DOI: 10.61782/fa.2025.0544

FORUM ACUSTICUM EURONOISE 2025

MASKER TYPE AND SPEAKER VOICE CUES DETERMINE MASKING OF SPEECH BY A SINGLE TALKER

Laura Rachman[1,2*] **Deniz Başkent**[1]
[1] Department of Otorhinolaryngology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[2] Pento Speech and Hearing Centers, Apeldoorn, The Netherlands

ABSTRACT

Speech-on-speech perception relies on perceptual and cognitive mechanisms. Differences in voice cues, such as mean fundamental frequency (F0) and vocal-tract length (VTL), and intensity levels (target-to-masker ratio, TMR), help discriminate between target and masker speech and facilitate speech-on-speech perception. Here, using Dutch Child-friendly Coordinate Response Measure (CCRM) sentences with numbers and colors as keywords, we show that masking patterns from a single-talker masker vary depending on the masker type. We used scrambled sentence maskers and intact full sentence maskers, where keywords from the target and masker speech largely overlapped. In addition, the effect of voice cue differences between target and masker speech was assessed for the full sentence condition only, by manipulating F0 and VTL of the masker speech. The scrambled sentence masker showed a weak monotonic masking effect as a function of varying TMR. The full sentence masker showed a non-monotonic masking effect, which was affected by TMR and voice differences. Hence, masking from single-talker speech highly depends on acoustic factors and overlapping keywords. The different patterns of speech perception scores as a function of masker type and TMR has implications on methods commonly used to quantify speech-on-speech perception, such as when determining speech reception thresholds using adaptive procedures.

[*] ***Corresponding author***: l.rachman@rug.nl.

Keywords: *speech-on-speech perception, single-talker speech masker, scrambled sentence masker, full sentence masker, voice cues.*

1. INTRODUCTION

Speech perception in the presence of competing speech (speech-on-speech perception) relies on both perceptual and cognitive mechanisms, and hence can be sensitive to development, aging, and linguistic factors. Studies have shown that even in listeners with normal hearing, speech-on-speech perception improves in children during school-age years [1] and becomes more challenging again in older adulthood [2]. Research paradigms targeting speech-on-speech perception range from multi-talker speech maskers [3] to variations of single-talker speech maskers, for example with competing speakers of the same or different genders [4], with native or non-native speech content [5], familiar or unfamiliar speakers [6], or time-reversed [7] or scrambled speech [8]. In all these paradigms, a common area of interest is listeners' ability to segregate target and masker speech based on the differences in the intensity level cues between the two speech streams, expressed as target-to-masker ratio (TMR) in decibels (dB). In addition, the use of single-talker speech maskers is appropriate for the assessment of listeners' ability to segregate target and masker speech based on voice cues, such as mean fundamental frequency (F0) and vocal-tract length (VTL). Yet, different types of single-talker speech maskers (e.g., full sentences vs. scrambled speech) may still have different effects on how listeners make use of intensity level and voice cues, making a direct comparison between different paradigms difficult.

In this study, we used the Dutch Child-friendly Coordinate Response Measure (CCRM), a closed-set test, in two experiments to characterize the masking effects of two types of masker speech: scrambled sentences and full sentences. We assessed the effect of various TMRs using scrambled sentence maskers without voice cue differences between target and masker speech. This version of the CCRM paradigm has been used in previous studies on childhood development of speech-on-speech perception [1] and on the effects of hearing loss [8], [9] or musical training [10]. In addition, we assessed the effect of differing TMRs in combination with F0 and VTL differences between target and masker speech with full sentence maskers and overlapping keywords.

2. MATERIALS AND METHODS

2.1 Participants

Twenty Dutch adult participants between 21 and 32 years old (M=25.2 years, SD=3.0) were recruited and reimbursed via the online testing-platform "Prolific". All participants were native Dutch speakers and reported to have normal or correct vision, normal hearing, and an absence of neurological or psychiatric illness. Participants' hearing status was confirmed with an online version of the digits-in-noise test (DIN) [11] for 18 out of 20 participants (sufficient: $n = 18$, insufficient: $n = 2$, poor: $n = 0$, all participants were included for data analysis). The study received ethical approval by the Medical Ethical Committee of the University Medical Center Groningen (METc 2018/427). Participants provided informed consent by completing an online consent form at the start of the online experimental session and received financial compensation according to Prolific and departmental guidelines.

