTS OF DUMPLINGS	8	NEW YEARS EVE PARTIES	4
AS SOON AS I LIE DOWN I'M OUT	9	THIS ONE MAN	3

Alku, P., Vintturi, J., and Vilkman, E. (**2001**). "The use of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal, and loud phonation," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EuroSpeech 2001), Aalborg, Denmark*, Vol. 2, pp. 919–922.

ANSI S3.5 (R2012). (**1997**). *American National Standard: Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York), 1969(R 1986), 1–35.

Beechey, T., Buchholz, J. M., and Keidser, G. (**2018**). "Measuring communication difficulty through effortful speech production during conversation," Speech Commun. **100**(April), 18–29.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**(3), 108–112.

Best, V., Keidser, G., Buchholz, J. M., and Freeston, K. (**2015**). "An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment," Int. J. Audiol. **54**(10), 682–690.

Best, V., Keidser, G., Buchholz, J. M., and Freeston, K. (**2016**). "Development and preliminary evaluation of a new test of ongoing speech comprehension," Int. J. Audiol. **55**(1), 45–52.

Boersma, P., and Weenink, D. (**2018**). "Praat: Doing phonetics by computer [computer program]," http://praat.org (Last viewed September 13, 2019).

Branigan, H. P., Catchpole, C. M., and Pickering, M. J. (**2011**). "What makes dialogues easy to understand?," Lang. Cognit. Process. **26**(10), 1667–1686.

Bronkhorst, A. W., Brand, T., and Wagener, K. (**2002**). "Evaluation of context effects in sentence recognition," J. Acoust. Soc. Am. **111**(6), 2874–2886.

Brungart, D. S., Sheffield, B. M., and Kubli, L. R. (**2014**). "Development of a test battery for evaluating speech perception in complex listening environments," J. Acoust. Soc. Am. **136**(2), 777–790.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., and Kiessling, J. (**1994**). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**(4), 2108–2120.

Castellanos, A., Benedí, J.-M., and Casacuberta, F. (**1996**). "An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect," Speech Commun. **20**(1-2), 23–35.

CHABA (**1988**). "Speech understanding and aging," J. Acoust. Soc. Am. **83**, 859–895.

Cipriano, M., Astolfi, A., and Pelegrín-García, D. (**2017**). "Combined effect of noise and room acoustics on vocal effort in simulated classrooms," J. Acoust. Soc. Am. **141**(1), EL51–EL56.

Cooke, M., Mayo, C., and Villegas, J. (**2014**). "The contribution of durational and spectral changes to the Lombard speech intelligibility benefit," J. Acoust. Soc. Am. **135**(2), 874–883.

Cord, M., Baskent, D., Kalluri, S., and Moore, B. C. J. (**2007**). "Disparity between clinical assessment and real-world performance of hearing aids," Hear Rev **14**, 22–26.

Davis, C., Kim, J., Grauwinkel, K., and Mixdorff, H. (**2006**). "Lombard speech: Auditory (A), Visual (V) and AV effects," Proc. Speech Prosody **2**(V), 361–365.

Dawson, P. W., Hersbach, A. A., and Swanson, B. A. (**2013**). "An adaptive Australian Sentence Test in Noise (AuSTIN)," Ear Hear. **34**(5), 592–600.

Diehl, R. L., Lindblom, B., Hoemeke, K. A., and Fahey, R. P. (**1996**). "On explaining certain male-female differences in the phonetic realization of vowel categories," J. Phonetics **24**(2), 187–208.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (**2001**). "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment (Ruidos ICRA: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditiv"), Int. J. Audiol. **40**(3), 148–157.

Ernestus, M., and Warner, N. (**2011**). "An introduction to reduced pronunciation variants," J. Phonetics **39**(3), 253–260.

Geissler, G., and Arweiler, I. (**2014**). "Speech reception threshold benefits in cochlear implant users with an adaptive beamformer in real life situations," Cochlear Implants Int. **16**(2), 69–76.

Gibson, E. (**1998**). "Linguistic complexity: Locality of syntactic dependencies," Cognition **68**(1), 1–76.

Gifford, R. H., Olund, A. P., and DeJong, M. (**2011**). "Improving speech perception in noise for children with cochlear implants," J. Am. Acad. Audiol. **22**(9), 623–632.

