Analyzing Whisper's Representation Learning of Prosodic Stress Patterns

Samuel S. Sohn, Kavindya Dalawella, Sten Knutsen, Karin Stromswold
Department of Psychology & Center for Cognitive Science, Rutgers University – New Brunswick

1 Introduction

Prosody serves as a critical cue in human communication, disambiguating meaning, signaling intent, and shaping linguistic interpretation. It guides listeners' resolution of ambiguous sentences while modulating garden-path effects (Carlson, 2009), just as speakers unconsciously modulate these features to convey nuanced meaning (Ferreira, 1993; Beach, 1991; Snedeker and Trueswell, 2003). Given this dual role in human speech processing, prosody should be equally vital for computational systems aiming to achieve human-like understanding. Recent advances in automatic speech recognition (ASR), particularly OpenAI's Whisper large-v2 (Radford et al., 2023), now enable accurate, scalable prosodic analysis. Sohn et al. (Sohn et al., 2025a,b,c,d) demonstrated that a fine-tuned Whisper model achieves prosodic stress annotation accuracy comparable to human performance. Besides its annotation utility, Whisper's internal representations can be analyzed to gain insight into the mechanisms underlying the perception of phrasal, lexical, and contrastive stress in neural systems. This not only provides a tool for linguistic research but also offers insights into the mechanisms underlying stress perception in neural systems.

2 Methods

The fine-tuning dataset used by Sohn et al. (Sohn et al., 2025b) is from an experiment by Knutsen and Stromswold (Knutsen and Stromswold, 2024) assessing the prosodic ability of 36 native English-speaking college students (18 men and 18 women) from the mid-Atlantic U.S. Participants were tasked with producing prosodic stress to distinguish meaning as specified by the Online Profiling Elements of Prosody in Speech Communication (Knutsen et al., 2023) on 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., "His <wetsuit/wet suit> is on the floor" in Table 1). For lexical stress, they produced 16 words differing only in stress pattern (e.g., "<*dis*card/dis*card*>"). For contrastive

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 all minimal pairs by prosodic stress type.

Proceedings of *AMP 2025*

stress, they listened to 16 sentences in which either a color or animal did not match a picture (e.g., "The blue cow has the ball" while looking at a red cow) and corrected the error both lexically and prosodically (e.g., "The <*red* cow/red *cow*> has the ball"). All utterances were truncated to the bracketed regions.

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)†
Human Coders	91.9% (1.6)	88.8% (1.6)	91.6% (1.5)

Table 2: Accuracy of Whisper trained on control stress, single-stress, all-stress, and human coders as reported in (Sohn et al., 2025c). $^{\dagger}p < .01$ $^{*}p < .05$ vs. Control

Five fine-tuned models have been provided by Sohn et al. for prosodically annotating utterances: a minimally trained control model; 3 single-stress models trained on either phrasal, lexical, or contrastive stress; and a model trained on all stress types. Table 2 reports the prosodic annotation accuracy of each model on the three stress types. By comparing these models, Sohn et al. found a clear dissociation: while lexical stress training shows strong positive transfer to/from contrastive stress (Δ = 27.8%/32.4%), phrasal stress demonstrates negative transfer to/from contrastive stress (Δ = -7.7%/-11.5%) and weak transfer to/from lexical stress (Δ = 9.2%/3.9%). Notably, joint training on all stress types resolves these conflicts, yielding near-human accuracy by discovering non-interfering patterns. In this paper, we focus on *how* Whisper organizes phrasal, lexical, and contrastive stress by analyzing their latent representations at the final decoding layer.

