WhiSPA: Semantically and Psychologically Aligned Whisper with Self-Supervised Contrastive and Student-Teacher Learning

Anonymous ACL submission

Abstract

Current speech encoding pipelines often rely on separate processing pipelines between text and audio, not fully leveraging the inherent overlap between these modalities for understanding human communication. Language models excel at capturing semantic meaning from text that can complement the additional prosodic, emotional, and acoustic cues from speech. This work bridges the gap by proposing WhiSPA (Whisper with Semantic-Psychological Alignment), a novel audio encoder trained with a contrastive student-teacher learning objective. Using over 500k speech segments from mental health audio interviews, we evaluate the utility of aligning Whisper's audio embeddings with text representations from an SBERT encoder and text-based assessments of psychological dimensions: emotion and personality. Over self-supervised and downstream mental health tasks, WhiSPA surpasses state-of-the-art speech models, achieving an average error reduction of 73.4% on the segment-level self-supervised objective and 83.8% on 11 psychological downstream tasks. WhiSPA demonstrates that cross-modal alignment can increase the amount of text-semantic and psychological information captured in audio-only encoder models.

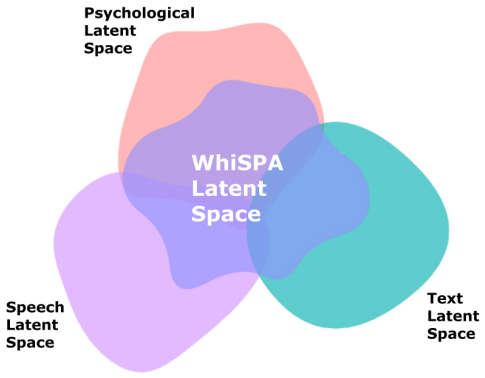


Figure 1: The goal of WhiSPA is to align Whisper's (speech-based) representations to better capture the semantic and psychological characteristics of communications that are currently best captured from text-based language models. Our approach enables speech-based models to enrich audio representations, which already capture prosodic cues, with stronger semantic and psychological dimensions.

1 Introduction

Human communication is inherently multimodal, combining semantic, prosodic, and psychological cues to convey meaning. However, AI integration of modalities is often fragmented (Lazaro et al., 2021; Gu et al., 2017), limiting models' ability to encode both acoustic and semantic dimensions. Speech models excel at capturing prosodic features but lack the nuanced semantic understanding of the text that language models provide (Soubki et al., 2024; Sriram et al., 2017).

To bridge this gap, we introduce an audio encoding model, WhiSPA (**Whis**per with **S**emantic and **P**sychological **A**lignment), by attempting to address the semantic and psychological representation disparity between Whisper-based representations and text-based LLMs. We align a pre-trained speech transcription model, Whisper (Radford et al., 2022), with latent dimensions from SBERT (Reimers and Gurevych, 2019) and language-based models of psychological attributes (V Ganesan et al., 2022; Park et al., 2014), intended to carry deeper semantic and psychological dimensions. Such alignment not only reduces computational and memory inefficiencies (from not needing to run a second LLM encoder on the transcripts from an audio model), but also enables a unified understanding of cross-modal dependencies between speech and language models which are often trained on much larger corpora.

While text-based language models effectively capture semantics and some degree of human affective information, audio inherently contains addi-

tional acoustic information. For example, prosodic elements such as tone of voice and pause duration can alter the meaning or better convey affective (i.e. as pertaining to human emotion or tone) aspects of language — transforming a statement into a question, conveying sarcasm instead of sincerity, or shifting a serious remark into insincere humour. Without cross-modal integration, language models, while semantically robust, remain incomplete in representing the full spectrum of human expressions (Zhang et al., 2023; Lian et al., 2023). Still, since text is derivable from speech, speech should intrinsically be mappable to the same rich semantic embeddings from language models, making an additional encoder redundant. By addressing this gap, our work streamlines the speech analysis pipeline and enhances its usability.

