Pure linguistic interference during comprehension of competing speech signals

Bohan Dai, James M. McQueen, Peter Hagoort, and Anne Kösem

Citation: The Journal of the Acoustical Society of America **141**, EL249 (2017); doi: 10.1121/1.4977590
View online: http://dx.doi.org/10.1121/1.4977590
View Table of Contents: http://asa.scitation.org/toc/jas/141/3
Published by the Acoustical Society of America

Pure linguistic interference during comprehension of competing speech signals

Bohan Dai,[1,a),b)] James M. McQueen,[2,c)] Peter Hagoort,[1,b)] and Anne Kösem[1,b)]

[1]*Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands*
[2]*Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6500 HB Nijmegen, The Netherlands*
bohan.dai@mpi.nl, j.mcqueen@donders.ru.nl, peter.hagoort@mpi.nl, a.kosem@donders.ru.nl

Abstract: Speech-in-speech perception can be challenging because the processing of competing acoustic and linguistic information leads to informational masking. Here, a method is proposed to isolate the linguistic component of informational masking while keeping the distractor's acoustic information unchanged. Participants performed a dichotic listening cocktail-party task before and after training on 4-band noise-vocoded sentences that became intelligible through the training. Distracting noise-vocoded speech interfered more with target speech comprehension after training (i.e., when intelligible) than before training (i.e., when unintelligible) at −3 dB SNR. These findings confirm that linguistic and acoustic information have distinct masking effects during speech-in-speech comprehension.

[DDO]
Date Received: November 4, 2016 **Date Accepted:** February 14, 2017

1. Introduction

In a multi-talker environment, the intelligibility of a target talker against a background of concurrent talkers can be degraded as a consequence of masking. This phenomenon, the cocktail-party problem, highlights that ignored sound streams can, in certain conditions, alter the perception of attended speech (Carlile, 2014; Cherry, 1953). In general, the ignored sound interferes with the target speech in two ways: through energetic masking and through informational masking. The spectro-temporal overlap between the target speech and the distractor, leading to the degradation of the target speech input signal, results in energetic masking. The interference in cognitive processing between the target and distractor is generally referred to as informational masking.

Knowing the diversity in features encoded in speech, informational masking could occur at different levels of the speech processing hierarchy (Brungart, 2001; Evans and Davis, 2015; Hoen *et al.*, 2007; Mattys *et al.*, 2012; Rhebergen *et al.*, 2005). Of particular interest is how to disentangle the effects related to acoustic processing from those related to linguistic analysis. Some evidence suggests that distracting signals have dissociable masking effects on target speech comprehension depending on their amount of linguistic information. For example, distracting speech impairs the processing of target speech more strongly than does other unintelligible noise [e.g., noise-vocoded (NV) speech, rotated speech, time-reversed speech] (Brungart, 2001; Calandruccio *et al.*, 2013; Hoen *et al.*, 2007; Rhebergen *et al.*, 2005); native distracting speech is a stronger distractor than non-native or unknown speech (Brouwer *et al.*, 2012; Rhebergen *et al.*, 2005; Van Engen and Bradlow, 2007).

While these results have been interpreted as the consequence of competition between the linguistic information in the target and unattended speech channels, it cannot be fully excluded that these observations originate from acoustic masking effects. Intelligible speech and unintelligible speech-like signals or foreign speech differ both in linguistic content and in the spectro-temporal properties of the acoustics (Tong *et al.*, 2006). Thus, intelligible speech could be a stronger distractor than unintelligible speech for two reasons, either because it presents closer acoustic information to the target speech, or because it carries competing linguistic information. We present a new

[a)]Author to whom correspondence should be addressed.
[b)]Also at: Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6500 HB Nijmegen, The Netherlands.
[c)]Also at: Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands.

method to distinguish these two components (referred to as acoustic masking and linguistic masking from here on), by manipulating the linguistic content of the masker while keeping its acoustic properties constant.

To do so, participants completed a dichotic listening cocktail-party task with NV speech sentences as distracting signals before and after they were given training on understanding the distracting NV speech. Crucially, the NV speech sentences were initially poorly intelligible but could be understood after training (Davis *et al.*, 2005; Sohoglu and Davis, 2016). Hence, before and after training, the NV speech would have the same acoustic information but not carry the same linguistic information. By comparing the intelligibility reports of the target speech before and after training, we could isolate the contributions of the linguistic masking on target speech comprehension.

