Harnessing Whisper for Prosodic Stress Analysis

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

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. Models can be found at github.com/SSSohn/ProsodyBench.

1 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 and 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 and 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.

2 Preliminaries

At its core, Whisper leverages deep learning to analyze audio waveforms, extract patterns aligned with human speech, and decode these patterns into

Findings of the Association for Computational Linguistics: ACL 2025, pages 25931–25942
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

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.

2.1 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.

2.2 Fine-tuning

The fine-tuning dataset is based on an experiment (Knutsen and 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 Online Profiling Elements of Prosody in Speech Communication test (Peppé and McCann, 2003; Knutsen et al., 2023).[1] No participants reported any issues with vision, hearing, language abilities (spoken or written), learning, or other neuropsychological conditions. For phrasal stress, participants produced 16 compound word and adjective-noun 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").

Stress	Minimal Pair Transcription
Phrasal	The <greenhouse / green house> spoils the view.
Phrasal	There's a <darkroom / dark room> in this house.
Phrasal	The <whiteboard / white board> needs cleaning.
Phrasal	That <hotdog / hot dog> is under the table.
Phrasal	A <blackbird / black bird> just flew past.
Phrasal	His <wetsuit / wet suit> is on the floor.
Phrasal	That <bluebell / blue bell> is pretty.
Phrasal	The <bullseye / bull's eye> 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.

To facilitate model training, transcriptions are capitalized to reflect canonical English stress patterns. All minimal pair transcriptions have been listed in Table 1 ordered as <stress on the first syllable / stress on the final syllable>. 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 and unique participants, iteratively training on four sub-

[1] The study was approved by Rutgers University's Institutional Review Board (IRB Number: Pro2019003032-MOD2025000804). All participants provided informed consent and were compensated with course credits.

sets and testing on the held-out fifth (De Rooij and Weeda, 2020). This aligns with findings from Xie et al. (Xie et al., 2021), which demonstrate that talker-specific variability is the primary source of ambiguity in prosodic meaning. Their work shows that prosodic cues (e.g., stress) vary significantly across speakers, requiring listeners to learn speaker-specific distributions for accurate disambiguation. By splitting data across participants, we directly evaluate Whisper's ability to handle this real-world variability, which is central to robust stress perception.

3 Related Work

As a baseline for model comparison, we use the Knutsen and Stromswold study (Knutsen and 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.

3.1 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 and Mertz, 2018), which after re-analysis by Knutsen and Stromswold showed that amplitude and duration play crucial roles for both 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.

3.2 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.

4 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

Acoustic Context	**Phrasal Stress** (SD)			**Lexical Stress** (SD)			**Contrastive Stress** (SD)		
	All	Men	Women	All	Men	Women	All	Men	Women
Syl. 1	81.1% (6.6)	84.5% (5.9)	77.2% (8.7)	73.3% (3.5)	71.5% (5.1)	73.6% (7.9)	76.7% (8.2)	76.1% (12.6)	77.0% (5.9)
Syl. 2	87.7% (3.0)	89.2% (4.0)	85.2% (6.0)	79.8% (5.3)	82.0% (7.3)	77.7% (11.4)	86.3% (2.9)	86.5% (5.5)	86.4% (3.2)
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, trained human coders, and RFC models.

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. Even within the ROI, the RFC uses limited hand-crafted features, i.e., the differences in mean F0, mean amplitude, and duration between the first and second syllable/morpheme. 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 there would otherwise be more conspicuous *syntactic* differences than prosodic differences. Hence, we could not analyze the role of context for lexical stress. 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.

We show that the first syllable/morpheme (e.g., "IN-", "green-", and "BLACK") and the second syllable/morpheme (e.g., "-sult", "-house", and "cow") hold different amounts of prosodic information. For both men and women, the second syllable/morpheme holds significantly more information than the first (Table 2), which is most prominent in contrastive stress accuracy (∼10 percentage point increase). The second-syllable accuracy is worse than the RFC's accuracy for lexical stress, comparable to the RFC for phrasal stress, and superior to the RFC for contrastive stress —most notably for women's contrastive stress trials by 4 percentage points. This suggests that the acoustic features in the second syllable/morpheme are better separated, and thereby more discriminative of the minimal pairs.

