Clear Speech Improves Word Recognition, but may Increase Listening Effort

Kirsten Meemann & Rajka Smiljanić

The University of Texas at Austin
kirsten.meemann@utexas.edu, rajka@austin.utexas.edu

ABSTRACT

Listener-oriented clear speech (CS) improves word recognition in noise and auditory memory. Using a dual-task paradigm, the current study examined whether this processing benefit is related to the effect of intelligibility-enhancing CS on listening effort in the presence of energetic and informational maskers. Native English listeners repeated sentences produced in a conversational and clear speaking style and mixed with either two-talker babble or speech-shaped noise. They simultaneously performed a visual task on the computer screen and their response times were measured.

The results showed that hearing clear speech sentences increased response times on the visual task in the presence of an energetic masker. The response times on the visual task in informational masker were not different between the two speaking styles. This suggests that listeners direct their attentional resources to the more salient hyperarticulated clear speech and that a decrease in listening effort may not underlie the CS intelligibility benefit.

Keywords: listening effort, clear speech, speech perception, dual task

1. INTRODUCTION

When an acoustic signal is degraded or inaccessible, accurate sound and word recognition becomes considerably more challenging resulting in a decrease in intelligibility. Even when word recognition is high, processing speech under challenging listening conditions can increase cognitive load and listening effort [1], [2]. The task demands, such as presence of noise or unfamiliar accent, can increase listening effort, or "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task, with listening effort applying more specifically when tasks involve listening" (Framework for Effortful Listening (FUEL), [1, p. 5S]; Ease of Language Understanding (ELU) model, [3]). According to these models, listeners have limited cognitive capacity and they draw on the available resources accordingly when performing several tasks simultaneously [4]. The more effort listeners expend to understand degraded speech, the fewer resources they have left to complete simultaneous tasks or remember spoken information [1], [5]. Effortful listening can lead to decreased motivation and task disengagement [6] and it can have long-term health consequences such as increased stress, fatigue, anxiety and social withdrawal [7].

One way in which speakers can increase intelligibility is through a listener-oriented hyperarticulated clear speech (CS) [8]. In addition to enhancing word recognition, CS aids various linguistic and cognitive processes, such as speech segmentation and auditory memory [9]. Using pupillometry, Borghini & Hazan [10] recently showed that the presence of the enhanced acoustic-phonetic cues present in CS improved intelligibility and significantly reduced listening effort. However, the relationship between intelligibility and effort is non-monotonic [11]. Furthermore, different measures of listening effort may assess different aspects of listening effort (e.g., [12]). The FUEL indicates that behavioral measures may be most sensitive to listening demands, while physiological measures may be more sensitive to both listening demands and motivation [1]. The current study examined how variation in intelligibility through speaking style modifications and different masker types affects listening effort using a dual-task paradigm. Previous work showed that maskers that vary in the degree of informational (IM) and energetic masking (EM) impact the intelligibility benefit listeners obtain from CS [13]. Here, we focus on investigating whether listener-oriented speaking style modifications affect listening effort differently in EM and IM ([14] for a review).

Native English listeners performed a word recognition task in speech-shaped noise (SSN) and 2-talker (2T) babble and a visual response task first separately and then simultaneously. Self-report measures have additionally been used to quantify subjective listening effort (e.g., [15]). We hypothesized that 1) processing easier-to-understand CS will reduce listening effort, 2) listening effort will be greater in 2T babble, where the speech signal is masked by both EM and IM, than in SSN, which imposes only EM, and 3) the CS benefit on effort will be smaller in 2T babble, where word recognition accuracy is expected to be higher than in SSN [13]. These hypotheses are consistent with the FUEL and ELU models in that more intelligible CS will reduce

cognitive load leaving additional processing resources for the secondary visual task. The aim was to provide a more nuanced understanding of the CS intelligibility benefit and its association with listening effort.

2. METHODS

2.1. Participants

Fifty-one native English listeners (average age: 21.0 years, SD: 4 years, range: 18-34 years) participated in the study. All participants were recruited at the University of Texas at Austin. They provided written consent to use their responses anonymously. All listeners passed hearing screening bilaterally at 25 dB HL at 500, 1000, 2000, and 4000 Hz. They received class credit or a small monetary compensation for participation.