2. Methods

2.1 Participants

Twenty-four participants (15 female; mean ± standard deviation = 24 ± 2 yrs) from the Max Planck Institute for Psycholinguistics participant panel were tested. All participants were native Dutch speakers and were right-handed. Subjects had no known history of neurological, language, or hearing problems. Written informed consent was obtained from all participants prior to measurement and the study received ethical approval from the local reviewing committee "CMO Arnhem Nijmegen." All participants were paid for their participation.

2.2 Materials

The speech stimuli were selected from a corpus with meaningful conversational Dutch sentences, digitized at a 44 100 Hz sampling rate and recorded at the VU University Amsterdam (Versfeld *et al.*, 2000) by a male or female speaker. Each speech stimulus consisted of a combination of two sentences of the corpus uttered by the same speaker, separated by a 300-ms silence gap (average duration = 4.15 ± 0.13 s).

The target speech stimuli consisted of intact sentence pairs spoken by one of the two speakers. The distracting speech stimuli were NV versions of distinct sentence pairs taken from the same corpus and spoken by the other speaker (opposite sex). Noise-vocoding (Davis *et al.*, 2005) was performed using either 4 (main condition) or 2 (control) frequency bands logarithmically spaced between 50 and 8000 Hz. These values were selected based on pilot data, which showed that the intelligibility of 4-band vocoded speech could be improved through training, while 2-band NV speech remained unintelligible. One trial consisted of the presentation of the target speech with the interfering NV speech. The signals were delivered dichotically to the two ears and the target speech reached randomly one of them. The stimuli were presented at a comfortable listening level (overall average level of 70 dB sound pressure level) with three signal-to-noise ratio (SNR) conditions: 0, −3, and −6 dB.

2.3 Procedure

Participants were tested individually in a sound proof booth in a behavioral experiment lab at the Max Planck Institute for Psycholinguistics. The experimental design was implemented using Presentation software (Version 16.2, www.neurobs.com). All stimuli were played from a sound card (X-Fi Extreme) in a computer (HP Z400) through headphones (Sennheiser HD-520). Participants' responses were recorded by a microphone (Sennheiser K6 + ME64) with a sampling rate of 44 100 Hz.

The experiment included three phases: pre-training, training, and post-training (Fig. 1). In the pre- and post- training phases [Fig. 1(B)], the participants performed the dichotic listening task, in which they were instructed to listen to the presentation of one intact speech channel and one unintelligible NV speech channel and pay attention to the intact target speech only. After the presentation, the participant's task was to repeat the sentences of the target speech. Trials were divided into blocks of 24 trials per SNR (0, −3, −6 dB) and NV condition (4- and 2-band). The target speech differed across trials and all conditions (hence a target stimulus was only presented once during the whole experiment). The distracting speech differed between the two NV conditions and each stimulus was repeated three times (one stimulus per SNR condition). In each block, the target stimuli in half of the trials were spoken by the female speaker, and the rest were spoken by the male speaker. The distracting speech (prior to noise-vocoding) was always spoken by the other speaker. A total of 144 trials were tested in each phase.

In the training phase [second phase of the experiment, Fig. 1(C)], participants were trained to understand the 4-band NV speech. The training phase included three parts: (a) pre-test: the participants were tested on their ability to understand the

