Prosody in the Age of AI: Insights from Large Speech Models

Samuel S. Sohn (samuel.sohn@rutgers.edu), Sten Knutsen, Karin Stromswold
Department of Psychology & Center for Cognitive Science,
Rutgers University – New Brunswick

Abstract

Prosody affects how people produce and understand language, yet studies of how it does so have been hindered by the lack of efficient tools for analyzing prosodic stress. We fine-tune OpenAI Whisper large-v2, a state-of-the-art speech recognition model, to recognize phrasal, lexical, and contrastive stress using a small, carefully annotated dataset. Our results show that Whisper can learn distinct, gender-specific stress patterns to achieve near-human and super-human accuracy in stress classification and transfer its learning from one type of stress to another, surpassing traditional machine learning models. Furthermore, we explore how acoustic context influences its performance and propose a novel black-box evaluation method for characterizing the decision boundaries used by Whisper for prosodic stress interpretation. These findings open new avenues for large-scale, automated prosody research with implications for linguistic theory and speech processing.

Keywords: prosody; speech recognition; stress perception; black-box evaluation

Introduction

Prosody plays a crucial role in spoken language comprehension and production. It influences how listeners interpret words, sentences, and the pragmatic import of utterances, guiding syntactic disambiguation and affecting sentence processing efficiency. For example, prosodic cues can bias interpretations of syntactically ambiguous sentences and either strengthen or weaken garden paths, where listeners initially favor an incorrect interpretation before reanalyzing the sentence structure (Beach, 1991; Snedeker & Trueswell, 2003; Carlson, 2009). Beyond comprehension, prosody is also integral to speech production, as speakers unconsciously modulate their intonation, rhythm, and stress to convey different meanings (Ferreira, 1993; Pierrehumbert, 1990).

Despite its importance, the study of prosody in both language processing and production remains relatively underdeveloped, largely due to the difficulty of analyzing prosodic features efficiently. Traditional prosodic analysis relies on trained human annotators who manually label stress patterns in speech data (see (Knutsen & Stromswold, 2024)), which is a time-consuming and resource-intensive process that lacks scalability. This bottleneck limits large-scale investigations into prosodic variation and its interaction with lexical, syntactic, and discourse structures.

A scalable and automated approach to prosodic analysis is therefore needed to advance our understanding of prosody and how it interfaces with other aspects of language. In this study, we explore the potential of OpenAI's Whisper large-v2 model (Radford et al., 2023), a state-of-the-art automatic speech recognition (ASR) system, to recognize and analyze prosodic stress. Although Whisper was not originally trained for prosodic annotation, we demonstrate that fine-tuning it with a small, carefully curated dataset of stress-annotated utterances enables it to recognize different types of prosodic stress (i.e., phrasal, lexical, and contrastive stress) and transfer learned acoustic patterns between them. We further investigate the relationships between stress types based on how they facilitate or impede such transfer and how, for individual stress types, broader acoustic context can improve prosodic annotation to a super-human level for both men and women. Finally, we propose a novel black-box evaluation methodology for identifying acoustic decision boundaries that distinguish stress patterns, shedding light on how prosodic stress conveys meaning for men and women.

Preliminaries

At its core, Whisper leverages deep learning to analyze audio waveforms, extract patterns aligned with human speech, and decode these patterns into transcriptions (Radford et al., 2023). It is based on a Transformer architecture (Vaswani, 2017) trained through large-scale weak supervision to generalize across diverse acoustic environments, speakers, and linguistic contexts.

Pre-training

Whisper has been pre-trained on 680,000 hours of labeled audio data, providing an extensive and diverse foundation for robust speech recognition. This dataset comprises 64% English transcriptions, 17% transcriptions from 96 non-English languages, and 18% X→English translations (Radford et al., 2023). The scale and diversity of this corpus enable Whisper to develop a highly flexible one-to-many mapping between text and the vast range of acoustic variations in spoken language. These variations include differences in speaker identity, accent, speech rate, background noise, and prosodic features such as phrasal, lexical, and contrastive stress.

