INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Quantifying Informational Masking due to Masker Intelligibility in Same-talker Speech-in-speech Perception

Mingyue Huo[1], *Yinglun Sun*[1], *Dan Fogerty*[2], *Yan Tang*[1,3]

[1]Department of Linguistics, University of Illinois Urbana-Champaign, USA
[2]Department of Speech and Hearing Science, University of Illinois Urbana-Champaign, USA
[3]Beckman Institute for Advanced Science and Technology, USA

{mhuo5, yinglun2, dfogerty, yty}@illinois.edu

Abstract

Intelligibility of the competing speech plays a significant role in causing informational masking (IM) to the target speech during speech-in-speech perception, especially in same-talker conditions where the target and the masker share a large number of similarities in acoustics. Few studies have quantitatively measured IM as a function of intelligibility of competing speech. Evidence shows that voiced segments are robust cues for speech intelligibility. In this study, the contribution of masker intelligibility to IM was studied by adjusting the voice-to-noise ratio (VNR) on voiced segments of the competing speech, while maintaining energetic masking (EM) at different target-to-masker ratios. Although model estimations suggested that the intelligibility due to EM converged when VNR$<$0 dB, listener performance showed that more release from IM was received with a further decrease in VNR. It was projected that masker intelligibility could lead to target intelligibility decreased by 50%.

Index Terms: informational masking, intelligibility, speech-in-speech perception

1. Introduction

Speech intelligibility can be negatively impacted in the presence of competing speech, which can cause two main categories of masking effects – energetic masking (EM) and informational masking (IM). Both effects may interfere with listeners' perception of the target speech. While EM is a consequence of the physical interactions between the target and the masking signal, IM occurs when listeners are able to hear both utterances but have difficulties in assigning acoustic elements in the mixture to the target and the masker. Compared to EM, the severity of IM can be complicated by factors such as the sex, signal intensity level relative to the target speech, spatial information, number of competing talkers, listeners' familiarity with the language, the linguistic content of the competing speech, and intelligibility [1, 2, 3, 4]. IM is known to be involved in multiple cognitive stages in the auditory pathway, including perceptual grouping, source segregation, attention, memory, and central cognitive processing and resource allocation [5]. Due to the nature of IM, its impact on intelligibility is often confounded with that of EM. Though it is difficult to accurately quantify the effect of IM by isolating it from EM, it has received growing attention in recent years.

Several studies have shown that the intelligibility of *competing speech* has a significant effect on IM. Acoustically, when the competing speech is locally reversed, noise-vocoded, or an unfamiliar or low-proficiency language to the listener, the listener tends to receive more releases from IM [6, 7, 4]. One explanation could be that the acoustic distance between the target and the competing speech may improve the listener's efficiency in segregating and re-grouping different sound sources [8]. For acoustic cues pertinent to speech intelligibility, voiced segments, including vowels, sonorants, and voiced consonants, are the places where many of them are located, due to their high energy and long durations. Therefore, acoustic cues on voiced segments are usually more robust to masking than those from elsewhere. The amount of spectro-temporal (S-T) regions that escape from energetic masking was reported to be highly correlated with listener speech recognition performance in noise [9], especially those on voiced segments [10]. It was also found that voiced segments in the range of 540-1700 Hz have a higher correlation with speech intelligibility than other frequencies [11]. This suggests that in general corrupting the acoustic cues located in voiced segments by decreasing the voice-to-noise ratio (VNR) could lead to reduced intelligibility of competing speech. Despite a growing body of research on IM, few studies have investigated the relationship between the VNR of voiced segments in competing speech and IM.

Attempts have been made to isolate the effect of IM from EM when studying the contribution of voiced segments to IM. From using globally reversed speech [12] to locally time-reversed speech [7] as the competing sources, the change of the EM effect was better controlled. Similar approaches include manipulating the formant structure [13] or pitch contour [14] of the competing speech, in order to minimize the change of S-T information. However, the effect of IM or EM was not quantitatively reported in any of the studies.

