INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

R^{2}SPIN: Re-recording the Revised Speech Perception in Noise Test

Lauren Ward[1,2], *Catherine Robinson*[3], *Matthew Paradis*[4], *Katherine M. Tucker*[5], *Ben Shirley*[1]

[1]Acoustics Research Centre, University of Salford, Manchester, UK
[2]BBC R&D, North Lab, MediaCityUK, Salford, UK
[3]BBC Cardiff, Wales, UK
[4]American Academy of Dramatic Arts, Los Angeles, USA
[5]BBC R&D, South Lab, London, UK

L.Ward7@edu.salford.ac.uk

Abstract

Speech in noise tests are an important clinical and research tool for understanding speech perception in realistic, adverse listening conditions. Though relatively simple to implement, their development is time and resource intensive. As a result, many tests still in use (and their corresponding recordings) are outdated and no longer fit for purpose. This work takes the popular Revised Speech Perception In Noise (RSPIN) Test and updates it with improved recordings and the addition of a female speaker. It outlines and evaluates a methodology which others can apply to legacy recordings of speech in noise tests to update them and ensure their ongoing usability. This paper describes the original test along with its use over the last four decades and the rationale for re-recording. The new speakers, new accent (Received Pronunciation) and recording methodology are then outlined. Subjective and objective analysis of the new recordings for normal hearing listeners are then given. The paper concludes with recommendations for using the R^{2}SPIN.

Index Terms: speech intelligibility, speech in noise tests, speech perception, adverse listening conditions

1. Introduction

Tests of intelligibility, or hearing in noise tests, are an important clinical tool for understanding how an individual's hearing loss affects them in everyday scenarios. They are also a vital research tool for understanding how humans perform auditory scene analysis and navigate 'cocktail party effect' type scenarios. They are also utilized to explore a wide range of psychoacoustic phenomena including the effect of linguistic context [1] and spatial release from masking [2]. A large variety of these tests have been developed and they can be segregated into two main types: everyday sentence tests, using meaningful real-life sentences [3, 4] (e.g. the sentence 'The clown has a funny face' [5]) and matrix sentence tests, which follow a strict sentence form for each stimuli and require training [6, 7] (e.g. 'Thomas wants nine cheap beds' [8]). These tests, regardless of type, generally have sentences which are phonetically balanced across the test. These tests often use an adaptive paradigm where the SNR is varied until the 'Speech Reception Threshold', where 50% of the speech is intelligible, is determined [6, 9, 10]. The alternate approach is to utilize a static signal to noise ratio and quantify performance by the percentage of words correctly identified [1, 11]. These tests use a variety of masking noises, commonly multi-talker babble [1] or speech shaped noise [11]. There can also be an advantage if the test includes visuals of the speaker, as this allows for characterization of the effect of lip-reading and other multi-modal cues.

Whilst highly valuable research tools, these tests require substantial time and resource to develop and validate. For this reason a single test, once developed, is often utilized for many decades [1, 12]. This facilitates comparison of current results to previous research and the normative values for the original test [13]. However if not regularly assessed and revised, these tests have the potential to become invalid (because of assumptions about target population) or for recordings to become obsolete.

1.1. The (Revised) Speech Perception in Noise test

The Speech Perception in Noise (SPIN) test consists of phonetically balanced sentences spoken by a male speaker in American English. It is presented in multi-talker babble mixed from 12 speakers. Sentences end with a monosyllabic noun, the keyword. These were selected from words with frequencies of use in the range of 5 to 150 per million [14]. Respondents are scored on whether they correctly identify this keyword.

The original SPIN stimuli were developed to evaluate both top-down and bottom-up processes involved in understanding speech in noise [1]. This was achieved by controlling the predictability of the sentences, either giving the listeners clues to the keyword (e.g. *'Stir your coffee with a **spoon**'* and termed *high predictability* - **HP**) or no clues (e.g. *'Bob could have known about the **spoon**'* and termed *low predictability* - **LP**). Recognition of the keyword in these LP sentences relies entirely on receiving the acoustic signal of the keyword correctly. The HP stimuli differ in that the surrounding sentences allows for the use of top-down processing: any ambiguity in the keyword's acoustic signal can be resolved using knowledge of the English language and the contextual information in the sentence.

