INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploring the mutual intelligibility breakdown caused by sculpting speech from a competing speech signal

Martin Cooke[1], *María Luisa García Lecumberri*[2]

[1]Ikerbasque, Bilbao, Spain
[2]University of the Basque Country, Vitoria-Gasteiz, Spain

m.cooke@ikerbasque.org, garcia.lecumberri@ehu.es

Abstract

Passing noise through a binary mask representing speech leads to remarkably intelligible speech. However, if the mask input is a competing speech signal, both the competing speech and the target speech represented by the mask are rendered unintelligible. The current study considers potential explanations for this abrupt breakdown. Competing speech was modified to reduce the influence of properties that may have interacted adversely with those of the target, including speaker, language, F0 and spectral detail. Properties were modified by noise-vocoding, envelope substitution and preservation of temporal modulations. The outcome of a listening experiment indicated that the impact of competing speech is largely due to conflicting formant-scale spectral detail and the absence of sufficient energy in specific temporal epochs, while conflicting F0 plays no role. These findings contribute to a broader understanding of the minimal representational basis that underlies speech perception.

Index Terms: speech perception, intelligibility

1. Introduction

Much has been learnt about speech perception over the last century through the deployment of acoustic signals that bear some relationship to speech, but which have been modified or distorted [1, 2, 3, 4, 5, 6, 7]. Recently, attention has been focused on processes that result in sparse representations of the spectro-temporal energy pattern of speech, following the finding that high levels of intelligibility can be obtained when a significant fraction of time-frequency regions are missing [8, 9, 10, 11].

One extreme hypothesis is that speech might be encodable solely by a spectro-temporal pattern of ones and zeros – a binary time-frequency mask. The binary masking concept was developed primarily to allow robust automatic speech recognisers to operate in the face of missing data [12], but in its extreme form it is the mask itself that constitutes the entire information about the speech signal available to listeners. One study [13] tested the hypothesis that the mask alone is all that is required to support speech perception by processing a speech-shaped noise signal through a binary mask. The synthesis procedure involved passing the noise through a filterbank, switching the noise on and off in each channel corresponding to the locations of ones in the mask, and summing the resulting filtered/gated signals. Using a 32-channel gammatone filterbank, listeners in [13] identified simple Danish sentences with a fixed syntactic structure at levels close to ceiling. A later study [14] replaced the speech-shape noise signal with cafeteria or factory noise, with similar findings. These studies provided an initial indication that a binary mask pattern alone may be sufficient to represent an intelligible speech signal.

However, a subsequent study [15] (see also dataset [16]) demonstrated that achieving a high level of intelligibility from a binary mask depends critically on the signal that is passed through the filterbank. That study referred to the signals processed by the filterbank as *substrates*, and the resulting output as *sculpted* speech, terms we will also adopt in the rest of this paper. Using somewhat more complex sentences than those employed by [13], the study in [15] found that sculpting speech using speech-shaped noise substrates led to listeners recognising around 4 words in every 5 correctly, relative to using the target speech signal as the substrate (note that since mask sparsity generally leads to some errors, using the target speech as the substrate signal represents ceiling performance for a given mask). Listeners identified 3 in 5 words when the substrate was a wide-spectrum music signal. However, the most intriguing finding in [15] was that using a different speech signal as the substrate led to a complete breakdown in intelligibility. These outcomes indicate that when the binary mask is held constant, intelligibility of sculpted speech depends on the nature of the substrate.

Why passing a ‘competing’ speech signal through a binary mask representing the target speech signal should destroy the target so effectively is not known. Possible hypotheses include:

H1. *Linguistic patterning or speaker idiosyncracies.* English sentences were used in [15] for the substrate speech, while the target sentences were Spanish. Conflicting linguistic patterning in the English sentences (e.g. stress-timed rhythm, different vowel space) may have led to insufficient Spanish-like structure available in the substrate to support correct perception of the target sentences. Although the substrate and target speakers were of the same gender, other talker differences (e.g. in F0 range or speech rate) may also have contributed to the unsuitability of the substrate.

H2. *Temporal modulations.* Unlike speech-shaped noise, competing speech has a time-varying temporal envelope, reflecting syllable-related energy variations in the speech signal. In general, these modulations will conflict with those of the target sentence as represented by the mask.