The current study aimed to quantify the IM effects in same-talker speech-in-speech perception. The degree of IM was controlled by altering the local VNR on the competing source, while maintaining EM under three target-to-masker ratios (TMR). The intelligibility of the target speech under the pure EM caused by the competing speech was estimated using a glimpse-based intelligibility model [10] in an attempt to isolate the IM effect in further steps. We hypothesize that the lower VNR, the more release from IM.

2. Method

2.1. Stimuli

The speech materials were the Harvard sentences [15], uttered by a male native American English speaker. Both target speech and competing speech were drawn from the same corpus, in which case a speech-on-speech informational masking (IM) may be maximized [1]. In order to provide the listener with an essential cue to discern the target speech from the computing speech, the sentences starting with the prompt word "the" were chosen as the target sentences; the competing counterpart did not start with the prompt word.

10.21437/Interspeech.2023-1122

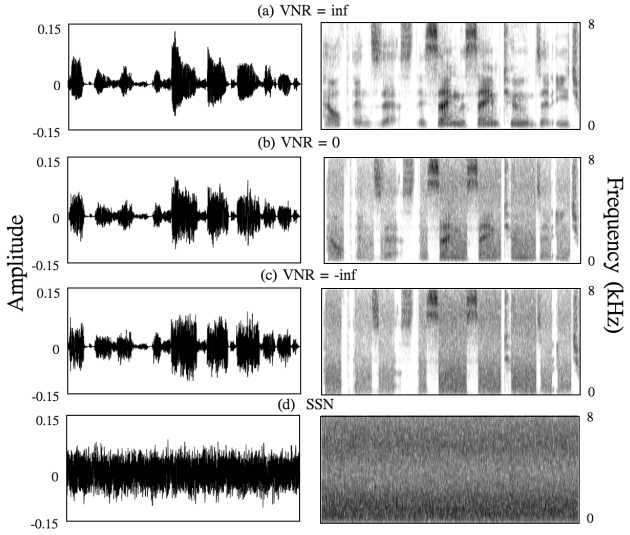

Figure 1: *Waveforms (left) and spectrograms (right) of a competing speech at three VNRs (a, b, c) and SSN (d)*

We used a two-stage control on signal intensity when manipulating the *competing speech* in order to isolate the IM effect from the energetic masking (EM) effect. At the segment level, all the voiced segments were corrupted by speech-shaped noise (SSN) that had the long-term average spectrum of the corpus. The noise was added at 11 voice-to-noise ratios (VNRs) – from -18 to 30 dB taking a step of 6 dB, "**inf**" (when no noise was added) and "**-inf**" (when all voiced segments were replaced by SSN) – in order to elicit the IM effect of different degrees. After adding the noise, the intensity of each noise-corrupted voiced segment was adjusted to align with the original segment RMS to minimize the change in the EM effect. In addition, SSN as an individual masker was also included as a reference condition of pure energetic masking ("**SSN**"), resulting in a total of 12 conditions that varied in the level of IM. Fig. 1 shows the waveform and spectrogram of the manipulated competing speech at four representative VNRs. At the utterance level, the RMS of the noise-corrupted competing speech was further equalised to that of the original signal.

The target speech and competing speech were mixed at TMRs of -9 dB, -4.5 dB or 0 dB, approximately leading to listener word recognition rates (WRRs) of 15%, 40% or 70% in SSN [16]. Previous studies [1, 17] found that the greatest IM occurred in a speech-on-speech masking condition when the TMR was at 0 dB, and that a better release from IM was received with the decrease or increase of TMR (i.e. a greater discrepancy in the levels of the target and masker). When TMR is greater than 0 dB, intelligibility is likely to converge quickly, hence we only tested at the negative TMRs to avoid the ceiling effect. The experiment consisted of 36 conditions (3 TMRs $\times$ 12 VNRs). The competing speech in each mixture was randomly extracted from a concatenated sequence of the same talker where VNR had been adjusted to a specific level. The same target-masker pair was used in all three TMRs in order to maintain the same S-T masking pattern on the target speech. All masking speech preceded and tailed the target speech for 250 ms. All signals were WAV files sampled at a rate of 16 kHz.

