Human Speech Perception and Feature Extraction

Bryce E. Lobdell (lobdell@illinois.edu),
Jont B. Allen (jontalle@illinois.edu),
Mark Hasegawa-Johnson (jhasegaw@illinois.edu)
Research Interest

The error patterns exhibited by humans are a consequence of the structure of the learning machine used by humans for phone classification.
Intelligibility Predictors

The Articulation Index predicts intelligibility in noise and filtering.

\[
AI_k = \frac{1}{3} \log_{10} \left( 1 + c^2 \frac{\sigma_{s,k}^2}{\sigma_{n,k}^2} \right) 
\]

\[
AI = \overline{AI_k}
\]

Filters were chosen to fit empirical data, and happen to have tuning proportional to human neural tuning.

The phone error rate is

\[
P_c(AI) = 1 - e^{\frac{AI}{c}}
\]

\( c \) and \( e_{\min} \) are parameters, fit to empirical data.
Intelligibility Predictors

Most surprising because of their generality over a variety of masking noise spectra:

Phone recognition accuracy.
Red = in speech spectrum noise, Blue = in white noise
Extending the Articulation Index

Prediction

more specific

Measurement (AI)

more specific

All Speech
"classic" AI

Phone Class

Predict Token from time/freq specific measure

Token

Pattern Recognition
Objective

• Determine the extent to which feature representation plays a roll in recognizer performance, and in similarity to human behavior.

• Determine whether an AI-based representation gives rise to generality with masking noise spectra.
Experiment

- Classify (consonant) phones using a pattern recognizer which makes as few assumptions as possible about the statistics of speech. A K-nearest neighbors classifier.

- Classify phones with a mis-match between the test and training noise spectra, which were speech spectrum noise and white noise.

- Data set: 16 consonants x 500 tokens = 8000 utterances
Features

Power Spectra-based Features

$$\log s_k(t)$$

Articulation Index-based Features

$$a_k(t) = \frac{1}{3} \log_{10} \frac{\mu_{n,k}^2 + (s_k(t) - \mu_{n,k})^2}{\mu_{n,k}^2}$$

$$\mu_{n,k} = \text{expected value of noise in channel } k$$

Spectral Subtraction-based Features

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2$$

$$\hat{S}(\omega) = (|Y(\omega)|^2 - |D(\omega)|^2)^{\frac{1}{2}} e^{j \angle Y(\omega)}$$
Similarity of confusions and accuracy vary by consonant.

Confusion patterns in white noise. Solid = human, dashed = power spectra, dotted = AI.
Generality with Masking Noise Spectra

(a) Trained and tested in white noise.

(b) Trained in speech spectrum noise, tested in white noise.

(c) Trained in white noise, tested in speech spectrum noise.

(d) Trained and tested in speech spectrum noise.
Conclusions

• Feature type plays a role even in a KNN recognizer
• The Articulation Index-based features provide the highest accuracy when SNR is low and the testing/training noise spectra are mismatched.

Questions and Future Work

• Why do AI-based features (sometimes) perform better?
• Why is HSR recognition accuracy related to the AI?
• What can be done with the mountain of confusion data?
• Can they be employed in a practical recognition system.