Despite Whisper's broad pre-training, it is not explicitly trained to recognize fine-grained prosodic phenomena. Instead, it learns to associate multiple prosodic variations with the same textual representation, effectively collapsing distinctions that are critical for nuanced prosody analysis. To

accurately distinguish phrasal, lexical, and contrastive stress, Whisper requires fine-tuning on a curated dataset where stress distinctions are explicitly annotated and linked to *unique* transcriptions. Such fine-tuning enables the model to differentiate stress patterns based on acoustic cues such as pitch, duration, and amplitude, rather than treating them as interchangeable variations of the same speech signal.

Fine-tuning

The fine-tuning dataset is based on an experiment (Knutsen & Stromswold, 2024) with 36 native English-speaking college students (18 men and 18 women) from the mid-Atlantic U.S., who were tasked with producing prosodic stress to distinguish meaning using the web-based platform FindingFive (FindingFive Team, 2019). No participants reported any issues with vision, hearing, language abilities (spoken or written), learning, or other neuropsychological conditions. For phrasal stress, participants produced 16 adjective-noun and compound word minimal pairs embedded in sentences (e.g., "The green house/greenhouse spoils the view"). For lexical stress, they produced 16 words differing only in stress pattern (e.g., "*in*sult" vs. "in*sult*"). For contrastive stress, they listened to 16 sentences in which either a color or animal did not match a picture (e.g., "The red cow has the ball" with an image of a black cow with a ball) and corrected the error both lexically and prosodically (e.g., "The *black* cow has the ball").

To facilitate model training, transcriptions are capitalized to reflect canonical English stress patterns. All minimal pair transcriptions have been listed in Table 1. The minimal pairs for phrasal stress are not capitalized because their distinct meanings are already encoded in their orthographic forms. Each instance of the Whisper model is fine-tuned for 5 epochs using default hyperparameters, and for a given transcription dataset, model performance is averaged over 5 instances using 5-fold cross-validation. This cross-validation protocol partitions the participant data into 5 equal subsets with balanced gender representation, iteratively training on four subsets and testing on the held-out fifth (De Rooij & Weeda, 2020). By ensuring that each data point is used for both training and validation across different iterations, cross-validation prevents data leakage and guarantees that the model is not memorizing specific training examples. This safeguards against overfitting and ensures that reported performance reflects generalizable transcription accuracy rather than an artifact of the training set.

Stress	Minimal Pair Transcription
Phrasal	The <green house / greenhouse> spoils the view.
Phrasal	There's a <dark room / darkroom> in this house.
Phrasal	The <white board / whiteboard> needs cleaning.
Phrasal	That <hot dog / hotdog> is under the table.
Phrasal	A <black bird / blackbird> just flew past.
Phrasal	His <wet suit / wetsuit> is on the floor.
Phrasal	That <blue bell / bluebell> is pretty.
Phrasal	The <bull's eye / bullseye> is red.
Lexical	<DIFfer / deFER>
Lexical	<DIScard / disCARD>
Lexical	<DIScount / disCOUNT>
Lexical	<INcrease / inCREASE>
Lexical	<INdent / inDENT>
Lexical	<INsert / inSERT>
Lexical	<INsight / inCITE>
Lexical	<INsult / inSULT>
Contra.	The <BLACK cow / black COW> has the ball.
Contra.	The <BLACK sheep / black SHEEP> has the ball.
Contra.	The <BLUE cow / blue COW> has the ball.
Contra.	The <BLUE sheep / blue SHEEP> has the ball.
Contra.	The <RED cow / red COW> has the ball.
Contra.	The <RED sheep / red SHEEP> has the ball.
Contra.	The <WHITE cow / white COW> has the ball.
Contra.	The <WHITE sheep / white SHEEP> has the ball.

Table 1: A list of minimal pairs by stress type.

