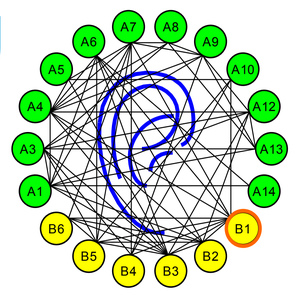
Speech intelligibility and detection in adverse monaural masking conditions: experimental and model data

Wiebke Schubotz (wiebke.schubotz@uni-oldenburg.de), Thomas Brand, and Stephan D. Ewert



This work is funded by the SFB/TRR 31 "The Active Auditory System"

Introduction

Speech intelligibility in everyday life is to a great extent ruled by the background noise (masker) in which a speech signal is perceived. There are three "classical"characteristics used when describing speech perception: **energetic** (**EM**), **amplitude modulation** (**AMM**), and **informational masking** (**IM**), see [1], [2], and [3]. It is thought that all occur to a different degree within a background noise, but it is not easy to entangle the individual contributions. While EM and AMM are determined directly by the relation between the speech and masker signal within the auditory filters (in terms of SNR or modulation frequency), informational masking supposedly occurs outside the periphery of the auditory system. [4] states that IM is influenced by the similarity between target and masking background, as well as uncertainty within the masker. It is thought that IM is most prominent in speech-on-speech masking. This study investigated the influence of the different masking characteristics to speech intelligibilty and detection and compared the obtained speech reception thresholds to the outputs of four speech prediction models ([7], [8], [9], and [10]).

Methods

Speech material:

We used the Oldenburger Satztest (OLSA, [5]) with a male and female speaker. This is a matrix sentence test with a fixed (non-sense) sentence structure (see Fig.1).

Name	Verb	Zahlwort	Adjektiv	Objekt
Peter	bekommt	drei	große	Blumen.
Kerstin	sieht	neun	kleine	Tassen.
Tanja	kauft	sieben	alte	Autos.
Ulrich	gibt	acht	nasse	Bilder.
Britta	schenkt	vier	schwere	Dosen.
Wolfgang	verleiht	fünf	grüne	Sessel.
Stefan	hat	zwei	teure	Messer.
Thomas	gewann	achtzehn	schöne	Schuhe.
Doris	nahm	zwölf	rote	Steine.
Nina	malt	elf	weiße	Ringe.

Fig.1: Sentence matrix (figure taken from the OLSA hand book).

Masking backgrounds:

We used eight different maskers, four speech shaped noise (SSN)- based and four speech-like maskers. All had the same long-term energy spectrum, but changed consecutively in the spectro- temporal domain to address the individual masking characteristics. **The maskers varied in their:**

- coherence of the applied modulations (co-modulation vs. across-frequency shifted (AFS) modulations)
- regularity of the modulations (8Hz (SAM) vs. broadband (BB) speech modulation)
- number of interfering talkers (ISTS [6] vs. single talker (ST))
- presence of fundamental frequency information (intact vs. noise-vocoded (NV) speech-like masker)

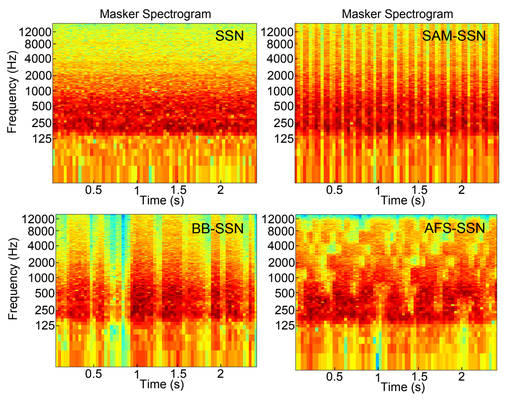


Fig.2: Spectro-temporal representation of the SSN-based maskers.

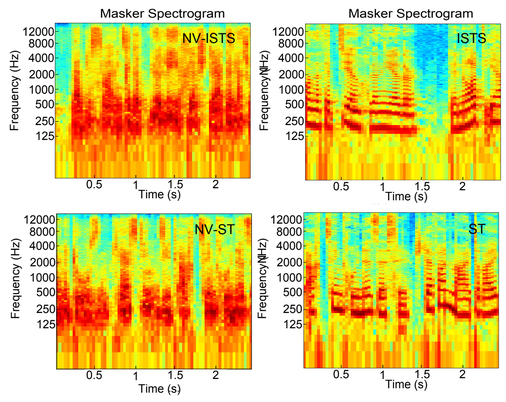


Fig.3: Spectro-temporal representation of the original and noise-vocoded speech-like maskers.