2.2 Stimuli

Speech-on-speech perception was measured using the Dutch and child-friendly version of the CRM corpus [1], [12], [13]. All stimuli were recorded by a Dutch female speaker with a mean F0 value of 242 Hz, and an estimated VTL of 13.6 cm based on the speaker's height of 166 cm [14]. Participants were presented with short sentences with a call sign and two key words, a color and a number, e.g., *Laat de hond zien waar de groene* (color) *drie* (number) *is* [Show the dog where the green (color) three (number) is]. The call sign was either *dog* or *cat* for the target and masker sentences, respectively. The key words consisted of six colors, all disyllabic words in Dutch: *blauwe, gele, groene, rode, witte, zwarte* [blue, yellow, green, red, white, black] and eight numbers, all monosyllabic words in Dutch (1–10, excluding *zeven* (seven) and *negen* (nine), which are disyllabic words in Dutch), such that the set of target sentences contained a total of 48 sentences. A second set of 48 masker sentences had the same structure as the target sentences, except that the call sign was *kat* (cat) instead of *hond* (dog). Both target and masker sentences were recorded by the same female speaker and the full CCRM-NL corpus is available online[1].

In the full sentence masker condition, masker sentences were randomly selected for each trial, making sure that the color and number key words did not overlap between the target and masker sentences. In the scrambled sentence masker condition, masker sentences were created on a trial-by-trial basis by scrambling sentence chunks derived from the masker sentences with *kat* (cat) as call sign and with colors and numbers that did not match the key words presented in the target sentence. Sentence chunks ranging from 150 to 300 ms were then randomly selected and concatenated after applying 50-ms raised cosine ramps. The scrambled sentence masker started 750 ms before the target sentence and ended 250 ms after the target. No voice difference was introduced for the scrambled sentences.

For the full sentence masker condition, three voice conditions were created by shifting the F0 and VTL of the masker voice by a number of semitones (st) [(ΔF0,ΔVTL): no voice difference (0,0); small voice difference (-6,+1.8); and large voice difference (-12,+3.6)]. These values for the F0 and VTL shifts were based on previous work showing that a decrease of F0 by 12 st and an increase in VTL by 3.6 st reliably changes the perceived gender of a speaker's voice from woman to man for adult listeners with normal hearing [15]. Stimuli were resynthesized using the PyWorld wrapper [16] for the WORLD vocoder [17], implemented in the Voice Transformation Server [18]. To avoid an effect of potential artifacts due to the resynthesis procedure, sentences for the "no voice difference" were resynthesized as well, using the same procedure. Note that the scrambled sentence masker condition was only presented without any voice cue differences between target and masker speech.

[1] CCRM-NL corpus: 10.5281/zenodo.4700993

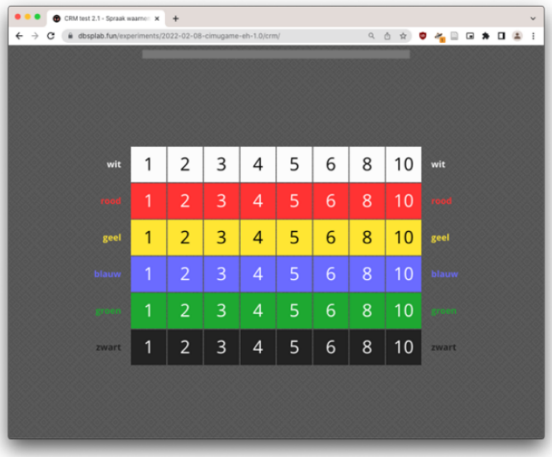

Figure 1. Online interface of the Dutch CCRM task. Participants had to click on the color-number combination that they heard in the target sentence *Show the dog where the* [color] [number] *is*, in Dutch.

