AN INFORMATION-THEORETIC FRAMEWORK FOR AUTOMATED DISCOVERY OF PROSODIC CUES TO CONVERSATIONAL STRUCTURE

Kornel Laskowski[1,2] and Anna Hjalmarsson[3]

[1] Carnegie Mellon University, Pittsburgh PA, USA
[2] Voci Technologies, Inc., Pittsburgh PA, USA
[3] KTH Speech, Music and Hearing, Stockholm, Sweden

Goals

PROPOSE & EVALUATE A **METHOD** FOR:

- **Quantifying** the relationship

  past prosody $\longrightarrow$ incipient (turn) structure

- in an **automated** fashion
  - manual labeling of turn structure unnecessary
  - automatically computable prosodic features
- which provides **fast and flexible** techniques
  1. to measure the **global** predictive power of a feature
  2. to measure the **local** predictive power of a feature
  3. to identify instants **when** a feature is most operative
  4. to **compare** features at global and local levels

Approach

1. **Discretize the speech activity of each conversant** at a framing frequency f.

instant t	1	2	3	4	5	6	7	8	9	⋯
Conversant 1:	■	■	■	■	□	□	□	□	□	⋯
Conversant 2:	□	□	□	□	□	□	■	□	■	⋯
Conversant 3:	□	□	□	■	■	■	■	□	□	⋯

2. Model the **probability of a conversant speaking at t**, conditioned on what they and their interlocutors were doing just before t.
3. Measure the error E between what the model predicts and what actually happens; $f \cdot E$ is the **cross-entropy rate** in bits per second.
4. Measure the difference ΔE of the the error E, with and without a feature of interest; $f \cdot \Delta E$ is the **conditional mutual information rate** in bits per second.

Findings

1. For signal energy:
   - a correlate of **speaking loudness**
   - the proposed framework indicates a **considerable effect**
   - scientific literature is much in agreement that loudness is relevant
2. For Mel-spectral flux (MSF):
   - a correlate of **speaking rate**
   - the proposed framework indicates **much weaker effect**
   - scientific literature is not in agreement that rate is relevant
3. The methodology is able to **quantify the global and local differences** between the utilities of the two features.
4. **Does not require manual annotation** of conversation structure.
   - Only per-frame, per-participant, binary speech/non-speech classification

Potential Impact

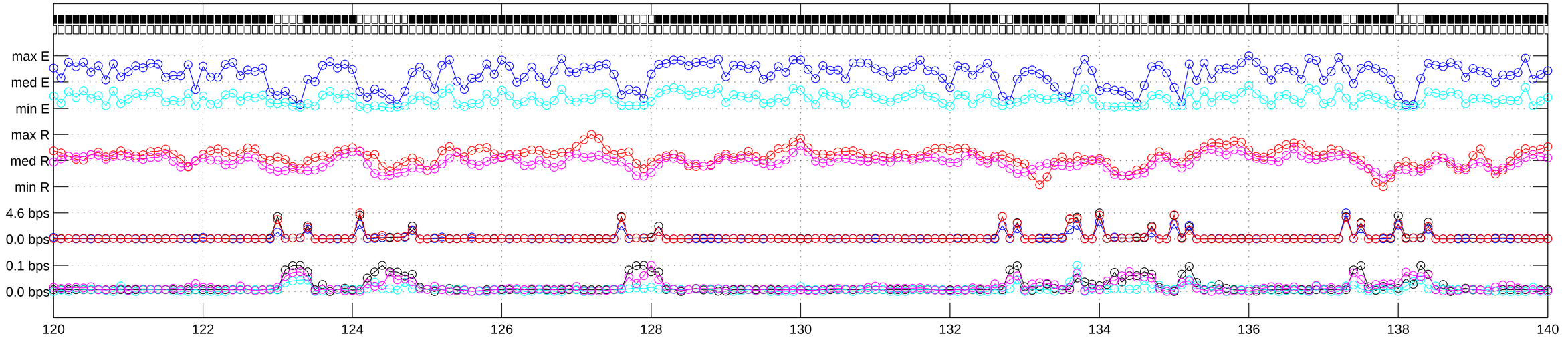
I. No annotation $\rightarrow$ can perform analysis for **very large speech corpora**, cheaply and **at all instants in time**.

II. Results are **theory-agnostic** — do not rely on definition of what a "turn" might be.

III. Can compare **prosodic practice across speech domains** within a language.

IV. Can compare **prosodic practice across languages**.

An Example: Dialogue 3161 from Switchboard Release 1 Version 2 (neither speaker observed during model training)



kornel@cs.cmu.edu, annah@speech.kth.se