Table 2 also 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 ∼9.7 percentage points.

5 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 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.

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, single-stress, all-stress, coders and RFCs. $^{\dagger}p < .01$ $^{*}p < .05$

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). Models were also re-trained with less noisy data by removing participant recordings that were coded as the opposite category by a majority of coders were removed (which could be considered as incorrect productions by participants). For phrasal stress, $56/575$ recordings were removed; $65/575$ recordings were removed for lexical stress; and $73/573$ recordings were removed. Table A.5 demonstrates that both the single- and all-stress models achieve significantly higher accuracy on their respective stress types when applied to this subset of data compared to the full dataset. Transfer accuracy between lexical and contrastive stress also improved, though the gains were more modest. The table reports precision and recall for stress-first and stress-final categories across each stress type, using a symmetric evaluation approach where each category alternately served as the positive class while the other was treated as negative.

Table 3 shows that the 2×2 matrix of lexical and contrastive results for single-stress models and the phrasal$\rightarrow$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 and Stromswold, 2024).

6 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 can be applied to any stress type and any prosodic annotation model. For each stress type, the minimal pairs can be separated into two categories based on their canonical stress patterns: stress-first or stress-final (Table 1). We select stress-first as the starting distribution (but either works) and we systematically perturb the acoustic features of every recording in that category such that they progressively capture some of the other target category's acoustic feature distribution. The perturbations create a surface in the feature space over which we can observe changes in Whisper's start-category accuracy. If the accuracy decreases along an axis of the surface, this indicates that the decision boundary between the starting and target categories (where accuracy = 50%) is sensitive to the corresponding perturbation.

We use the stress-first category for each stress type and perturb its pitch (**P1**), amplitude (**P2**), and pause duration (**P1** and **P2**). In order to ground these perturbations, we first consider

Stress	Metric	Word	Percentile				
			10th	30th	50th	70th	90th
Phrasal	Pitch Shift (semitone)	1 2	−1.46 −2.02	−0.55 −0.63	−0.11 0.08	0.37 0.92	1.13 2.15
	Amp. Shift (proportion)	1 2	0.58 0.53	0.75 0.77	0.91 0.99	1.15 1.33	1.62 1.77
	Pause Shift (s)	N/A	−0.07	−0.03	0.00	0.01	0.06
Lexical	Pitch Shift (semitone)	1 2	−2.95 −7.96	−1.16 −2.02	0.00 −0.04	1.25 1.97	3.18 9.68
	Amp. Shift (proportion)	1 2	0.36 0.39	0.64 0.63	0.9 0.95	1.23 1.4	1.85 2.04
	Pause Shift (s)	N/A	−0.01	0.00	0.00	0.00	0.01
Contrastive	Pitch Shift (semitone)	1 2	−1.68 −10.57	−0.48 −4.79	0.65 −2.28	1.68 0.77	2.72 6.01
	Amp. Shift (proportion)	1 2	0.59 0.35	0.84 0.48	1.03 0.65	1.31 0.88	1.76 1.57
	Pause Shift (s)	N/A	−0.05	−0.01	0.00	0.01	0.07

Table 4: A summary of pitch, amplitude, and pause duration shifts from the stress-first category to the stress-final category measured on participant data.

the distributions of pitch and amplitude shifts that participants produce for the first and second syllables/morphemes of their minimal pairs as well as pause duration shifts across both syllables/morphemes (Table 4). For either syllable/morpheme, we measure pitch shift as the semitone difference (equivalent to frequency quotient) in average pitch from the stress-first category to the stress-final category, which is extracted from the ranges of 80 to 450 Hz for women and from 30 to 400 Hz for men (Knutsen and Stromswold, 2024). Amplitude shift is measured as the stress-final category's mean amplitude divided by the stress-first category's mean amplitude, and pause duration shift is the difference in pause duration between the categories. This process results in 5 distributions from which we remove outliers using the 1.5 IQR rule and select 5 representative percentiles (Table 4). According to these distributions, the stress-first category of each stress type has its pitch and pause duration shifted by **P1** and its amplitude and pause duration shifted by **P2**. When absolute pitch shift is greater 4 semitones, the perturbation exhibits artifacts, so we replace such distributions (i.e., 2nd syllable of lexical stress and 2nd morpheme of contrastive stress) with [-4,-2,0,2,4]. Only positive shifts in pause duration are considered for computational simplicity, and for lexical stress, pause duration is not perturbed because its distribution is rarely non-zero. If any perturbation causes Whisper not to recognize a recording as either category, it is not considered in the accuracy (Figures A.3–A.7).