The original version by Kalikow contained 10 lists, each consisting of 50 sentences; 25 HP and 25 LP. Each half list (25 sentences) was further constrained to at least 12 sentences of each predictability level and edited so that their RMS levels were within 0.1dB of each other. Eight of these lists were shown to give equal performance for young normal hearing listeners at a 0dB speech to noise ratio (SNR).

In 1984 Bilger revised Kalikow's SPIN to give the lists balanced performance for hard of hearing listeners, terming it the Revised SPIN (RSPIN). In doing so he removed two lists and redistributed the remaining sentences using the psychometric data from 128 elderly listeners with sensorineural hearing loss. Validation of the new lists was performed with 32 of the original listeners, who had a mean performance of 76% and 37% for the HP and LP sentences, respectively, at 8 dB SNR. Over the ensuing decades the RSPIN test has been utilized widely with normal hearing as well as hard of hearing listeners [13, 15, 16, 17, 18, 19, 20]. It has had a variety of modifications

http://dx.doi.org/10.21437/Interspeech.2019-1281

and inclusions to the stimuli and test procedure. The following gives a limited review of these.

Beyond investigating the difference between young and old listeners performance in noise, R-SPIN has been used by Pichora-Fueller with an additional working memory task [20]. This involved the respondent remembering the last n words they had identified in order to evaluate the effect of adverse listening conditions on how working memory is allocated for both young and older listeners. Wilson restructured the test into a multiple signal to noise ratio paradigm, to allow for a speech reception threshold for hearing impaired listeners to be identified [16]. The test has also been utilised in broadcast research [13, 15, 17, 18]. Shirley modified the stimuli to introduce the comb filtering effects of a phantom centre speaker and demonstrated its adverse affect on intelligibility [15]. Ward modified the stimuli to include non-speech sounds (sound effects) which provide the same level of context to the listener as the HP sentences and utilised with both normal hearing [13] and hard of hearing participants [18].

1.2. Reasons for re-recording

The R-SPIN test has provided a valuable research tool for a number of decades, however it has limitations. The original recording was made onto magnetic tape, resulting in the speech having a limited bandwidth. Furthermore there is the presence of high frequency tape hiss on the recordings. The stimuli has only a male speaker, making the stimuli an inherently unbalanced tool. The speaker is speaking in American English which, whilst suitable for research in America, limits applicability in other English speaking countries.

2. Methodology

To address some of the issues presented by the original recording, the re-recording contains both a male and female speaker speaking British Received Pronunciation. Whilst Received Pronunciation does not resolve the issue of a single accent, it broadens the options available to researchers. Utilization of modern recording techniques increases the recording quality. Beyond addressing the limitations of the original recording, R^2SPIN adds a a binaural recordings of the stimuli, which will be described in subsequent work. This widens the application of the tool to evaluation of spatial hearing in noise.

2.1. Validity and Phonetic Balance

In the original test to ensure word familiarity the words were selected from [14], which contain the 30,000 most frequently used words. Whilst some sentences do show their age (*My TV has a 12 inch screen*), given their simplicity they were still deemed to be familiar and interpretable to modern listeners. Furthermore, retaining all the same sentence text helped to maintain the phonetic balance described in the following section.

Phonetic analysis was conducted to ensure that the change of accent did not significantly change the phonetic balance of the lists. Sentences were transcribed into Standard American Pronunciation and Received Pronunciation by the authors. All transcriptions are available with the recordings.

2.2. Speakers and Audio Recording

Two native British English speakers, one male and one female, with extensive experience in broadcast and radio were selected. Each speaker recited all 400 sentences. The speakers were instructed to use a neutral tone and pace. The audio was recorded in a quiet room at BBC R&D, Cardiff. They were recorded using a Neumann TLM 193 microphone at a distance of 0.25m from the speaker. They were recorded into a Sadie digital audio workststation at a sample rate of 48kHz and bit depth of 32bit and saved as uncompressed .wav files.