Contributions. We hypothesize that aligning text and audio latent spaces can significantly improve audio-based representations. We test this hypothesis and make the following specific contributions: (1) We propose and train a novel audio encoder, WhiSPA (Whisper with Semantic and Psychological Alignment), which demonstrates significantly greater performance on both speech-to-text self-supervised tasks and downstream psychological tasks; (2) We systematically evaluate variants on the proposed alignment architecture, finding (a) aligning with text-based representations of both semantics and psychological features drastically improves audio representations, (b) aligning with emotional and personality factors result in SotA person-level psychological assessments, (c) a contrastive learning objective results in a superior point of convergence for encoding semantic and psychological dimensions of audio, and finally, (d) for downstream tasks, we found marginal benefit in adding SBERT representations to WhiSPA, suggesting that WhiSPA already captures nearly all the information provided by its text-based teacher model.

WhiSPA integrates acoustics with language-derived semantics and psychological features, enabling a more context-aware multimodal AI system to encode semantic and psychological dimensions from speech.

2 Background

Transformer architectures have revolutionized natural language processing (NLP) (Vaswani et al., 2017), enabling significant advances in speech processing and cross-modal learning by setting new benchmarks for understanding language and multimodal data.

Speech Recognition. Whisper, OpenAI's state-of-the-art automatic speech recognition (ASR) model, was pre-trained on $680,000$ hours of multilingual, multitask audio data. Leveraging an encoder-decoder architecture, it performs speech transcription and translation (Radford et al., 2022). Its autoregressive decoder predicts the next most probable token while attending to prior tokens and cross-attending to the audio representations from the encoder. Whisper's intermediary layers encode high-level audio representations, capturing both phonetic and semantic content, enabling contextual understanding beyond sound patterns (Radford et al., 2022).

The encoder processes Mel filterbank representations into embeddings encapsulating rich acoustic and linguistic features, such as phonemes, intonation, and broader context. The decoder refines this understanding, generating text even in ambiguous or noisy conditions. This capability mirrors human speech perception, where phonetic accuracy and contextual understanding are seamlessly integrated (Radford et al., 2022).

Audio Encoders. Audio-specific models like Wav2Vec have advanced the encoding of audio signals into high-dimensional representations. Wav2Vec (Schneider et al., 2019), developed by Facebook AI, introduced a framework for learning latent features from raw audio using convolutional neural networks (CNNs) and self-supervised contrastive loss, achieving meaningful representations generalizable across tasks.

Wav2Vec 2.0 (Baevski et al., 2020) improved on this with a masked prediction objective inspired by BERT (Devlin et al., 2019), enabling better modelling of long-range dependencies in audio. With proper fine-tuning, its transformer-based architecture and contrastive loss enable state-of-the-art results in speech-processing tasks, especially emotion recognition (Baevski et al., 2020).

Semantics. Language models prioritize semantic understanding through contextual mechanisms. For instance, BERT employs bidirectional attention to consider both preceding and succeeding tokens, learning fine-grained contextual representations through a masked language modelling objective (Devlin et al., 2019). Sentence-BERT (SBERT) ex-

tends BERT by mapping input sentences into dense vector spaces with a Siamese network, enabling semantic similarity comparisons via metrics like cosine similarity (Reimers and Gurevych, 2019).

However, SBERT is limited to textual data and cannot capture affective or prosodic nuances inherent to spoken language (Mohebbi et al., 2021). This limitation underscores the need for multimodal approaches that integrate speech and text modalities to bridge the semantic and affective gaps in communication (Zhang et al., 2023; Lian et al., 2023).

Multimodal Learning. Cross-modal alignment embeds data from different modalities into shared spaces to capture their relationships. Techniques like contrastive learning align related inputs (e.g., audio and text segments) while separating unrelated pairs (Ye et al., 2022). Efforts to align text and audio include SpeechBERT (Chuang et al., 2020), which adapted BERT's framework to paired speech-text data, and SLAM (Speech-Language Aligned Models) (Bapna et al., 2022), which optimized joint embedding spaces to improve downstream tasks like speech recognition and audio-text retrieval. Similarly, models like HuBERT (Hsu et al., 2021) have shown promise in bridging text and audio through hierarchical feature learning. However, many models struggle with capturing prosodic and affective nuances, often focusing heavily on semantic alignment.

