Speech separation: human single-channel and spatial performance

A.W. Bronkhorst

TNO Human Factors, The Netherlands

Human speech separation

Hearing impairment
(Plomp, Pavlovic)

Dip listening
(Festen)

Contextual information
(Boothroyd, Bronkhorst)

Talker characteristics
(Florentine & Buus, Bradlow, van Wijngaarden)

Masking
(Miller, French & Steinberg)

CP effect
(Cherry, Carhart)

Segregation, streaming
(Bregman, Darwin, Brokx & Nooteboom)

Room acoustics
(Houtgast & Steeneken)

Binaural unmasking
(Licklider, Levitt & Rabiner)

Informational masking
(Carhart, Kidd, Brungart, Freyman)

Attentional resources
(Cherry, Broadbent, Treisman)

Outline

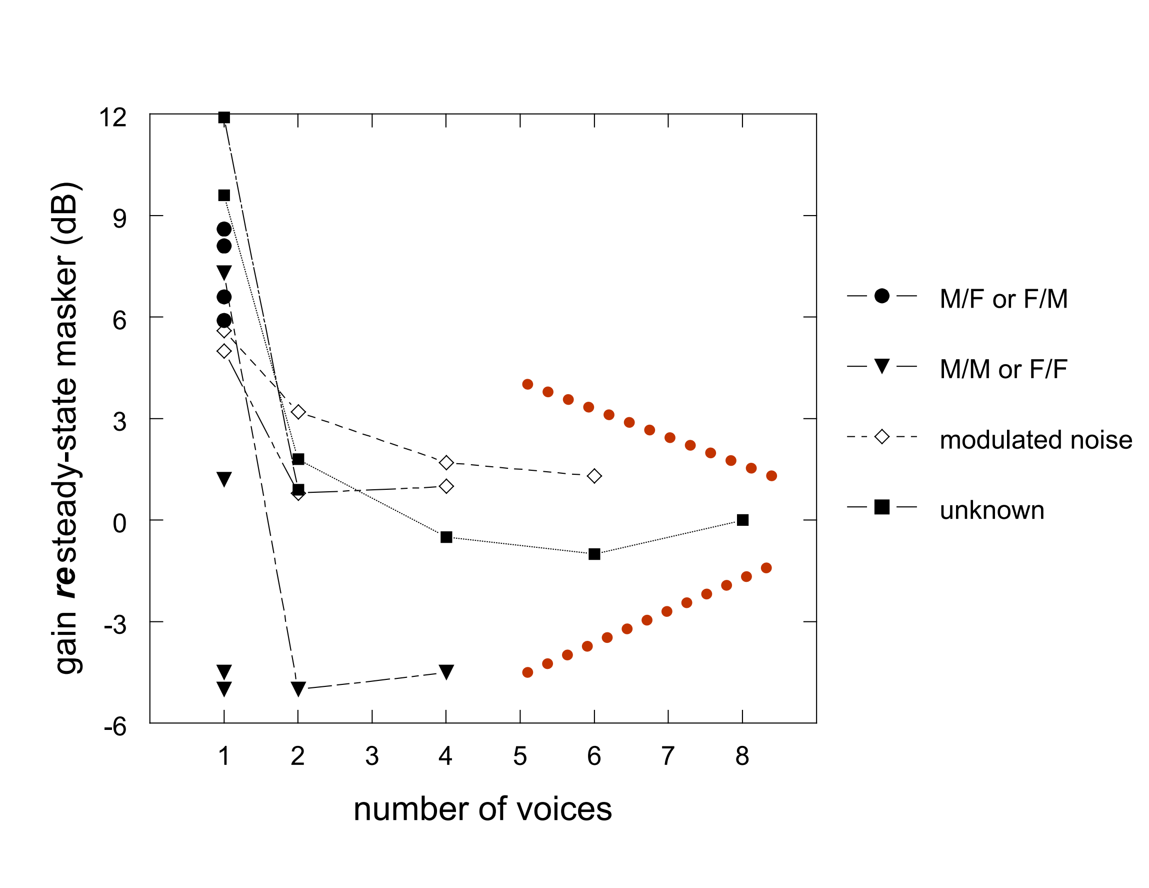
- **How can factors be modeled?**
  - *Prediction of speech intelligibility, often no useful for machine separation*
- **Single-channel speech separation**
  - *Type of interference*
  - *Energetic vs. informational masking*
  - *Reverberation, talker characteristics*
- **Spatial performance**
  - *Single source*
  - *Multiple sources*
  - *Informational masking*
- **Conclusion**

Single-channel speech separation (1)

- **Interference is noise**
  - *Old line of research, resulted in Articulation index*
    - Contribution in frequency band is proportional to SNR
    - Frequency bands can be combined in weighted sum
      - depends on speech material
    - Nonlinear relationship between AI and % correct
      - depends on speech material (e.g. contextual information)
  - *Recent developments*
    - Prediction for low-bitrate channels (PESQ, Beerends, $$$)
    - Improvement of prediction for non-smooth noise spectra
      - Modified STI (Steeneken); Speech Recognition Sensitivity (SRS) model of Müsch & Buus
    - Modeling of context effects
      - SRS model, context model of Bronkhorst et al.