2.2. Materials

Primary Task

One 21-year-old female native speaker of American English read 60 semantically meaningful sentences (e.g., *Mice like to eat cheese*; [16]) first in conversational (CO) and then in clear (CS) speech. For CO speech, the talker was told to speak in a casual manner as if she was talking to a friend or family member. For CS, the talker was instructed to speak as if she was communicating with someone who has a low proficiency in English and who has difficulty following them conversationally (following [17]). All sentences were normalized for root-mean-square amplitude and mixed with competing speech and with SSN at -5dB signal-to-noise ratio (SNR). The competing speech masker consisted of 2-talker (2T) babble (2 female talkers) [18]. The speech-shaped noise masker had been generated by shaping white noise to match the long-term average spectrum of 6-talker (6T) babble (3 male, 3 female talkers; [18]). 12 additional sentences produced in CO and CS were used as practice stimuli for the word recognition task. They were mixed with 2T babble and SSN at -5dB SNR.

Secondary Task

Visual stimuli for the secondary non-linguistic visual task consisted of two square boxes (approximately 5 cm across), one on the left and one on the right side of the screen. At each trial, a left- or a right-pointing arrow appeared in one of the two boxes.

Subjective Effort

We adapted four questions from the original NASA-TLX [21] related to exerted effort, mental demand, perceived performance (error rate), and frustration to assess subjective listening effort [12].

2.3. Procedure

The experiment began with the primary word recognition task alone followed by the secondary visual stimulus response task alone presented in SuperLab 5 (Cedrus). The word recognition single task consisted of 12 sentences, alternating between speaking style and masker combinations, CO+2T, CO+SSN, CS+2T, CS+SSN, three times. The order of presentation in the single task was fixed. Sentences were presented at set intervals of 4000 ms. For each trial, the noise started 100 ms before the target sentence and stopped 100 ms after the target ended. Participants were instructed to listen to the sentences, one at a time, and to repeat each target sentence verbally. Oral responses were recorded and scored off-line afterwards.

For the secondary visual task, a left- or right-facing arrow appeared in the left- or the right-hand side box (at 500-2000 ms intervals). In the visual single-task condition, 40 trials were presented. Participants were instructed to press ‘z’ if a left-pointing arrow appeared in either of the boxes on the screen, and to press ‘m’ if a right-pointing arrow appeared on the screen. Accuracy and response times were recorded for each trial.

In the dual task following the two single tasks, participants heard and verbally repeated target sentences while simultaneously completing the visual task described above. They were told to complete both tasks to the best of their ability, but to prioritize the word recognition task [19]. In the dual task, participants were presented with four 15-sentence blocks of speaking style (CO and CS) and masker (2T and SSN) combinations, one combination per block. With three keywords per sentence, there were 45 target words for each of the four speaking style and masker combinations. All listeners heard sentences produced in both speaking styles and presented with both types of masking noise. The speaking style and noise masker combination was counterbalanced across blocks such that each listener heard a different style and masker combination order. The order in which the blocks of sentences were presented was pseudorandomized. Listeners never heard the same sentence twice. In the primary word recognition task, sentences were presented at fixed intervals of 4000 ms, and in the concurrent secondary visual task, arrows remained on the screen until participants responded or after 2500 ms had elapsed. The

presentation of visual and auditory stimuli in the dual-task condition did not systematically coincide. Response times to correct responses in the secondary visual task were collected to quantify listening effort.

After each of the four dual-task blocks, subjective ratings of listening effort were collected. The order of the five questions after each block was consistent. For each question, participants were told to click on a blank square on an unnumbered 21-point scale that corresponded to their experience during the most recent dual-task block. The full experimental session took about 30 minutes to complete.

3. ANALYSES

To assess effort, response times (RTs) for answers to the visual task were analyzed for speaking style and masker combinations. Only RTs for correct responses (96.57% of stimuli) were included in the analysis [15]. RTs slower than 2500 ms and faster than 200 ms were excluded from the data analysis.

Data were analyzed with linear mixed effects regression using the *lme4* package in R. Model comparisons were done via likelihood ratio tests, where models differed only in fixed effects. All models included by-participant and by-item (word) random intercepts. Significant results were further analyzed with estimated marginal means (*emmeans* package in R [20]).