2.2. Participants

Twenty-one native American English speakers (13 females, 8 males) between the ages of 18 and 31 years (an average of 21) were recruited in this study. A hearing screening showed that all the participants had normal hearing.

2.3. Procedure

The listening experiments took place in a sound-attenuating audio-metric booth. Stimuli were presented to listeners diotically over a pair of open-back headphones. The presentation level of the target speech was calibrated to approximately 69 dB SPL and the masker level was adjusted to achieve the target TMR. A listener heard 5 sentences in each of the 36 conditions, leading to 180 unique sentences in total. With a balanced design, under each condition, no two listeners heard the same set of five sentences. All 180 sentences were blocked by TMR. All listeners heard three TMR levels in a random order; the stimuli in a TMR block were also randomized. The listeners were instructed to type down the words from the target utterance that started with the prompt word "the". They were also prompted that the target utterance was no louder than the competing utterance in all conditions. All listeners completed a practice session with 15 sentences before the main experiment. The listeners could only hear each stimulus once.

3. Results

3.1. Energetic masking

The intelligibilities of the stimuli under EM in all the conditions were first predicted using the High-Energy Glimpse Proportion (HEGP [10]). HEGP makes predictions by quantifying the number of S-T regions on speech with local energy above the average level of the speech-plus-noise mixture and with a local SNR above a given threshold (e.g., 3 dB). It only solely accounts for the EM effect on intelligibility in principle. The higher the HEGP score, the less EM, hence the better intelligibility. HEGP predictions have been reported to be highly correlated ($r^2 > 0.80$) with listener intelligibility in many temporally stationary and fluctuating noises. Fig. 2 shows the average HEGP scores across all the sentences in each condition: at all three TMRs, HEGP as a function of VNR has a similar pattern. First, the SSN masker always causes the strongest EM, while the "**inf**" masker (the original competing speech) results in the least of EM due to the modulation dips in both the temporal and spectral domains, resulting in more opportunities for listeners to glimpse the target. At the same TMR, decreasing VNR leads to a reduction of S-T modulation, HEGP falls as a consequence

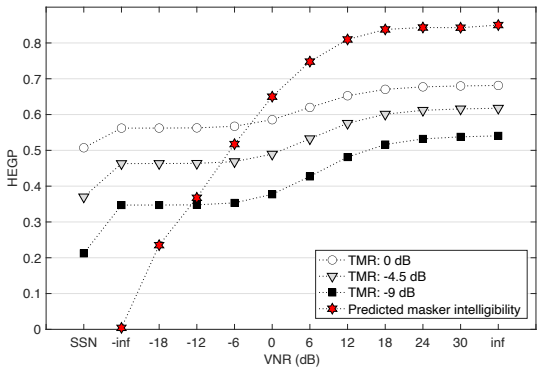

Figure 2: *Mean HEGP of target speech as a function of VNR at three TMR conditions. Mean HEGP of competing speech for each VNR is shown as red hexagrams.*

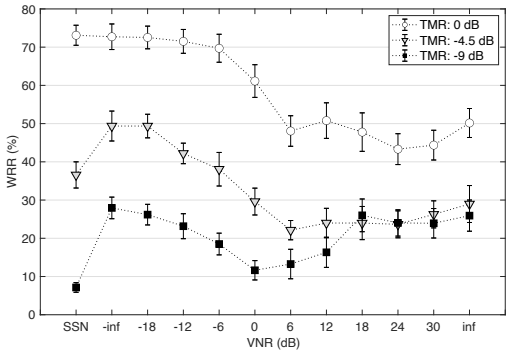

Figure 3: *Average listener WRR as intelligibility in 12 VNRs at three TMRs. Error bars indicate* ± 1 *standard error.*

