Leveraging Automatic Speech Recognition for Prosodic Stress Analysis

Samuel S. Sohn (samuel.sohn@rutgers.edu), Sten Knutsen, Karin Stromswold
Rutgers University – New Brunswick

1 Introduction. Prosody affects how people process spoken sentences. For example, prosody can bias the way people interpret ambiguous sentences and can strengthen or weaken garden paths [1, 8, 2]. Some studies suggest people (unconsciously) alter their prosody depending on the meaning they want to convey [3, 6]. Despite this, the role of prosody in sentence processing and production remains understudied because of the difficulty in quickly assessing prosody. What is needed is an efficient way of accurately characterizing the prosody of utterances. To this end, we used OpenAI Whisper large-v2 [7], a state-of-the-art automatic speech recognition model. Whisper uses deep learning to analyze audio waveforms, identify patterns that align with human speech, and then decode these patterns into a transcription. Although it was not originally trained to recognize prosodic stress, in this paper, we demonstrate that using a small number of carefully constructed, stress-annotated transcriptions, Whisper can be fine-tuned to recognize phrasal, lexical, and contrastive stress and investigate the acoustic similarities and differences among these different types of prosodic stress.

2 Methods. The fine-tuning dataset is based on an experiment with 36 native English-speaking college students (18 men and 18 women) from the mid-Atlantic U.S. [4]. For phrasal stress, participants produced 16 adjective-noun and compound word minimal pairs embedded in sentences (e.g., "The white board/whiteboard is dirty"). For lexical stress, they produced 16 words differing only in stress pattern (e.g., "*re*cord" vs. "re*cord*"). For contrastive stress, they listened to 16 sentences in which either a color or animal did not match a picture (e.g., "The black sheep has the ball" with an image of a red sheep with a ball) and corrected the error both lexically and prosodically (e.g., "The *red* sheep has the ball"). The transcriptions were capitalized to reflect the placement of stress as it would typically occur in English. The model was fine-tuned using 5-fold cross validation to ensure robustness (Table 1).

3 Results. In order to assess the transfer of acoustic patterns between different types of stress, we first fine-tune a Control model using all types of stress from a single random participant (Table 2). This equips Whisper with the minimum knowledge needed to learn the transcriptions in our fine-tuning dataset. (Phrase stress accuracy is higher for the Control because the prosodic difference between AdjN vs. compound word is implicitly included in the orthography of Whisper's pre-training lexicon.) For phrasal stress, we fine-tune Whisper on the control data and the phrasal stress of the training subset, producing the Phrasal model that is then tested on all types of stress in the *testing* subset. This is repeated for lexical stress, contrastive stress, and the combination of all three. Table 2 shows that all non-phrasal results for single-stress training (except phrasal$\rightarrow$ phrasal) have a statistically significant improvement in accuracy over the Control model. Phrasal and contrastive stress models learn slightly conflicting acoustic patterns in isolation, worsening their transfer accuracy significantly, but in the all-stress model, new non-conflicting patterns are learned. Three trained native English-speaking research assistants (coders), who were blind to the utterances' stress type, hand-coded the stress of each utterance. When fine-tuning on all stress, we achieve near-human accuracy w.r.t. the coders, and higher average accuracy across phrasal, lexical, and contrastive stress compared to single-stress Random Forest Classifiers (RFCs) [4].

4 Discussion. The successful application of Whisper to prosodic stress analysis enables large-scale studies of spoken language processing and production that account for prosody's role in communication. Namely, Whisper's ability to achieve near-human accuracy in identifying multiple types of stress patterns while simultaneously transcribing audio allows researchers to explore these questions more comprehensively than with RFCs. The observed transfer effects between different types of stress provide compelling evidence for shared acoustic patterns in stress production [5]. The strong lexical$\leftrightarrow$contrastive transfer suggests similar acoustic patterns between word-level and discourse-level prosodic phenomena. The weak transfer to and from phrasal stress is consistent with RFC findings that indicate lexical and contrastive stress are signaled by a combination of frequency, amplitude and duration, whereas phrasal stress is signaled almost exclusively by duration [4]. By identifying shared acoustic patterns across stress types, Whisper can support the investigation of theoretical frameworks in human sentence processing, such as how prosody integrates with syntax and semantics during sentence processing.

Training Instance	Training Epochs	Fine-tuning Dataset				
		Fold 1 0-20%	**Fold 2** 20-40%	**Fold 3** 40-60%	**Fold 4** 60-80%	**Fold 5** 80-100%
1	5	Test	Train	Train	Train	Train
2	5	Train	Test	Train	Train	Train
3	5	Train	Train	Test	Train	Train
4	5	Train	Train	Train	Test	Train
5	5	Train	Train	Train	Train	Test

Table 1: For cross-validation, 5 instances of a model are trained for 5 epochs each using default hyperparameters. Each instance uses the 4 non-diagonal folds for training and the 1 diagonal fold for testing. Reported accuracies have been averaged over all 5 instances, and each fold has equal gender representation.

Training Stress	Metric	Testing Stress		
		Phrasal (SD)	**Lexical** (SD)	**Contra.** (SD)
Control	%	70.7% (4.2)	39.5% (3.6)	49.7% (2.6)
Phrasal	Δ%	19.5% (2.9)†	9.1% (9.1)	−7.7% (4.7)*
Lexical	Δ%	3.9% (3.7)	47.1% (3.6)†	27.8% (4.8)†
Contra.	Δ%	−11.4% (2.8)†	32.4% (3.1)†	38.9% (2.5)†
All	%	90.2% (2.5)†	86.6% (2.3)†	88.7% (4.1)†
Coders	%	91.9% (1.6)	88.8% (1.6)	91.6% (1.5)
RFCs	%	86.4% (0.2)	83.9% (0.3)	83.7% (0.3)

Table 2: This table reports the accuracy (%) of the control-stress model, all-stress model, coders and RFCs, as well as the residual accuracy (Δ%) of single-stress models with respect to the control-stress model. $^{\dagger}p < .01$ $^{*}p < .05$

References.

[1] Beach, C. M., 1991. "The Interpretation of Prosodic Patterns at Points of Syntactic Structure Ambiguity: Evidence for Cue Trading Relations". *Journal of Memory and Language*.

[2] Carlson, K., 2009. "How Prosody Influences Sentence Comprehension". *Language and Linguistics Compass*.

[3] Ferreira, F., 1993. "Creation of Prosody During Sentence Production." *Psychological Review*.

[4] Knutsen, S. and Stromswold, K., 2024. "Gender Differences in the Acoustic Realization of Stress". *Penn Working Papers in Linguistics*.

[5] Ladd, D. R., 2008. "Intonational Phonology". *Cambridge University Press*.

[6] Pierrehumbert, J., 1990. "The Meaning of Intonational Contours in the Interpretation of Discourse". *Intentions in Communication/Bradford Book*.

[7] Radford, A. et al., 2023. "Robust Speech Recognition Via Large-Scale Weak Supervision". *International Conference on Machine Learning*.

[8] Snedeker, J. and Trueswell, J., 2003. "Using Prosody to Avoid Ambiguity: Effects of Speaker Awareness and Referential Context". *Journal of Memory and Language*.