Energetic and Semantic Models of Intelligibility

Juan-Pablo Ramirez, Alexander Raake

Deutsche Telekom Laboratories, Berlin Institute of Technology

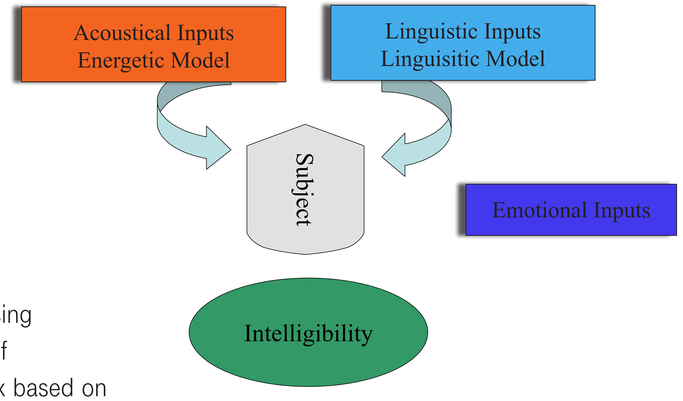
juan-pablo.ramirez@telekom.de

Deutsche Telekom Laboratories
An-Institut der Technischen Universität Berlin

Introduction

In realistic sceneries, listening is generally impaired by the presence of non-stationary masking sources. Traditional models of intelligibility hardly address to the temporal aspects of masking. Additionally, for competing speech the energy of the masker is unequally distributed in time. The present study some observations based on the predictions of an automatic speech recognizer (ASR-p) in an attempt to isolate the energetic component of masking.

A relevant index of energetic masking is necessary before addressing the question of informational cues. The study proposes a corpus of variable semantic complexity evaluated on subjects and an index based on N-grams measured on the world wide web which is believed to correlate with semantic complexity.



Acoustical Inputs Energetic Model

- MFCC-based ASR on speech
- $p(q_t^c \mid x_t)$ is the score ot the correct phone (reference) at a time-frame t,
- the EM-model ASR-p is the average probabilities across the N time frames

$$ASR-p = \frac{1}{N}\sum_{t=1..N} p(q_t^{\,c} \mid x_t)$$

ASR-p onto the SII-scale for stationary Speech-Shaped Noise (SSN)

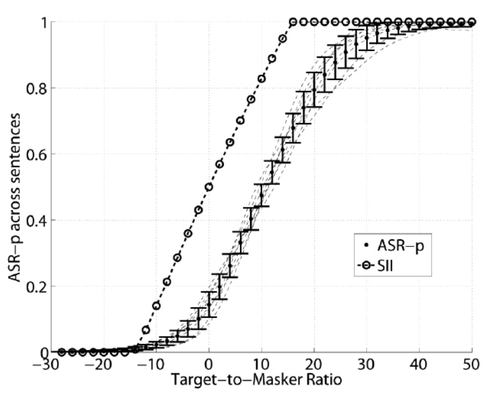


Figure 1: In open circles is the SII. Dots show the average ASR-p.

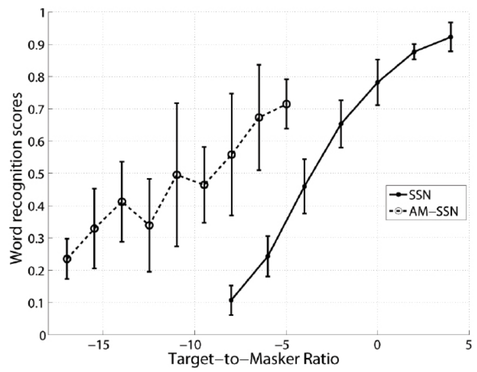


Figure 2: Intelligibility scores for speech masked by resp. SSN and AM-SSN.