3 Analysis

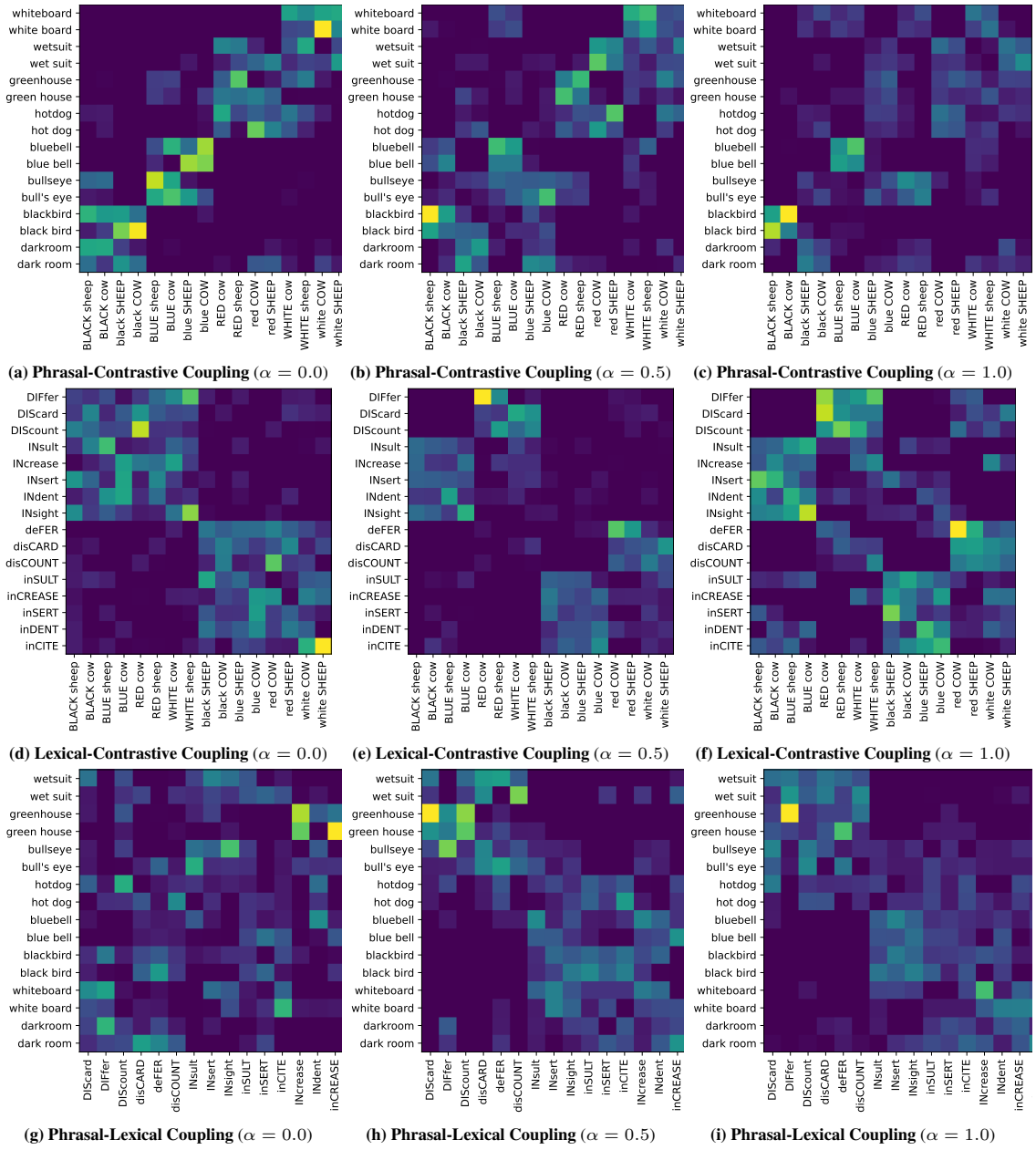
We labeled representations using 2 label sets. The 8 minimal pair id's per stress type act as *segmental* labels because they have full phonemic overlap, and the 2 canonical stress patterns shared among stress types become *suprasegmental* labels (e.g., 1 : {*wet*suit, *dis*card, *red* cow, ...} and 2 : {wet *suit*, dis*card*, red *cow*, ...}). The silhouette coefficient (SC) (Rousseeuw, 1987) measures cluster separation quality, where 1.0 indicates perfect separation and 0.0 indicates no separation. For each stress type, we consider a label set as clusters (either 8 segmental clusters or 2 suprasegmental clusters) and compute the silhouette coefficient for every model's internal representation. Table 3 reports coefficients averaged over 5-fold cross validations and fold-paired t-tests with respect to the Control model. For segmental representations, the Control model showed highest separation in phrasal stress (SC = 0.081), likely reflecting the inherent phonemic diversity of its minimal pairs, while contrastive stress exhibited the most similar segmental patterns (SC = 0.015) likely due to shared semantic categories (colors/animals).