2.3 Procedure

Data collection took place through a remote testing procedure on a web-based platform that was developed using the JavaScript framework JsPsych [19]. Participants completed the experiment on their own computers and all participants were requested to use headphones and to be placed in a quiet room during the test. Informed consent and demographic information were provided at the start of the experiment. Furthermore, participants were asked to complete an online DIN test [11] via https://www.hoortest.nl to verify their hearing status (sufficient: n=18, insufficient: n=2, poor: n=0).

Target and masker sentences were presented at six target-to-masker ratios (TMRs): -12, -8, -4, 0, +4, and +8 dB. The experiment consisted of eight items for each TMR, masker condition, and voice condition, resulting in a total number of 192 trials (8 items · 6 TMRs · 1 voice condition = 48 trials for the scrambled sentence masker; 8 items · 6 TMRs · 3 voice conditions = 144 trials for the full sentence masker). Sentences were presented in two consecutive blocks of 24 trials for the scrambled sentence masker and six consecutive blocks of 24 trials for the full sentence masker. The order of the masker type condition was randomized across participants. Within each block, the different TMRs and, if applicable, voice conditions were presented in a randomized order. The total test session lasted about 20–25 minutes.

3. RESULTS

The results were analyzed using generalized additive models (GAM; binomial distribution, logit link function). First, performance in the "no voice difference" condition was assessed for the two masker types:

```
score ~ masker + s(TMR, by = masker)
```

The model showed a significant effect of masker type (Figure 2, $\chi2(1) = 215$, $p<0.001$), as well as a significant effect of TMR for both scrambled sentence maskers ($\chi2(2.1) = 56.3$, $p<0.001$) and full sentence maskers ($\chi2(3.1) = 56.1$, $p<0.001$).

For the full sentence masker condition without voice cue differences, we used the fitted GAM to investigate at which TMR accuracy was lowest, based on the estimated derivative of the spline. An evaluation of the slope of the spline revealed that scores were lowest at a TMR of -3.3 dB [95% CI: -5.8,-1.1].

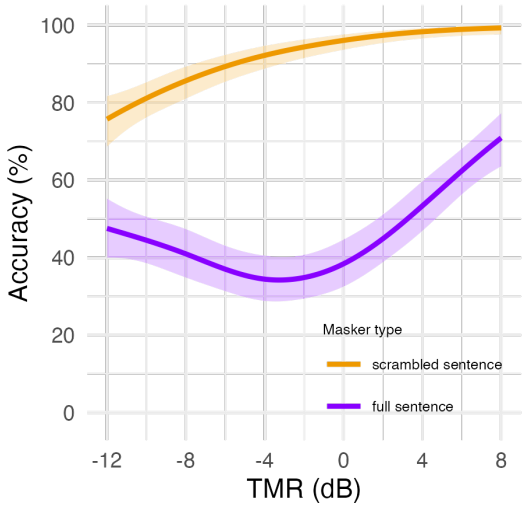

Figure 2. Relation between CCRM scores (accuracy in %) and TMR for full sentence maskers (purple) and scrambled sentence maskers (orange).

In a second analysis, the effect of voice differences was assessed in the full sentence masker condition:

```
score ~ voice + s(TMR, by = voice)
```

The model showed a significant effect of TMR in each voice condition (Figure 3; no voice difference: $\chi2(2.9) = 56.6$, $p<0.001$; small voice difference: $\chi2(1.3) = 15.3$, $p<0.001$; large voice difference: $\chi2(1.7) = 31.2$, $p<0.001$). Further, the model showed a significant effect of voice condition, where accuracy scores in the no voice difference condition were significantly lower compared to the small ($\chi2(1) = 319$, $p<0.001$) and large ($\chi2(1) = 312$, $p<0.001$) voice difference conditions across the complete TMR range tested in this study (-12 to + 8 dB). The accuracy scores in the condition with small voice differences were also significantly lower than in the condition with large voice difference ($\chi2(1) = 8.08$, $p=0.004$), but this difference was only significant in the TMR range between -7.5 and +6.0 dB.