3 Data & Tasks

Audio Datasets. We utilize two psychological, mental health-focused datasets for training and evaluation: **WTC-Segments** (WTC) (Kjell et al., 2024) and **HiTOP-Segments** (HiTOP) (Kotov et al., 2022). WTC recordings were completed by patients in a clinic for WTC responders who came for a health monitoring visit. HiTOP interviews were completed by outpatients with psychiatric diagnoses who were recruited by the study team to complete a research interview. Both datasets consist of paired audio-text data, ensuring alignment between spoken content and its corresponding textual transcription.

From its source, WTC was curated from ∼6 minute interview recordings, on average, of patients responding to both personal and general questions in a structured manner (Kjell et al., 2024). Contrarily, HiTOP followed a semi-structured format, where patients described experiences on set topics while also organically conversing with the interviewer. Once filtered for audio segments solely spoken by patients, interviews generally ranged from 45 to 90 minutes, yielding a voluminous and broadened set of audio segments (Kotov et al., 2022). The recordings were diarized using NVIDIA NeMo and transcribed with `openai/whisper-large-v2`.

Dataset	WTC	HiTOP
Total Segment Duration (*min*)	15,087	28,460
Mean Segment Duration (*s*)	5.86	2.99
Total Audio Segments	154,586	571,420
Total Participants	1,396	524

Table 1: Audio dataset metadata (after preprocessing and filtering for participant-only speech).

Psychological Assessments. For each dataset, psychological measures were collected for each user. For WTC, each subject completed the self-reported PTSD CheckList (PCL), yielding scores for four specific subscales: Re-experiencing (REX), Avoidance (AVO), Negative Alterations in Mood (NAM), Hyperarousal (HYP). For HiTOP, trained interviewers provided ratings for the following six psychopathology scales: Internalizing (INT), Disinhibition (DIS), Antagonism (ANT), Somatoform (SOM), Thought-Disorder (THD), and Detachment (DET) (Kotov et al., 2022, 2024).

To evaluate the encoding ability of WhiSPA for any given audio segment, we manually annotate a small subset from both datasets for valence and arousal dimensions expressed in their speech. Three random audio segments containing more than 5 uttered words from each user were sampled from each dataset and were annotated by two individuals with a background in psychology using the affective circumplex scale. This resulted in 300 audio segments, equally split between the two datasets.

Self-Supervised PsychEmb. For each audio/text pair in our datasets, we extract theoretically derived psychological features using pre-trained lexica (V Ganesan et al., 2022), which we refer to as PsychEmb. PsychEmb broadly covers three domains of psychological constructs measured at different temporal granularity: (a) states, which reflect the emotional condition of the person at a point in time; (b) dispositions, which are slightly more stable than states and reflect the tendencies of humans to behave in certain ways and finally (c) the traits, which are long term stable charac-

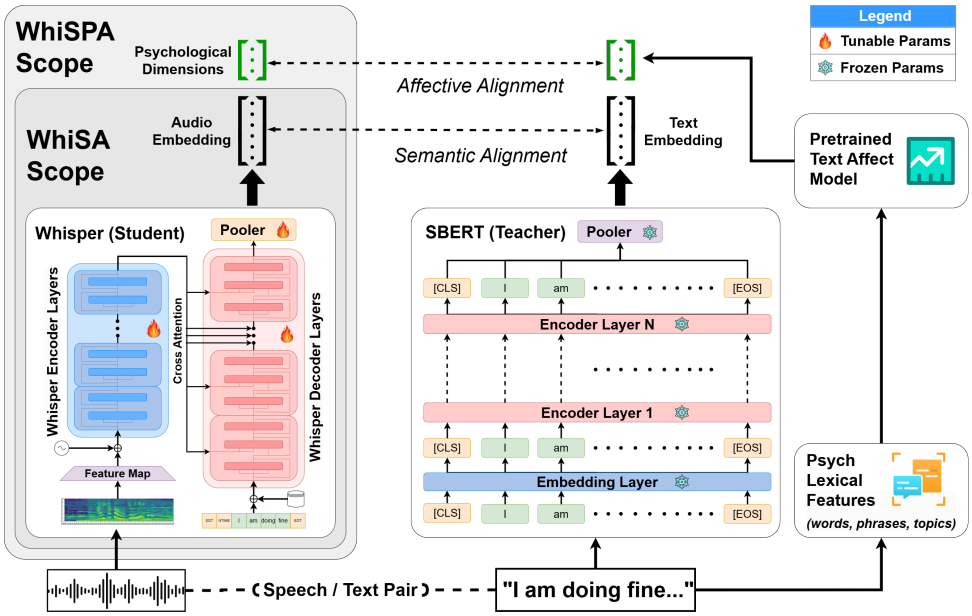


Figure 2: Diagram of WhiSA and WhiSPA training procedure involving a student-teacher model paradigm. Whisper (left) is semantically aligned to the ground truth embeddings encoded by SBERT (right). When PsychLex's dimensions are included in the alignment function, the WhiSPA framework semantically and psychologically aligns the corresponding embeddings with contrastive loss criteria.