Silhouette Coefficients	**Stress Type**	**Training**				
		Control	Phrasal	Lexical	Contrastive	All
Segmental	Phrasal	0.081 (0.014)	0.144 (0.029)†	0.061 (0.024)*	0.061 (0.021)*	0.158 (0.048)*
	Lexical	0.059 (0.022)	0.094 (0.026)*	0.122 (0.040)*	0.108 (0.028)*	0.123 (0.019)†
	Contrastive	0.015 (0.017)	0.036 (0.027)	0.018 (0.026)	0.096 (0.045)*	0.059 (0.016)*
Supra-segmental	Phrasal	0.005 (0.003)	0.026 (0.005)†	0.008 (0.002)	0.015 (0.011)	0.039 (0.016)*
	Lexical	0.006 (0.005)	0.029 (0.016)*	0.200 (0.028)†	0.106 (0.046)†	0.225 (0.023)†
	Contrastive	0.005 (0.003)	0.042 (0.026)*	0.115 (0.044)†	0.324 (0.038)†	0.324 (0.028)†

Table 3: Segmental/suprasegmental silhouette coefficients across models. $^{\dagger}p < .01$ $^{*}p < .05$ vs Control.

Stress-specific training consistently enhanced segmental separation of the training stress type(s) versus the Control model, especially for phrasal stress. However, learning lexical or contrastive stress *reduced* the segmental silhouette coefficients for phrasal stress on average, while learning phrasal stress *increased* the segmental silhouette coefficients for lexical and contrastive stress. Similar to the Control model, the segmental silhouette coefficients for contrastive stress were consistently lower than the coefficients for phrasal and lexical stress across all models except for the Contrastive model. Suprasegmental analysis revealed near-zero silhouette coefficients for all stress types in the Control model, indicating a lack of systematic

stress encoding. Phrasal training improved its own suprasegmental silhouette coefficient significantly less ($\Delta = 0.021$) than lexical and contrastive training (lexical: $\Delta = 0.194$; contrastive: $\Delta = 0.319$), which also improved each others' silhouette coefficients. These asymmetries explain the lack of phrasal transfer and the bidirectional lexical-contrastive transfer observed in task performance by Sohn et al., as lexical and contrastive stress types organize suprasegmental features similarly in the latent space.



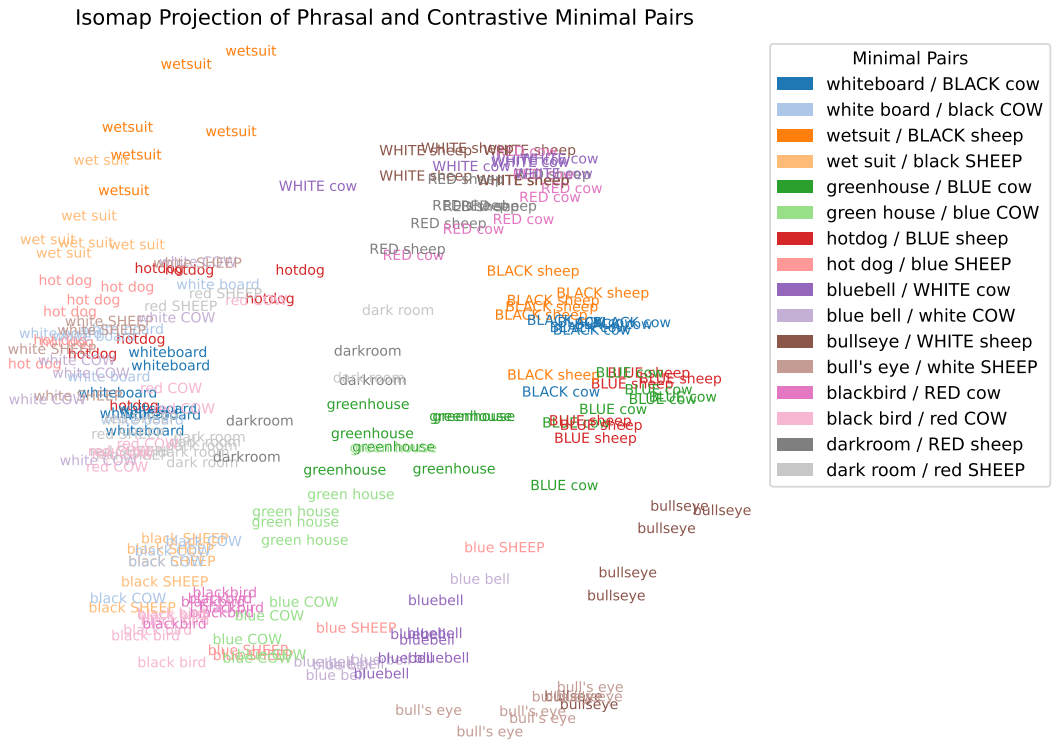
(a) Phrasal-Contrastive Coupling ($\alpha = 0.0$) **(b) Phrasal-Contrastive Coupling** ($\alpha = 0.5$) **(c) Phrasal-Contrastive Coupling** ($\alpha = 1.0$)