Related Work

As a baseline for model comparison, we use the Knutsen and Stromswold study (Knutsen & Stromswold, 2024), from which the fine-tuning dataset was derived. They examined gender differences in the acoustic realization of phrasal, lexical, and contrastive stress, addressing a gap in prior research on prosodic variation between men and women. Acoustic features (including pitch, amplitude, and duration) were extracted and analyzed using Bayesian ANOVAs, Random Forest Classification (RFC), and Bayesian mixed-effects regression to determine their relative importance in signaling stress. Their results indicate that while both men and women employ pitch (measured by fundamental frequency F0), amplitude, and duration to mark stress, their reliance on these features differs systematically.

Stress Patterns

Knutsen and Stromswold found that phrasal stress was predominantly marked through durational differences, where adjective-noun morphemes were often longer and had more pause between them. Subtle gender-based distinctions also emerged according to RFC results: pitch had a slightly higher importance score than amplitude for men, and amplitude had a higher importance score than pitch for woman by a similar margin. This finding enriches prior work from Plag (Plag, 2006), which found that F0 differences in compound words were more pronounced for women than men.

For lexical stress, Knutsen and Stromswold's RFC results and regression analyses revealed that women use amplitude, duration, and pitch, with amplitude being most important, whereas men primarily rely on amplitude and duration. This aligns with the data from Koffi and Mertz (Koffi & Mertz, 2018), which after re-analysis by Knutsen and Stromswold showed that amplitude and duration play crucial roles for both

Acoustic Context	**Phrasal Stress** (SD)			**Lexical Stress** (SD)			**Contrastive Stress** (SD)		
	All	Men	Women	All	Men	Women	All	Men	Women
None	90.1% (3.2)	93.0% (2.4)	89.8% (4.9)	87.1% (4.6)	88.0% (6.2)	86.1% (6.3)	89.9% (3.1)	90.8% (5.0)	88.2% (4.3)
Front	91.3% (1.9)	92.4% (2.8)	90.7% (3.5)	N/A	N/A	N/A	90.6% (2.8)	92.5% (3.1)	88.6% (4.8)
Back	92.0% (2.6)	92.0% (4.0)	92.4% (3.4)	N/A	N/A	N/A	93.1% (2.5)	95.1% (4.4)	92.2% (3.9)
Full	92.6% (2.2)	92.6% (3.3)	93.0% (1.8)	N/A	N/A	N/A	92.8% (2.3)	95.1% (4.0)	91.1% (3.6)
Coders	91.9% (1.6)	92.9% (1.1)	90.8% (1.3)	88.8% (1.6)	89.3% (1.5)	88.3% (1.5)	91.6% (1.5)	92.1% (1.2)	91.1% (1.6)
RFC	86.4% (0.2)	90.3% (0.3)	84.3% (0.5)	83.9% (0.3)	84.8% (0.4)	80.8% (0.5)	83.7% (0.3)	85.5% (0.4)	82.4% (0.5)

Table 2: Accuracy of Whisper models trained on phrasal, lexical, and contrastive stress using different types of acoustic context.

genders, but pitch is more relevant for women than for men.

For contrastive stress, Knutsen and Stromswold found that both men and women relied on all three acoustic features, utilizing pitch, amplitude, and duration. The RFC analysis showed that the features had similar importance for both genders. However, the regression analysis showed that women used pitch to signal contrastive stress, while men did not.

Benchmarks

Machine learning analyses using RFC models revealed that men's speech was classified with greater accuracy than women's, suggesting that men's use of acoustic features is more consistent and less variable. This finding is particularly noticeable in lexical stress, where the RFC model correctly classified 84.8% of men's utterances and 80.8% of women's utterances. A similar trend was found for phrasal and contrastive stress, where women's more variable use of pitch may have contributed to the lower classification accuracy. Bayesian regression analyses further confirmed that pitch was a significant predictor of stress accuracy for women but not for men, reinforcing the notion that women employ a more complex, multi-dimensional approach to stress marking.