teristics (Park et al., 2014). The ten dimensions of PsychEmb are Valence (VAL), Arousal (ARO), Openness (OPE), Consciousness (CON), Extraversion (EXT), Agreeableness (AGR), Neuroticism (NEU), Anger (ANG), Anxiety (ANX), and Depression (DEP), each represented with scalar values. Once the self-supervised PsychEmb dimensions were extracted for each segment across both datasets, we perform a 80:10:10 (train/val/test) split.

4 Methodology

Model Architecture. We begin with `Whisper-tiny` pre-trained encoder-decoder model (Radford et al., 2022) as our initial point prior to alignment. As seen in the Whisper (Student) portion of Figure 2, once the last non-padding token is predicted (<EOT>), we apply a mean pooling layer to the last hidden state of Whisper's decoder yielding a singular representation for the input audio. This representation is then pooled using a learnable dense layer, and the output serves as our embedding during alignment. This aggregated representation is aligned to the pooled representations from pre-trained SBERT for semantic alignment and the PsychEmb's dimensions for psychological alignment. We denote the pre-trained Whisper results in the paper with **Whisper-384**, referring to the number of hidden dimensions.

4.1 Alignment Objective

The alignment objective aims to improve the semantic and psychological information encoded in Whisper (student) with the help of the representations from a strong text encoding teacher model like SBERT and PsychEmb. In this work, we explore two suitable candidate objective functions to align speech representations with text, which are described below in detail.

4.1.1 Cosine Similarity Loss *(CS)*

The success of the cosine similarity-based approach in building geometrically robust representations in SBERT motivated its use as an alignment objective in this work. We apply cosine similarity loss to the pooled audio embeddings and pooled SBERT embeddings[1], given by the following equation:

$$\mathcal{L}^{CS} = \sum_{i \in \mathcal{I}} \mathcal{L}_i^{CS} \quad (1)$$

$$\mathcal{L}_i^{CS} = 1 - \text{sim}(\mathbf{A}_i, \mathbf{T}_i)$$

$$\text{where sim}(\mathbf{A}_i, \mathbf{T}_i) = \frac{\mathbf{A}_i \cdot \mathbf{T}_i}{||\mathbf{A}_i||\,||\mathbf{T}_i||}$$

where $i \in \mathcal{I} \equiv \{1...N\}$ refers to the index of audio/text pair in a batch of N samples. $\mathbf{A}_i$ refers

[1]SBERT version: `all-MiniLM-L12-v2`

to the source audio embedding, $\mathbf{T}_i$ refers to its corresponding target text embedding, and sim() computes the cosine similarity between audio and text embeddings which produces a scalar value between $[-1, 1]$. This loss can also be interpreted as the cosine diversity of the two embeddings. To align the embedding spaces, we aim to maximize the cosine similarity between corresponding embedding pairs (Reimers and Gurevych, 2019; Sanh et al., 2020), and hence decrease $\mathcal{L}^{CS}$.

4.1.2 Noise Contrastive Estimation Loss *(NCE)*

The Noise Contrastive Loss (Equation 2) is optimized to increase the cosine similarity between a pair of audio embedding and its corresponding text embedding while simultaneously increasing the differentiation between the audio embedding and randomly sampled text embedding in that batch (Ye et al., 2022).

