A glimpsing model with a variable “tile” size

Christopher Conroy, Virginia Best, Todd R. Jennings, & Gerald Kidd, Jr.
Department of Speech, Language & Hearing Sciences
Hearing Research Center

BACKGROUND

- Ideal time-frequency segregation (ITFS or “glimpsing”) has been used in previous studies to separate the peripheral and central components of speech-on-speech (SOS) masking (e.g. Brungart et al., 2006).

- A core assumption underpinning the use of ITFS in this way is that it roughly emulates the effects of the peripheral component under a given stimulus configuration. Thus, the increase in masking observed for unprocessed stimuli relative to ITFS processed (“glimpsed”) stimuli can be attributed to the central component of masking.

- This difference, measured as the dB increase in target-to-masker ratio (T/M) at threshold for unprocessed stimuli relative to glimpsed stimuli under the same masking conditions, has been termed “additional masking” (Kidd et al., 2016).

- While this has proven a viable approach, a straightforward interpretation of the findings from previous studies is complicated by the fine time-frequency (T-F) resolution that was used during ITFS (i.e., a relatively “small” analysis “tile”), which may be incompatible with the internal T-F resolution of human observers. For example, Brungart *et al.* (2006) and Kidd *et al.* (2016) both used 128 frequency analysis bands (*n*=128) and 20-ms time windows (*m*=20) with 10-ms overlap.

- In the present study, we systematically varied the number and spectral width of the frequency analysis bands and the duration of the temporal windows that were used to generate glimpsed speech. **We asked:**
  - ○ How does reduced spectral and/or temporal resolution during ITFS influence the intelligibility of glimpsed speech?
  - ○ How does reduced spectral and/or temporal resolution during ITFS influence estimates of additional masking? What might changes in additional masking tell us about the internal “glimpse resolution” of human observers?

METHODS

Observers

- 6 normal hearing, including the first author (2 females; 18-29 years; mean=22 years)

Task

- Identification of a glimpsed target sentence

Stimuli

- Speech materials derived from BU Corpus (closed-set of 40 monosyllabic words organized into five syntactic categories with eight words in each category).
- Only female talkers used
- **Target:** Five-word sentence, cue word “Sue”; fixed at 55 dB SPL
- **Masker:** One of two types depending on condition
  - ○ *Speech maskers:* Two competing five-word sentences.
  - ○ *Noise maskers:* Speech-shaped, speech-envelope modulated noise; single-channel broadband envelope derived from an unused speech masker (i.e., 1-channel vocoder).

Conditions

- 12 total conditions
  - ○ Three *number of bands* conditions where *n*=8, 32, or 128
  - ○ Two *duration of time windows* conditions where *m*=20- or 80-ms
  - ○ Two *masker type* conditions where the masker was either speech or noise

Glimpsing model

- In general, same as Brungart *et al.* (2006) and Kidd *et al.* (2016)
- Each signal (target and masker) passed through a bank of *n* Gammatone filters (where *n*=8, 32, or 128), linearly spaced between 80 and 8000 Hz with overlapping passbands. Signal within each band further subdivided using *m*-ms time windows (where m=20 or 80) with 50% overlap. Result: two-dimensional matrix of tiles that varied in their spectro-temporal extent with condition.
- Ideal binary mask calculated using a local criterion (LC) of 0 dB S/N. S/N of each tile compared to LC. Target dominated tiles (S/N>LC) retained. Masker dominated tiles (S/N<LC) discarded. Retained tiles resynthesized to constitute experimental stimulus.