In addition to the RFC baseline, this study presented a human benchmark, i.e., the gold standard. This benchmark used three trained native English-speaking research assistants (coders), who were blind to the target utterance, to mark the perceived stress in each trial. Coders used Praat to mark morpheme (for phrasal and contrastive stress) or syllable (for lexical stress) boundaries. They also marked whether phrasal stress trials contained an adjective-Noun or compound word, whether the first or second syllable was stressed in lexical stress trials, and whether the color or animal was stressed in contrastive stress trials.

Extent of Acoustic Context

The RFC baseline is limited in that the sentence-embedded minimal pairs for phrasal and contrastive stress do not leverage acoustic features outside the minimal pair, which is the region of interest (ROI) bracketed in Table 1. This was likely done for methodological simplicity, since the embedding sentences vary not only in length but also lexically depending on how participants produced them. For instance, both "No, now the black COW has it" and "The black COW has it" were responses for a contrastive stress trial. Unlike the RFC model, Whisper automatically processes variable-length audio using its Transformer architecture, which can handle different input lengths while keeping track of word order. This makes the analysis of acoustic context more practicable than with an RFC model.

For lexical stress trials, only the ROIs were uttered by participants (e.g., "<INsult>"), because in the lexical stress trials, the words were said in isolation, we could not analyze the role of context.. To determine the extent to which acoustic information outside of the ROI includes information about what element is stressed, for phrasal and contrastive stress, we compared Whisper's performance when given only the ROIs (e.g., "<greenhouse>", "<BLACK cow>"), the ROIs and preceding context (i.e., front context: e.g., "the <greenhouse>", "No, the <BLACK cow>"), the ROIs and following context (i.e., back context: e.g., "<greenhouse> spoils the view", "<BLACK cow> has the ball") and the full sentence.

Table 2 shows that for phrasal and contrastive stress, having more acoustic context tends to improve Whisper's performance. Namely, the front acoustic context is much shorter than the back acoustic context, and this difference is reflected proportionally by the improved accuracy over having no acoustic context. For contrastive stress, this effect is much more pronounced (2.5 percentage points) than for phrasal stress (0.7 percentage points). An exception to this trend is the phrasal stress produced by men, which results in the highest Whisper accuracy when there is no acoustic context. While Whisper is unable to beat the average accuracy of human coders (i.e., the gold standard) for lexical stress without acoustic context, it is able to surpass the gold standard for phrasal and contrastive stress from both men and women. Furthermore, across *all* gender-stress combinations, Whisper's performance exceeds RFC models by an average of 6.6 percentage points. The most improvement is observed for contrastive stress, where the average improvement over men and women is $\sim$9.7 percentage points.

Transfer Between Stress Types

In addition to handling varying acoustic contexts with ease, Whisper is able to generalize its learnings across diverse acoustic environments, speakers, and linguistic contexts. Given its superior performance to RFC models (Table 2), it

Training Stress	Testing Stress		
	Phrasal (SD)	**Lexical** (SD)	**Contra.** (SD)
Control	70.7% (4.2)	39.5% (3.6)	49.7% (2.6)
Phrasal	90.2% (2.6)†	48.7% (6.0)	42.0% (6.3)*
Lexical	74.6% (3.0)	86.6% (1.2)†	77.5% (6.8)†
Contra.	59.2% (1.6)†	71.9% (4.9)†	88.7% (4.5)†
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 3: Accuracy of control stress, all-stress, coders and RFCs, and residuals for single-stress. $^{\dagger}p < .01$ $^{*}p < .05$

follows that Whisper is learning more valuable features that also generalize as evidenced by cross validation. However, this generalization is within stress type. We hypothesize that Whisper's pre-training has implicitly learned relationships between the acoustic patterns of stress types that can be uncovered through fine-tuning.

