Democratizing Prosodic Stress Recognition Across Genders and Neurotypes

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

1 Introduction. The accurate transcription of speech recordings for diverse populations is crucial to enhancing inclusivity in communication technologies. Prosody affects how both people and Automatic Speech Recognition (ASR) models process spoken sentences [2, 10, 4, 8]. For instance, phrasal stress is vital for differentiating compound words from their adjective-noun counterparts (e.g., "greenhouse" vs. "green house"). This work focuses on leveraging OpenAI's Whisper large-v2 model [9], a state-of-the-art ASR system, to demonstrate through the lens of phrasal stress how equitable outcomes can be achieved across 4 groups: neurotypical males (NT-M), neurotypical females (NT-F), males with Autism Spectrum Disorder or ASD (ASD-M), and females with ASD (ASD-F). To this end, we used a fine-tuning dataset based on an experiment with 66 native English-speaking college students (18 NT-M, 18 NT-F, 12 ASD-M, and 18 ASD-F) from the mid-Atlantic U.S. [7]. Participants in the ASD groups scored above 28 on the Autism-Spectrum Quotient test [3] and/or reported a formal ASD diagnosis. All participants were tasked with producing 16 Adj-N and compound word minimal pairs embedded in sentences (e.g., "The white board/whiteboard is dirty").

2 Classification. Although Whisper was intended for transcription, we demonstrate that it can be fine-tuned on the dataset (into the proposed Whisper-C model) to *classify* recordings into NT-M, NT-F, ASD-M, ASD-F, and an "unknown" class (UK) for ambiguous recordings. Whisper-C was trained 5 separate times for 5-fold cross-validation, each time using default hyperparameters and a different 20% partition of the dataset for testing [5]. Cross-validation was preferred over a paired t-test with random data splits, as it provides a more robust estimate of model performance averaged over its splits [1, 6, 5]. Table 1 shows that Whisper-C achieved near-perfect precision and moderate recall across all known categories except ASD-M. Using just one recording with a mean duration of 1.7 (SD=0.4) seconds, 55.4% of all cases were correctly classified. Ambiguous inputs were effectively categorized as UK, with this fallback mechanism capturing 42.6% of all cases and safeguarding the reliability of the classification pipeline. We aggregated Table 1 into separate 2×2 confusion matrices for M-F and NT-ASD by considering all UK cases as misclassifications (for a worst-case analysis). Fisher's exact test yielded p-values of $<$ 0.00001 for M-F and $<$ 0.0042 for NT-ASD, indicating highly significant results. We partly attribute the low precision and recall for ASD-M to there being 33.3% less data than other classes. This coincidental class imbalance is dramatically amplified in large-scale datasets used to train ASR models [8].

3 Transcription. Using Whisper-C, we can deploy a class-specific *transcription* model that best suits an individual. Six transcription models (4 versions of Whisper fine-tuned on individual classes, 1 pre-trained version, and a composite of the other 5) were cross-validated in the same way as Whisper-C. Table 2 presents test accuracy results, highlighting that in this small context within phrasal stress, a model trained on class X is not always most accurate when tested on class X. While interesting for prosodic analysis, this finding on 0.5 hours of controlled audio data does not reflect the bias of ASR models trained on over 500,000 hours against individuals with disabilities [8]. Nevertheless, the Composite model (Table 2) demonstrates that switching from pre-trained to class-trained transcription when that class is detected by Whisper-C can significantly improve performance for that class. In practice, under-represented groups (e.g., ASD-M and ASD-F) are expected to see larger gains in performance from this adaptive method since the class imbalance already favors neurotypical individuals [8].

4 Discussion. This research bridges psycholinguistic exploration with practical societal applications by addressing the unique challenges posed by diverse speech patterns. By enabling precise classification and transcription tailored to specific user groups, this approach holds the potential to transform education, healthcare, and advocacy efforts. For example, tailored transcription models can empower educators and therapists working with neurodivergent individuals, providing insights into speech patterns and supporting effective intervention strategies. This study not only advances ASR technology but also exemplifies how cutting-edge research can address pressing societal challenges. By focusing on inclusivity, it lays the groundwork for accessible communication technologies that benefit diverse populations, marking a significant step toward equitable technological solutions.

Ground Truth Label	Predicted Label				
	NT-M (SD)	NT-F (SD)	ASD-M (SD)	ASD-F (SD)	UK (SD)
NT-M	64.2% (12.4)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	35.8% (12.4)
NT-F	0.0% (0.0)	65.5% (11.4)	0.0% (0.0)	0.4% (0.8)	34.1% (11.4)
ASD-M	8.5% (7.9)	0.0% (0.0)	26.0% (8.4)	0.0% (0.0)	65.4% (7.7)
ASD-F	0.0% (0.0)	1.6% (2.4)	0.0% (0.0)	55.1% (11.6)	43.3% (11.2)

Table 1: Whisper-C Classification Accuracy. This table reports the classification accuracy of Whisper-C after 5-fold cross-validation. Its class-wise precision is near-perfect and recall is moderate for all classes except ASD-M, which has 33.3% less data.

Training Class	Testing Class			
	NT-M (SD)	NT-F (SD)	ASD-M (SD)	ASD-F (SD)
NT-M	92.4% (2.8)	91.6% (3.7)	92.1% (5.5)	90.8% (3.5)
NT-F	92.0% (3.3)	90.6% (2.6)	93.1% (5.2)	89.2% (4.8)
ASD-M	89.1% (5.0)	84.9% (10.0)	91.5% (7.5)	85.4% (3.2)
ASD-F	93.7% (3.8)	91.5% (4.2)	93.1% (5.7)	88.6% (4.9)
Pre-trained	76.4% (3.8)	75.2% (7.5)	80.2% (3.8)	76.1% (3.1)
Composite	87.6% (3.6)	86.8% (5.0)	83.3% (3.5)	86.9% (3.6)

Table 2: Prosodic Transcription Accuracy. This table reports the transcription accuracy for 6 models that were all 5-fold cross-validated. The four class-specific models showed an asymmetric generalization, where no training class had the best average accuracy on its own test set. The Composite model, combining pre-trained and class-trained transcription using Whisper-C, was significantly more accurate than the pre-trained model in every class except ASD-M, which had the least data and the least accurate class-specific transcription model.

References.

[1] Arlot, S. and Celisse, A., 2010. "A Survey of Cross-Validation Procedures for Model Selection". *Statistics Surveys*.

[2] 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*.

[3] Broadbent, J., Galic, I., and Stokes, M. A., 2013. "Validation of Autism Spectrum Quotient Adult Version in an Australian Sample". *Autism Research and Treatment*.

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

[5] De Rooij, M. and Weeda, W., 2020. "Cross-Validation: A Method Every Psychologist Should Know". *Advances in Methods and Practices in Psychological Science*.

[6] Dietterich, T. G., 1998. "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms". *Neural Computation*.

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

[8] Ngueajio, M. K. and Washington, G., 2022. "Hey ASR System! Why Aren't You More Inclusive?" *International Conference on Human-Computer Interaction*.

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

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