RESULTS

Threshold analysis

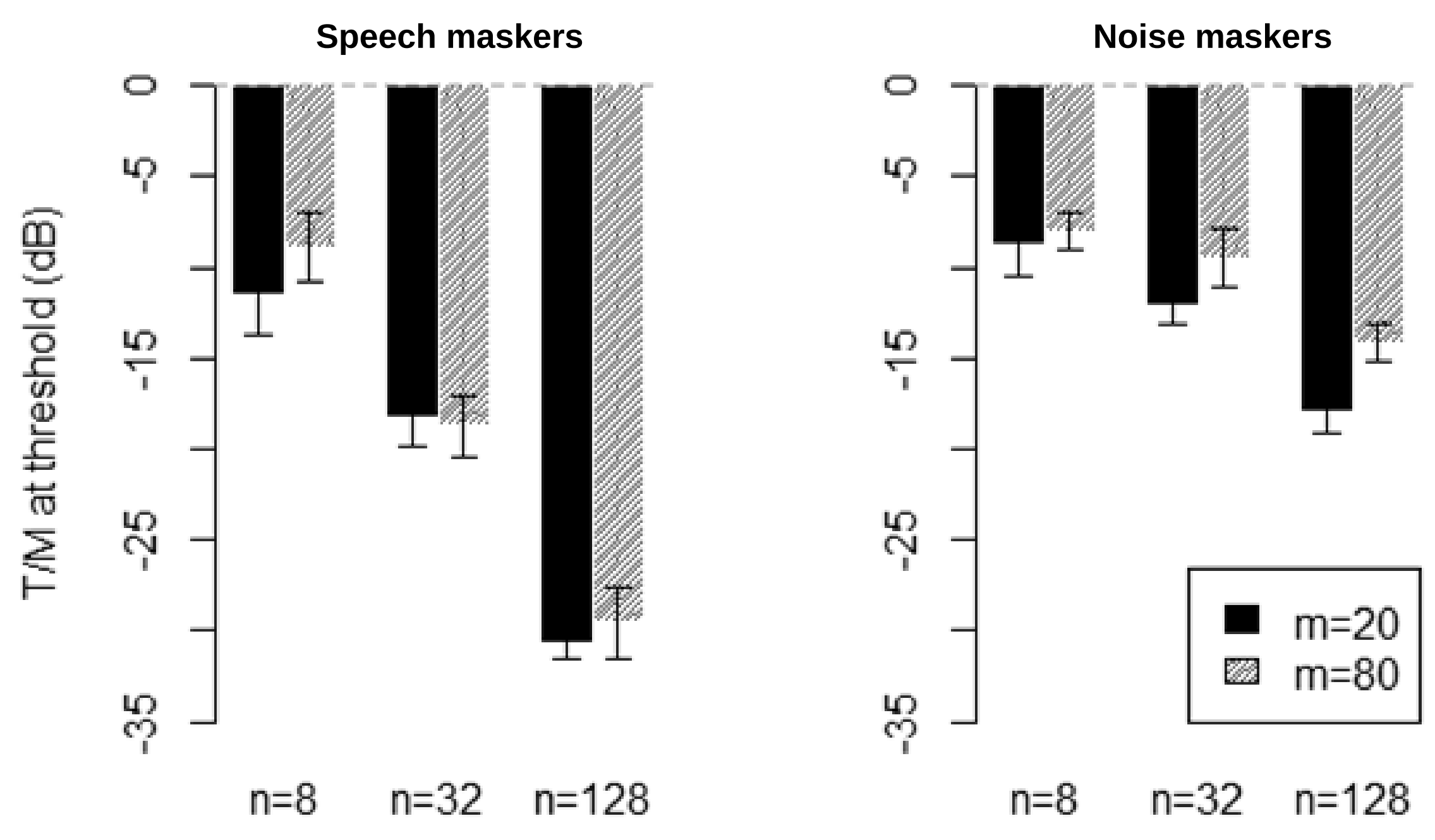


Figure 1. Group mean T/Ms at threshold for all conditions. Error bars are standard errors of the means.

- **For speech maskers:** Thresholds increased with each successive reduction in *number of bands* (p<.0001). No significant effect of *duration of time windows*.
- **For noise maskers:** Significant negative main effect of *number of bands* (p<.0001) and *duration of time windows* (p<.01). No significant effect of duration of time windows for any *n*<128.

Increase in masking re. reference condition & additional masking

Two additional measures of performance were calculated:
- ○ **Increase in masking**: Defined as an observer’s T/M at threshold in a given condition minus their T/M at threshold in the reference condition (defined as *n*=128, *m*=20).
- ○ **Additional masking**: Defined as an observer’s T/M at threshold in a given condition minus the group mean T/M at threshold in an unprocessed condition from Kidd *et al.* (2016, 2019).