To this end, we first fine-tune a Control model using all types of stress from a single random control participant. This equips Whisper with the minimum knowledge needed to learn the unique lexicons in our fine-tuning dataset (i.e., the capitalization of stressed syllables). For consistency between stress types, we only use the ROIs (Table 1). The Control model's accuracy for phrasal stress is significantly higher than for lexical and contrastive stress (Table 3), because the prosodic difference between adjective-noun vs. compound word is implicitly included in Whisper's pre-training lexicon (e.g., "green house" vs. "greenhouse"). The control participant's data then becomes part of the fine-tuning data for 3 single-stress models and an all-stress model. For phrasal stress, Whisper is fine-tuned on the superset of control data and phrasal stress data, producing the Phrasal model that is then tested on *all* types of stress in the *testing* subset (Table 3, row 2). This is repeated for each fold in the cross-validation, and the entire process is repeated for lexical stress, contrastive stress, and the combination of all three (Table 3, rows 3-5).

Table 3 shows that the 2×2 matrix of lexical and contrastive results for single-stress models and the phrasal→phrasal result have a statistically significant improvement in accuracy over the Control model. Phrasal and contrastive stress models learn partially conflicting acoustic patterns in isolation, worsening their transfer accuracy significantly (-11.4% and -7.7%), but in the all-stress model, new non-conflicting patterns are learned. When fine-tuning on all stress types, we achieve near-human accuracy compared to the coders and higher average accuracy compared to the RFC models across phrasal, lexical, and contrastive stress reported in (Knutsen & Stromswold, 2024).

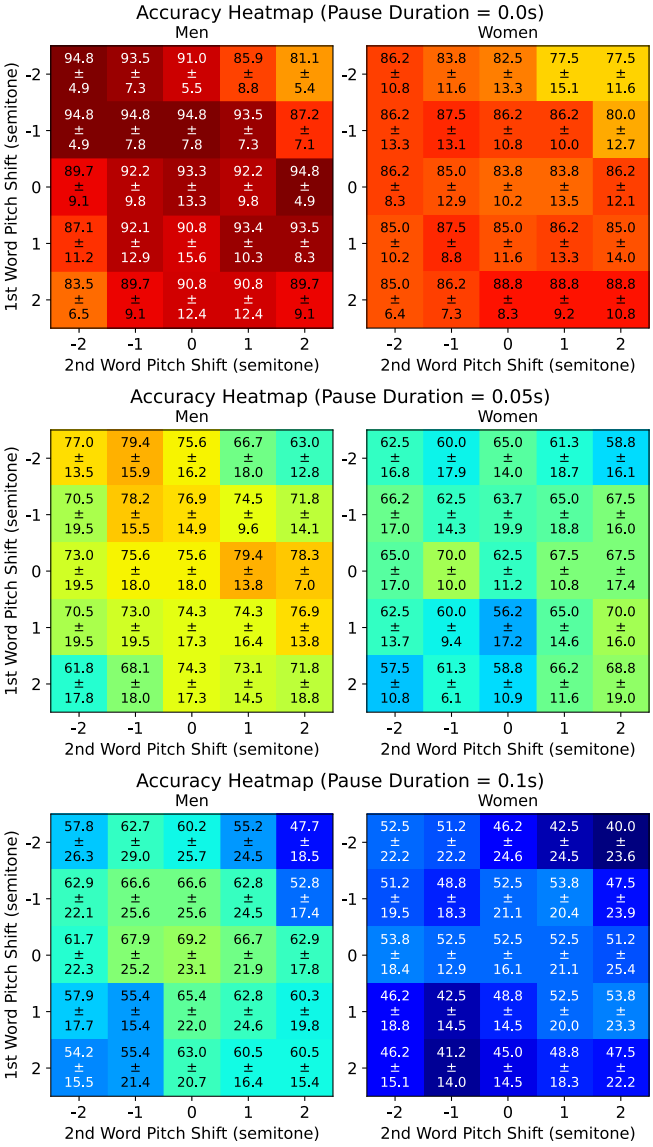

Figure 1: Heatmaps of Whisper's compound-word accuracy showing that as pitch shift diverges or pause duration increases, accuracy decreases because compound words are being recognized as two separate words. Hotter colors indicate higher accuracy.