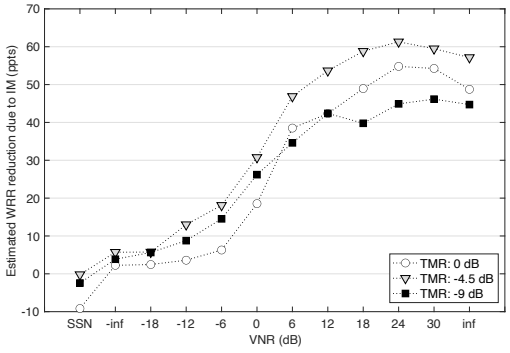

Figure 4: *Average estimated reduction in WRR due to IM in 12 VNRs at three TMRs.*

until *the EM effect appears to start to converge below VNR= -6 dB*. At "**-inf**", the masker can be viewed as modulated noise, but with more glimpsing opportunities during the unvoiced segments (see subplot (c) in Fig. 1) – there is at least a further 0.05 HEGP increase from "**SSN**" to "**-inf**" across TMRs. The greater increase from "**SSN**" to "**-inf**" at the lower TMR indicates the unvoiced segments and temporal modulation in "**-inf**" may provide a considerable release from EM. Fig. 2 further confirms that noise maskers may vary in EM even under the same overall TMR due to the spectral and temporal modulations within voiced segments [18, 19]. In addition to the stimuli intelligibility, the masker intelligibility was also estimated as HEGP; corrupting the voiced segment indeed started resulting in a dramatic reduction in masker intelligibility when VNR$<$12 dB.

3.2. Listener intelligibility

Listener intelligibility was measured as word recognition rate (WRR) in each condition. During scoring, homonyms and typos were treated as correct responses. Fig. 3 shows the mean performance across all the participants. Overall, listener performance exhibits a rather opposite pattern to HEGP predictions based on pure EM: while HEGP has suggested that target intelligibility decreases with the decrease of VNR, listener performance appears to increase, especially when VNR goes below 6 dB. Intriguingly, despite EM reaching the ceiling at VNR=-6dB, a further release from masking is observed when VNR further decreases until "**-inf**". This release led to an increase in WRR of 3.0, 11.3 and 9.4 percentage points (ppts) under the TMRs from high to low. Further losing both spectral and temporal modulation from the unvoiced segments in "**SSN**" resulted in a further reduction of 12.8 and 20.8 ppts in WRR from "**-inf**" at -4.5 and -9 dB TMR, respectively, but not at all at 0 dB TMR. When VNR is above 12 dB, intelligibility first decreases for approximately 20.7 ppts on average when TMR decreases from 0 dB to -4.5 dB, but further decreasing TMR hardly leads to any intelligibility loss with a change of merely 0.7 ppts in WRR, despite a substantial increase of EM.

A two-way repeated-measures ANOVA found both the main effects, TMR $[F(2, 40) = 275.343,\ p < .001,\ \eta^2 = .570]$ and VNR $[F(11, 220) = 17.311,\ p < .001,\ \eta^2 = .222]$, significantly affect listener performance in speech-in-speech perception. A significant two-way interaction $[F(22, 440) = 6.655,\ p < .001,\ \eta^2 = .160]$ suggested that the release from IM at various VNRs was affected differently by the overall TMR. Post-hoc pairwise comparisons using Fisher's Least Significant Difference (FLSD) with a within-subject design were further conducted[1]. With an FLSD of 8.4 ppts, it was confirmed that when VNR$>$0 dB, listeners received similar intelligibility at 0 and -4.5 dB TMR [$\forall\Delta < 8.4$], despite EM levels suggested by HEGP predictions in Fig. 2. The same insensitivity of WRR to VNR was also observed at -9 dB TMR but only when VNR$>$12 dB, where the WRRs for all VNRs across -4.5 dB and -9 dB TMR were similar [$\forall\Delta < 8.4$], suggesting a great release of IM when TMR reduced from -4.5 dB and -9 dB. The result is broadly consistent with the findings in [1] that discrepancy in signal intensity between the target and competing speech may lead to increased intelligibility. When comparing "**inf**" to "**-inf**", where IM is supposed to be the strongest and the weakest, respectively, and when the temporal modulation in the maskers is still largely retained, the latter led to a significant intelligibility gain at 0 dB ($\Delta = 22.6 > 8.4$) and -4.5 dB ($\Delta = 20.4 > 8.4$) TMR, while maintaining the intelligibility to a comparable level at -9 dB TMR ($\Delta = 2.0 < 8.4$). From where EM starts converging (VNR=-6 dB) to "**-inf**", the further release from IM is also significant at the two lower TMRs ($\Delta > 8.4$), but not at 0 dB TMR ($\Delta = 3.0 < 8.4$).