H3. *Formant structure.* The formant structure of the speech substrate will typically be in moment-by-moment conflict with the corresponding structure in the target.

H4. *Fundamental frequency (F0).* Similarly, the F0 contours of the substrate and target may also interact.

H5. *Insufficient spectro-temporal energy.* While the mask indicates where in time and frequency the target speech should be synthesised, there may be too little energy in the speech substrate at those points.

The current study explores these hypotheses. Concerning H1, the same talker speaking the same language was used for

10.21437/Interspeech.2023-63

both the target sentences and the substrate speech. To address H2-H5 the speech substrate was altered to make it less speech-like, and the effect of the consequent modifications on intelligibility was assessed in a listening experiment.

2. Methods

2.1. Sculpting procedure

Generation of sculpted speech requires a mask and a sculpting signal. The first step in producing the mask is to compute, independently, auditory spectrograms for the target speech signal and a speech-shaped noise masker, after normalising the noise to produce a specific signal-to-noise ratio (SNR). Here, 0 dB SNR was used; this value is not critical, as the goal is simply to produce a mask pattern that is broadly representative of the utterance and that leads to high intelligibility when sculpted using a noise substrate. Auditory spectrograms are log-transformed, 10 ms downsampled Hilbert envelopes at the output of a 55-channel gammatone filterbank with centre frequencies in the range 50-7500 Hz (see examples in figs. 1 and 2). A binary mask is then formed by setting each time-frequency cell to 1 if the auditory spectrogram for the speech exceeds that of the masker in the corresponding cell, and 0 otherwise.

To produce sculpted speech, the substrate signal is processed by the same gammatone filterbank. The output of each filter is weighted by a 20 ms triangular filter centred on each time frame where there is a 1 in the mask in the corresponding frequency band. For frames with a 0 in the mask, the filter output is set to zero. The outputs of all frequency bands processed in this manner are then summed. To remove phase artefacts caused by the group delay function of the filterbank, the output signal is reversed, refiltered, then reversed again.

2.2. Experimental conditions

Eight types of substrate were tested in the current study (Tab. 1). Substrates were derived from Spanish sentences produced by the male talker of the Sharvard Corpus [17]. All sentences were sampled at 16 kHz.

Table 1: *Substrates used in the current study.*

Substrate		Hypothesis
SSN	speech-shaped noise	baseline
SPEECH	unmodified competing speech	H1
SMN	speech-modulated noise	H2
NV 5	5-channel noise-vocoded	H3
NV 15	15-channel noise-vocoded	H3, H4
NV 30	30 channel noise-vocoded	H4
ENV SPEECH	envelope from speech	H5
ENV SSN	envelope from noise	H5

The SSN condition replicates the equivalent speech-shaped noise condition of [15] and serves as a baseline for comparison. The SSN condition was generated by passing uniform random noise through a filter representing the steady-state spectrum of the male Sharvard talker.

The SPEECH condition is similar to the competing speech case of [15] but differs in that here, speech comes from same talker/language as the target speech. We hypothesise [H1] that if language and/or talker differences are responsible for intelligibility breakdown, scores in the SPEECH condition will be higher than in the corresponding condition of [15].

To test the hypothesis [H2] that temporal modulations in the competing speech are responsible for the intelligibility breakdown, a speech modulated noise (SMN) condition was generated by replacing the envelope of speech-shaped noise by the short-term envelope of the competing speech. We predict scores in the SMN condition will be similar to those in the SPEECH condition if modulations of the substrate interfere with perception of the target speech.

To elucidate any role for formant and harmonic structure, a number of noise vocoding conditions were constructed. Noise vocoding [2] involves filtering speech into a number of contiguous bands, then replacing the temporal fine structure at the output of each band with a random noise signal. The amount of spectral detail available is governed by the number and placement of frequency bands. Here, in the NV 5 condition, 5 log-spaced bands in the range 50-7500 Hz were used, while the NV 15 and NV 30 conditions employed 15 or 30 bands in the same frequency range. Formant structure is weak in the NV 5 condition, but well represented in the NV 15 and NV 30 conditions; harmonic structure is not present in NV 5, weak in NV 15, but strong in NV 30. These properties are illustrated in Fig. 1. If the formant structure of the substrate interferes with correct reception of the target [H3], we predict lower scores in the NV 15 condition compared to NV 5. Similar considerations for harmonic structure [H4] result from a comparison of the NV 15 and NV 30 substrates.