Masking variability and the ASR-p

- ASR-p on conditions similar to Figure 2.
- Fluctuations observed.
- ASR-p speaker-independant

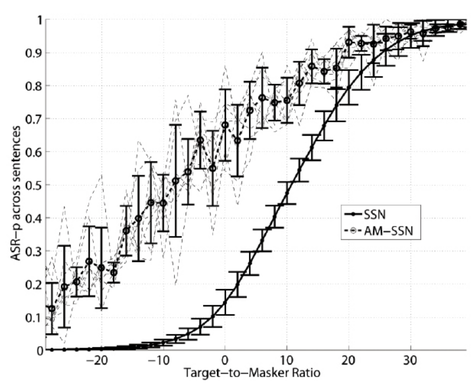


Figure 3: ASR-p for speech masked by resp. SSN and AM-SSN

Linguistic Inputs Linguisitic Model

The Web-inspired Sentence Complexity Index

- Calcultation of N-grams based Google search results

$$WiSCI(w_1 \ldots w_N) = \sum_{i=1..N-2} \log_{\mathbf{0}} \left[c_{web}(w_i w_{i+1} w_{i+2})\right]$$

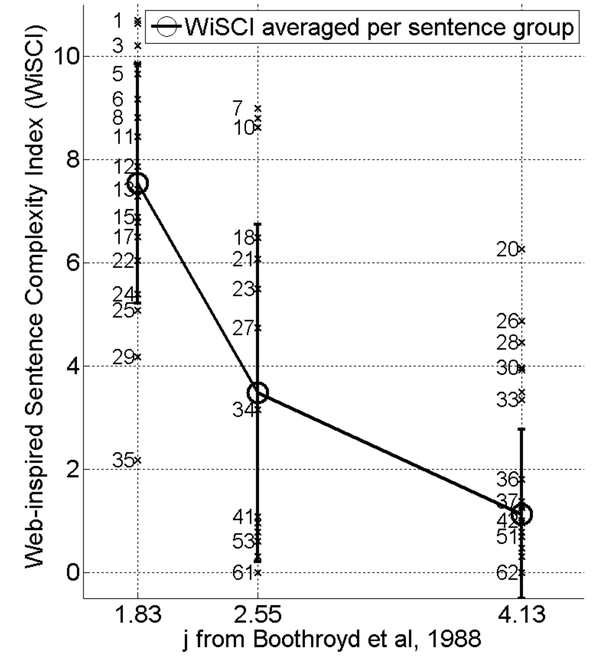


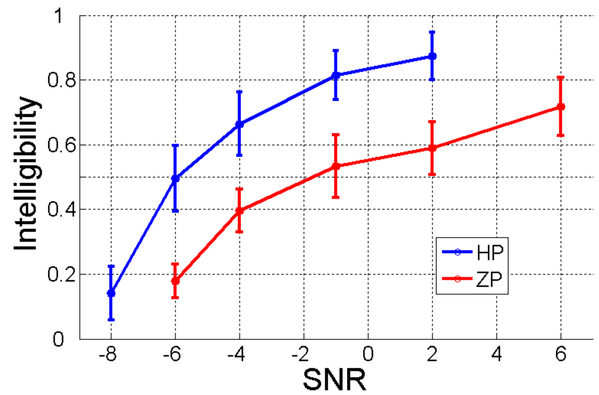
Figure 3: Google-based WiSCI for the corpus of Boothroyd.

Set of Semantically Controlled Sentences HP/ZP

- Inspired from Boothroyd and al. Jasa, 1988
- 60 Semantically Predictable HP
- 60 non-predictable sentences ZP
- 4 CVC each of English
- Manual phonetic annotation
- ASR trained on TIMIT database

Subjective test for AM-SSN

- 10 subjects for SSN
- 8 subjects for AM-SSN
- Corpus of German semantically unpredictable sentences
- 4 keywords per sentence
- 3 trials per levels



Subjective test HP/ZP

- 25 american subjects
- Masking by SSN
- j_{hp} of 3
- j_{zp} of 4

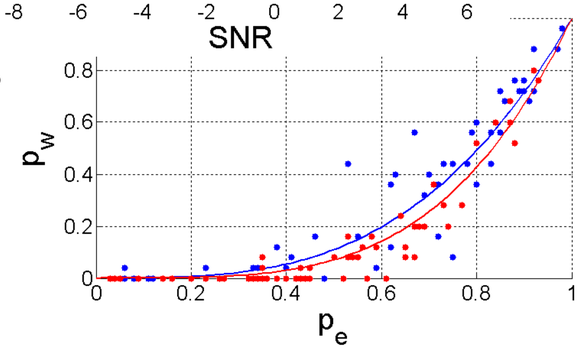


Figure 4a) and 4b): Intelligibility scores and semantic index j for HP/ZP corpus.

Conclusion

Our on-going work aims at predicting subjects performances in complex listening scenari for non-stationary masking and for sources of controled semantic similarity.