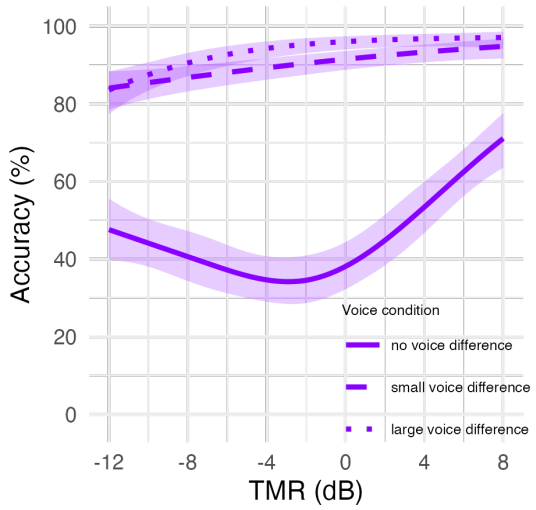

Figure 3. Relation between CCRM scores (accuracy in %) and TMR for the full sentence masker condition with no voice differences (solid line), small voice differences (dashed line), and large voice differences (dotted line).

4. DISCUSSION

The results of this study show that the effect of TMR on speech-on-speech perception was affected by the type of speech masker that was presented when the voice of the target and masker speaker was the same. For scrambled sentence maskers, TMR had a monotonic effect on speech-on-speech perception scores, with increasing scores for larger TMRs. For full sentence maskers, TMR had a non-monotonic effect, where scores were estimated to be lowest at a TMR of about -3.3 dB. This effect of masker type has implications for methods commonly used to quantify speech-on-speech perception, such as when determining a speech reception threshold (SRT) using an adaptive procedure (e.g., [20]) . Given the non-monotonic effect of TMR when using full sentence maskers, the measured SRT may be highly dependent on specific parameters used in the adaptive procedure, such as the starting value or step size.

The full sentence masker condition showed a large effect of F0 and VTL differences between target and masker speech for all TMRs. While the accuracy scores showed a non-monotonic effect of TMR when there were no voice cue differences between target and masker speaker, accuracy increased as a function of TMR when there were both small ($\Delta F0=-6$, $\Delta VTL=+1.8$) and large ($\Delta F0=-12$, $\Delta VTL=+3.6$) voice differences between the target and masker speaker. Moreover, at large intensity level differences between target and masker speech (TMR: -12, -8, and +8 dB), large voice differences do not seem to provide any added benefit over small voice differences.

These results show that perceptual mechanisms of speech-on-speech perception may differ for speech maskers consisting of scrambled vs. intact full sentences and demonstrate that the perceptual system makes use of available cues (differences in intensity level or voice cues), based on what cue is most salient. The different effects of masker type, voice cue difference, and TMR on speech perception with single-talker speech maskers urge researchers to carefully consider their experimental design. This becomes particularly relevant when comparing different listener groups (e.g., with normal hearing and with hearing loss) or exploring effects of age, especially when large differences in performance can be expected.

5. ACKNOWLEDGMENTS

This work was funded by a VICI Grant (No. 918-17-603) from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw). The authors would like to thank Dr. Etienne Gaudrain for advice on data analysis.

6. REFERENCES

[1] L. Nagels, E. Gaudrain, D. Vickers, P. Hendriks, and D. Başkent, "School-age children benefit from voice gender cue differences for the perception of speech in competing speech," *J. Acoust. Soc. Am.*, vol. 149, no. 5, pp. 3328–3344, May 2021, doi: 10.1121/10.0004791.

[2] T. Goossens, C. Vercammen, J. Wouters, and A. van Wieringen, "Masked speech perception across the adult lifespan: Impact of age and hearing impairment," *Hear. Res.*, vol. 344, pp. 109–124, 2017, doi: 10.1016/j.heares.2016.11.004.