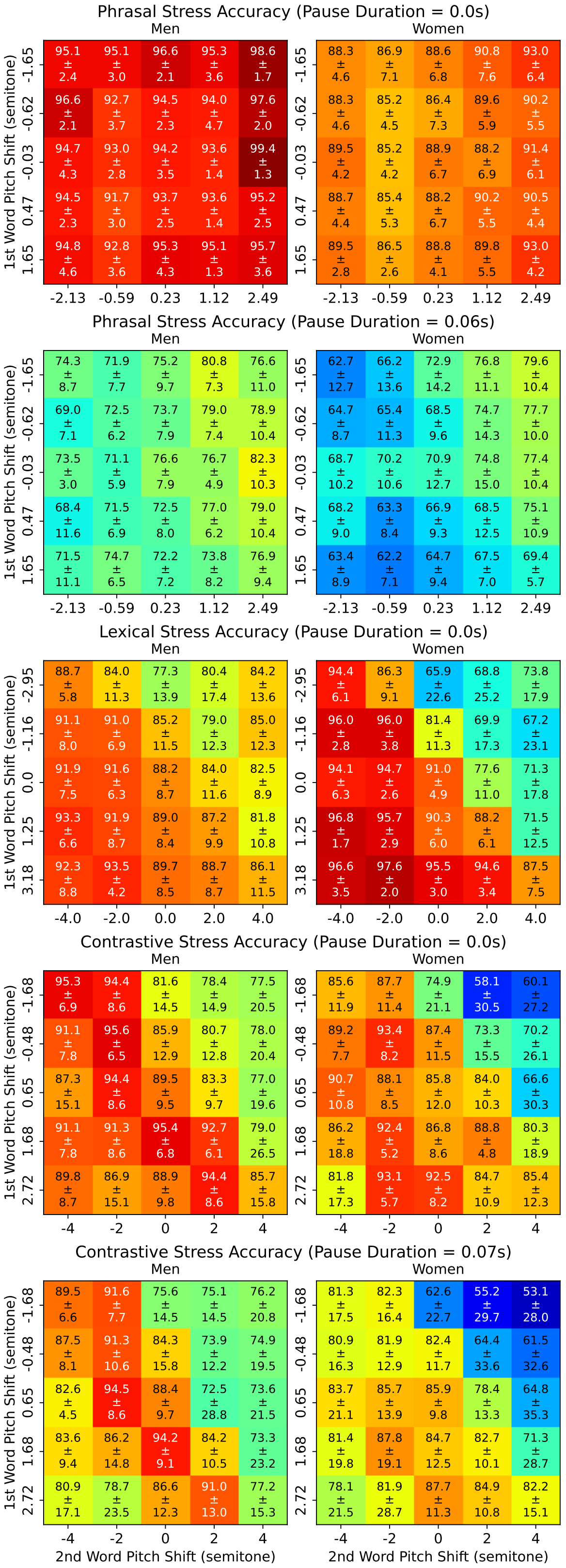

Figure 1: Heatmaps of Whisper's stress-first accuracy showing that decision boundaries for lexical and contrastive stress are sensitive to pitch shift divergences and for phrasal stress is sensitive to pause duration shift. Hotter colors indicate higher accuracy.