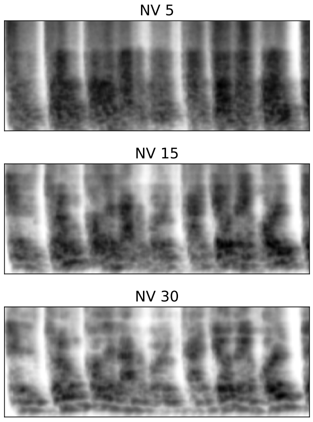

Figure 1: *Auditory spectrograms for the 3 noise-vocoded conditions. The sentence is 'El tronco del árbol cayó de golpe en la calle ['The tree trunk fell suddenly in the street'].*

The final two conditions address the hypothesis [H5] that there is insufficient energy in the substrate in some spectro-temporal regions of the mask i.e., there is no material to sculpt out the relevant acoustic features. For these conditions the sculpting procedure described in Sec. 2.1 was modified to apply a weighting to each spectro-temporal region. In the ENV SSN condition, the sculpting signal was SPEECH, but the energy came from the speech-shaped noise signal in that region; conversely, the ENV SPEECH case was a speech-shaped noise signal with energies replaced by those of the SPEECH signal. In essence, the temporal fine structure of the speech substrate is retained in the ENV SSN condition, while the envelope is retained in the ENV SPEECH case (see [18] for a review of envelope and fine structure cues in speech perception). The ENV SSN condition ensures that there is some energy in all spectro-temporal regions where the mask is present; consequently, a finding that intelligibility for ENV SSN improves significantly compared to the SPEECH substrate would point to the importance of this factor. Likewise, any significant reduction in intelligibility compared to SSN in the ENV SPEECH condition would reinforce the importance of having sufficient energy in the masked regions.

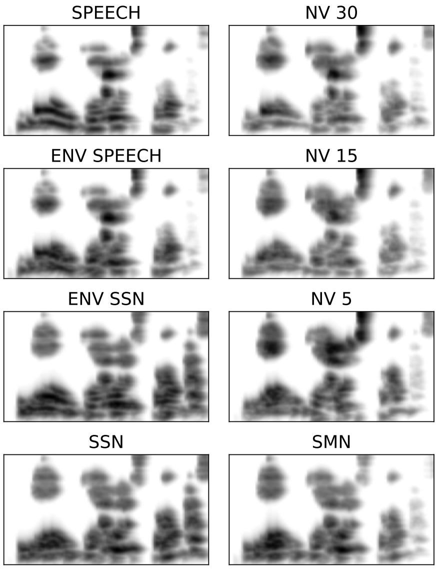

Figure 2: *Example auditory spectrograms for speech sculpted using the substrates of the current study. The sentence fragment is 'limpiar las botas' ['clean the boots'].*

Fig. 2 shows spectrographic examples of each form of sculpted speech. Spectrograms appear broadly similar because the mask involved in their production was identical in each case. However, there are subtle yet discernible differences due to the differing substrates. For example, the SPEECH and ENV SPEECH conditions have a similar spectro-temporal energy pattern by design, but the harmonic structure in the SPEECH case is stronger since the ENV SPEECH substrate is derived from a noise signal. The converse is the case for the SSN and ENV SSN pair. The SMN case shows a different energy distribution across time than its SSN counterpart, reflecting the temporal energy distribution of the SPEECH substrate. The differences are clearest in the final part of the spectrogram where a weaker signal is evident in the SMN case. The harmonic pattern in the noise-vocoded cases approximates that of the SPEECH substrate as the number of channels increases from 5 to 30.

2.3. Stimuli

Stimuli were constructed from 240 utterances (sentences 461-700) of the Sharvard Corpus in each of the 8 sculpted speech conditions. Sentences were normalised to the same root-mean-square level prior to presentation.