Single-channel speech separation (2)

- **Interference is speech(like)**
  - *Strong effect of type of masker*
    - noise/voice
    - same/different sex
  - *Interaction with number of maskers*

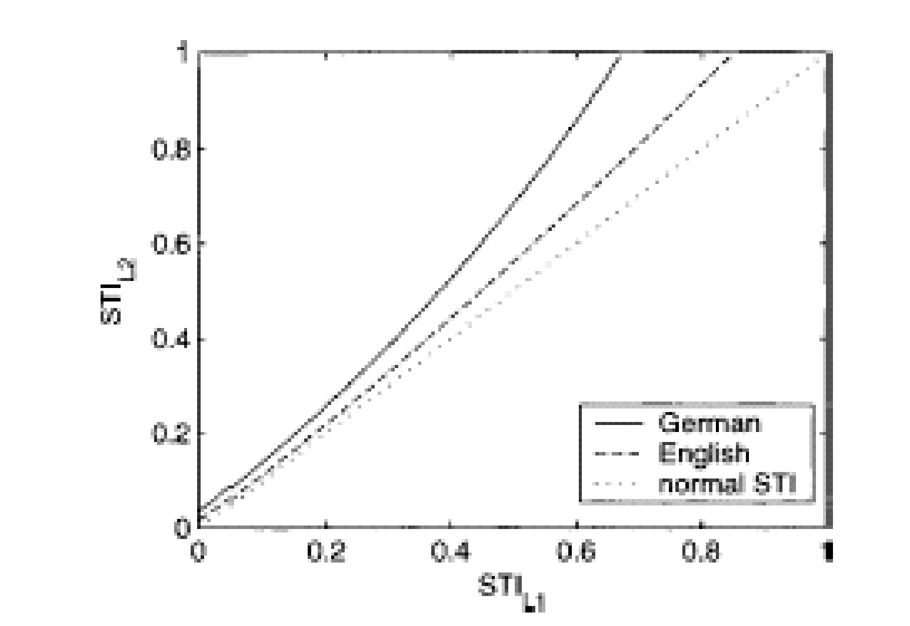


Single-channel speech separation (3)

- **Energetic vs. informational masking**
  - *Energetic masking*
    - Occurs during encoding, cannot be resolved by an “ideal” listener
    - Can be modeled using current knowledge of auditory system
      - problem: dip listening / contextual information
  - *Informational masking*
    - “The rest”
      - stimulus and/or masker uncertainty
      - at different processing levels (phonetic, semantic)
    - Occurs only when target and interferer are similar
      - studies use very specific material
    - Large inter-individual differences, effects of training and a-priori information
    - Shallow psychometric functions
    - Difficult to model

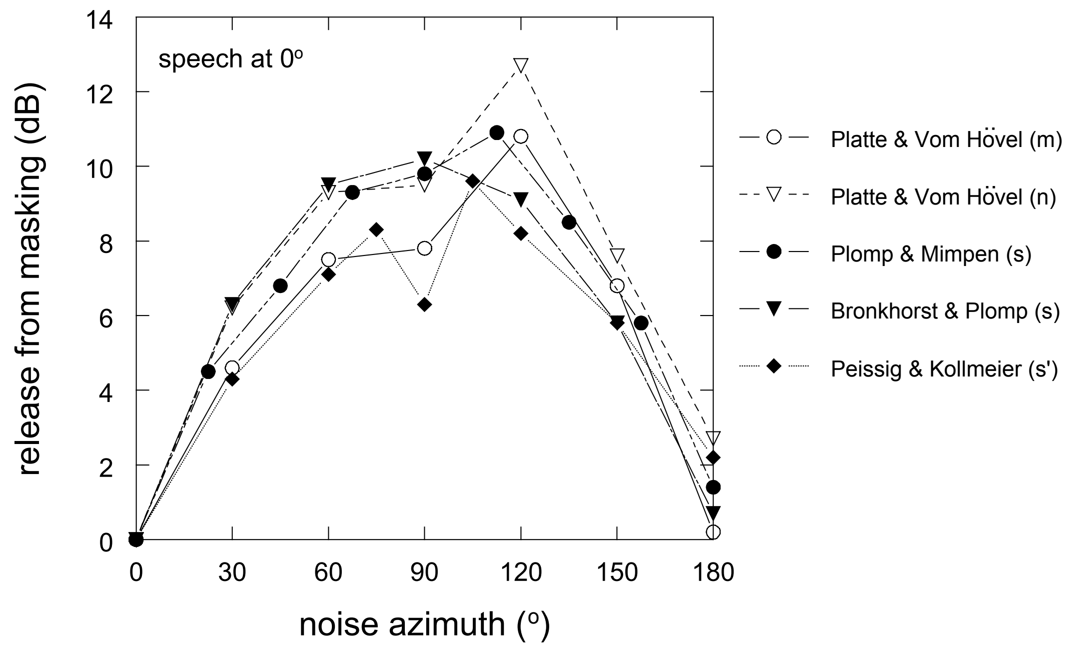
Single-channel speech separation (4)