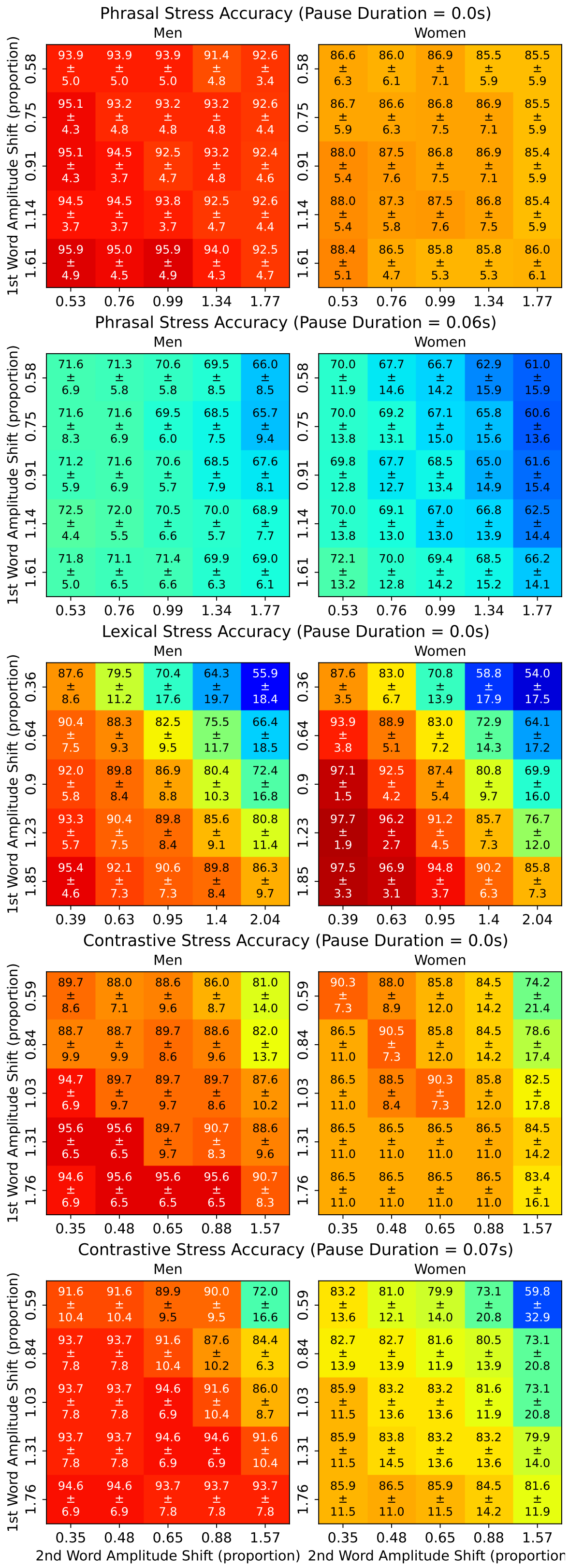

Figure 2: Heatmaps of Whisper's stress-first accuracy for phrasal, lexical, and contrastive stress, which tends to decrease as the first syllable's amplitude decreases and the second amplitude increases.

Figure 1 depicts the stress-first accuracies of Whisper trained on phrasal, lexical, and contrastive stress and tested on **P1**. For phrasal stress, there is an amorphous decision boundary that is apparent for men and women when pause duration increases by 0.06s. However, when the pause shift is 0s, it is no longer evident. We interpret the boundary's sensitivity to pitch shift as an artifact from overfitting that exacerbates at distributional extremes. Otherwise, the decision boundary is clearly sensitive to pause duration. For lexical stress, we observe a linear decision boundary along the pitch divergence axis (from bottom-left to top-right) at a pause duration shift of 0s, which is representative of the feature distribution (Table 4). The boundary is much stronger for women than men to the extent that decreasing first pitch and increasing second pitch lowers accuracy for women by as much as 25.1 percentage points and the inverse shifts raise accuracy up to as high as 97.6%. Contrastive stress exhibits a quadratic decision boundary along the same axis as lexical stress, but within the distribution, the boundary is effectively linear in the same way as lexical stress. Unlike lexical stress, which does not vary much in pause duration, contrastive stress has a similar distribution of pause durations as phrasal stress (Table 4), but is much less sensitive to pause duration than phrasal stress.

For **P2** (Figure 2), we observe a linear decision boundary along the axis of *amplitude* divergence (bottom-left to top-right) for lexical and contrastive stress, where stress-first accuracy decreases as the first amplitude decreases and second amplitude increases. The decision boundary is strongest for lexical stress and moderately strong for contrastive stress with low sensitivity to pause duration. Similar to pitch shift, the linear decision boundary is only apparent for phrasal stress when pause duration increases by 0.06, which reduces its credibility. However, its high sensitivity to pause duration still holds true.