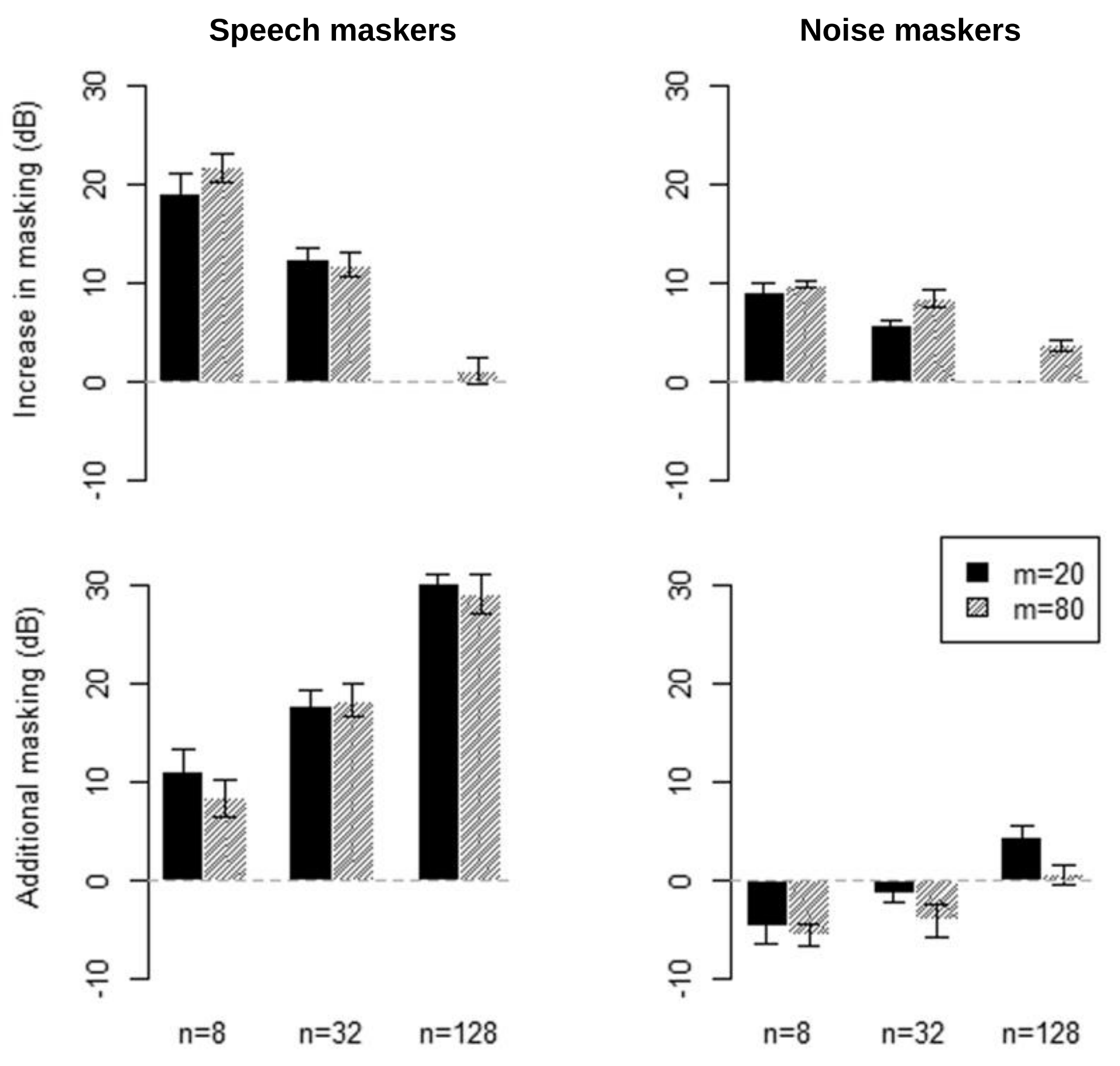


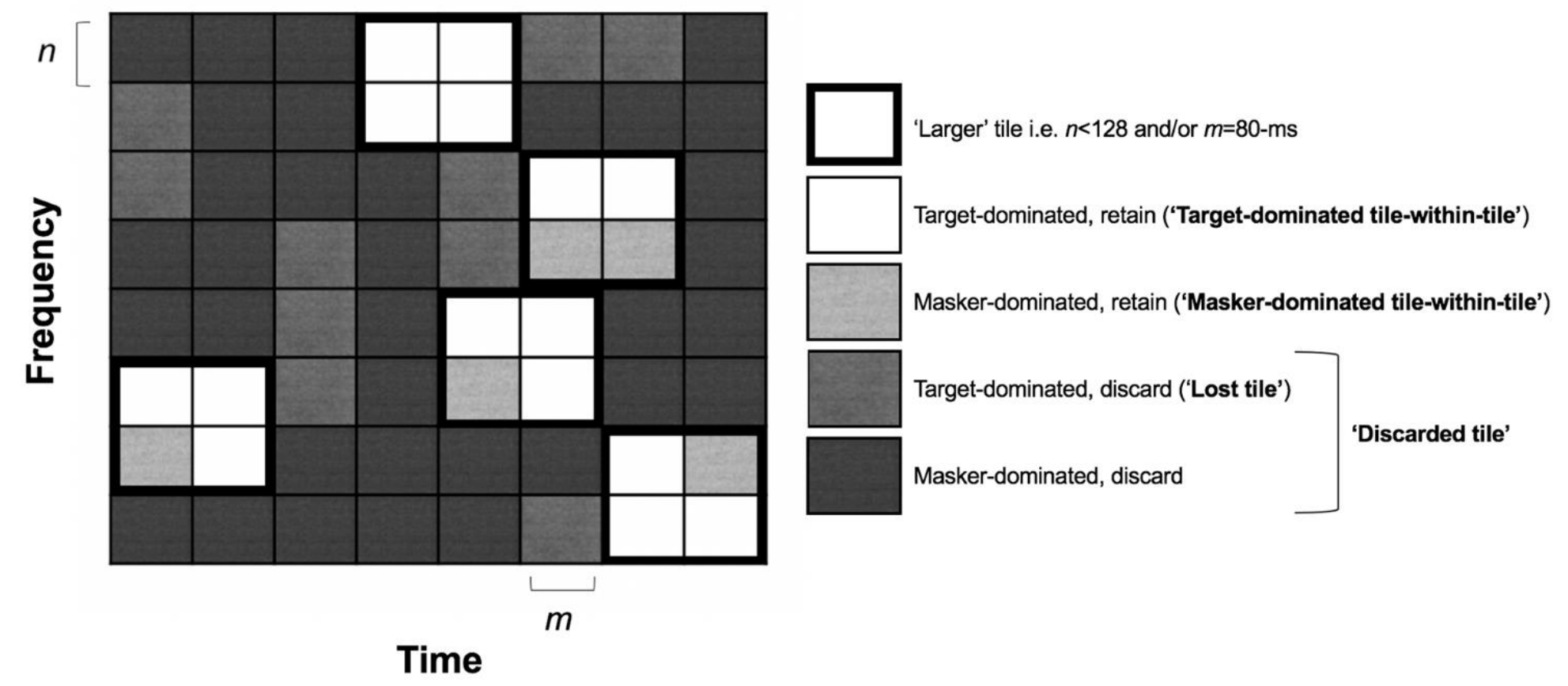
Figure 2. Group mean increase in masking (top row) and additional masking (bottom row)

- **Increase in masking:** There was a clear detrimental effect (increase in thresholds) of degraded spectral resolution when the masker type was speech and a relatively more restricted effect of both spectral and temporal resolution when the masker type was noise.
- **Additional masking**: Estimates of additional masking decreased with increasing tile size. For noise maskers, a large analysis tile resulted in “negative additional masking” indicating poorer performance than in unprocessed masked speech mixtures.

ACOUSTIC ANALYSIS

Why did thresholds increase? Informational masking or information loss?

Figure 3. Schematic showing hypothetical distribution of target and masker energy across T-F plane and how changes in tile size might affect within-tile distributions of target, masker, or both