2.4. Participants

A total of 36 participants (33 female; mean age 19.1, range 18–24, st. dev. 1.25 years) took part in the listening experiment. All were students in their second year of study at the University of the Basque Country (Alava Campus, Spain). No listener reported hearing impairment, and all had Spanish or Spanish and Basque as their first language(s). All listeners were paid for participation.

2.5. Procedure

Listeners responded to 30 sentences in each of 8 conditions. No sentences were repeated. Condition order was balanced across listeners following a Latin square design. Sentence order in each condition block was randomised. The experiment took place using an online platform described in [19], implemented using Flask [20] and Howler [21]. That study validated the online approach via a series of experimental replications of traditional lab experiments, one of which was a replication of the sculpted speech study [15]. All listeners were familiar with the online platform as they had been using it extensively for classes in English Phonetics prior to the experiment. Listeners heard 5 practice sentences chosen from a subset of experimental conditions prior to the main experiment. Listeners completed the experiment one block at a time, and were able to take a break between blocks.

2.6. Postprocessing

As part of the design of the Sharvard Corpus, five keywords in each sentence were preselected for scoring purposes. Scores were based on the number of such keywords identified correctly, producing an integer value in the range 0-5. Prior to scoring, participants' responses were subjected to the following set of normalising processes: (i) vowel stresses were removed, since participants were told that indicating stress was optional; (ii) all non-alphanumeric characters were removed; (iii) all numbers represented with characters in the range 0-9 were replaced by lexical equivalents (e.g. 10 was replaced by 'diez'); (iv) any extraneous digits were then removed; (v) common typos/orthographic errors from a list of 80 such errors identified in prior experiments were replaced (e.g. the non-word 'silvar' was replaced by its homophone 'silbar' ['to whistle']).

Outlier analysis (based on identifying values more than 1.5 times the inter-quartile range below the first quartile or above the third quartile) performed on mean per-subject scores across all conditions led to the removal of data from one participant (mean 5% versus cohort mean of 36% keywords correct). Subsequent analysis was based on the remaining 35 participants.

3. Results

Fig. 3 depicts the key outcomes of the listening experiment. Scores in the SPEECH and SSN conditions were compared to those obtained in the online replication [19] of the earlier sculpted speech study [15] in which listeners identified 3.3% and 69.0% of keywords in the SPEECH and SSN conditions respectively. Due to lack of normality, Mann-Whitney rank-sum tests were used to compare scores for the SSN and SPEECH substrates in the two studies. These tests indicated that while the SSN conditions were not statistically-different [$U = 863, p = 0.71$], the SPEECH substrates were [$U = 370, p < .001$]. Since the SSN substrate conditions were identical in the two studies, the similarity in keyword scores provides some confidence that the samples in the two studies are comparable. The difference between the studies for the SPEECH condition may stem from the fact that the competing speech substrates differed in talker and language (Hypothesis H1; see Discussion).

A generalised linear mixed-effects model, implemented using the `glmer` function of the `lme4` package [22] in R [23] was used to predict the proportion of keywords recognised correctly. This model had SUBSTRATE as a fixed effect, with random intercepts and per-substrate slopes for each participant, and random intercepts for each sentence. Model comparisons using the

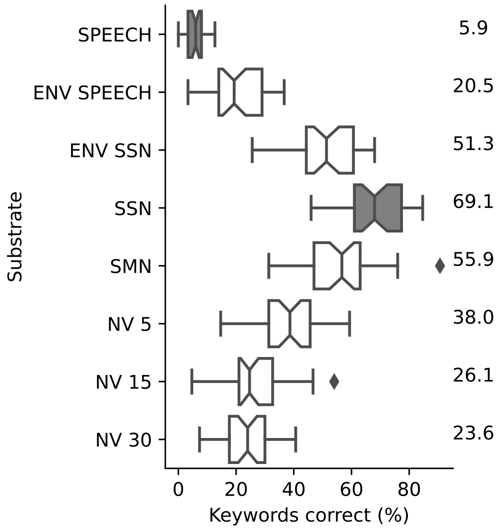

Figure 3: *Keyword scores for target sentences sculpted from the eight substrates of the experiment. Notches in the boxplots indicate 95% confidence intervals. Grey boxes denote baseline conditions. Mean scores are indicated in the right hand column.*

`Anova` function from the *car* package [24] confirmed a clear main effect of substrate [$\chi^2(7) = 1453, p < .001$] that is evident in Fig. 3. Post-hoc analyses made use of the `emmeans` function from the package of the same name [25], using Tukey HSD corrections for multiple comparisons. This analysis indicated that keyword recognition rates were different for all but three pairs of substrates at the $p < .001$ level, while two pairs – ENV SSN vs. SMN and ENV SPEECH vs. NV 30 – differed at the $p < .05$ level. The only pair that did not differ statistically were two of the noise-vocoded conditions, viz. NV 15 and NV 30 [$p = .89$].

