Crowdsourced and Automatic Speech Prominence Estimation

Max Morrison, Pranav Pawar, Nathan Pruyne, Jennifer Cole, Bryan Pardo
Northwestern University
ICASSP 2024

Goal

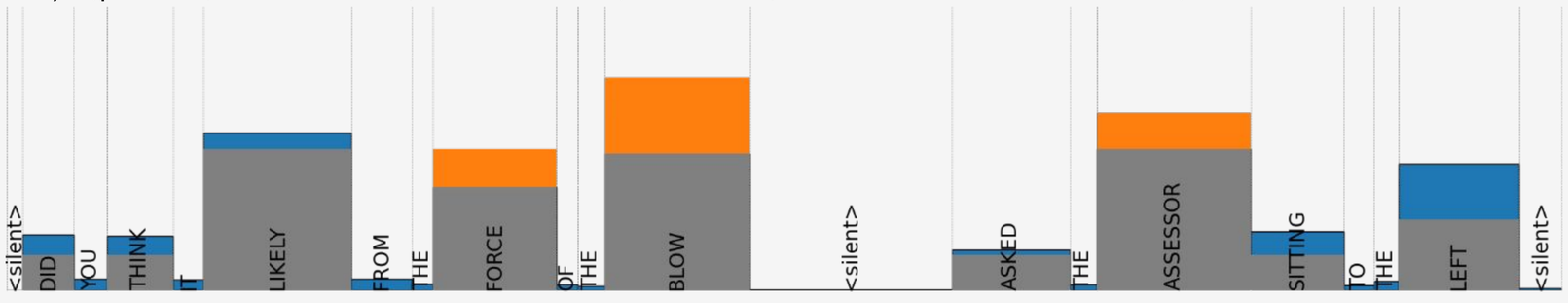
Phoneme- and word-level representation of, e.g., prominence (this paper), falsetto, or vocal fry

Challenges

Data scarcity -> We open-source human prominence annotations and tools
Annotation is expensive -> We show how to reduce annotation costs for a fixed budget
Resolution mismatch -> We show variable-rate downsampling within neural networks

What is prominence?

Prominence is a multi-factorial, continuous representation of speech emphasis or salience.
Factors include *prosody* (pitch, duration, loudness) and *information structure* (novel information in the discourse).



*Crowdsourced (**blue**) and automatic (**orange**) speech prominence estimation; **gray** indicates agreement*

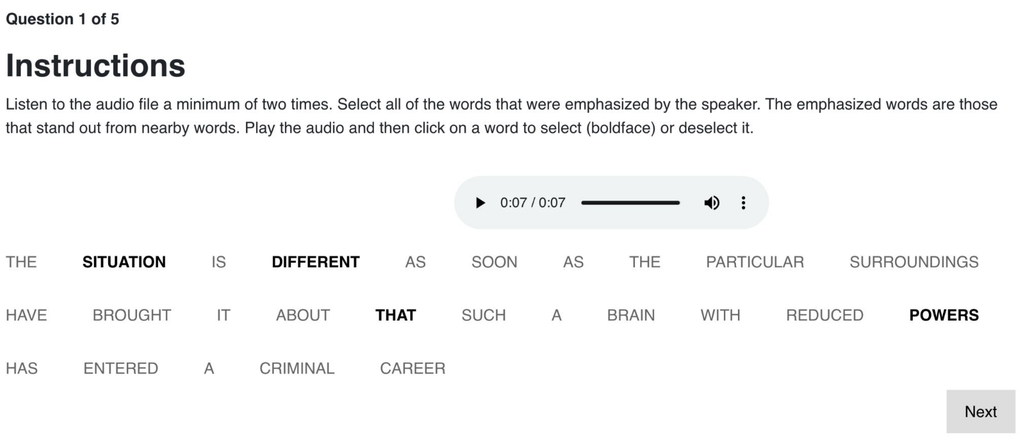
We represent the scalar projection of prominence m_i as one Bernoulli distribution per word i.
The corresponding random variable e_i is the binary status indicator of whether a word is emphasized.

$$p(e_i = 1) = m_i$$

Prominence annotations enable emphasis-controlled TTS and analysis tasks such as emotion recognition

Crowdsourced estimation

We open-source a tool for human annotation of word- or phoneme-resolution features

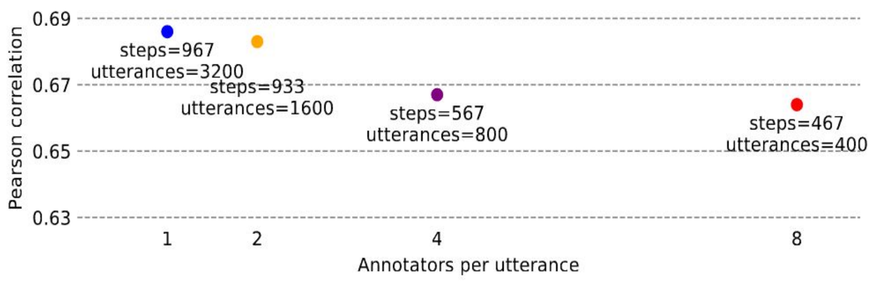


Crowdsourced human emphasis annotation interface
Annotators listen to speech and click the words they perceive as emphasized
github.com/reseval/reseval