Characterizing Decision Boundaries

Unlike RFC models, which provide interpretable feature importance scores, Whisper's learned features are more challenging to extract. To this end, we propose a black-box evaluation methodology, which we demonstrate on phrasal stress, but which can be applied to any stress type and any model. For each stress type, the minimal pairs can be separated into two categories based on their canonical stress patterns (Table 1). We select one of these categories as the starting distribution and we systematically perturb the acoustic features of every recording in *only* that category such that they progressively capture some of the other target category's acoustic

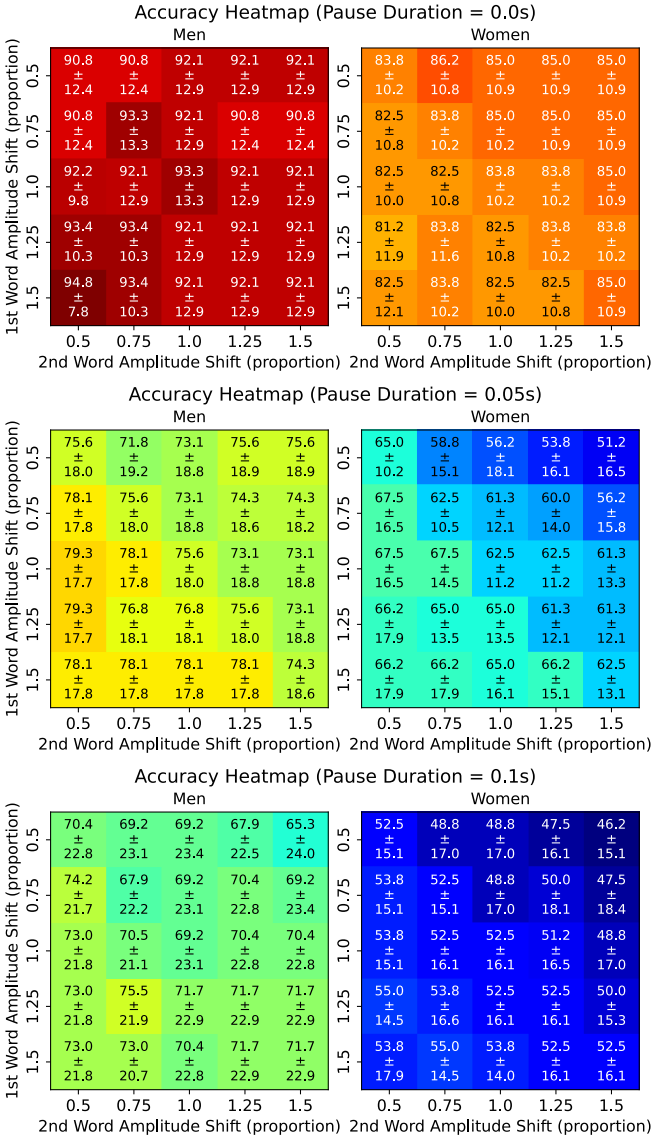

Figure 2: Heatmaps of Whisper's compound-word accuracy, which decreases either as the first amplitude decreases and the second amplitude increases or as pause duration increases.

feature distribution. The perturbations create a surface in the feature space over which we can observe continuous changes in Whisper's compound-word accuracy. If the accuracy decreases along an axis of the surface, this indicates that the decision boundary between the starting and target categories is sensitive to the corresponding perturbation. As this is a computationally costly procedure, we have limited the data to recordings from 10 men and 10 women, which decreased the overall accuracy for women.