$$\mathcal{L}^{NCE} = \sum_{i \in \mathcal{I}} \mathcal{L}_i^{NCE} \tag{2}$$

$$\mathcal{L}_i^{NCE} = -\log \frac{\exp(\text{sim}(\mathbf{A}_i, \mathbf{T}_i)/\tau)}{\sum_{b \in B(i)} \exp(\text{sim}(\mathbf{A}_i, \mathbf{T}_b)/\tau)}$$

where $\mathcal{L}_i^{NCE}$ refers to contrastive loss criteria in which pairwise cosine similarities are calculated for each audio embedding with all text embeddings in that batch. Hence, there is only one positive text embedding that pairs with an audio embedding, while the remaining text embeddings from the batch serve as contrastive samples. Let $B(i) \in \mathcal{I}$, where $B(i)$ represents all other SBERT text embeddings in the batch such that $\mathbf{T}_b \neq \mathbf{T}_i$ (Ye et al., 2022; Chen et al., 2020; Khosla et al., 2021). The variable $\mathbf{T}_b$ denotes the index of an arbitrary, negative SBERT text embedding sample and τ, temperature, represents a trainable scalar parameter which is set to a default of 0.1.

4.2 Whisper Semantically Aligned (WhiSA-384)

WhiSA leverages a student-teacher model paradigm (Hinton et al., 2015; Sanh et al., 2020) to align Whisper's audio-based embeddings with SBERT's text-based embeddings, which serve as the teacher model. SBERT encodes corresponding text sentences into semantically rich embedding vectors, which serve as $\mathbf{T}$ in the above equations during training. Whisper's embeddings ($\mathbf{A}$ in the above equations), derived from its decoder's last hidden state, are aligned to these SBERT embeddings using the loss functions described above. This process is aimed at WhiSA to learn robust semantic representations directly from audio inputs by minimizing the cosine distance between Whisper and SBERT embeddings as shown in Figure 2.

4.3 Adding Psychological Alignment (WhiSPA)

WhiSPA extends the WhiSA framework by augmenting PsychEmb dimensions into Whisper's. While maintaining the semantic alignment objective, WhiSPA injects the PsychEmb dimensions into the SBERT embeddings under two settings: (1) **with replacement**: We adopted a naive strategy of replacing the first ten dimensions of SBERT's embedding with the PsychEmb dimensions to maintain the same number of latent dimensions between both models. We use **WhiSPA-384**$_r$ to refer to this. (2) **with projection**: We concatenate the PsychEmb dimensions to the text embedding from SBERT. Consequently, this requires a 384×10 learnable projection matrix, P, to transform Whisper embeddings of dimensionality 384 to 394, which is then transformed using a $TanH$ activation. This model goes by the name **WhiSPA-394**. To address the numerical instability issues from modelling the PsychEmb dimensions in its absolute range, we standardize and scale them to match their distribution to that of SBERT's embeddings. Refer to Appendix subsection A.2 for more information on training.

5 Results & Discussion

We consider two popular speech models as baselines, Wav2Vec 2.0 (Baevski et al., 2020) and whisper-tiny (Radford et al., 2022), which are referred to as Wav2Vec-768 and Whisper-384 respectively. We measured the effectiveness of these embeddings by computing Pearson correlation coefficient (r) and mean squared error (mse) over a 10-fold cross-validated ridge regression model for each task. For each model variant in Table 2, we encode audio segments for each participant and aggregate them with a statistical mean to represent person-level embeddings for tasks in Table 3.

Alignment improved the models' ability to capture psychological dimensions from language. We evaluated the speech-based models' ability to capture the psychological dimensions of lan-