7 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. Between local features in the first and second syllables/morphemes, we found that both hold information about the categories of each stress type, but the latter consistently holds more information. Accordingly, we recommend against the prior practice of only using relative feature values between the two syllables/morphemes (Knutsen and Stromswold, 2024), because there is useful information in their absolute values. Beyond the region of interest, 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 ∼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 and 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 and 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. Figures 1 and 2 show that lexical and contrastive stress are sensitive to perturbations in pitch (**P1**) and amplitude (**P2**), while phrasal stress is sensitive to perturbations in pause duration. More specifically for lexical and contrastive stress, the decision boundaries between the stress-first and stress-final categories are approached as the first syllable of the stress-first category decreases in pitch/amplitude and the second pitch/amplitude increases. On the other hand, the inverse perturbations greatly exaggerate the stress-first pattern in lexical stress, *increasing* its recognition well beyond the original recordings. This aligns with canonical stress patterns for lexical and contrastive stress (Solé Sabater, 1991) and makes visually clear the continuum between the stress-first and stress-last categories with respect to pitch, amplitude, and pause duration. These findings reveal more nuance to the 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.

8 Conclusion

Whisper demonstrates near-human and superhuman capabilities for recognizing prosodic stress, harnesses variable-length acoustic context with ease, and transfers learned acoustic patterns between broader stress types, greatly surpassing prior work in accuracy 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 nuanced interplay of acoustic features, which importance scores only convey coarsely, and it makes no assumptions about the model, meaning that all models can be evaluated in a standardized manner. 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 weights (trained on the full dataset and the less noisy subset for all stress types) are available at github.com/SSSohn/ProsodyBench.

9 Limitations

There are several limitations of this work to consider. First, the 24 total minimal pairs are based on the well-studied Profiling Elements of Prosody in Speech Communication (PEPS-C) test (Peppé and McCann, 2003), but the acoustic patterns learned on these pairs have not yet been evidenced to generalize to other examples within the same types of stress. Our cross-validation procedure divided participants into different splits, but all splits had equal and full representation of the corresponding minimal pairs. However, our investigation of acoustic pattern transfer between stress types lends some credibility to the completeness of the PEPS-C test from a modeling standpoint. Second, regarding the transfer between stress types, the control participant was not varied, because the fine-tuning of 5 models (with 5-fold cross-validation) is conditioned on their data. Evaluating all participants in the same way would require a total of 900 model instances to be fine-tuned. Finally, regarding the characterization of decision boundaries, the same perturbations were applied between both men and women for methodological simplicity. However, this simplification caused some recordings to fall out of the distribution of real recordings. Figures A.3–A.7 show how certain perturbations resulted in recognition percentages to drop. Although the accuracies reported in Figures 1 and 2 only consider recordings that are within distribution, this recognition issue reveals that the pitch, amplitude, and pause duration shift operations (1) could be improved and applied with more nuance (for instance, gender-wise) and (2) do not fully capture the acoustic differences between the minimal pair categories.

References

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

K. Carlson. 2009. How prosody influences sentence comprehension. In *Language and Linguistics Compass*.

M. De Rooij and W. Weeda. 2020. Cross-validation: A method every psychologist should know. In *Advances in Methods and Practices in Psychological Science*.

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

Sten Knutsen, Sue Peppe, and Karin Stromswold. 2023. Online profiling elements of prosody in speech communication (o-peps-c).

Sten Knutsen and Karin Stromswold. 2024. Gender differences in the acoustic realization of stress. *University of Pennsylvania Working Papers in Linguistics (PWPL)*, 31.1.

Ettien Koffi and Grace Mertz. 2018. Acoustic correlates of lexical stress in central minnesota english. *Linguistic Portfolios*, 7(1):7.

D Robert Ladd. 2008. *Intonational phonology*. Cambridge University Press.

Sue Peppé and Joanne McCann. 2003. Assessing intonation and prosody in children with atypical language development: the peps-c test and the revised version. *Clinical Linguistics & Phonetics*, 17(4-5):345–354.

J. Pierrehumbert. 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication/Bradford Book*.

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

Alec Radford et al. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*. PMLR.

