Columbia: Recent + Future

- More information
  - FDLP / PLP2 features
- Better classifiers
  - MI-based broad-class experts
- Reducing variability
  - Temporal variation
  - Formant “automatic gain control” (AGC)
- Signal model
  - “Deformable spectrograms”
**Broad-Class Experts**
Patricia Scanlon

- MI-based feature masks make superior class-specific classifiers (vowels, stops...)
- smaller models: good for data-limited case
- Apply to ASR by ‘patching in’ probabilities via separate broad-class center detector
MI-Based Class Experts

- **Idea:** Different speech sounds have different information distribution
  - ...as identified by MI to phone | class

- **Good for reducing model complexity**
  - benefits disappear given enough data

- **Not measuring joint MI**
  - quick hack: checkerboard
Broad-Class Detector

- Expert gives $\text{Pr}(\text{phone} \mid \text{class}, \text{features})$
  - still need $\text{Pr}(\text{class} \mid \text{features})$
- Repeat same approach
  - separate detectors for each broad class
  - measure MI from TF cell to class
  - train MLP from those features
- False accept/false detect tradeoff
  - try to detect only center of phone
  - reasonable vowel recognition with 10% insertions (6.3% deletions) of centers
Overall System

- ‘Patch in’ expert’s posteriors:
  \[ P(q_i|X) = \sum_{\text{class}} P(q_i|\text{class}, X) \cdot P(\text{class}|X) \]
  - ‘non-expert’ MLP for when \( P(\text{class}|X) \) are small

- Still looking at:
  - using more experts
  - better \( P(\text{class}|X) \)

TIMIIT phone err rate
Baseline: 28.4%
Oracle P(VC): 26.9%
Real P(VC): 28.0%
Vowels+Fric: 27.6%
Temporal Variation

- **Idea:** Normalize phone durations by averages
  - .. to reveal per-speaker bias
  - .. and timing variation within phrases

- **Focus on vowels**
  - per-phone deviations are very noisy

- **Use to vary sampling/modeling?**
Formant AGC

- **Hypothesis:**
  Casual speech has ‘compressed’ formant motion
  - can we ‘enhance’ formant motions
to make speech more canonical / read-like?
Read vs. Spontaneous

- Speaker-dependent means, vars of PLP pole locations in read vs. spontaneous speech

- Variance of angle of pole 3 discriminates well for red and green speakers - but opposite changes!
Deformable Spectra

- Accurate spectral modeling in conventional HMMs requires 1000s of states
  - cumbersome, especially transition matrices

- Observation:
  Speech spectra undergo minor deformations
  - suggests a different generative model:

\[
\begin{align*}
N_P &= 5 \\
N_C &= 3 \\
T &= \begin{bmatrix}
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\end{align*}
\]
States + Transformation Model

- Time-frequency state grid
- State →
  - explicit prototype
  - or a transformation on prior frame
- Infer underlying states

![Diagram](image)

**a) Signal**

- Green: Identity transform
- Yellow/Orange: Upward motion (darker is steeper)
- Blue: Downward motion (darker is steeper)

**b) Transformation Map**
Two-layer model

- Source-filter decomposition
  - pitch and formants have different dynamics
- Apply transformation models for both
  - log-spectra:
    - sum of excitation & filter
  - inference does separation