Procedure:

Speech intelligibility was measured with an adaptive procedure to obtain the speech reception thresholds $SRT_{50\%}$ and $SRT_{80\%}$. Speech Detection was measured with an 1up-2down alternative forced choice method.The stimuli were presented monaurally at a masker level of 65 dB to normal hearing listeners. We used different combinations of target and masker spectrum (male/female) to investigate how sensitive the measured thresholds are to such spectral changes. According to [11], speech detection is mostly ruled by EM, so comparing detection and intelligibility thresholds provides insight on the influence of AMM and IM.

References

[1] Durlach, N. I., Mason, C. R., Kidd, G., Jr., Arbogast, T. L., Colburn, H S., Shinn-Cunningham, B. G. (2003a). “Note on informational masking,” J. Acoust. Soc. Am. 113, 2984-2987.
[2] Dubbelboer, F., Houtgast, T. (2008). “The concept of signal-to-noise ratio in the modulation domain and speech intelligibility,” J. Acoust. Soc. Am. 124, 3937-3946.
[3] Micheyl, C. and Oxenham, A. J. (2010). "Pictch, harmonicity, and concurrent sound segregation: Psychoacoustical and neurophysiological findings," Hear. Res. 266, 36-51.
[4] Lutfi, R. A., Gilbertson, L., Heo, I., Chan, A., and Stamas, J. (2013). “ The information-divergence hypothesis of informational masking,” J. Acoust. Soc. Am. 134 (3), 2160-22170.
[5] Wagener, K., Brand, T., and Kollmeier B. (1999). „Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Design, Optimierung und Evaluation des Oldenburger Satztests,“ Z. Audiol. 38 (3), 86-95.
[6] Holube, I., Fredelake, S., Vlaming, M, and Kollmeier, B. (2010), "Development and analysis of an International Speech Test Signal (ISTS)", Int. J. Audiol., 49, 891-903.
[7] ANSI. (1997). ANSI S3.5-1997. Methods for the Calculation of the Speech Intelligibility Index (American National Standards Institute, New York).
[8] Rhebergen, K.S., Versfeld, N.J., and Dreschler, W.A. (2006), “Extended speech intelligibility index for the prediction of the speech reception threshold in uctuating noise,” J. Acoust. Soc. Am. 120, 3988-3997.
[9] Jörgensen, S., Ewert, S.D., Dau, T. (2013), "A multi-resolution envelope-power based model for speech intelligibility", J. Acoust. Soc. Am, 134, 436-446.
[10] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J.(2011).\An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech.", IEEE, Vol.19, No.7, 2125-2136.
[11] Arbogast, T. L., Mason, C., R., and Kidd, G. Jr. (2005). "The effect of spatial separation of informational masking of speech in normal-hearing and hearing-impaired listeners,"J. Acoust. Soc. Am. 117 (4), 2169-2180.

Results

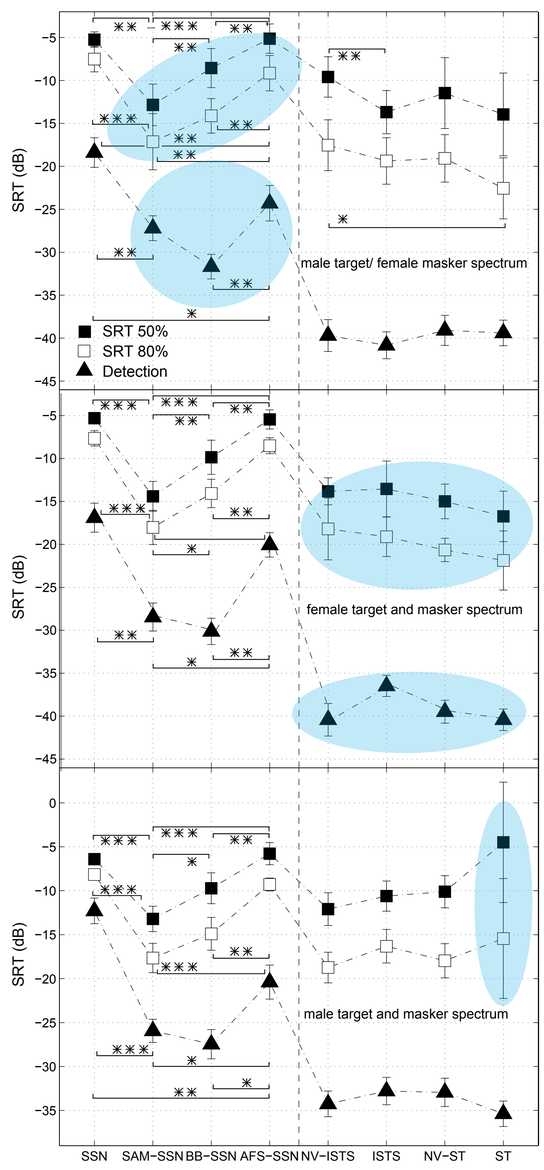


Fig 4: Measured speech intelligibility and detection thresholds with corresponding standard deviations for the three different combinations of target and masker spectrum. Comparison between detection and intelligibility data gives insight on the role of energetic masking. Stars indicate statistical significances of p<0.05 (*), p<0.01 (**), and p<0.001 (***).

