Connecting Information Theory, Auditory Modeling, and Perception of Speech

*The Channel Capacity-gram (CCgram)*

Bryce Lobdell and Jont B. Allen

Univ. of IL,

Beckman Inst., Urbana IL
Neural Correlates of Perceptual Cues

Distorted Speech (speech + noise) → Human Listener → catagorical perception confusion matrix $P_{\text{heard}=i,\text{spoken}=j}(SNR)$

Closed Set of consonant–vowel nonsense speech sounds → Auditory Model → Detection Analysis

CCgram → "information density" in neural domain → Neural Correlates of Perceptual Categories

Goals

- Quantify neural time-frequency information using the CCgram.
- Quantify perceptual confusions in the form of a confusion matrix, as a function of SNR ($P_{\text{spoken}=i,\text{heard}=j}(SNR)$).
- Connect these two information theoretic measures.

Tough questions:

- What detection tasks are easy to test on humans and resemble speech?
- What is the brain doing in these detection tasks beyond the level of auditory nerve firing rates?
Filter widths are chosen based on the human critical ratio (Fletcher), with the intent that they will provide the correct tone detection threshold.
Parallel Processing Detection Model

\begin{itemize}
\item Given a model of tone detector (Munson 1947), determine $\alpha_{tone}$ corresponding to threshold.
\item Assuming the detection threshold ($\alpha_{speech}$) for narrowband speech in noise detection is the same as $\alpha_{tone}$, we may determine the threshold SNR in each band.
\end{itemize}
Left: The probability density at (vi) for tone+noise near the detection threshold for a 1 kHz tone. The ML criteria was used to implement the tone detector after (v) and (vi), resulting in the right figure.

Right: Comparison of model performance at (v) and (vi) with human performance for a 1AFC tone detection task.
Scaling in JNDs

AGN Channel Capacity:
\[ R = \int f \log_2(1 + \text{SNR}^2(f)) df \]

Articulation Index (AI):
\[ \text{AI} = \int f \, w(f) \log(1 + c^2 \text{SNR}^2(f)) df \]

- The JND depends on the noise variance (spread of the signal PDF due to the noise).
- The AI is the integral over frequency of the time-average number of JNDs available, computed from the SNR (French and Steinberg 1947, Allen 2005).
- The channel capacity (of an AGN channel) takes a similar form to the AI, and can be interpreted as the integral over frequency of the number of “resolvable signal levels”, converted to bits.
Samples below $\alpha$ are discarded (stage (iii)).

Stage (iv) is scaled by the noise standard deviation $\sigma$, giving the AN rate in JNDs, then $\log_2(1 + a_i(t)/\sigma)$ converts JNDs to bits.

Stages (iii) to (vi) are independent of auditory model.
Discrimination of Speech Sounds

We have:

- The CCgram provides us with neural time-frequency information.
- The confusion matrix as a function SNR quantifies human detection of perceptual categories.

Goal: Connect these two information theoretic measures.
Examples

- /t/ is confused primarily with /p/ and /k/.
- The event (perceptual feature) which distinguishes /t/ from /p/ and /k/ is visible in the CCgram.
- /t/ in many cases will “morph” into /p/ at around 0 dB SNR. The disappearance of the bar highlighted in red occurs at roughly the same SNR where the identity of the sound changes.

Demo

- Hear truncation of the red bar, which also changes the identity of the sound from /t/ to /p/.
- Hear recordings of /t/ above and below the “morph” threshold SNR.
Example 1

\( P(h|s=/Taa/) \)

SNR = 0.00 dB

perceived as pa

SNR = 6.00 dB

perceived as ta

SNR = 12.00 dB

perceived as ta
Example 2

SNR = −2.00 dB
perceived as t(vowel)

SNR = −10.00 dB
perceived as p(vowel)

SNR = −16.00 dB
perceived as p(vowel)
Conclusions

An event (perceptual feature) has been identified via the confusion matrix and CCgram methodology.

Further experiments are necessary to prove the validity of the model events.

A rate $$r(t) = \int f \log_2 (1 + \text{SNR}(f, t)) df < 0.5 \text{ bits/cs}$$ seems to be the lowest recognition-useful rate.