For phrasal stress, these two categories are the compound words (starting) and the adjective-nouns (target), and we perform two sets of perturbations for pitch (**P1**) and amplitude (**P2**) to the *starting* category. In order to ground these perturbations, we first consider the distributions of pitch and amplitude shifts that participants produce for the first and second constituent words of their minimal pairs as well as pause duration shifts (Table 4). For a given constituent word, we measure pitch shift as the semitone difference (equivalent to frequency quotient) in average pitch from the starting category to the target category, which is extracted from the ranges of 80 to 450 Hz for women and from 30 to 400 Hz for men (Knutsen & Stromswold, 2024). Amplitude shift is measured as the target category's mean amplitude divided by the starting category's mean amplitude, and pause duration shift is the difference in pause duration between the categories. This process results in 4 distributions from which we remove outliers using the 1.5 IQR rule (5.9% removed) and select 5 representative values (Table 4). According to these distributions, **P1** shifts the pitch of each constituent in the compound word by -2, -1, 0, 1, and 2 semitones, and **P2** changes each constituent's amplitude by the proportions of 0.5, 0.75, 1.0, 1.25, and 1.5. Both **P1** and **P2** apply their respective shifts in conjunction with 3 changes to the duration of pause at the morpheme boundary of the compound word: 0.0, 0.05, and 0.1s, which deliberately exaggerates the pause durations to determine the most robust morphological characteristics of the decision boundaries.

Metric	Word	Percentile				
		10th	30th	50th	70th	90th
Pitch Shift (semitone)	1	−1.46	−0.55	−0.11	0.37	1.13
	2	−2.02	−0.63	0.08	0.92	2.15
Amp. Shift (proportion)	1	0.58	0.75	0.91	1.15	1.62
	2	0.53	0.77	0.99	1.33	1.77
Pause Shift (s)	N/A	0.00	0.00	0.00	0.01	0.03

Table 4: A summary of pitch, amplitude, and pause duration shifts from compound word to adjective-noun measured on participant data for phrasal stress.

Figure 1 depicts the compound-word accuracies of Whisper trained on phrasal stress and tested on **P1**. For all pause duration and gender combinations (except for recordings from women with a pause of 0.0s), we observe a strong quadratic decision boundary. In the exception case, the scope of semitones may be too small to observe a quadratic decision boundary, but a strong linear decision boundary is still evident along the same axis. As the pitch shifts of the constituent words diverge to either (-2, 2) or (2, -2), Whisper's accuracy decreases by as much as 21.5%, meaning that it is starting to recognize compound words as adjective-nouns. This quadratic decision boundary is consistent as pause duration increases, but the accuracies decrease globally with respect to pitch. For **P2**, Figure 2 shows a linear decision boundary when the pause duration is greater than 0.0s, where a decrease in the first word's amplitude and an increase in the second word's amplitude decreases Whisper's accuracy and the inverse improves its accuracy. When pause duration is 0.0s, no strong linear decision boundary is evident for either

men or women.

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.

Acoustic Context. An analysis of acoustic context revealed that stress interpretation depends on broader sentential prosody, not just local features. The superior performance of models with full and back acoustic context and the asymmetric contributions of front and back context, especially in contrastive stress, provides evidence for anticipatory and retrospective planning. A unique exception to this trend is the phrasal stress produced by men, for which acoustic context was not found to improve Whisper's performance. Nevertheless, Whisper's performance surpassed the gold standard of human performance for both men and women on phrasal and contrastive stress. Without acoustic context for lexical stress, Whisper's performance was near-human, but unable to exceed it. Relative to the RFC models, Whisper improved accuracy for women more than men for phrasal and lexical stress and for both women and men by $\sim$9.7 percentage points for contrastive stress. We conclude that Whisper is learning not only superior features than RFC models in general, but also better discriminatory features between men and women compared to RFC models. This is beneficial toward gender equity because it improves accuracy despite gender imbalances in pre-training data. Whisper offers an efficient and accurate alternative to the labor-intensive process of manual coding, enabling larger-scale prosodic studies that were previously unfeasible.

Stress Transfer. Unlike the RFC models, which as yet have only been applied to singular types of stress (Knutsen & Stromswold, 2024), we have demonstrated that Whisper can learn multiple types of stress in tandem. Furthermore, the RFC models cannot be used for transcription outside of the specific classification problem they were trained for, while Whisper (in this work) is being applied to classification *through* transcription, preserving its ASR capability. This works to Whisper's advantage when transferring acoustic patterns learned from one type of stress to another, because Whisper is relying on its extensive pre-training.