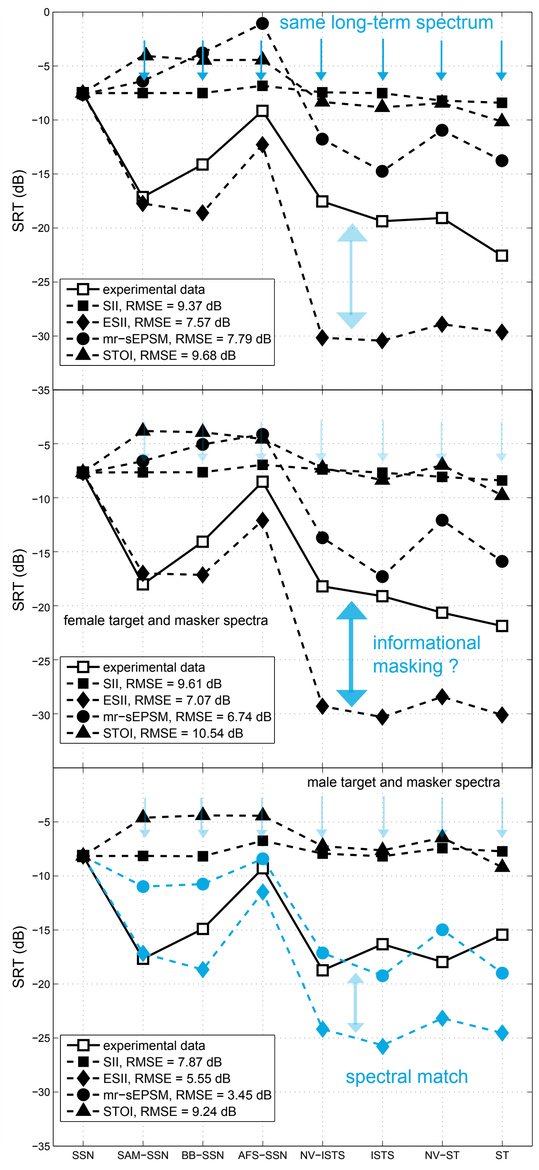


Fig 5: Experimental data and model predictions for the $SRT_{50\%}$, including the model's root mean square errors. Experimental data is depicted with open symbols, model data with filled symbols. All predictions are matched to the SSN masking condition.

- Highest masking thresholds arise for SSN and AFS-SSN.
- Modulations introduce a release from masking that grows with coherence and regularity of the applied modulations.
- All speech-like maskers show similar intelligibility thresholds, suggesting a similar amount of IM.
- Thresholds for the three spectral combinations are very similar.
- Detection thresholds are up to15 dB lower than intelligibility thresholds.
- The course of the intelligibility and detection thresholds is different, suggesting indeed an influence of AMM and IM for the intelligibility measurement.

- All model predictions are similar across the three combinations of spectra.
- SII predictions hardly vary according to the same long-term spectrum for all maskers.
- ESII shows an offset for speech-like maskers (no prediction of IM ?).
- Accuracy of the predictions by the mr-sEPMS model increases with spectral match of target and masker.
- STOI predictitions generally overestimate all masking effects.
- Predictions are generally better for the case of matched target and masker spectrum.

Summary & Conclusion

- Speech detection thresholds are well below intelligibility thresholds and show a similar, but not identical shape. Those differences suggest that AMM and IM work in addition to EM.
- Energetic masking seems to have the highest masking effect. If a hierarchy of the masking characteristics would be drawn, it is EM, AMM and IM.
- A release from masking is found for the modulated SSN-based maskers. For the BB- and AFS-SSN conditions this could be a co-modulation masking release effect.
- Informational masking thresholds show no influence of F0 removal or number of interfering talkers. This could be caused by an interplay of uncertainty within the masker and similarity between target and masker. Both aspects would work contrarily, but add up to the same amount of IM.

- There is no "Top Model" that represents all experimental data well. Moreover, more complex models do not yield more precise predictions (despite many free parameters).
- SII predictions are a "proof of concept" for the masker manipulations.
- ESII predictions fit the listerner's data well, suggesting that the auditory system performs a short-time analysis of an auditory signal. Predictions do not account for IM.
- Despite claims in [10], STOI does not work well in additive noise conditions.
- Other reference frames than SSN are well possible. This very much improves the accuracy of the predictions by [9].