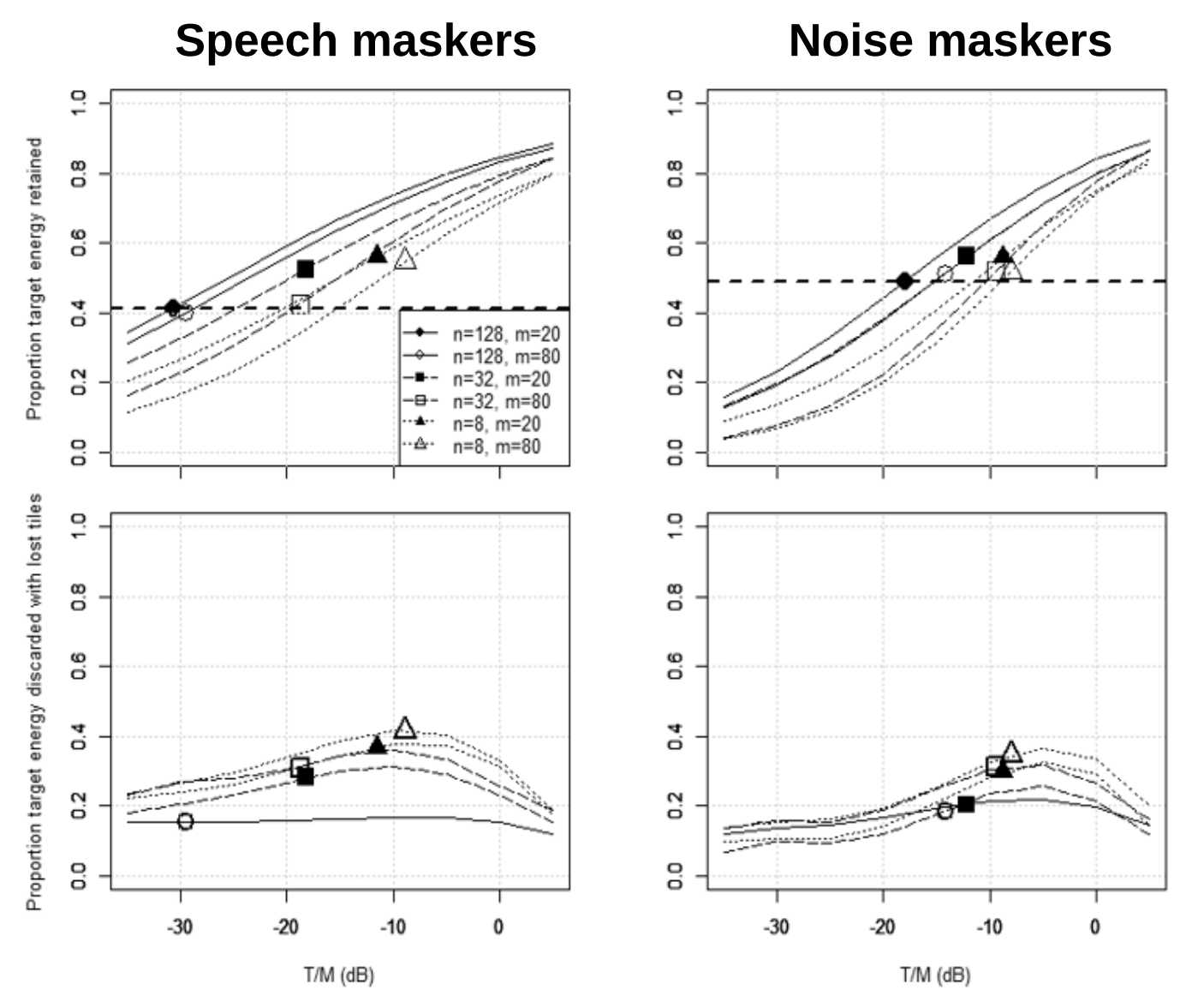

n
Frequency
m
Time
‘Larger’ tile i.e. n<128 and/or m=80-ms
Target-dominated, retain (‘Target-dominated tile-within-tile’)
Masker-dominated, retain (‘Masker-dominated tile-within-tile’)
Target-dominated, discard (‘Lost tile’)
Masker-dominated, discard
‘Discarded tile’

- Increased thresholds with increased tile size may reflect increased informational masking (e.g., interference from “masker-dominated tiles-within-tiles”) or increased energetic masking, when energetic masking is conceptualized as masking *imposed* by ITFS (e.g., “lost tiles”)
- We did not find evidence for increased informational masking (e.g., masker confusions at chance)
- Acoustic analysis (below) suggested that a loss of target information was responsible for increased thresholds

Increased tile size resulted in a loss of target information

Figure 4. Proportion of target energy retained (top row) and discarded with “lost tiles” (bottom row). The ordinate gives the ratio of energy retained/discarded to the energy of the unprocessed target. Symbols are located at the behavioral group mean T/M at threshold for the condition corresponding to that line.