Dataset	Speech Model	Traits										States				Dispositions					
		OPE		CON		EXT		AGR		NEU		VAL		ARO		ANG		ANX		DEP	
		$r(\uparrow)$	$mse(\downarrow)$	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse
HiTOP	Wav2Vec-768	.61	.15	.60	.15	.59	.13	.47	.11	.59	.14	.40	.001	.49	.001	.34	.04	.42	.02	.48	.04
	Whisper-384	**.74**	**.11**	.80	.08	.69	.10	.76	.06	.78	.08	.71	.001	.82	**.000**	.53	**.03**	.61	.01	.65	.03
	WhiSA-384	.71∗	**.11**	.81∗	.08	.70∗	.10	.77∗	.06	.78∗	.08	.73∗	.001	.83∗	**.000**	.59†	**.03**	.61	**.01**	.61†	.04
	WhiSPA-384$_r$	**.74**∗	**.11**	**.83**†	**.07**	.70∗	.10	**.79**†	**.05**	.79†	**.07**	**.78**†	**.000**	**.85**†	**.000**	.59†	**.03**	.61†	**.01**	**.66**∗	**.03**
	WhiSPA-394	.72∗	**.11**	**.83**†	**.07**	**.72**∗	**.09**	**.79**†	**.05**	**.82**†	**.07**	.76†	**.000**	.84∗	**.000**	**.62**†	**.03**	**.65**†	**.01**	.63∗	**.03**
WTC	Wav2Vec-768	.26	.56	.45	.59	.21	.69	.31	.48	.25	.68	.24	.004	.47	.006	.14	.23	.00	.16	.17	.16
	Whisper-384	.57	.43	.70	.37	.68	.38	.64	32	.67	.40	.56	.003	.82	**.002**	.54	.16	.46	.13	.45	.13
	WhiSA-384	.70†	.31	.82†	.24	.75†	.32	.76†	.23	.77†	.30	.67†	**.002**	.85†	**.002**	.66†	.13	.61†	.10	.61†	.10
	WhiSPA-384$_r$	.71†	.29	.82†	.24	.74†	.30	.76†	.20	.76†	.27	.68†	**.002**	.85†	**.002**	.67†	.01	.61†	**.09**	.61†	**.09**
	WhiSPA-394	**.72**†	**.28**	**.83**†	**.22**	**.76**†	**.29**	**.79**†	**.19**	**.79**†	**.26**	**.70**†	**.002**	**.86**†	**.002**	**.69**†	**.11**	**.64**†	**.09**	**.66**†	**.09**

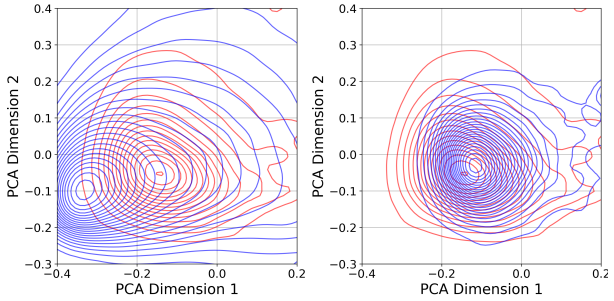
Table 2: Pearson r correlations and mean squared errors for self-supervised prediction of person-level affect/personality scores with audio models (10-fold cross-validation with ridge regression). **Bold** indicates the best metric for the psychological scale in the respective dataset. ↑ *implies higher is better.* ↓ *implies lower is better.* ∗ *indicates statistically significant* ($p < .05$) *predictions compared to Wav2Vec-768.* † *indicates statistically significant* ($p < .05$) *predictions compared to Whisper-384.*

Audio Model	WTC										HiTOP											
	PCL		REX		AVO		NAM		HYP		INT		DIS		ANT		SOM		THD		DET	
	$r(\uparrow)$	$mse(\downarrow)$	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse	r	mse
Wav2Vec-768	.23	128.44	.19	11.87	.15	3.93	.21	10.68	.20	16.62	.37	.20	.36	.23	.22	.12	-.02	.24	.21	.11	.11	.22
Whisper-384	.23	128.85	.21	11.77	.06	4.00	.19	10.87	.23	16.41	.39	.19	.33	.24	.33	.11	.07	.23	.28	.11	.29	.20
WhiSA-384	.29†	119.68	.27†	11.26	.19†	3.90	.26†	10.12	.28†	15.56	.55†	.16	.53†	**.19**	**.43**†	**.10**	.22†	.23	.37†	**.10**	.33†	**.18**
WhiSPA-384$_r$	.34†	119.24	**.30**†	11.23	.17	3.88	.31†	**10.08**	**.32**†	15.54	.56†	**.15**	.53†	**.19**	.42†	**.10**	**.23**∗	**.22**	**.39**†	**.10**	**.39**†	.19
WhiSPA-394	**.35**†	**118.91**	**.30**†	**11.18**	**.20**	**3.85**	**.32**†	10.09	**.32**†	**15.48**	**.57**†	**.15**	**.54**†	**.19**	**.43**†	**.10**	.22†	**.22**	.37†	**.10**	.38†	.19