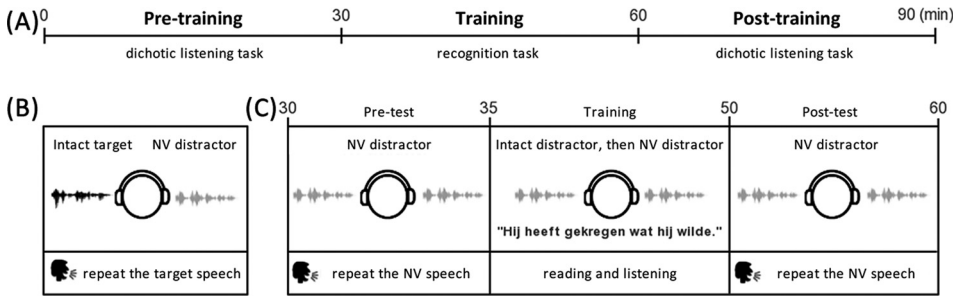

Fig. 1. Experimental design. (A) The experiment consisted of three phases. In the first and third phases of the experiment, participants performed a dichotic listening cocktail-party task, in between the two dichotic listening tasks participants were trained to understand 4-band NV speech. (B) During the dichotic listening tasks, participants listened to the presentation of one intact target with one NV distractor (either 2- or 4-band) and were asked to repeat the intact target speech. (C) During the training of the 4-band NV sentences, participants listened to the distractor once in the intact and then once in the NV version. At the same time, they read the text of the sentences on the screen. We tested the intelligibility of the 4-band NV sentences before (pre-test) and after (post-test) the training by asking participants to listen to and repeat the NV sentences.

4-band NV speech; they were presented with the interfering speech binaurally and were asked to repeat it afterwards; (b) training: they were presented one token of an intact version of an NV stimulus followed by one token of the NV version of that stimulus; at the same time, they could read the content of the NV speech on the screen; (c) post-test: they performed the intelligibility test again. Untrained 4-band sentences were added in the last section to test the generalization of the training procedure. To ensure no learning of 2-band NV speech, 2-band NV sentences were not presented in the training session. In total, 96 trials were tested in this phase.

3. Results and discussion

The intelligibility of speech was measured by calculating the percentage of correct content words (excluding function words) in participants' reports for each speech sequence. Words were regarded as correct if there was a perfect match (correct word without any tense errors or singular/plural form changes). The percentage of correct content words was chosen as a more accurate measure of intelligibility based on acoustic cues than percentage correct of all words, considering that function words could be guessed based on the content words (Brouwer *et al.*, 2012; Hustad, 2006). But we also performed the analyses with all correct words; these analyses yielded similar results.

3.1 Training improved the intelligibility of 4-band NV speech

As shown in Fig. 2(A), the training phase significantly improved the perception of 4-band NV speech. The intelligibility of the NV sentences was low before training [mean = 26.07%, standard error (SE) = 1.91] and improved after training (mean- = 56.34%, SE = 2.59). The intelligibility of untrained 4-band NV sentences also improved (relative to the matched pre-training sentences) after training (mean- = 40.50%, SE = 2.10), which suggests that the training could generalize to new speech utterances, in line with previous reports (Davis *et al.*, 2005; Sohoglu and Davis, 2016). Both improvements were significant [pre- vs post-trained: $t(23) = 18.34$, $p < 0.001$; pre- vs post-untrained: $t(23) = 10.14$, $p < 0.001$, Bonferroni corrected], but intelligibility after training was stronger for the trained sentences compared to the untrained sentences [$t(23) = 10.64$, $p < 0.001$].

3.2 NV speech was a stronger masker when intelligible

We predicted that the distracting NV signal would interfere more with target speech comprehension in the dichotic listening task when intelligible, that is, after training. In addition, the training phase should have no influence on the performance in the dichotic listening task when the distracting signal is untrained 2-band NV speech. A three-way repeated-measures analysis of variance (ANOVA) was performed to assess the contribution of the three factors: Time (pre-/post-training), NV (trained 4-band/untrained 2-band), and SNR (0, −3, −6 dB). In line with our predictions, the interference on target speech through training was observed [Fig. 2(B)]: intelligibility scores dropped significantly after training [main effect of Time: $F(1, 23) = 13.509$, $p = 0.001$]. NV speech (4-band) was in general a stronger distractor than 2-band NV speech [main effect of NV: $F(1, 23) = 9.879$, $p = 0.005$]. No main effect of SNR was observed [$F(2, 46) = 2.298$, $p = 0.112$].