(d) Lexical-Contrastive Coupling ($\alpha = 0.0$) **(e) Lexical-Contrastive Coupling** ($\alpha = 0.5$) **(f) Lexical-Contrastive Coupling** ($\alpha = 1.0$)

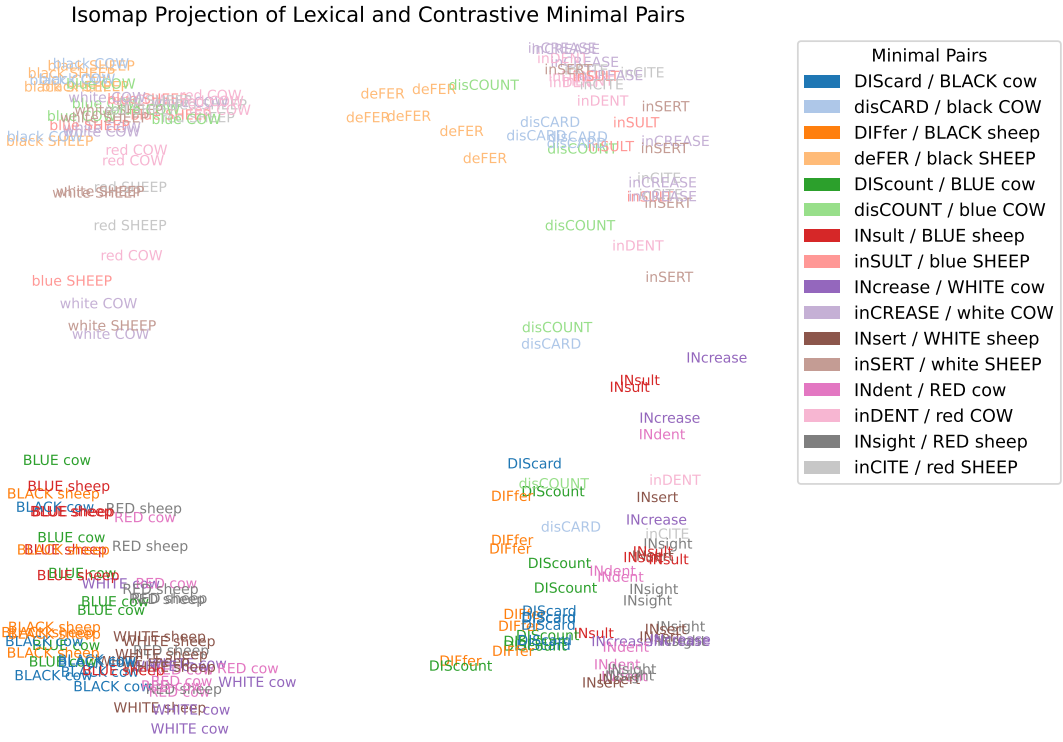
(g) Phrasal-Lexical Coupling ($\alpha = 0.0$) **(h) Phrasal-Lexical Coupling** ($\alpha = 0.5$) **(i) Phrasal-Lexical Coupling** ($\alpha = 1.0$)

Figure 1: Fused Gromov–Wasserstein couplings for every pair of prosodic stress types with $\alpha \in \{0.0, 0.5, 1.0\}$.