For intelligibility, each sentence used in the primary task contained four target words which were scored as correct (1) or incorrect (0). Data were analyzed with logistic mixed-effects regression using the *glmer* function in R [21]. All models had word recognition score as a dependent variable with a binomial link function, and random intercepts for listeners and words. Likelihood ratio tests were conducted to determine the models with the best fit. Further post-hoc analyses were done via estimated marginal means tests.

Subjective listening effort data were analyzed via mixed effects modeling (*lmer*). Self-report scores for the effect of speaking style and noise masker were analyzed on only the effort question. All models included by-subject random intercepts.

4. RESULTS

Response Times

Descriptive statistics for response time data in the dual task are shown in Figure 1. Results revealed that RTs to visual stimuli were significantly slower in the dual task condition (mean = 652.8 ms; SD = 296.2 ms;) than when performing the visual task alone (mean = 566.1 ms; SD = 186.2 ms) (z = 13.39, p < .001).

Post-hoc analyses of a two-way interaction between speaking style and masker ($\chi^2(1) = 9.14$, p < .01) showed that RTs were significantly shorter for sentences produced in CO than in CS when presented in SSN (z = -4.71, p < .001). There was no difference in RTs between speaking styles when presented in 2T babble (z = -0.41, p = .68). Lastly, listeners responded significantly more slowly to CO sentences presented in 2T babble than in SSN (z = 3.65, p < .001), but not CS sentences (z = -0.57, p = .57).

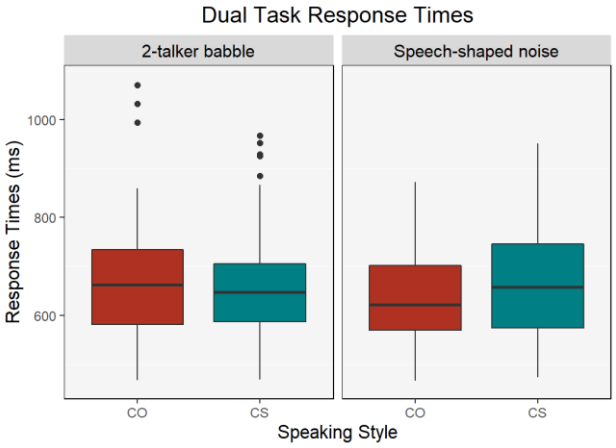

Figure 1. Listeners' response times in Conversational speech (CO) and Clear speech (CS) in 2-talker (2T) babble (left panel) in speech-shaped noise (right panel).

Speech Intelligibility

A two-way interaction between speaking style and noise masker ($\chi^2(1) = 31.50$, p < .001) revealed that intelligibility was significantly better for CS than CO sentences in both masker conditions (SSN: z = 8.37, p < .001; 2T babble: z = 2.62, p < .001). The odds of successful word recognition were 90% higher for CS than CO speech in SSN (OR = 1.90, 95% CI, 1.64-2.21, and 22% higher in 2T babble (OR = 1.22, 95% CI, 1.05-1.42). Additionally, post-hoc comparisons showed that listeners correctly identified significantly more words when target sentences were presented in SSN than in 2T babble, but only when the sentences were produced in CS (z = 4.45, p < .001). Specifically, listeners were 41% more likely to recognize CS-produced words in SSN than in 2T babble (OR = 1.41, 95% CI, 1.21-1.64).

Subjective Effort

To evaluate subjective effort, we examined participants' ratings for how effortful they found the task. A model with main effects of speaking style and masker yielded the best fit ($\chi^2(1) = 7.70$, p < .01). Results showed that listeners perceived CO sentences as more difficult than CS sentences ($t(79) = 2.79$, p < .01) and sentences presented in 2T babble as more

effortful than sentences presented in SSN ($t(79) = 3.67$, $p < .001$).

5. DISCUSSION

In contrast to our hypothesis, the results of the dual-task experiment showed that CS did not reduce listening effort despite increasing intelligibility. Listening effort, as measured by the RTs in response to the visual task, did not differ between CS and CO speaking styles in 2T babble, and it even increased in CS compared to CO when speech was masked by SSN. Additionally, listening effort was increased in 2T babble compared to SSN when sentences were produced in CO, but there was no difference between masker types when speech was produced in CS.