2.3. Quality Verification and Post-processing

The quality of the recorded sentences was assessed to ensure recording quality, speech clarity and correct pronunciation. This assessment was undertaken by five experienced listeners validating the sentences over headphones. They were instructed to identify any of the following problems: mispronunciations (compared with the target sentences), rushed or slow delivery, speaking too loudly or softly, editing and recording artifacts as well as any miscellaneous problems. One assessor conducted the validation without the target sentence list to ensure the effects of priming did not prevent error identification. This was performed for both the male and female sentences and at least two of the assessors were different for each iteration. All problematic sentences identified were re-recorded using the same equipment and conditions described in Section 2.2. For sentences with mispronunciations, specific error notes were fed back to the speakers.

Post-processing was performed on the sentences using Adobe Audition software. Silences between sentences were removed manually. The sentences were aligned with the keyword of the original speech, such that the babble noise for all keywords (old and new) were as identical as practically possible.

2.4. Pilot

A pilot study with 6 listeners was undertaken to determine the appropriate SNR for Sections 3 and 4. An appropriate SNR is defined as one where the HP sentences do not saturate at the top of the psychometric curve and similarly the LP sentences don't saturate at the floor. This is achieved by having an overall word recognition rate of $\approx 50\%$. Initially -2dB SNR was used with 3 pilot participants, as in previous work [13]. These were seen to have a mean word recognition rate of 74.5%, averaged across both HP and LP sentences. This was reduced to -4dB SNR, for a further three pilot participants, reducing the mean value to 56.3% across all sentences.

3. Objective intelligibility analysis

Objective intelligibility measures analyse signal and masker interactions and how they affect word level intelligibility (e.g. [21, 22]). This gives an initial insight into the possible intelligibility differences between lists, without requiring large scale subjective evaluation. Both the original and new stimuli were analyzed using an objective intelligibility metric called the glimpse proportion (GP) [23]. The GP quantifies the number of time-frequency regions of speech which survive energetic masking and reflects the local audibility of speech in noise.

3.1. Methodology

All sentences were normalized to -23 LUFS using the ITU-R BS.1770 specification [24]. This was also done for the multitalker babble signal, allowing the -4dB SNR selected in the pilot (Section 2.4) to be set. This was done for both the original and new speakers, though only the new speakers were used for the subjective evaluation (Section 4). The GP was calculated in

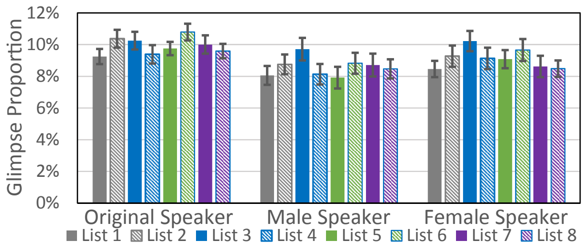

Figure 1: *Glimpse proportion for each speaker and list, with standard error shown*

Table 1: *Mean Glimpse Proportions by list*

List	Key Word	Preceding Speech
1	8.6%	15.3%
2	9.5%	15.4%
3	10.1%	15.6%
4	8.9%	16.0%
5	8.9%	15.9%
6	9.8%	15.2%
7	9.1%	15.7%
8	8.9%	16.2%

Table 2: *Mean Glimpse Proportions by speaker*

	Key Word	Preceding Speech
Original	9.9%	15.3%
Male	8.6%	17.9%
Female	9.1%	13.8 %

Matlab and calculated twice for each sentence: once over the keyword only and once over the preceding speech only. This allows analysis of the effect of energetic masking on the keyword itself and the speech preceding the keyword (given its importance in the HP sentences). Keywords were aligned as described in Section 2.3, so that direct comparisons could be made. Due to differences in accent and pacing, the remainder of the sentences were not aligned.