To quantify cross-stress relationships, we compute Fused Gromov-Wasserstein (FGW) distances between phrasal, lexical, and contrastive representations from the all-stress model (Figure 1). With $\alpha = 0.0$, FGW couples utterances based solely on Whisper's representation similarity, while $\alpha = 0.5$ jointly optimizes for representation similarity and distribution alignment. We have included results using $\alpha = 1.0$, but omitted them from our analysis as $\alpha = 1.0$ discards cross-stress acoustic features entirely, which is unsuitable for prosodic analysis. The phrasal-contrastive coupling with $\alpha = 0.0$ (Figure 1a) reveals a hierarchical mapping that is broadly segmental and finely suprasegmental (e.g., *black*bird $\leftrightarrow$ *black* cow/sheep and

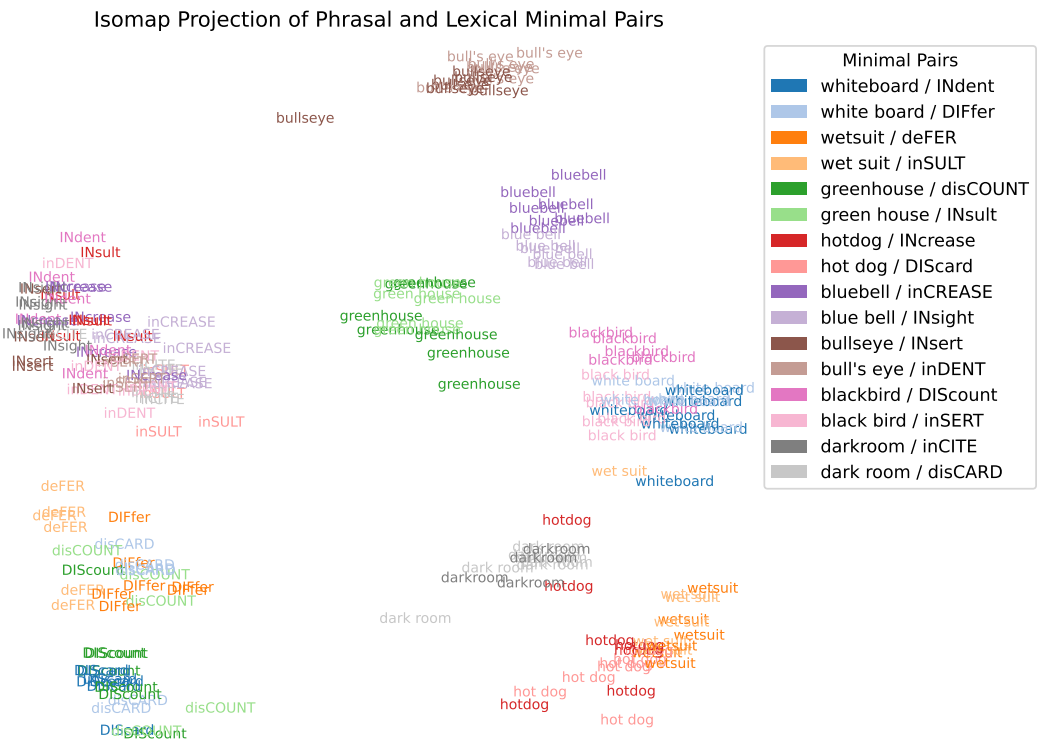


(a)



(b)

Figure 2: Isomap projections of Whisper's internal representations of pairs of stress types.



(c)

Figure 2: Isomap projections of Whisper's internal representations of pairs of stress types.

black *bird* ↔ black *cow*/*sheep*). At $\alpha = 0.5$ (Figure 1b), the diffuse mapping indicates poor distribution matching, confirming the asymmetries observed in the silhouette analysis. In contrast, lexical-contrastive couplings show inverse behavior: $\alpha = 0.0$ yields near-perfect *suprasegmental* alignment (evident in quadrant-structured mapping in Figure 1d), with segmental features playing a weak role. At $\alpha = 0.5$ (Figure 1e), the coupling separates further along phonemic boundaries (e.g., *in-* vs. *dis-* maps systematically to *bl-* vs. *r*/*w-*[1]), demonstrating that lexical and contrastive stress share suprasegmental representation similarity and segmental distribution alignment. Phrasal-lexical coupling (Figure 1g-h) exhibits extremely weak segmental and suprasegmental representation similarity ($\alpha = 0.0$) and distribution alignment ($\alpha = 0.5$), which we attribute to its significantly worse phonemic overlap compared to phrasal-contrastive couplings.