- **Other factors**
  - *Reverberation*
    - Can be adequately modeled by STI
      - treatment of frequency domain similar to AI
      - Modulation Transfer Function (MTF) integrates effects of noise and reverberation
  - *Talker characteristics*
    - Effects are difficult to model
    - Speech perception in noise (SRT) can be used as measure of talker proficiency
    - Can be incorporated in STI (van Wijngaarden et al., 2004)



Spatial performance (1)

- **Single noise source**
  - *Combination of best-ear (ILD) and binaural (ITD) listening*
  - *Can be modeled quite well (vom Hövel, 1984; Zurek, 1990)*
  - *Strong effect of acoustic environment*



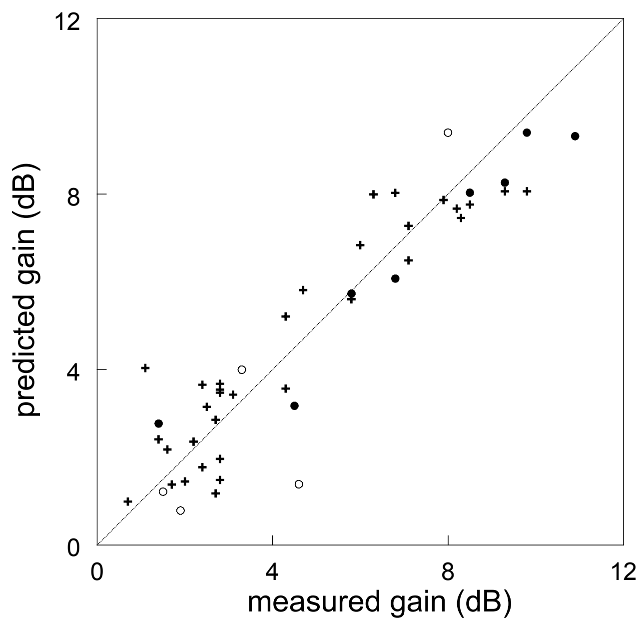
Spatial performance (2)

- **Multiple noise sources**
  - *Binaural gain generally decreases, depending on source configuration*
  - *Modeling: extended single-source model*
- **Multiple speech(like) sources**
  - *Same effects as in single-channel case*
    - dip listening
    - strong influence of type of interferer
  - *Indication that binaural release is largest for 2-3 interferers (Hawley et al., 2004)*

Simple descriptive model (Bronkhorst, 2000)

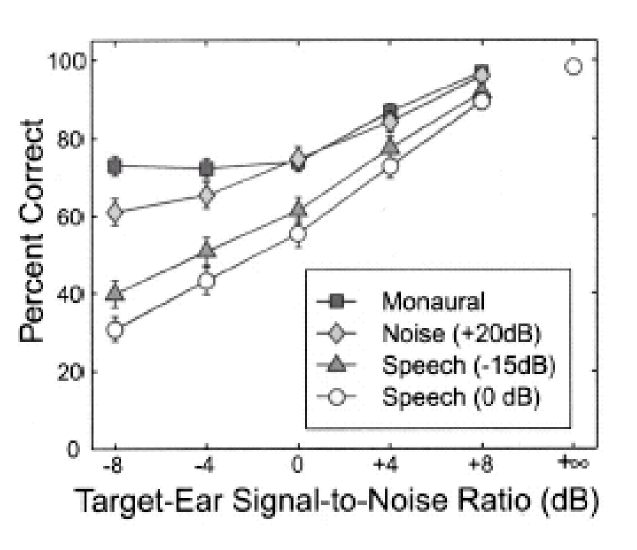
$\alpha = 1.4; \beta = 8$

$$R=\left[\alpha\left(1-\frac{1}{N}\sum_{i=0}^{N}\cos\theta_i\right)+\beta\frac{1}{N}\left|\sum_{i=0}^{N}\sin\theta_i\right|\right].$$



Spatial performance (3)

- **Informational masking**
  - *Spatial release from masking*
    - Can be much larger than the release for energetic masking (Arbogast et al., 2002)
    - Can occur in conditions where there is no release from energetic masking
      - due to a difference in perceived location (Freyman et al., 1999, 2001, 2004)
  - *Limited attentional resources*
    - Demonstrated in “classical” shadowing experiments (e.g. Wood & Cowan, 1995)
    - Large effect of contralateral distracter in CRM task (Brungart & Simpson, 2002)
    - Better monaural than binaural performance in speaker recognition task (Drullman & Bronkhorst, 2000)



Conclusion

Good progress

Dip listening
(Festen)

Masking
(Miller, French & Steinberg)

Room acoustics
(Houtgast & Steeneken)

Binaural unmasking
(Licklider, Levitt & Rabiner)

CP effect
(Cherry, Carhart)

Difficult

Contextual information
(Boothroyd, Bronkhorst)

Talker characteristics
(Florentine & Buus, Bradlow, van Wijngaarden)

Segregation, streaming
(Bregman, Darwin, Brokx & Nooteboom)

Informational masking
(Carhart, Kidd, Brungart, Freyman)

Attentional resources
(Cherry, Broadbent, Treisman)

No problem for machines