3.2. Results

Table 1 indicates some inter-list difference between the mean keyword GP. The distribution of the means across the lists is reasonably constant (i.e. the lists with higher GP are roughly similar across the speakers). The GP results indicate which lists and speakers are likely to be more difficult due to higher energetic masking (i.e. lower GP). Tables 1 and 2 show the mean GP for keyword and preceding speech for each speaker and list respectively. Table 1 shows that List 3 has the highest mean GP for the keyword, followed by List 6. The list with the lowest keyword GP is List 1. For the preceding speech the lists with the highest GP are 4 and 8, with List 6 being the lowest. Table 2 shows that the original speaker has the highest mean GP and easiest level of energetic masking. For the preceding speech this order changes with the male speaker (most difficult for the keyword) having the highest GP for the preceding speech. The female speaker's preceding speech shows the lowest average GP. The keyword GP of the keywords are smaller and more variable, as in previous studies [13]. This results from the short time window of the keyword which makes it more vulnerable to the fluctuations in the masker.

A two-way ANOVA was performed to determine whether these apparent differences were significant and was performed separately for the keyword and preceding GPs. A highly significant difference between speakers was shown for both keywords $[F = 9.99, p < 0.001]$ and preceding speech $[F = 185, p < 0.001]$. Post-hoc testing showed that for the keywords the male and female speakers were significantly different to the original speaker, but not to each other. For the preceding speech, all speakers were significantly different. The lists showed a weak significant difference for both the keywords $[F = 2.09, p < 0.05]$ and the preceding speech $[F = 2.13, P < 0.05]$. Interaction effects were not significant.

4. Subjective intelligibility analysis

This section's small scale subjective analysis complements the objective analysis with a human normal hearing population (as validation of all possible experimental combinations with a large human population is impractical). The subjective analysis only evaluates the new speakers and uses the same stimuli as prepared for Section 3.

4.1. Methodology

The sentence lists were presented to the participants in a pseudo-random order. Each participant received all eight lists, four with a male speaker and four with female and each gender using a keyword only once. Lists with the same keyword were separated as much as practical to reduce learning effects. The experiment was broken into four parts each containing two lists, one of each gender, and after each part participants were offered breaks to avoid fatigue. The order in which the lists were presented was also pseudo-randomized with each of the sixteen possible orders being presented once.

Stimuli were presented to the listener over a set of Sennheiser HD 800 headphones. Tests were undertaken in listening rooms or sound-proofed studio environments at two locations; BBC R&D and the University of Salford. They were presented with speech and noise co-located at 0° and reproduced at a level of 69dBA (calibrated using pink noise). Participants used pen and paper to record each keyword.

4.2. Participants

Only participants with normal hearing (self-identified) and with English as their first language were recruited. Participants were naive listeners, defined here as having participate in less then 10 previous listening experiments. Participants were under the age of 35 years old to avoid the possibility of un-diagnosed age-related hearing loss. 5 females and 11 males meeting these criteria were recruited in the age range 19-35, with a median age of 23 years old and mean age of 24.

4.3. Results

The word recognition rate (WRR) results can be seen in Figure 2. For all except the female speaker on list three, the characteristic R-SPIN improvement of 30-40% between LP and HP sentences is maintained. Variation exists between the lists and speakers, which is to be expected given Section 3. We can see that for all except the male HP sentences in list 1, the female speaker has a higher WRR and this is particularly pronounced for the LP sentences. Given that the female speaker has a higher

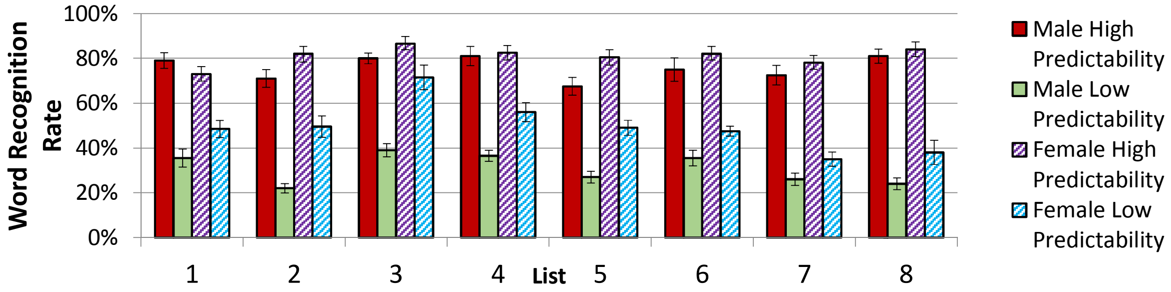

Figure 2: *Word recognition rate for each list and speaker, with standard error shown*