3.3. Quantifying the effect of IM on intelligibility

Linear regression was first performed on the WRRs in "**SSN**" and "**-inf**", where IM was the least, in order to establish the relationship between HEGP and WRR as shown in Fig. 5. The formula in Fig. 5 expresses the predicted WRR affected by EM only, WRR', as a function of HEGP. By further subtracting WRR' from the measured listener WRR (in Fig. 3), the WRR reduction in ppts due to IM for each condition was then estimated in Fig. 4. The impact of IM reduces along with the decrease of VNR in general, with a steeper slope between -6 and 6 dB VNR. This pattern is contrary to that of EM in Fig. 2, where EM increases otherwise. When VNR$>$18 dB where IM is approaching its maximum, it may lead to a WRR reduction of approximately 55.0 ppts at least and further up to 61.3 ppts. Interestingly, while one would expect decreasing TMR could lead to better release from IM, the estimation indicates that IM accounted for a higher WRR reduction in -4.5 dB than in 0 dB TMR. In high VNRs, IM at the lowest TMR (-9 dB) is indeed expected to result in the lowest WRR reduction, but not in low VNRs. Across all VNRs, IM is estimated to have caused an intelligibility loss of 26.0, 34.2 and 25.8 ppts in WRR in the TMRs from highest to lowest, respectively. Finally, IM tends

[1]For comparison, if the difference in the means in any two conditions, Δ, is smaller than a given FLSD, the difference is then insignificant.

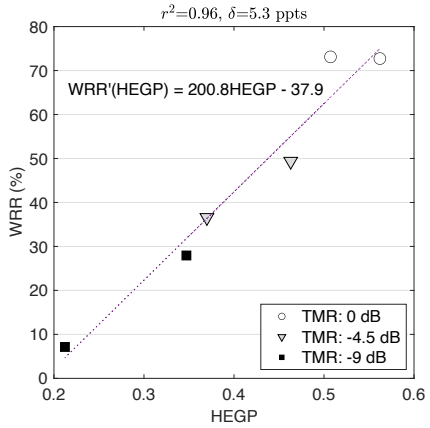

Figure 5: *Listener intelligibility (WRR) vs HEGP predictions in the least EM conditions ("***SSN***" and "***-inf***"). Linear regression was performed to model WRR as a function of HEGP. r^2 and δ are the Pearson correlation coefficient and the maximum root-mean-square error in WRR' after a linear transformation is performed.*

to diminish to its minimum after VNR goes below -12 dB at all TMRs.

4. Discussion

One of the key findings from this study is that the release from informational masking (IM) in same-talker speech-in-speech perception indeed improves with the decrease of masker intelligibility controlled by voice-to-noise ratio (VNR) (Fig. 4). Such release from IM was estimated to compensate up to 57 ppt in WRR. We also observed interactions between energetic masking (EM) (Fig. 2) and IM (Fig. 4): when target-to-masker ratio (TMR) was 0 or -4.5 dB and VNR was above 6 dB, listener WRRs plateaued despite those conditions varying in both EM and IM, as shown in Fig. 3. The convergence of listener performance may be explained by the interaction between worsened EM and alleviated IM resulting from decreased masker intelligibility. Second, at -9 dB TMR a significant drop in WRR was seen when VNR decreased from 18 to 12 dB, presumably due to EM becoming dominant and the release from IM not being sufficient to compensate for the negative EM effect. An estimated more drastic increase in EM can also be seen in Fig. 2: the slope between -6 and 18 dB VNR for -9 dB (0.0072) TMR appears to be somewhat greater than for -4.5 (0.0059) and 0 dB (0.0046) TMRs. This is consistent with the previous finding that IM is the dominant masking effect in speech-on-speech masking [1].