The results suggest that the well-established CS intelligibility benefit does not arise from a decrease in listening effort. This finding is in contrast with Borghini and Hazan [10] who showed smaller pupil dilation for CS than CO, suggesting a decrease in listening effort. Behavioral measures (e.g., response times) are hypothesized to be most sensitive to listening demands, while physiological measures, such as pupillometry, are sensitive to both listening demands and motivation (cf., [1]). This is also in line with other recent studies that found a lack of correlation between different measures of listening effort [22], [23].

Dual-task paradigms assume limited cognitive capacity, such that the difficulty of the primary task can determine how many resources are left for simultaneously performing a secondary task: The more difficult the primary task, the more cognitive resources are directed to completing that task and the fewer resources are left to perform the secondary task, leading to a performance decrease in the secondary task [24]. Consequently, this suggests that RTs in a dual task can reflect listeners' direction of attention, regardless of instructions to prioritize the primary task.

Our results suggest that it is not only the difficult primary task that may draw listeners' attention. Listeners seem to direct their attentional resources to the more acoustically salient and easier-to-process clear speech compared to conversational speech (cf., [25]). This finding is in line with Brown and Strand [15] who reported increased use of cognitive resources for processing speech with visual cues than without, despite better word recognition in the audiovisual modality. CS may thus be more intelligible not because of lower processing cost but because of the allocation of cognitive resources. Under this view, listeners are automatically drawn to the exaggerated acoustic-phonetic cues of the listener-oriented clear speech, and to the presence of visual cues, resulting in enhanced intelligibility via the engagement of additional resources.

Listeners' subjective effort ratings seem to stand in contrast to the dual-task findings. Listeners perceived CS and SSN as less effortful than CO and 2T babble. However, the combined findings show that objectively measured listening effort was larger in the speaking style and masker combination that listeners subjectively perceived as easier. This further supports the notion that different measures tap into different kinds of effort, such as the effort listeners exert during a task versus the emotional response to their performance and exerted effort.

Regarding masker type, word recognition accuracy was better in the purely energetic masker (SSN) than in the predominantly informational masker (2T babble) only for clearly produced speech. Conversely, RTs did not differ between the two masker types in CS, but they were longer in 2T babble than in SSN for CO speech. This indicates that listeners spent more effort to understand conversationally produced speech masked by 2T babble compared to SSN to achieve similar recognition accuracy. In terms of subjective effort, listeners rated 2T babble as more effortful compared to sentences presented in SSN. Taken together, the objective and subjective measures of effort show that informational masking (IM) from 2T babble is more effortful than energetic masking from SSN. This is consistent with previous studies that reported an association between IM and increased objectively measured listening effort as indicated by pupillometry [26] and increased subjective listening effort [27]. The finding that objectively measured effort and subjective effort aligned here suggests that suppressing irrelevant lexical intrusion from IM is more effortful. It seems that listeners correctly assessed that they were working harder in order to attend to the target speech. This further suggests that the subjective measures of effort may not be equally sensitive to the challenges presented by speaking style variation vs. acoustic degradation through signal masking. The link between the subjective ratings and various types of listening challenges warrants further exploration.

In summary, this study provides evidence that the well-established CS processing benefits may not be arising from a decrease in listening effort. Instead, listeners seem to direct their cognitive resources to more attention-grabbing, salient speech. The results also showed that speech degraded predominantly by informational masking increased listening effort. This research has the potential of improving the quality of treatments for hearing-impaired individuals by relying on insights from both intelligibility and listening effort.

REFERENCES

[1] M. K. Pichora-Fuller *et al.*, "Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)," *Ear Hear.*, vol. 37, pp. 5S-27S, 2016.

[2] A. A. Zekveld, S. E. Kramer, and J. M. Festen, "Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response.," *Ear Hear.*, vol. 32, no. 4, pp. 498–510, 2011.

[3] J. Rönnberg *et al.*, "The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances," *Front. Syst. Neurosci.*, vol. 7, no. July, pp. 1–17, 2013.

[4] D. Kahneman, *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall, 1973.

[5] P. M. A. Rabbitt, "Channel-capacity, intelligibility and immediate memory," *Q. J. Exp. Psychol.*, vol. 20, no. 3, pp. 241–248, 1968.

[6] M. Richter, "The moderating effect of success importance on the relationship between listening demand and listening effort," *Ear Hear.*, vol. 37, pp. 111S-117S, 2016.