Table 3: *First Order Significant Factors from GEE Analysis*

Factor	Odds	95% Conf.	
	Ratio	Interval	
Retake	0.73	±1.05	**
List	0.96	±0.22	*
Speaker Gender	0.26	±0.73	***
Predictability	2.41	±1.10	***
Keyword GP	1.10	±0.10	***

$^*p < .05$, $^{**}p < .01$, $^{***}p < .001$

GP on average than the male speaker over the keywords (which are likely to dominate the result for LP sentences) this could be attributed to lower energetic masking. To investigate this effect further the 8 female lists were repeated with three listeners at a more challenging -6dB SNR. This reduced the average value of the WRR from 81.1% to 65.5% and 49.4% to 31.8% for HP and LP respectively. For list 3, which Figure 2 shows has a particularly high WRR for the LP sentences, the WRR reduces from an average of 71.5% to 52.0%.

Given the dichotomous outcome variable (right or wrong), a standard ANOVA could not be performed. More traditional methods for dichotomous outcome variables, such as logistic regression, also could not be used due to the repeated measures design which violates the assumption of the independence of errors. Generalized estimating equations (GEE) are used here instead, as they follow a similar form to logistic regression but additionally utilize robust standard error estimates to account for random or repeated factors. The data was tested for multicollinearity and complete separation, and was found not to violate these assumptions. In addition to the two design factors of the experiment (predictability and speaker gender), further predictors were investigated: *list*, *order of presentation*, *GP* (keyword and preceding speech), *whether the sentence was the original re-recording or a retake made after Section 2.3* and *participant age* and *gender*. A model containing up to second order interactions, to ensure interpretable results given the large number of factors, was developed using the package geepack in R [25]. Wald's test was then used to determine which factors offered significant improvements to the power of the model.

Table 3 shows significant first order factors and confirms that *predictability*, the factor under evaluation in RSPIN, is significant and has the largest effect size (largest odds ratio). *speaker gender* and *keyword GP* are also both highly significant with *speaker gender* also having a large effect size. This, given the results in Section 3, is probably caused by the speaker based differences in energetic masking. *Differences between lists* was also significant, though only weakly and with a small effect size. *Take/retakes* were also significantly different, though given the large confidence interval of the odds ratio, it is likely this is capturing the variation inherent in the speech rather than the variation due to re-recording. The interactions with high significance ([$P < 0.01$]) were: list*predictability, participant gender*age, age*order, GP*order, GP*predictability, list*keyword GP and speaker gender*keyword GP. These interactions had odds ratios in the range of 0.85 to 1.08, indicating their effects are small.

5. Recommendations for Use

Section 4 shows that R^{2}SPIN continues to be a useful research tool for analyzing the effect of top-down processing on speech understanding in adverse listening conditions. The difference Section 3 highlights between the energetic masking of the different speakers is to be expected given the different spectral characteristics of the speakers due to gender (new speakers) and recording techniques (new and original speaker). From Section 4.3 it is recommended that the female speaker is set at a larger SNR than the male speaker, e.g. -6dB SNR, if using the original RSPIN noise or similar masker. Variation is to be expected between the lists and could easily be controlled by adjusting the alignment of the sentences relative to the noise to make the GPs more uniform.

It needs to be noted that, as with the original RSPIN, the normative data in this paper only holds for the exact timings relationship between the speech and noise used in this recording. This is due to the fluctuating nature of the multi-talker babble masker. Using other timing relationships, or other noise, will produce different results. This paper's results validate that the new recording still effectively evaluates the effect of predictability of speech in noise. Furthermore, it provides an important tool for those utilizing the stimuli to better understand the behaviour of the different lists and speakers. Finally, we have presented a methodology which others can apply to legacy recordings of speech in noise tests to update them and ensure their ongoing value to the research community.

The recordings of R^{2}SPIN are openly available and can be accessed at: `https://github.com/bbc/r2spin`

6. Acknowledgment

The EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1), the BBC Audio Research Partnership and General Sir John Monash Foundation supported this work. The authors acknowledge the University of Illinois for their permission to re-record RSPIN and in obtaining original RSPIN documentation.