The masker intelligibility estimated in Fig. 2 displayed a high *negative* correlation ($r^2 = 0.87$) with the mean WRRs across the three TMRs in Fig. 3, confirming that masker intelligibility is strongly associated with target intelligibility in same-talker speech-in-speech perception. However, the masker intelligibility here was altered by adjusting VNR; the voiced cues of the masker at the acoustic level may also have an impact on the target intelligibility [13, 14]. Thus, further experiments are required to identify whether the entire release from IM here was due to the reduced masker intelligibility in the sense of impaired linguistic cue carried by the masker, the corrupted voiced cues (e.g., F0 and harmonicity), or the two effects combined.

The intensity difference between the target and competing speech is also a cue used by human listeners in same-talker speech-in-speech perception (e.g., [1]). As reported in [1], intelligibility in speech-on-speech masking decreased monotonically from high TMRs to 0 dB (i.e., when the target and masker have the same intensity level), but plateaued or increased slightly as TMR further decreased from 0 to -9 dB; introducing a level difference of about 3 dB to the target and competing speech could alleviate the IM effect and significantly improve intelligibility. However, such a pattern was not observed in the current study when the competing speech was unaltered (i.e. "**inf**") at the two higher TMRs. The intelligibility measured in this study for same-talker conditions was about 30-40 ppts higher on average than those reported in [1] and [13] at the 0 and -4.5 TMR, but similar to [3]. Comparisons between studies show that there appear to be more IM effects when the target and competing speech are drawn from the same closed set, where target words largely overlap with words in the competing speech [5]. This could explain the relatively better performance in this study, where an open set was used. Moreover, the performance at 0 and -4.5 dB TMRs did not inhibit a "slightly increasing" pattern as reported in [1]. This was possibly due to the difference in task difficulty: while only a limited number of tokens for number and colour were used as keywords in [1], listener performance was evaluated from the recognition of all non-article non-propositional words in the Harvard corpus. Consequently, the level cue at -4.5 dB TMR, even if exists, may be unable to introduce enough release from IM to compensate for the EM effect. The effect of the level cue, however, was indeed translated to enough intelligibility gain to reconcile the EM effect when further increasing the level difference from 4.5 to 9 dB, resulting in similar WRRs at the two lower TMRs.

High-Energy Glimpse Proportion (HEGP) demonstrated a robust predictive power ($r^2 = 0.96$ as in Fig. 5) when predicting intelligibility in the conditions where little IM effect was present. However, when extending the prediction to the conditions causing a large IM effect, HEGP failed with $r^2 = 0.20$, suggesting that the strong IM effect must be modelled carefully for better predictive accuracy. While computational models are proposed primarily to make intelligibility estimations by accounting for EM, only a handful of such models investigate the impact of IM on intelligibility in detail. In [20], Wu and Chen integrated a frequency-dependent penalty factor, which was calculated as the distance between the harmonic features extracted from the target speech and target-competing speech mixture, into the calculation of the Speech Intelligibility Index (SII). An improvement over the original SII was demonstrated when predicting intelligibility in 20 conditions described in [14]. Tang and Cox used the auditory salience score to weight the contribution of glimpses at different spectro-temporal regions, assuming that the IM effect on intelligibility is partly due to listeners' limited capacity for selective attention [21]. Models (e.g., [22, 23]) quantifying modulation masking could be another approach to this problem. However, ascribing IM and EM to modulation masking and modelling them as a whole is also debatable. It was demonstrated in [20] that only modelling modulation masking was not adequate to explain IM in intelligibility predictions. Further investigation on this topic is thus guaranteed.