Hagerman, B. (**1982**). "Sentences for testing speech intelligibility in noise," Scand. Audiol. **11**(2), 79–87.

Hochmuth, S., Kollmeier, B., Brand, T., and Jürgens, T. (**2015**). "Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests," Int. J. Audiol. **54**, 62–70.

Holmes, E., Folkeard, P., Johnsrude, I. S., and Scollie, S. (**2018**). "Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment," Int. J. Audiol. **57**(7), 438–492.

Jansen, S., Koning, R., Wouters, J., and Van Wieringen, A. (**2014**). "Development and validation of the Leuven intelligibility sentence test with male speaker (LIST-m)," Int. J. Audiol. **53**(1), 55–59.

Jerger, J. (**2009**). "Ecologically valid measures of hearing aid performance," Starkey Audiol. Ser. **1**, 1–14.

Keidser, G., Ching, T., Dillon, H., Agung, K., Brew, C., Brewer, S., Fisher, M., Foster, L., Grant F., and Storey, L. (**2002**). "The National Acoustic Laboratories' (NAL) CDs of speech and noise for hearing aid evaluation: Normative data and potential applications," Aust. N. Z. J. Audiol. **24**(1), 16–35.

Kelly, H., Lin, G., Sankaran, N., Xia, J., Kalluri, S., and Carlile, S. (**2017**). "Development and evaluation of a mixed gender, multi-talker matrix sentence test in Australian English," Int. J. Audiol. **56**(2), 85–91.

Kiessling, J., Pichora-Fuller, M. K., Gatehouse, S., Stephens, D., Arlinger, S., Chisolm, T., and von Wedel, H. (**2003**). "Candidature for and delivery of audiological services: Special needs of older people," Int. J. Audiol. **42**(Suppl. 2), 2S92–2S101.

Killion, M. C., Schulein, R., Christensen, L., Fabry, D., Revit, L., Niquette, P., and Chung, K. (**1998**). "Real-world performance of an ITE directional microphone," Hear. J. **51**, 1–6.

Laan, G. P. (**1997**). "The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style," Speech Commun. **22**(1), 43–65.

Lane, H., Tranel, B., and Sisson, C. (**1970**). "Regulation of voice communication by sensory dynamics," J. Acoust. Soc. Am. **47**(2B), 618–624.

Lombard, E. (**1911**). "Le signe de l'elevation de la voix," ("The sign of the raising of the voice") Ann. Mal. L'Oreille Larynx 101–119.

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

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099.

Oreinos, C., and Buchholz, J. M. (**2015**). "Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones," J. Acoust. Soc. Am. **137**(6), 3447–3465.

Oreinos, C., and Buchholz, J. M. (**2016**). "Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids," J. Am. Acad. Audiol. **27**(7), 541–556.

Pichora-Fuller, M. K. (**2007**). "Audition and cognition: What audiologists need to know about listening," in *Hearing Care for Adults*, edited by C. Palmer and R. Seewald (Phonak, Stäfa, Switzerland), pp. 71–85.

Schneider, B. A., Pichora-Fuller, M. K., and Daneman, M. (**2010**). "Effects of senescent changes in audition and cognition on spoken language comprehension," in *The Aging Auditory System*, edited by S. Gordon-Salant, R. D. Frisina, A. N. Popper, and R. R. Fay (Springer, New York), pp. 167–210.

Summers, W. Van, Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**(3), 917–928.

van Rooij, J. C. G. M., and Plomp, R. (**1991**). "The effect of linguistic entropy on speech perception in noise in young and elderly listeners," J. Acoust. Soc. Am. **90**(6), 2985–2991.

Van Wieringen, A., and Wouters, J. (**2008**). "LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands," Int. J. Audiol. **47**(6), 348–355.

Weisser, A., and Buchholz, J. M. (**2019**). "Conversational speech levels and signal-to-noise ratios in realistic acoustic conditions," J. Acoust. Soc. Am. **145**(1), 349–360.

Weisser, A., Buchholz, J. M., Oreinos, C., Badajoz-Davila, J., Galloway, J., Beechey, T., and Keidser, G. (**2019**). "The ambisonic recordings of typical environments (ARTE) database," Acta Acust. Acust. **105**(4), 695–713.