Speech recognition in noise

- **Spectral properties of the noise**: white, pink, speech-shaped, competing speech, speech babble
- **Temporal properties of the noise**: steady vs. modulated or interrupted

Effects of noise on vowel spectra

- Broadband noise tends to fill up the valleys between the formant peaks.
- **Spectral contrast** (peak-to-valley ratio) is reduced by the addition of noise.
- Because of the sloping long-term *spectrum* of speech, the upper formants (F3, F4, F5) are more susceptible to masking and distortion by the noise.

Vowel /i/ in quiet and in noise

Effects of noise on formant peaks

- (a) Vowel in quiet
- (b) Vowel + white noise
- (c) Vowel + multi-talker babble
- (d) Vowel + multi-talker babble
Resistance to distortion

- **Articulation score**: % items correct on spoken lists of syllables, words or sentences
- **Signal-to-noise ratio (SNR)**: when speech and noise have the same average rms level (SNR=0 dB), articulation scores are above 50% for listeners with normal hearing

Effects of noise on speech recognition

![Graph showing effects of noise on speech recognition](image1.png)

Source: Miller, Heise and Lichten, J. Exp. Psychol. 1951

Effects of context

- **Contextual cues lead to improved speech understanding in noise.**
  - Acoustic-phonetic context
  - Prosodic context
  - Semantic context
  - Syntax

![Graph showing effects of context](image2.png)

Miller, Heise & Lichten, 1951

Effects of filtering on speech

- **Low-pass filtering** to remove frequencies above 1800 Hz reduces intelligibility from near perfect to around 67%.
- **High-pass filtering** to remove components below 1800 Hz also produces about 67%.

![Graph showing effects of filtering on speech](image3.png)

Miller, Heise & Lichten, 1951
High-pass and low-pass filtering

Other frequency distortions

- Bandpass filtering with one-third octave filters centered 1500-2100 Hz produces better than 95% accuracy for high-predictability sentences (Warren et al., 1995; Stickney & Assmann, 2001).

Other frequency distortions

- Notch filtering to remove frequencies between 800 and 3000 Hz leads to consonant identification scores better than 90% (Lippman, 1996)
- Conclusion: speech cues are widely distributed

Perception of filtered speech

- Everyday English sentences filtered using narrow bandpass filters remain highly intelligible (>90% words correct)
  - one-third octave bandwidth, 1500 Hz center frequency, 100 dB/octave slopes

Warren et al. (Percept Psychophys 1995; JASA 2000)

Perception of filtered speech

Bashford et al. (2000)
Interrupted noise

- When the noise is **intermittent** rather than **continuous** there is a **release from masking**.
- The benefits of non-stationarity depend on the interruption rate and the **duty cycle** (on-off ratio) of the noise.

Interrupted noise

- When a noise masker is alternated with silence using a 50% duty cycle, there may be considerable masking release, compared to a continuous masker, especially with alternation rates between 1 and 200 per second (Miller and Licklider, 1950).

Interrupted speech

- In quiet, **speech** can be interrupted (turned on and off) periodically without loss of intelligibility (Miller and Licklider, 1950).
- Miller and Licklider found the worst intelligibility for interruption rates < 2 Hz, where large speech fragments (words, phrases) are missing.

Interrupted noise

1. At alternation rates between about 1 and 200 per second, listeners can “patch together” cues from the clean segments between the bursts of noise.
2. With slower interruption rates, entire words or phrases are masked; others are noise-free.
3. At rates > 200/sec the masking effect is the same as uninterrupted, continuous noise.
Interrupted speech

• Miller and Licklider found improved performance for interruption rates between 10 and 100 Hz. Why?
• For very high interruption rates (>1 kHz) the signal sounded continuous, and performance was near perfect.

16 Hz 128 Hz 512 Hz

Interrupted noise

1. At alternation rates between about 1 and 200 per second, listeners can “patch together” cues from the clean segments between the bursts of noise.
2. With slower interruption rates, entire words or phrases are masked; others are noise-free.
3. At rates > 200/sec the masking effect is the same as uninterrupted, continuous noise.

Interrupted speech

“This Picket-fence” effect

• Interrupted speech can have a harsh, distorted quality.
• But when speech and noise are alternated periodically, filling silent gaps with noise, the speech sounded smooth and continuous.
• Possibly, noise in the gaps enhances the listener’s ability to exploit contextual cues.

Howard-Jones and Rosen (1993)

• “Checkerboard” noise maskers
• Listeners can exploit asynchronous time-frequency glimpses, but only over broad frequency ranges
A glimpsing model of speech perception in noise

Martin Cooke


“Glimpsing” speech in noise

• “speech is a highly modulated signal in time and frequency, regions of high energy are typically sparsely distributed.”

Glimpsing speech in noise

• Can listeners take advantage of “glimpses”?  
  ■ direct attention to spectrotemporal regions where the S+N mixture is dominated by the target speech  
  ■ ASR system trained to recognize consonants in noise  
  ■ Maskers differed in “glimpse size”  
  ■ ASR model developed to exploit non-uniform distribution of SNR in different time-frequency bands  
  ■ Conclusion: model + listeners benefit from glimpsing.

Speech + noise mixtures

• Some regions dominated by target voice  
  ■ Local SNR varies across time and frequency  
  ■ Where the target voice dominates, the problem of source segregation is solved because the signal is effectively “clean” speech.  
  ■ Clean speech is highly redundant; it remains intelligible after 50% or more of its energy is removed by gating and/or filtering

Sparseness and redundancy

• Glimpses = spectrotemporal regions where signal exceeds masker by ~3 dB.

Results

FIG. 6. Model and listeners’ consonant identification rates for the glimpses-only model at a local S/N detection threshold of 6 dB. Listeners’ data here and in Fig. 9 are the same as that of Fig. 3.
Conclusions

- A higher “glimpse threshold” (e.g. local SNR > 0 dB) produces fewer glimpses, but this provide less distorted information than a lower threshold (e.g. -5 dB).

Speech source separation

- How do the ear and brain separate the target voice from the noise?
  - spatial cues
  - lip-reading
  - semantic context
  - auditory scene analysis (Bregman, 1990)
  - glimpsing and tracking

Speech + noise mixtures

- Some regions dominated by target voice
- Local SNR varies across time and frequency
- Where the target voice dominates, the problem of source segregation is solved because the signal is effectively “clean” speech.
- Clean speech is highly redundant; it remains intelligible after 50% or more of its energy is removed by gating and/or filtering

Periodicity and noise

- Hypothesis:
  - Periodicity of speech contributes to robustness
  - Harmonicity in the frequency domain
  - Across-frequency grouping of spectral features
  - Unvoiced sounds (e.g., whispered speech) are more susceptible to masking and interference by competing sounds