The observed transfer effects between different types of stress provide compelling evidence for shared acoustic patterns in stress production. Particularly noteworthy is the strong bidirectional transfer between lexical and contrastive stress (+32.4% and +27.8%), suggesting similar acoustic patterns between word-level and discourse-level prosodic phenomena. These findings quantitatively support theoretical frameworks proposing common acoustic patterns underlying different forms of prosodic stress (Ladd, 2008). In contrast, 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 (Knutsen & Stromswold, 2024).

Decision Boundaries. In prior work, RFC models have proven valuable for ranking acoustic feature importance, which is challenging to achieve for a black-box model such as Whisper. However, our proposed evaluation methodology for identifying decision boundaries in acoustic feature space bridges the gap in interpretability between Whisper and RFC models. Namely, the systematic perturbation of pitch, amplitude, and pause duration elicits changes in Whisper's accuracy that we can analyze. Figure 1 shows that when pitch perturbations **P1** diverge, Whisper recognizes compound words as adjective-nouns. When the first pitch decreases and the second pitch increases, the compound stress pattern approaches the phrasal stress pattern attributed to adjective-noun. On the other hand, when the first pitch increases and the second pitch decreases, we posit that the exaggeration of the compound stress pattern brings it closer to contrastive stress (e.g., "GREEN house"). Meanwhile, the amplitude perturbations **P2** only share one side of this effect: decreasing the first amplitude and increasing the second amplitude (in accordance with the canonical phrasal stress pattern) causes Whisper to detect adjective-nouns (Figure 2). Within the scope of amplitude changes being investigated, it appears that increasing the first amplitude and decreasing the second often *reinforces* the recognition of compound words (i.e., increases Whisper's accuracy) in spite of increases in the pause between words. These findings reveal a complex interplay between pitch, amplitude, and pause duration, which offers prosodic annotation models a new way to analyze their learned acoustic patterns beyond coarse feature importance scores.

Conclusion

Whisper demonstrates near-human and super-human capabilities for recognizing prosodic stress, harnessing variable-length acoustic context, and transferring learned stress patterns, greatly surpassing prior work in accuracy, accessibility, and robustness. The final gap between Whisper and prior work was in interpretability, which we have addressed with our black-box evaluation methodology. This method elucidates the complex interplay of acoustic features, which importance scores convey coarsely, and it makes no assumptions about the model, meaning that all models can be evaluated in a standardized. With proper fine-tuning using a very small, carefully curated dataset, Whisper could become a promising tool for cross-linguistic prosodic research, potentially illuminating questions about cross-language and language-specific patterns in stress. The fine-tuned Whisper model will be publicly released upon acceptance.

References

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. In *Journal of memory and language* (pp. 644–663).

Carlson, K. (2009). How prosody influences sentence comprehension. In *Language and linguistics compass.*

De Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. In *Advances in methods and practices in psychological science.*

Ferreira, F. (1993). Creation of prosody during sentence production. In *Psychological review.*

FindingFive Team. (2019). *FindingFive: A web platform for creating, running, and managing your studies.* NJ, USA. Retrieved from `https://www.findingfive.com`

Knutsen, S., & Stromswold, K. (2024). Gender differences in the acoustic realization of stress. *Penn Working Papers in Linguistics*.

Koffi, E., & Mertz, G. (2018). Acoustic correlates of lexical stress in central minnesota english. *Linguistic Portfolios*, *7*(1), 7.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Pierrehumbert, J. (1990). The meaning of intonational contours in the interpretation of discourse. In *Intentions in communication/bradford book.*

Plag, I. (2006). The variability of compound stress in english: structural, semantic, and analogical factors. *English Language & Linguistics*, *10*(1), 143–172.

Radford, A., et al. (2023). Robust speech recognition via large-scale weak supervision. In *Icml.*

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. In *Journal of memory and language.*

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.