J. Snedeker and J. Trueswell. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. In *Journal of Memory and language*.

Maria-Josep Solé Sabater. 1991. Stress and rhythm in english. *Revista alicantina de estudios ingleses, No. 04 (Nov. 1991); pp. 145-162*.

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

Xin Xie, Andrés Buxó-Lugo, and Chigusa Kurumada. 2021. Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, 211:104619.

A Appendix

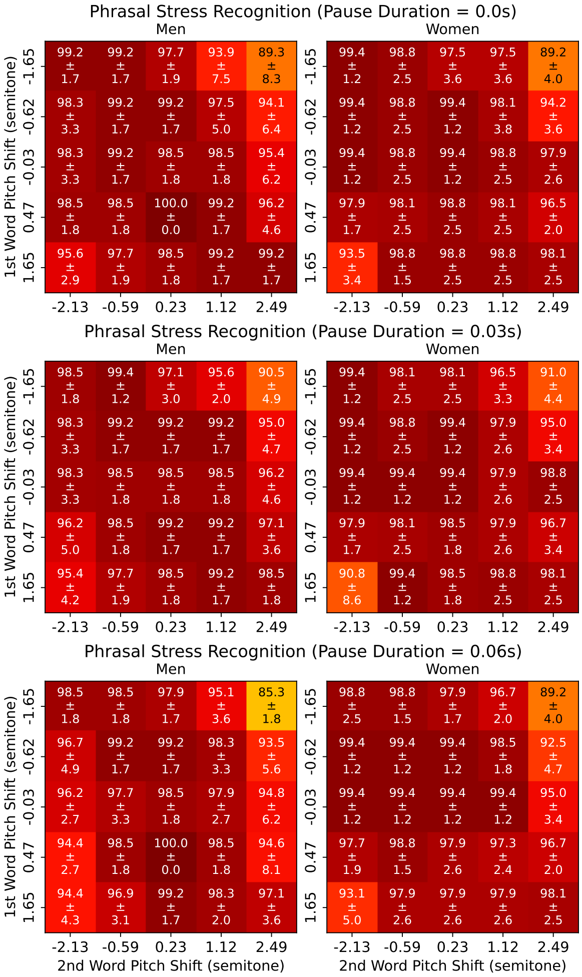

Figure 3: Heatmaps of Whisper's minimal pair recognition percentage for phrasal stress as a function of pitch and pause duration shift. Hotter colors indicate higher recognition.

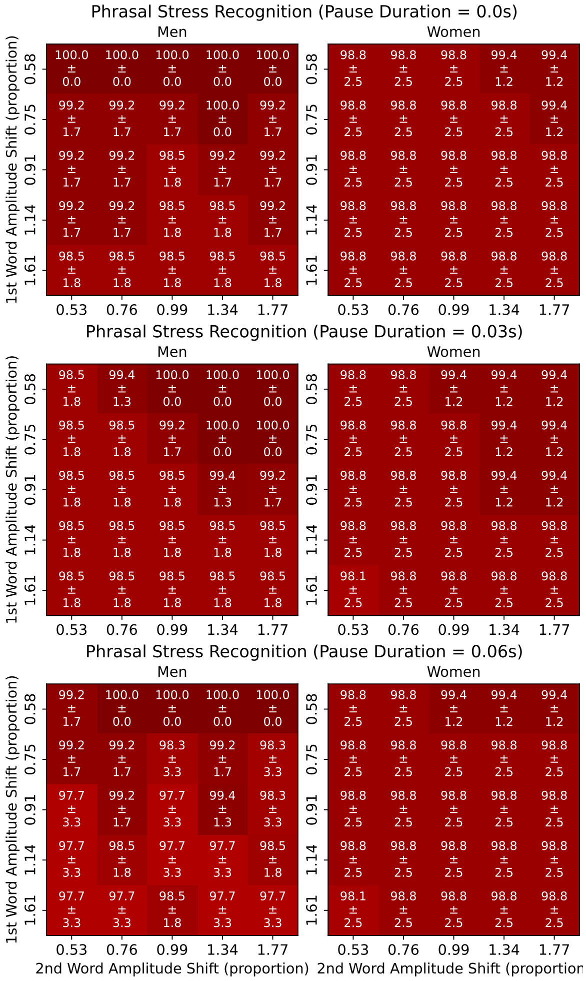

Figure 4: Heatmaps of Whisper's minimal pair recognition percentage for phrasal stress as a function of amplitude and pause duration shift. Hotter colors indicate higher recognition.