[7] B. W. Y. Hornsby, G. Naylor, and F. H. Bess, "A taxonomy of fatigue concepts and their relation to hearing loss," *Ear Hear.*, vol. 37, pp. 136S-144S, 2016.

[8] R. Smiljanić, "Clear Speech Perception: Linguistic and Cognitive Benefits," in *The Handbook of Speech Perception*, 2nd ed., D. B. Pisoni, R. E. Remez, L. C. Nygaard, and J. S. Pardo, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc, 2021, pp. 177–205.

[9] S. Keerstock and R. Smiljanić, "Clear speech improves listeners' recall," *J. Acoust. Soc. Am.*, vol. 146, no. 6, pp. 4604–4610, 2019.

[10] G. Borghini and V. Hazan, "Effects of acoustic and semantic cues on listening effort during native and non-native speech perception," *J. Acoust. Soc. Am.*, vol. 147, no. 6, pp. 3783–3794, 2020.

[11] M. B. Winn, D. Wendt, T. Koelewijn, and S. E. Kuchinsky, "Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started," *Trends Hear.*, vol. 22, pp. 1–32, 2018.

[12] J. F. Strand, V. A. Brown, M. B. Merchant, H. E. Brown, and J. Smith, "Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures," *J. Speech Lang. Hear. Res.*, vol. 61, no. 6, pp. 1463–1486, 2018.

[13] K. Meemann and R. Smiljanić, "Intelligibility of Noise-Adapted and Clear Speech in Energetic and Informational Maskers for Native and Nonnative Listeners," *J. Speech, Lang. Hear. Res.*, vol. 65, no. 4, pp. 1263–1281, 2022.

[14] J.-P. Gagné, J. Besser, and U. Lemke, "Behavioral assessment of listening effort using a dual-task paradigm: A review," *Trends Hear.*, vol. 21, pp. 1–25, 2017.

[15] V. A. Brown and J. F. Strand, "About Face: Seeing the Talker Improves Spoken Word Recognition but Increases Listening Effort," *J. Cogn.*, vol. 2, no. 1, p. 44, 2019.

[16] M. Fallon, S. E. Trehub, and B. A. Schneider, "Children's use of semantic cues in degraded listening environments," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2242–2249, 2002.

[17] A. R. Bradlow and J. A. Alexander, "Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners.," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2339–2349, 2007.

[18] K. J. Van Engen and A. R. Bradlow, "Sentence recognition in native- and foreign-language multi-talker background noise," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 519–526, 2007.

[19] V. A. Brown and J. F. Strand, "Noise increases listening effort in normal-hearing young adults, regardless of working memory capacity," *Lang. Cogn. Neurosci.*, vol. 34, no. 5, pp. 628–640, 2019.

[20] R. V. Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." R package version 1.3.1, 2018.

[21] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[22] C. Visentin, C. Valzolgher, M. Pellegatti, P. Potente, F. Pavani, and N. Prodi, "A comparison of simultaneously-obtained measures of listening effort: pupil dilation, verbal response time and self-rating," *Int. J. Audiol.*, vol. 61, no. 7, pp. 561–573, 2021.

[23] A. L. Francis, T. Bent, J. Schumaker, J. Love, and N. Silbert, "Listener characteristics differentially affect self-reported and physiological measures of effort associated with two challenging listening conditions," *Attention, Perception, Psychophys.*, vol. 83, no. 4, pp. 1818–1841, 2021.

[24] S. Fraser, J.-P. Gagné, M. Alepins, and P. Dubois, "Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues.," *J. speech, Lang. Hear. Res.*, vol. 53, no. 1, pp. 18–33, 2010.

[25] V. Boswijk, H. Loerts, and N. H. Hilton, "Salience is in the eye of the beholder: Increased pupil size reflects acoustically salient variables," *Ampersand*, vol. 7, p. 100061, 2020.

[26] S. Villard, T. Perrachione, S.-J. Lim, A. Alam, and G. Kidd, "Listening effort elicited by energetic versus informational masking," *Proc. Meet. Acoust.*, vol. 45, pp. 1–12, 2022.

[27] J. Rennies, V. Best, E. Roverud, and G. Kidd, "Energetic and Informational Components of Speech-on-Speech Masking in Binaural Speech Intelligibility and Perceived Listening Effort," *Trends Hear.*, vol. 23, pp. 1–21, 2019.