4. Discussion

In terms of the hypotheses, the above findings suggest the following:

H1. *Linguistic patterning or speaker idiosyncracies.* The fact that scores in the SPEECH condition were higher in the current study, which used the same talker and language as the target speech material, than in [19] which used a different language and talker, suggests that language and/or talker differences (which cannot be pulled apart here) may be responsible for part of the intelligibility breakdown. However, the difference of less than 3 percentage points in the SPEECH condition is modest.

H2. *Temporal modulations.* Processing the speech substrate to leave just its temporal modulation envelope intact (SMN) is able to restore most, but not all, of the intelligibility of the target sentence, arguing against such modulations being the main factor behind intelligibility breakdown in the SPEECH condition.

H3. *Formant structure.* Intelligibility dropped by 12 points as the number of bands in the noise vocoder increased from 5 to 15, supporting the hypothesis that information in the formant structure of the substrate interferes with target processing. Indeed, the 18 point reduction in keyword scores from the SMN condition to the NV 5 condition also supports this conjecture.

H4. *Fundamental frequency (F0).* The lack of a statistically-significant score difference between the NV 15 and NV 30 substrates indicates that harmonic structure differences between substrate and target play little if any role in the intelligibility breakdown of the latter.

H5. *Insufficient spectro-temporal energy.* When the spectro-temporal energy balance of the competing speech was altered to ensure sufficient energy in all parts of the mask (ENV SSN), keyword scores improved by around 45 points, providing a strong indication that lack of energy in the necessary places contributed to intelligibility breakdown, a finding reinforced by the drop of 49 points when a speech-shaped noise envelope was replaced by that of the speech (ENV SPEECH condition).

One limitation of the current study is that acoustic properties related to one or other of the hypotheses cannot always be modified in a way that is completely independent of changes to other acoustic properties. For example, the temporal energy modulations inherent in the SMN condition also impact the detailed spectro-temporal energy balance that is the subject of the ENV SPEECH and ENV SSN conditions. For this reason any conclusions about why a speech substrate has a negative impact when used to sculpt a different target speech signal must be regarded as tentative. Notwithstanding such limitations, when taken together the outcomes of the current study suggest that two factors – loss of spectral detail, and temporal energy redistribution – play a role in the impact that a speech substrate has on the sculpted target signal.

The role played by the first of these factors i.e. the amount of spectral detail available in the substrate, is evidenced by the recovery of intelligibility as spectral detail is progressively removed, from fully-present in the SPEECH condition, to partially-present in the NV 30, NV 15 and NV 5 cases, to fully lost in the SMN condition. While the noise-vocoding and SMN progression effectively broadens or fills in the *spectral* distribution of the substrate, it maintains the original *temporal* distribution of energy in the substrate. This property may be responsible for the finding that even in the SMN condition there is not full recovery of intelligibility (scores in the SMN condition fall short of those in the SSN condition by 13 points). Consequently, it appears that in addition to a blurring of spectral detail it is necessary to ensure that there is sufficient substrate energy in the appropriate temporal zones to 'activate' the target speech as represented by the mask, as accomplished by the temporal redistribution of substrate energy that is a consequence of conditions ENV SPEECH and ENV SSN.

5. Conclusions

When a competing speech signal is passed through a binary mask representing a different speech target, the intelligibility of both competing and target speech is largely destroyed. The current study suggests that intelligibility breakdown stems from a combination of conflicting spectral detail and insufficient energy at critical temporal epochs in the competing speech.

6. Acknowledgements

This work was supported by grant 110321NICK PES21/90 awarded to the first author. The authors would like to thank Dr O. Scharenborg for discussion that led to testing hypothesis H5.