7. References

[1] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1337–1351, 1977.

[2] T. L. Arbogast, C. R. Mason, and G. Kidd Jr, "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2169–2180, 2005.

[3] R. Bilger, *Speech recognition test development, In: E. Elkins ed. Speech recognition by the hearing impaired*, 1984, vol. 14, pp. 2–15.

[4] Etymotic, "BKB-SIN^{TM} Speech-in-Noise test Version 1.03," 2005.

[5] J. Bench, Å. Kowal, and J. Bamford, "The bkb (bamford-kowal-bench) sentence lists for partially-hearing children," *Br. J. Audiol.*, vol. 13, no. 3, pp. 108–112, 1979.

[6] K. Wagener, V. Kühnel, and B. Kollmeier, "Development and evaluation of a german sentence test i: Design of the oldenburg sentence test," *Zeitschrift Fur Audiologie*, vol. 38, pp. 4–15, 1999.

[7] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[8] HearCom, "Matrix sentence test," 2010. [Online]. Available: http://hearcom.eu/prof/DiagnosingHearingLoss/AuditoryProfile/SpatialHearing.html

[9] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1085–1099, 1994.

[10] R. A. McArdle, R. H. Wilson, and C. A. Burks, "Speech recognition in multitalker babble using digits, words, and sentences," *J. Am. Acad. Audiol.*, vol. 16, no. 9, pp. 726–739, 2005.

[11] K. R. Duncan and N. L. Aarts, "A comparison of the HINT and Quick Sin Tests," *J. Speech Lang. Path. Audiol.*, vol. 30, no. 2, p. 86, 2006.

[12] E. Skoe and K. Karayanidi, "Bilingualism and speech understanding in noise: Auditory and linguistic factors," *J. Am. Acad. Audiol.*, vol. 30, no. 2, pp. 115–130, 2018.

[13] L. Ward, B. Shirley, Y. Tang, and W. Davies, "The effect of situation-specific acoustic cues on speech intelligibility in noise," in *Proc. Interspeech 2017: 18th Annual Conf. of International Speech Communication Association*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 2958–2962.

[14] E. L. Thorndike and I. Lorge, "The teacher's word book of 30,000 words." 1952.

[15] B. Shirley, "Improving television sound for people with hearing impairments," Ph.D. dissertation, University of Salford, 2013.

[16] R. H. Wilson, R. McArdle, K. L. Watts, and S. L. Smith, "The revised speech perception in noise test (R-SPIN) in a multiple signal-to-noise ratio paradigm," *J. Am. Acad. Audiol.*, vol. 23, no. 8, pp. 590–605, 2012.

[17] A. Carmichael, "Evaluating digital "on-line" background noise suppression: Clarifying television dialogue for older, hard-of-hearing viewers," *J. Neuropsychol. Rehab.*, vol. 14, no. 1-2, pp. 241–249, 2004.

[18] L. Ward and B. Shirley, "Television dialogue; balancing audibility, attention and accessibility," in *Conference on Accessibility in Film, Television and Interactive Media*, York, UK, Oct. 2017.

[19] L. E. Humes, B. U. Watson, L. A. Christensen, C. G. Cokely, D. C. Halling, and L. Lee, "Factors associated with individual differences in clinical measures of speech recognition among the elderly," *J. Speech, Lang. Hear. Res.*, vol. 37, no. 2, pp. 465–474, 1994.

[20] M. K. Pichora-Fuller, B. A. Schneider, and M. Daneman, "How young and old adults listen to and remember speech in noise," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 593–608, 1995.

[21] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication*, vol. 49, no. 5, pp. 402–417, 2007.

[22] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Proc. Interspeech 2011: 12th Annual Conf. of International Speech Communication Association*. Florence, Italy: ISCA, 2011, pp. 345–348.

[23] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[24] ITU Recommendation, "ITU-R BS. 1770-2, Algorithms to measure audio programme loudness and true-peak audio level," 2011.

[25] U. Halekoh, S. Højsgaard, J. Yan *et al.*, "The r package geepack for generalized estimating equations," *J. Stat. Softw.*, vol. 15, no. 2, pp. 1–11, 2006.