Training Stress	**Testing Stress**	**Stress Placement**				
		All	Stress-First		Stress-Final	
		Accuracy (SD)	Precision (SD)	Recall (SD)	Precision (SD)	Recall (SD)
Control	Phrasal	0.684 (0.056)	0.724 (0.049)	0.613 (0.157)	0.669 (0.070)	0.754 (0.105)
Phrasal		0.934 (0.028)	0.929 (0.045)	0.945 (0.020)	0.942 (0.020)	0.922 (0.054)
Lexical		0.716 (0.080)	0.751 (0.097)	0.673 (0.109)	0.696 (0.081)	0.759 (0.125)
Contrastive		0.564 (0.021)	0.597 (0.028)	0.441 (0.073)	0.546 (0.018)	0.689 (0.077)
All		0.940 (0.022)	0.946 (0.041)	0.937 (0.008)	0.936 (0.007)	0.942 (0.047)
Control	Lexical	0.519 (0.047)	0.532 (0.057)	0.502 (0.073)	0.504 (0.041)	0.534 (0.063)
Phrasal		0.671 (0.076)	0.739 (0.117)	0.565 (0.080)	0.631 (0.056)	0.787 (0.096)
Lexical		0.933 (0.028)	0.942 (0.045)	0.928 (0.038)	0.927 (0.029)	0.939 (0.048)
Contrastive		0.739 (0.094)	0.706 (0.098)	0.862 (0.053)	0.799 (0.087)	0.612 (0.140)
All		0.931 (0.022)	0.927 (0.035)	0.940 (0.035)	0.938 (0.029)	0.921 (0.039)
Control	Contrastive	0.486 (0.059)	0.529 (0.173)	0.168 (0.054)	0.480 (0.042)	0.830 (0.084)
Phrasal		0.489 (0.046)	0.545 (0.162)	0.108 (0.067)	0.485 (0.031)	0.903 (0.053)
Lexical		0.801 (0.096)	0.896 (0.083)	0.720 (0.188)	0.763 (0.117)	0.896 (0.088)
Contrastive		0.960 (0.023)	0.961 (0.041)	0.964 (0.015)	0.960 (0.019)	0.956 (0.047)
All		0.974 (0.025)	0.972 (0.034)	0.979 (0.016)	0.977 (0.018)	0.969 (0.038)

Table 5: Accuracy of the control model, single-stress models, and the all-stress model as well as the precision and recall for both stress-first and stress-final categories within each stress type. All metrics have been averaged across 5 folds.

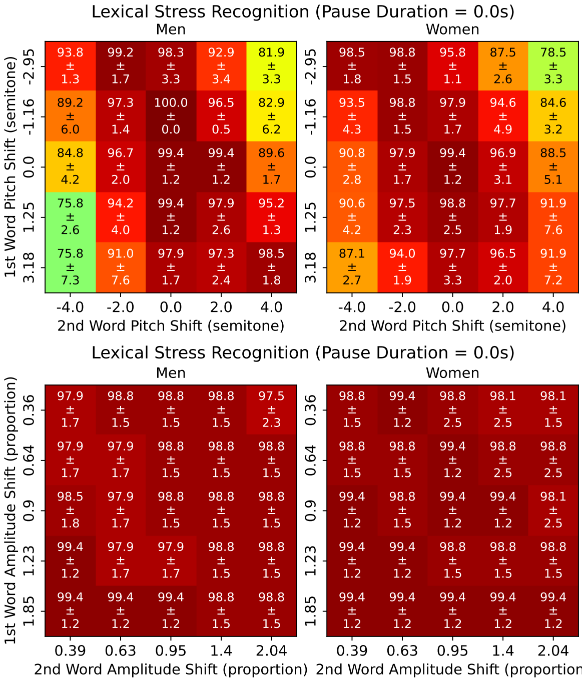

Figure 5: Heatmaps of Whisper's minimal pair recognition percentage for lexical stress as a function of amplitude and pause duration shift (top) and pitch and pause duration shift (bottom). Hotter colors indicate higher recognition.