[3] L. J. Leibold, A. Yarnell Bonino, and E. Buss, "Masked Speech Perception Thresholds in Infants, Children, and Adults," *Ear Hear.*, vol. 37, no. 3, pp. 345–353, May 2016, doi: 10.1097/AUD.0000000000000270.

[4] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, no. 3, pp. 1101–1109, 2001, doi: 10.1121/1.1345696.

[5] S. Brouwer, K. J. Van Engen, L. Calandruccio, and A. R. Bradlow, "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 1449–1464, Feb. 2012, doi: 10.1121/1.3675943.

[6] I. S. Johnsrude, A. Mackey, H. Hakyemez, E. Alexander, H. P. Trang, and R. P. Carlyon, "Swinging at a Cocktail Party: Voice Familiarity Aids Speech Perception in the Presence of a Competing Voice," *Psychol. Sci.*, vol. 24, no. 10, pp. 1995–2004, Oct. 2013, doi: 10.1177/0956797613482467.

[7] P. M. Johnstone and R. Y. Litovsky, "Effect of masker type and age on speech intelligibility and spatial release from masking in children and adults," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2177–2189, 2006, doi: 10.1121/1.2225416.

[8] N. El Boghdady, E. Gaudrain, and D. Başkent, "Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?," *J. Acoust. Soc. Am.*, vol. 145, no. 1, pp. 417–439, 2019, doi: 10.1121/1.5087693.

[9] L. Nagels, E. Gaudrain, D. Vickers, P. Hendriks, and D. Başkent, "Prelingually Deaf Children With Cochlear Implants Show Better Perception of Voice Cues and Speech in Competing Speech Than Postlingually Deaf Adults With Cochlear Implants," *Ear Hear.*, vol. 45, no. 4, p. 952, Aug. 2024, doi: 10.1097/AUD.0000000000001489.

[10] E. Harding *et al.*, "Effects of age and musical expertise on perception of speech in speech maskers in adults," *J. Acoust. Soc. Am.*, vol. 153, no. 3_supplement, p. A173, Mar. 2023, doi: 10.1121/10.0018564.

[11] C. Smits, P. Merkus, and T. Houtgast, "How we do it: The Dutch functional hearing–screening tests by telephone and internet," *Clin. Otolaryngol.*, vol. 31, no. 5, pp. 436–440, Oct. 2006, doi: 10.1111/j.1749-4486.2006.01195.x.

[12] L. Nagels, E. Gaudrain, D. Vickers, P. Hendriks, and D. Başkent, "CCRM-NL: Dutch Child-friendly Coordinate Response Measure Corpus." Zenodo. doi: 10.5281/zenodo.4700994.

[13] R. S. Bolia, W. T. Nelson, M. A. Ericson, and B. D. Simpson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, vol. 107, no. 2, pp. 1065–1066, 2000, doi: 10.1121/1.428288.

[14] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1511–1522, 1999, doi: 10.1121/1.427148.

[15] C. D. Fuller *et al.*, "Gender categorization is abnormal in cochlear-implant users," *J. Assoc. Res. Otolaryngol.*, vol. 15, no. 6, pp. 1037–1048, Aug. 2014, doi: 10.1007/s10162-014-0483-7.

[16] J. Hsu, *Python Wrapper for World Vocoder*. (2016). [Online]. Available: https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Trans. Inf.*, vol. E99-D, no. 7, pp. 1877–1884, Jul. 2016, doi: 10.1587/transinf.2015EDP7457.

[18] E. Gaudrain, "Voice Transformation Server." [Online]. Available: https://egaudrain.github.io/VTServer/index.html

[19] J. R. de Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," *Behav. Res. Methods*, vol. 47, no. 1, pp. 1–12, 2015, doi: doi:10.3758/s13428-014-0458-y.

[20] S. M. Saleh, S. R. Saeed, and D. Vickers, "Test-Retest Reliability of the Coordinate Response Measure in Adults with Normal Hearing or Cochlear Implants," *Audiol. Neurotol.*, vol. 28, no.

2, pp. 84–93, Apr. 2023, doi: 10.1159/000521466.