To visualize the all-stress model's internal representations of phrasal, lexical, and contrastive stress, we used Isomap (Balasubramanian and Schwartz, 2002) to reduce the dimensionality of the representations from 5120 to 2. Isomap was chosen over other methods because it preserves manifold distance (i.e., distance over the surface that the data lives on) better than t-SNE (Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018), and Isomap is better suited for nonlinearly structured data than PCA (Abdi and Williams, 2010). Figure 2a depicts phrasal and contrastive stress representations, showing strong suprasegmental separation of contrastive items along the northeast-trending diagonal axis. However, only a few phrasal items appear to conform with this axis (i.e., wetsuit, greenhouse, and bullseye). We also observe segmental proximity among {wetsuit, white cow, white sheep} and { blackbird, bluebell, bullseye, blue cow, blue sheep }. Figure 2b depicts the lexical and contrastive stress representations and showcases a clear hierarchical organization. Stress type is divided along the horizontal axis, where contrastive stress items are leftward and lexical stress items are rightward. The halves are both further suprasegmentally divided along the vertical axis, where stress-first items (e.g., *blue* cow, *red* sheep, *dis*card, and *in*sert) are below and stress-final items (e.g., blue

[1] This phonemic bimodality is attributed to our particular choice of lexical and contrastive stress items.

cow, red *sheep*, dis*card*, and in*sert*) are above. Within each quadrant, there are segmental separations mostly along the horizontal axis, which divide *dis-* items from *in-* items for lexical stress and *bl-* items from *r/w-* items for contrastive stress. In contrast, Figure 2c shows phrasal and lexical stress representations, which demonstrates a purely segmental organization of items from both stress types. We observe strong *in-* and *dis-* clusters for lexical stress, and several minimal pairs are well-separated for phrasal stress (e.g., bullseye, greenhouse, and bluebell).

4 Discussion

Whisper's representations emerge from an implicit optimization for the most statistically salient acoustic regularities in the stress productions. The observed patterns reflect this optimization: the high segmental and low suprasegmental separation of phrasal representations reflects the high phonemic diversity of phrasal items, which accounts for more acoustic variability than the prosodic realizations relying mainly on durational cues (Knutsen and Stromswold, 2024). Accordingly, phrasal and contrastive items become coupled ($\alpha = 0.0$) based on phonemic overlap, creating poor distribution alignment ($\alpha = 0.5$). Meanwhile, the high suprasegmental and low segmental separation of lexical and contrastive representations reflects consistent stress patterns (Knutsen and Stromswold, 2024) that account for more acoustic variability than the low phonemic diversity of their items. The FGW couplings directly manifest these statistical properties. Phrasal-contrastive pairs couple segmentally ($\alpha = 0.0$) but show poor distribution matching ($\alpha = 0.5$), whereas lexical-contrastive pairs achieve both suprasegmental alignment ($\alpha = 0.0$) and phonemic separation ($\alpha = 0.5$) through their complementary statistical structures. The silhouette coefficients also quantify these relationships, with phrasal stress's marginal suprasegmental gain ($\Delta = 0.021$) reflecting its variable realizations, and the strong bidirectional lexical-contrastive transfer ($\Delta > 0.19$) emerging from shared stable regularities. These results demonstrate that Whisper's representations emerge from principled statistical optimization, prioritizing the most reliable acoustic cues while suppressing noisy variations, mirroring the statistical learning mechanisms observed in human language processing (Pierrehumbert, 2003).

References

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Balasubramanian, M. and Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, 295(5552):7–7.

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*, pages 644–663.

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

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.

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

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

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, 46(2-3):115–154.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

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

Sohn, S. S., Knutsen, S., and Stromswold, K. (2025a). Democratizing prosodic stress recognition across genders and neurotypes. In *Human Sentence Processing*.

Sohn, S. S., Knutsen, S., and Stromswold, K. (2025b). Harnessing whisper for prosodic stress analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25931–25942.

Sohn, S. S., Knutsen, S., and Stromswold, K. (2025c). Leveraging automatic speech recognition for prosodic stress analysis. In *Human Sentence Processing*.

Sohn, S. S., Knutsen, S., and Stromswold, K. (2025d). Prosody in the age of ai: Insights from large speech models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.