CORPUS-INDEPENDENT HISTORY COMPRESSION FOR STOCHASTIC TURN-TAKING MODELS

Kornel Laskowski[1] & Elizabeth Shriberg[2]

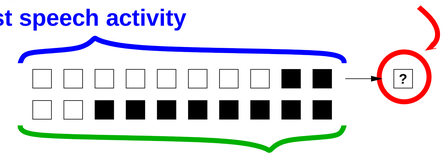
[1] Carnegie Mellon University, Pittsburgh PA, USA
[2] Microsoft Speech Labs, Mountain View CA, USA

Goals

Stochastic turn-taking models

predict a speaker's **future speech activity**

using that **speaker's past speech activity**



and their **interlocutor's past speech activity**

Q1. How **far back** should models look?
Q2. How to look that far back **tractably**?

Previously Known

Most frequently implemented as "conditionally independent (CI)" ***N*-gram models**.

"Unconditionally independent (UI)" variants that ignore the interlocutor are weaker.

Least recently (Jaffe, Feldstein & Cassotta, 1967):

- 300-ms frame step
- no back-off or interpolation
- **look back at most as far as 300 ms**

Most recently (Laskowski, Edlund & Heldner, 2011):

- 100-ms frame step
- recursive linear interpolation
- **look back at least as far as 1000 ms**

Conclusions

1. Speech activity as far back as **7-8 seconds** for both speakers is relevant to the prediction of future speech activity of either.
2. The context can be extended — without an increase in model complexity — via (non-linear) **projection onto a lower-dimensional space**.
3. Empirically determined projections **sacrifice precision of less recent events** for higher precision of more recent ones, yielding a quasi-logarithmic window profile.

Impact

I. Applications requiring prediction of speech activity should look at speech activity much further back than previously thought.

II. May lead to better SAD systems.

III. Lexical, prosodic, and other features may also benefit from quasi-logarithmic history compression, as may models for other prediction tasks in speech processing.

IV. The precise profile of temporal compression may have implications for models of human processing, particularly models of attention allocation.

Chronogram History Compression

In the uncompressed history baseline: use one conditioning context window per **frame**

To compress history:

- Use one conditioning context window per **sequence of frames**
- Grouping determined empirically using automated depth-first search

- Map **sequence of frame values** to **one window value**
- Here, using **majority-class voting** (also explored OR and AND)

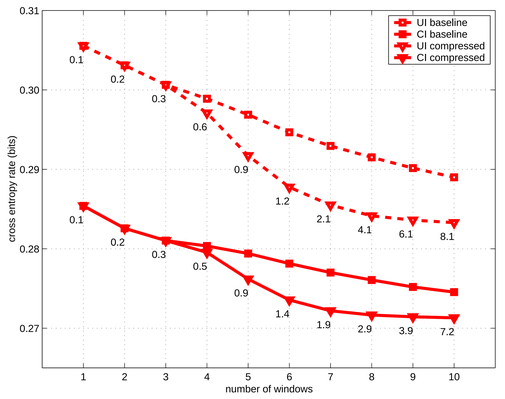
- From subsequent modeling point of view: data looks (nearly) the same as if it were not compressed
- Longer context durations achieved at same model orders

Effect of History Compression on Chronogram Cross Entropy Rate

Optimize history compression scheme by minimizing cross entropy rate for the Switchboard-1 TESTSET.

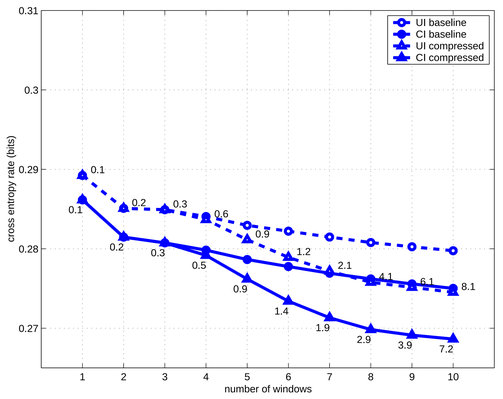
Switchboard-1
spontaneous, Americal English, telephone

TRAINSET: 762 dialogues, 2.64M frames
DEVSET: 227 dialogues, 0.88M frames
TESTSET: 199 dialogues, 0.81M frames



Spontal
spontaneous, Swedish, face-to-face

TRAINSET: 23 dialogues, 0.50M frames
DEVSET: 6 dialogues, 0.14M frames
TESTSET: 6 dialogues, 0.14M frames



kornel@cs.cmu.edu, elshribe@microsoft.com