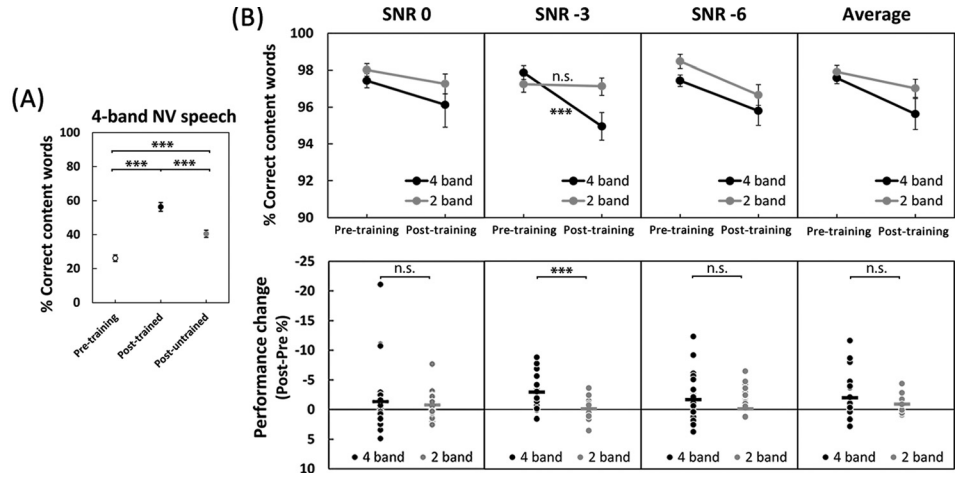

Fig. 2. Results. (A) Training phase. Intelligibility of 4-band NV speech before and after training. (B) Dichotic listening task. Top panel: Target speech intelligibility during the dichotic listening task before and after training, when the distracting speech is trained 4-band NV speech (black) or untrained 2-band NV speech (gray). Bottom: Relative performance change between pre-training and post-training. Each dot represents the difference of performance for one participant. ***$p < 0.001$, n.s., not significant.

The performance in post-training could have declined because of increased intelligibility of the distractor acquired via training, but also because of confounding effects such as fatigue. To test for this, we compared performance in the trained 4-band and the untrained 2-band NV distracting speech conditions. We expected that the training influences should only be observed for 4-band NV signals. In line with this prediction, the three-way interaction between Time, NV, and SNR was significant [$F(2, 46) = 4.494$, $p = 0.016$]. This suggested that training differentially influenced target speech perception depending on the NV masker. Although the two-way interactions were not significant [Time * NV: $F(1, 23) = 2.246$, $p = 0.148$; NV * SNR: $F(2, 46) = 0.058$, $p = 0.944$; Time * SNR: $F(2, 46) = 1.267$, $p = 0.291$], the three-way interaction indicates that the effect of training was dependent on SNR.

3.3 Linguistic masking effects were most prominent at SNR = −3 dB

To further investigate the effect of SNR on intelligibility reports, we performed a two-way ANOVA analysis for each SNR level. A significant interaction was found between NV and Time at SNR −3 dB only [$F(1, 23) = 24.08$, $p < 0.001$] [Fig. 2(B)]. At SNR = −3 dB, 4–band distracters interfered more strongly with the processing of target speech after training compared to before training, whereas intelligibility of the target speech did not vary significantly when the distracting signal was 2-band NV speech (4-band post vs pre: $p < 0.001$; 2-band post vs pre: $p = 0.269$, Bonferroni corrected for multiple comparisons). At the other SNRs, the interaction between NV and Time was not significant [SNR 0: $F(1, 23) = 0.35$, $p = 0.657$; SNR −6: $F(1, 23) = 0.23$, $p = 0.845$].

Hence, the effects attributable to linguistic masking were primarily found when the distracting speech was 3 dB louder than the target speech. One potential explanation for this is based on the observation that linguistic masking has distinct effects on target speech comprehension depending on SNR (Brungart *et al.*, 2005). Speech-like noise maskers have been shown to impair target speech intelligibility monotonically for low SNRs (≤−3 dB) (Brungart, 2001). In contrast, target speech comprehension with speech maskers also drops as the SNR drops but tends to plateau and even improve when SNR levels are below −3 dB (Brungart, 2001). As a consequence, the effect of linguistic masking may reverse as a function of SNR, starting from an increasingly distracting influence on target speech comprehension as SNRs drop down to at least −3 dB, but a facilitating effect at even lower SNRs (Brungart *et al.*, 2005). It is thus possible that −3 dB SNR is the level in this study at which linguistic interference is strongest (higher SNRs provide less interference, and lower SNRs provide facilitation because the distractors become easier to identify).

