Detecting Laughter and Filled Pauses Using Syllable-based Features

Guozhen An, David Guy Brizan, Andrew Rosenberg
CUNY Graduate Center / Queens College (CUNY)

2013-08-26

Interspeech 2013
Lyon, France

Problem: Frame-based decisions → frame-based errors

- Baseline classifier makes frame-level errors
  - Many times surrounding frames' labels contradict classifier decisions
  - Often probability of predicted label is close to some other class

```
'S1393.wav',188,garbage,0.700,0.065,0.235
'S1393.wav',188,garbage,0.575,0.093,0.332
'S1393.wav',190,garbage,0.474,0.126,0.400
'S1393.wav',191,filler, 0.425,0.113,0.462
'S1393.wav',192,garbage,0.478,0.091,0.430
'S1393.wav',193,garbage,0.534,0.080,0.385
'S1393.wav',194,garbage,0.692,0.069,0.239
'S1393.wav',195,garbage,0.783,0.054,0.163
'S1393.wav',196,garbage,0.792,0.072,0.136
```

Experiments

- Rescore within a region
  - Change frame prediction where appropriate
  - Consider surrounding frame labels

- Add prosodic features
  - With rescoring
  - Without rescoring

AuToBI for syllable Regions

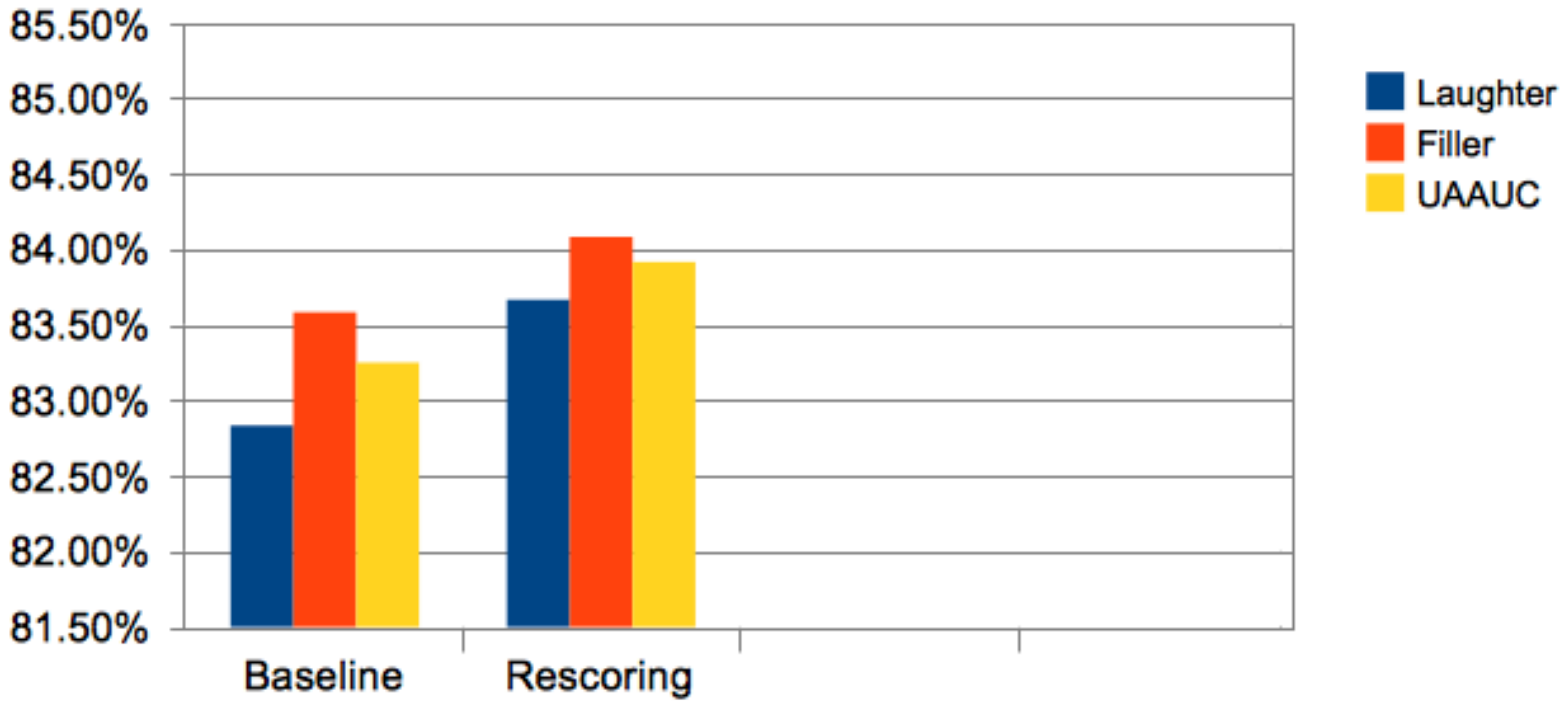
- Used for automatic detection of ToBI-labeled prosodic events:
  - Pitch Accent Detection
  - Pitch Accent Classification
  - Phrase Detection
  - Phrase Ending Classification

- Syllable *Regions*
  - Produced to generate Pitch Accent Classification
  - Generates *Region* hypotheses without input word segmentation (new feature)

Rescoring within a Region

- Determine the Region's label:
  - ***Majority Class***
    - If the total number of frames in the Region exceeds a certain fraction of all frames (0.9 for *Garbage*, 0.7 otherwise)
  - ***Average Segment***
    - If no *Majority Class* - OR - if unweighted average for the Region exceeds threshold (0.9 for *Garbage* / 0.7 otherwise)
- Determine whether to rescore a Frame:
  - **Label Exchange**
    - Change the frame's label if below a prediction threshold (0.9 for *Garbage*, 0.7 otherwise)