In conclusion, this study investigated the effect of IM on same-talker speech-in-speech perception by altering the voice-to-noise ratio (VNR) of the voiced segments in competing speech. With a glimpse-based model that estimates pure EM component, we quantitatively isolated IM as a function of masker VNR. It is found that the release from IM improved with the decrease of masker intelligibility controlled by VNR. Future studies on whether such release was due to impaired linguistic cue or corrupted voiced cues in the masker will be carried out.

5. References

[1] D. S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1101–1109, 2000.

[2] R. L. Freyman, U. Balakrishnan, and K. S. Helfer, "Spatial release from informational masking in speech recognition," *The Journal of the Acoustical Society of America*, vol. 109, no. 5, pp. 2112–2122, 2001.

[3] M. Cooke, M. Garcia Lecumberri, and J. Barker, "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," *The Journal of the Acoustical Society of America*, vol. 123, no. 1, pp. 414–427, 2008.

[4] B. Roberts, R. J. Summers, and P. J. Bailey, "The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes," *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1711, pp. 1595–1600, 2011.

[5] G. Kidd, C. R. Mason, V. M. Richards, F. J. Gallun, and N. I. Durlach, "Informational masking," *Auditory perception of sound sources*, pp. 143–189, 2008.

[6] K. Ueda, Y. Nakajima, W. Ellermeier, and F. Kattner, "Intelligibility of locally time-reversed speech: A multilingual comparison," *Scientific reports*, vol. 7, no. 1, pp. 1–8, 2017.

[7] K. Ueda, Y. Nakajima, F. Kattner, and W. Ellermeier, "Irrelevant speech effects with locally time-reversed speech: Native vs non-native language," *The Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. 3686–3694, 2019.

[8] S. Rosen, P. Souza, C. Ekelund, and A. A. Majeed, "Listening to speech in a background of other talkers: Effects of talker number and noise vocoding," *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 2431–2443, 2013.

[9] M. Cooke, "A glimpsing model of speech perception in noise," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[10] Y. Tang, M. Cooke *et al.*, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions," in *Proceedings of the Annual Conference of the International Speech Communication Association*. International Speech and Communication Association, 2016, pp. 2488–2492.

[11] K. Ueda and I. Matsuo, "Intelligibility of chimeric locally time-reversed speech: Relative contribution of four frequency bands," *JASA express letters*, vol. 1, no. 6, p. 065201, 2021.

[12] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler, "Release from informational masking by time reversal of native and non-native interfering speech," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1274–1277, 2005.

[13] B. Roberts and R. J. Summers, "Informational masking of speech by time-varying competitors: Effects of frequency region and number of interfering formants," *The Journal of the Acoustical Society of America*, vol. 143, no. 2, pp. 891–900, 2018.

[14] J. Chen, H. Li, L. Li, X. Wu, and B. C. Moore, "Informational masking of speech produced by speech-like sounds without linguistic content," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2914–2926, 2012.

[15] E. Rothauser, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, 1969.

[16] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[17] N. Iyer, D. S. Brungart, and B. D. Simpson, "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 2998–3010, 2010.

[18] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *The Journal of the Acoustical Society of America*, vol. 88, no. 4, pp. 1725–1736, 1990.

[19] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 585–592, 1995.

[20] X. Wu and J. Chen, "A computational model for assessment of speech intelligibility in informational masking," *Frontiers of Electrical and Electronic Engineering*, vol. 7, pp. 107–115, 2012.

[21] Y. Tang, T. J. Cox *et al.*, "Improving intelligibility prediction under informational masking using an auditory saliency model," in *Proceedings of the International Conference on Digital Audio Effects 2018*, 2018, pp. 113–119.

[22] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech communication*, vol. 41, no. 2-3, pp. 331–348, 2003.

[23] A. Edraki, W. Y. Chan, J. Jensen, and D. Fogerty, "A spectro-temporal glimpsing index (STGI) for speech intelligibility prediction," in *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. International Speech Communication Association, 2021, pp. 2738–2742.