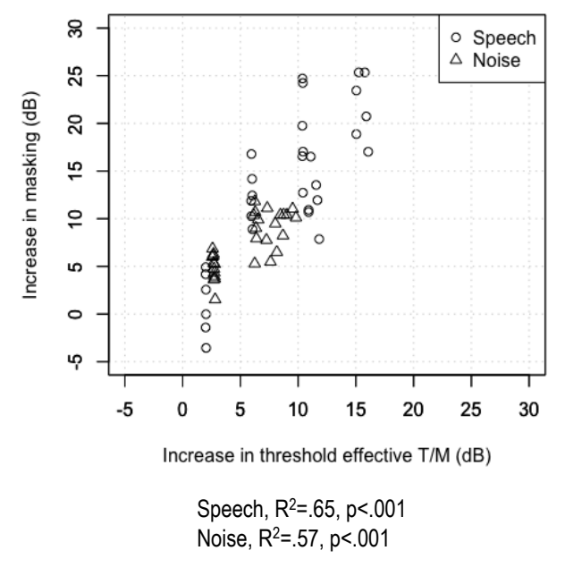
Top row: The point at which each line crosses the dashed horizontal line represents “threshold effective T/M”, the T/M in that condition at which the same proportion of target energy was retained as was retained at threshold in the reference condition (i.e., *n*=128, *m*=20). Note roughly 15 dB difference for tile size extremes.



Bottom row: Lost tiles represent discarded target information in each condition that *would have been retained* in the reference condition.

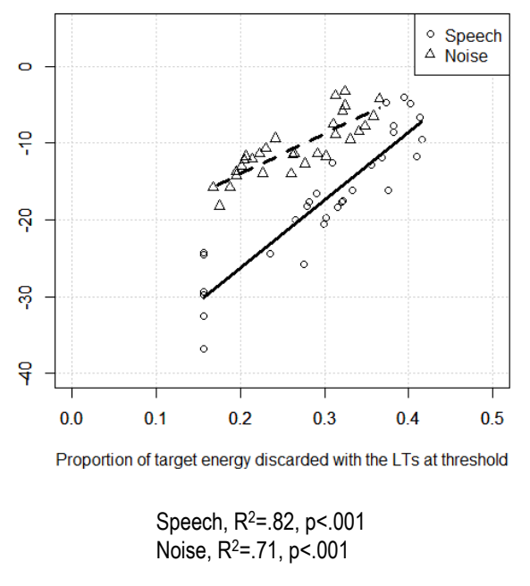
Loss of target information was predictive of performance

Figure 5. Individual increase in masking for each condition plotted against individual increase in threshold effective T/M in for the corresponding condition.



Speech, R^2=.65, p<.001
Noise, R^2=.57, p<.001

Figure 6. Individual thresholds for each condition plotted against individual estimates of proportion of target energy discarded with the lost tiles for the corresponding condition.



Speech, R^2=.82, p<.001
Noise, R^2=.71, p<.001

SUMMARY

- The number of spectral channels that are used during ITFS is the dominant tile size parameter determining the intelligibility of glimpsed speech. The duration of the temporal analysis window (within the bounds of those tested here) has a relatively minor influence on intelligibility.
- The application of ITFS using relatively few spectral channels (e.g., 8) and/or long temporal analysis windows can still provide substantial intelligibility benefits in high informational masking situations relative to unprocessed stimuli.
- Glimpsing models with slightly coarser T-F resolution may be more appropriate with respect to the internal resolution of human observers.
- These results may have implications for applications of ITFS to hearing impaired observers.

REFERENCES

- Brungart, D. S., Chang, P. S., Simpson, B. D., & Wang, D. (2006). Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am.*, 120(6), 4007-4018
- Kidd Jr, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *J. Acoust. Soc. Am.*, 140(1), 132-144.
- Kidd Jr, G., Mason, C.R., Best, V., Roverud, E., Swaminathan, J., Jennings, T., Clayton, K.K., & Colburn, H.S. (2019) Determining the energetic and informational components of speech-on-speech masking in listeners with sensorineural hearing loss. *J. Acoust. Soc. Am.* (in press)