We use our open-source tool to annotate part of LibriTTS

- 6.42 hours of `train-clean-100`
- 18 speakers (9 male; 9 female)
- 3,626 utterances; 69,809 annotated words
  - >= 1 annotation of 3,626 utterances
  - >= 2 annotations of 2,259 utterances
  - >= 4 annotations of 974 utterances
  - >= 8 annotations of 453 utterances
- zenodo.org/records/10402793

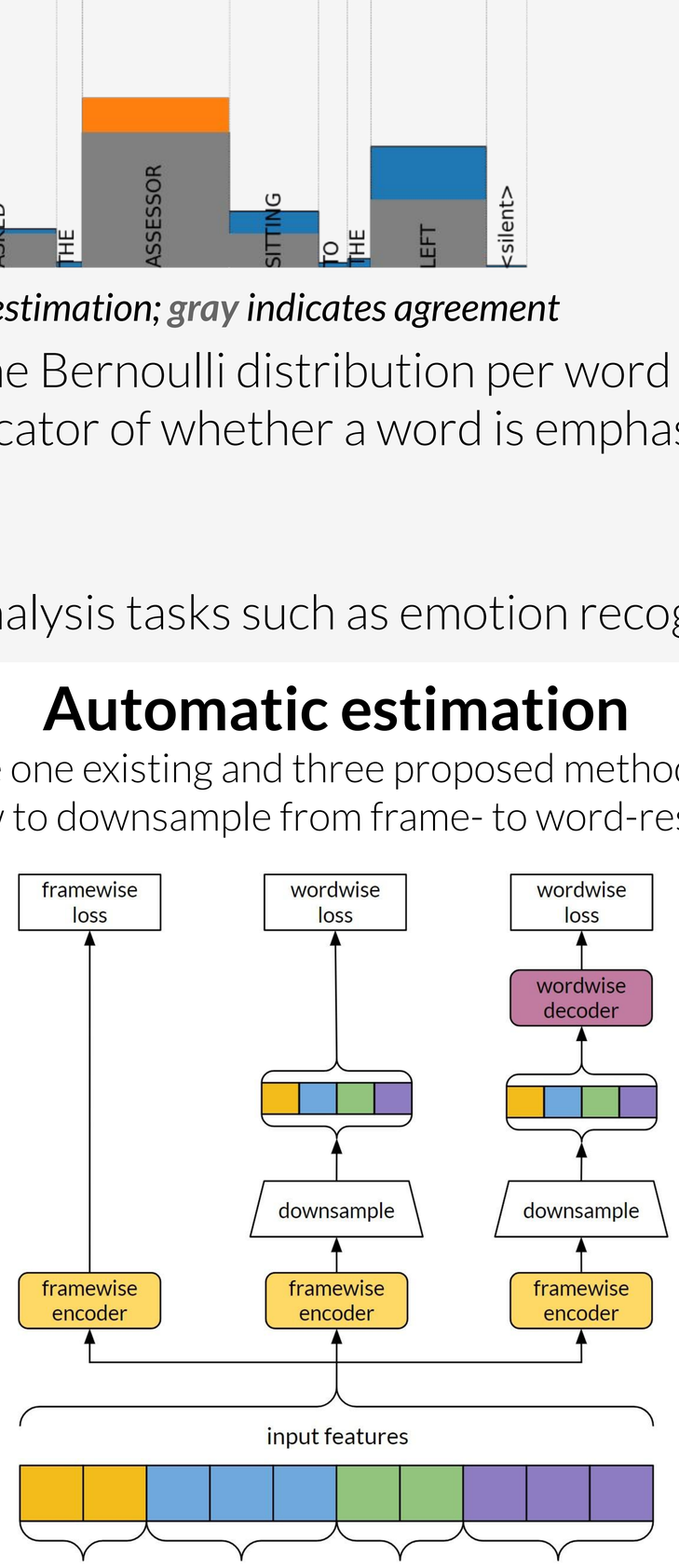
For training an automated annotator, how many human annotations per utterance are best?



Pearson correlations between automatic and crowdsourced prominence estimates when varying the number of human annotators on a fixed budget. Averages over three runs.

Automatic estimation

We compare one existing and three proposed methods for where and how to downsample from frame- to word-resolution



Proposed neural prominence estimation models
*We experiment with three proposed locations for downsampling from frame- to word-resolution: downsample during inference **(left; framewise)**, downsample just before the loss function **(center; posthoc wordwise)**, and downsample within the neural network **(right; intermediate wordwise)**.*

Downsample from frame- to word-resolution using per-channel summation within the neural network

	Downsampling location	Downsampling Method			
		Average	Center	Max	**Sum**
Proposed	Inference (framewise)	0.102	0.153	0.102	0.137
Proposed	**Intermediate (wordwise)**	0.656	0.438	0.674	**0.675**
Proposed	Posthoc (wordwise)	0.440	0.385	0.623	0.645
Existing	Prehoc [17] (wordwise)	0.670	0.471	0.670	0.656

Pearson correlations (higher is better) between estimated and ground truth prominence for various downsampling methods and locations. Averages over three runs.

Code, paper, dataset, and listening examples available at maxrmorrison.com/sites/prominence-estimation