7. References

[1] G. A. Miller and J. C. R. Licklider, “The intelligibility of interrupted speech,” *J. Acoust. Soc. Am.*, vol. 22, pp. 167–173, 1950.

[2] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science*, vol. 270, pp. 303–304, 1995.

[3] Z. M. Smith, B. Delgutte, and A. J. Oxenham, “Chimaeric sounds reveal dichotomies in auditory perception,” *Nature*, vol. 416, pp. 87–90, 2002.

[4] K. Kasturi, P. C. Loizou, M. Dorman, and T. Spahr, “The intelligibility of speech with ‘holes’ in the spectrum,” *J. Acoust. Soc. Am.*, vol. 112, pp. 1102–1111, 2002.

[5] P. A. Howard-Jones and S. Rosen, “Uncomodulated glimpsing in ‘checkerboard’ noise,” *J. Acoust. Soc. Am.*, vol. 93, pp. 2915–2922, 1993.

[6] R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker, “Spectral redundancy: intelligibility of sentences heard through narrow spectral slits,” *Perception and Psychophysics*, vol. 57, pp. 175–182, 1995.

[7] R. P. Lippmann, “Accurate consonant perception without mid-frequency speech energy,” *IEEE Trans. Speech and Audio Proc.*, vol. 4, pp. 66–69, 1996.

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

[9] V. Isnard, M. Taffou, I. Viaud-Delmon, and C. Suied, “Auditory sketches: Very sparse representations of sounds are still recognizable,” *PLoS ONE*, vol. 11, p. e0150313, 2016.

[10] G. Kidd, T. M. Streeter, A. Ihlefeld, R. K. Maddox, and C. R. Mason, “The intelligibility of pointillistic speech,” *J. Acoust. Soc. Am.*, vol. 126, pp. EL196–EL201, 2009.

[11] M. I. Mandel, S. E. Yoho, and E. W. Healy, “Measuring time-frequency importance functions of speech with bubble noise,” *J. Acoust. Soc. Am.*, vol. 140, pp. 2542–2553, 2016.

[12] M. Cooke, P. Green, and M. Crawford, “Handling missing data in speech recognition,” in *Proc. 3rd International Conference on Spoken Language Processing (ICSLP 1994)*, 1994, pp. 1555–1558.

[13] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, “Speech perception of noise with binary gains,” *J. Acoust. Soc. Am.*, vol. 124, pp. 2303–2307, 2008.

[14] U. Kjems, J. B. Boldt, M. S. Pedersen, T. Lunner, and D. Wang, “Role of mask pattern in intelligibility of ideal binary-masked noisy speech,” *J. Acoust. Soc. Am.*, vol. 126, pp. 1415–1426, 2009.

[15] M. Cooke and M. L. García Lecumberri, “Sculpting speech from noise, music, and other sources,” *J. Acoust. Soc. Am. Express Letters*, vol. 148, pp. EL20–EL26, 2020.

[16] M. Cooke, “Sculpted speech,” Apr. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3755383

[17] V. Aubanel, M. L. Garcia Lecumberri, and M. Cooke, “The Sharvard corpus: A phonemically-balanced Spanish sentence resource for audiology,” *Int. J. Audiology*, vol. 53, pp. 633–638, 2014.

[18] C. Lorenzi and B. C. J. Moore, “Role of temporal envelope and fine structure cues in speech perception: A review,” in *Auditory Signal Processing in Hearing-Impaired Listeners*, T. Dau, J. M. Buchholz, T. M. Harteand, and T. U. Christiansen, Eds. Centertryk A/S, 2008, pp. 263–272.

[19] M. Cooke and M. L. Garcia Lecumberri, “How reliable are online speech intelligibility studies with known listener cohorts?” *J. Acoust. Soc. Am.*, vol. 150, pp. 1390–1401, 2021.

[20] Flask, “Flask v1.1.2,” https://palletsprojects.com/p/flask/, 2021.

[21] Howler, “Howler v2.2.1,” https://howlerjs.com, 2021.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[24] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[25] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2023, R package version 1.5.5-1. [Online]. Available: https://CRAN.R-project.org/package=emmeans