Results (Test Set) Rescoring



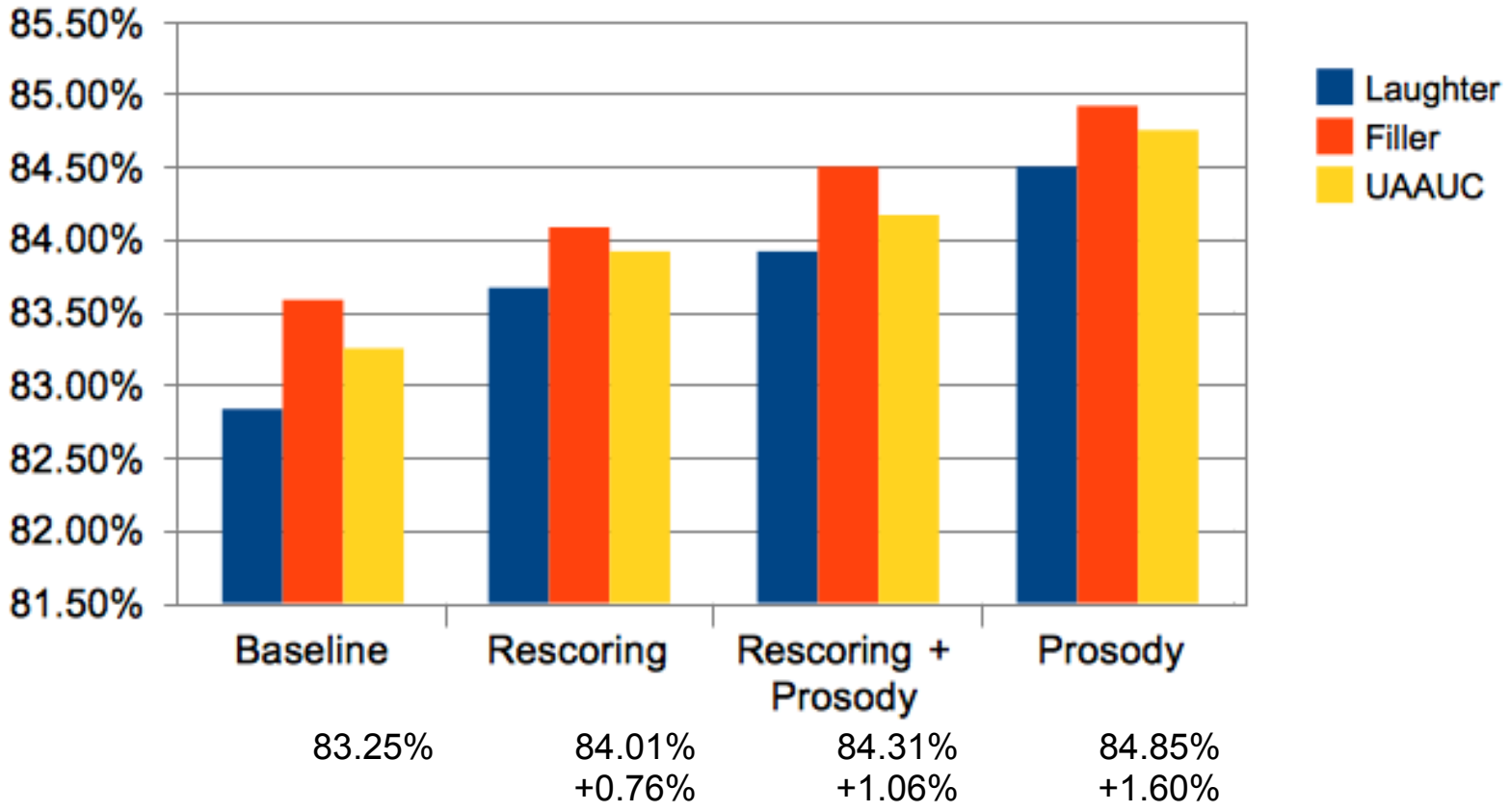
83.25%

84.01%
+0.76%

Adding Prosodic Features

- Added 28 AuToBI-based features to baseline for previous, current, subsequent *Region*:
  - *Region* sequence in file
  - Duration of *Region*
  - Range Normalized Intensity & its delta
  - Z-score Normalized Pitch (log Hz) and its delta
  - Mean Spectral Tilt and its delta
  - Length of preceding and following pauses
- Tested additional features with & without rescoring
  - Rescoring may be superfluous with new features

Results (Test Set) + Prosody



If we had more time, we would

- Contrast our approach with well-established algorithms
  - Viterbi for rescoring
- Model laughter, filler & garbage (each) as heterogeneous phenomena
  - There is no stereotypical *Laughter*
  - Different languages express *Filler* (filled pauses) differently

Thanks

The Speech Lab at Queens College
http://speech.cs.qc.cuny.edu/

Guozhen An (gan@gc.cuny.edu)
David Guy Brizan (dbrizan@gc.cuny.edu)
Andrew Rosenberg (andrew@cs.qc.cuny.edu)

Questions?

Results (Test Set)

Config	Laughter	Filler	UAAUC
Baseline	82.90%	83.60%	83.25%
Rescoring	83.80% +0.9%	84.22% +0.62%	84.01% +0.76%
Rescoring + Prosody	84.05% +1.15%	84.58% +0.98%	84.31% +1.06%
Prosody	84.64% +1.74%	85.06% +1.46%	84.85% +1.6%