Another plausible reason could be the existence of an interaction effect between linguistic and acoustic masking. By analogy to a previous discussion on interactions between informational masking and energetic masking (Kidd *et al.*, 2005; Neff and Jesteadt, 1996), the total amount of masking produced may be either dominated by acoustic or linguistic masking depending on contextual constraints. When the distractor was much louder than the target speech (i.e., at low SNR conditions), the

acoustic masking could dominate the overall informational masking effects. This would be in line with the idea that informational masking partly originates from the degradation of auditory representations at higher levels of the speech processing hierarchy (Durlach *et al.*, 2003). Auditory networks encoding the competing speech signals may only represent one of the stimuli if the SNR is too unbalanced. Acoustic masking could dominate at low SNRs due to this representational bias operating prior to linguistic analysis.

4. Summary

In this study, we investigated the origins of informational masking effects during speech-in-speech comprehension. We asked whether the competition between target and distracting speech results from the parallel processing of linguistic information, or from the competition of processed acoustic information. We proposed a new paradigm to test this, where distracting signals were NV speech that were initially unintelligible, and that became more intelligible via training. We show that unattended NV speech becomes more distracting when intelligible (after training) compared to when it is poorly understood (prior to training) and that these effects appear more prominent at SNR −3 dB. In line with past reports (Brouwer *et al.*, 2012; Brungart, 2001; Calandruccio *et al.*, 2013; Hoen *et al.*, 2007; Iyer *et al.*, 2010; Lecumberri and Cooke, 2006; Lecumberri *et al.*, 2010; Rhebergen *et al.*, 2005; Van Engen and Bradlow, 2007), the present results show that an intelligible distractor interferes more with the processing of target speech and suggest that speech-in-speech interference originates, to a certain extent, from the parallel processing of competing linguistic content.

Acknowledgments

We would like to thank Ole Jensen for providing insightful suggestions on this work; Maarten van den Heuvel and Vera van't Hoff for help with transcribing data. This research was supported by a Spinoza grant to P.H.

References and links

Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (**2012**). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," J. Acoust. Soc. Am. **131**(2), 1449–1464.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., and Kidd, G., Jr. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**(1), 292–304.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., and Bradlow, A. R. (**2013**). "Masking release due to linguistic and phonetic dissimilarity between the target and masker speech," Am. J. Audiol. **22**(1), 157–164.

Carlile, S. (**2014**). "Active listening: Speech intelligibility in noisy environments," Acoust. Australia **42**(2), 90–96.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**(2), 222–241.

Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking (L)," J. Acoust. Soc. Am. **113**(6), 2984–2987.

Evans, S., and Davis, M. H. (**2015**). "Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis," Cereb. Cortex **25**(12), 4772–4788.

Hoen, M., Meunier, F., Grataloup, C. L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., and Collet, L. (**2007**). "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension," Speech Commun. **49**(12), 905–916.

Hustad, K. C. (**2006**). "A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners," Am. J. Speech Lang. Pathol. **15**(3), 268–77.

Iyer, N., Brungart, D. S., and Simpson, B. D. (**2010**). "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," J. Acoust. Soc. Am. **128**(5), 2998–3010.

Kidd, G., Jr., Mason, C. R., and Gallun, F. J. (**2005**). "Combining energetic and informational masking for speech identification," J. Acoust. Soc. Am. **118**(2), 982–992.

Lecumberri, M. L. G., and Cooke, M. (**2006**). "Effect of masker type on native and non-native consonant perception in noise," J. Acoust. Soc. Am. **119**(4), 2445–2454.

Lecumberri, M. L. G., Cooke, M., and Cutler, A. (**2010**). "Non-native speech perception in adverse conditions: A review," Speech Commun. **52**(11–12), 864–886.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cogn. Process. **27**(7–8), 953–978.

Neff, D. L., and Jesteadt, W. (**1996**). "Intensity discrimination in the presence of random-frequency, multi-component maskers and broadband noise," J. Acoust. Soc. Am. **100**(4), 2289–2298.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**(3), 1274–1277.

Sohoglu, E., and Davis, M. H. (**2016**). "Perceptual learning of degraded speech by minimizing prediction error," Proc. Natl. Acad. Sci. U.S.A. **113**(12), E1747–E1756.

Tong, R., Ma, B., Zhu, D., Li, H., and Chng, E. S. (**2006**). "Integrating acoustic, prosodic and phonotactic features for spoken language identification," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1, p. I.

Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–26.

Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**(3), 1671–1684.