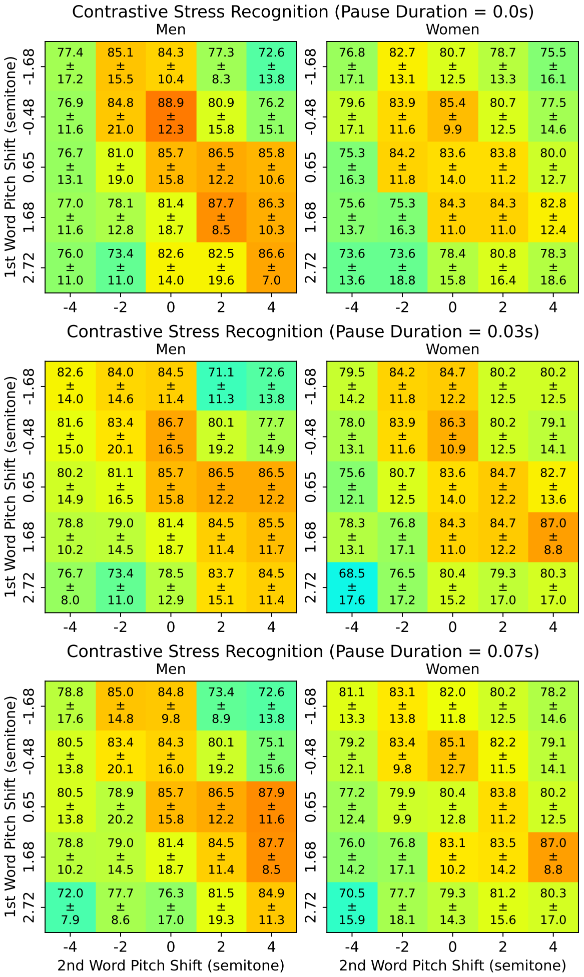

Figure 6: Heatmaps of Whisper's minimal pair recognition percentage for contrastive stress as a function of pitch and pause duration shift. Hotter colors indicate higher recognition.

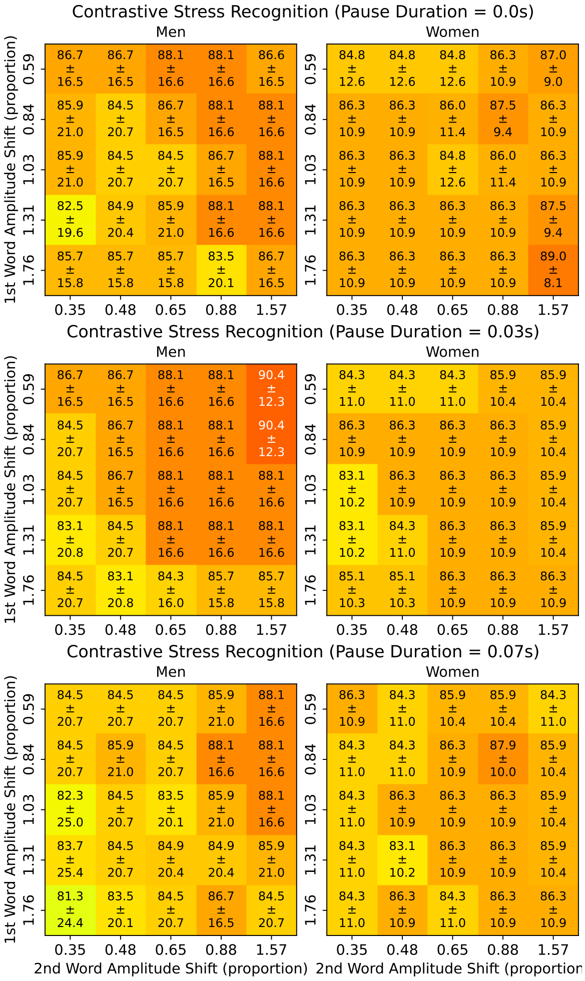

Figure 7: Heatmaps of Whisper's minimal pair recognition percentage for contrastive stress as a function of amplitude and pause duration shift. Hotter colors indicate higher recognition.