Table 3: Pearson r correlations for predicting self-reported/annotated person-level psychological scales with audio models (10-fold cross-validation with ridge regression). **Bold** indicates the best metric for the psychological scale in the respective dataset. ↑ *implies higher is better.* ↓ *implies lower is better.* ∗ *indicates statistically significant* ($p < .05$) *predictions compared to Wav2Vec-768.* † *indicates statistically significant* ($p < .05$) *predictions compared to Whisper-384.*

guage by comparing our models' predictions to PsychEmb derived values at the segment level. As summarized in Table 2, we found that both semantic (WhiSA) and psychological alignments (WhiSPA) significantly outperformed traditional speech-based models (Wav2Vec and Whisper) across all ten dimensions on both metrics. Compared to Whisper, which was evidently a stronger baseline than Wav2Vec2 ($Avg\Delta = 36$ Pearson points for WTC & 21 points for HiTOP), Our semantic alignment method showed a marked improvement in performance, with an average of 11 in Pearson points for WTC and 2 in HiTOP. A paired t-test was used to confirm that all improvements over Wav2Vec and all improvements over Whisper, except for 4 outcomes in HiTOP, were statistically significant ($p < .05$). This result highlighted our alignment methods improved the speech model's ability to capture psychological dimensions in language (PsychEmb).

Between semantic and semantic-psychological alignment, the latter offered marginally better performance than the former, increasing by 1-5 Pearson points on both datasets. Interestingly, deriving



(a) Before Alignment (b) After Alignment

Figure 3: Bivariate KDE contour plot of dimensionally reduced speech/text embeddings with PCA. Speech representations in blue. Text representations in red.

psychological estimates from semantic dimensions (WhiSPA-394) was consistently better than the replacement (WhiSPA-384$_r$) of 10 semantic dimensions with PsychEmb dimension. This shows the importance of curating the semantic dimensions before replacing them with different embeddings.

We also observed that the alignment increased the overlap between the latent space of the speech and text embeddings, as shown in Figure 3. Before alignment (Figure 3a), speech and text embeddings

Model	Loss	Self-Supervision Tasks		Downstream Tasks	
		Pearson r (↑)	MSE (↓)	Pearson r (↑)	MSE (↓)
WhiSA-384	*CS*	.72	.11	.34	15.26
	NCE	.72	.11	.36	14.63
WhiSPA-384$_r$ *(with replacement)*	*CS*	.72	.12	.34	15.08
	NCE	.73	.11	.36	14.68
WhiSPA-394 *(with projection)*	*CS*	.72	.11	.34	15.21
	NCE	**.74**	**.10**	**.37**	**14.59**

Table 4: Comparison of loss functions on self-supervised and downstream tasks. The reported Pearson r's and MSE's are averaged across all outcomes. **Bold** indicates the best metric when comparing loss functions across different models. ↑ implies *higher is better*. ↓ implies *lower is better*.

show distinct contours with very little overlap in their dense regions, highlighting a clear modality gap and a lack of shared contextual meaning. After alignment (Figure 3b), the contours exhibit greater overlap, indicating a unified embedding space with reduced variance. Figure 3 demonstrates that the alignment process effectively bridges the semantic gap between the two modalities.

Semantic-Psychological alignment is SotA for person-level psychological assessments. Table 3 shows that the improvements brought by our aligned models over traditional models were preserved even when evaluated on a spectrum of downstream psychological assessment tasks. In particular, the alignment showed a stark increase in capturing deeper psychological conditions such as Internalizing ($\geq$ 16 Pearson points) and Disinhibition ($\geq$ 20 Pearson points) from very long durations of speech data. Consistent with behaviours exhibited with PsychEmb dimensions, in Table 2, semantic-psychological alignment from semantically-derived psychological dimensions (WhiSPA-394) performed the best, followed by semantic-psychological alignment from replacement (WhiSPA-384$_r$) and finally semantic-only alignment (WhiSA-384). For these tasks, we averaged the segment-level representations of the interview audio file to produce a person-level embedding. These embeddings were used to perform 10-fold cross-validation with a ridge regression model, and its performance was measured using Pearson correlation coefficient (r) and mean squared error (mse).

The success of WhiSPA-394 can be attributed to its integration of psychological feature alignment, which complements semantic alignment by explicitly encoding affective dimensions such as valence and arousal. The improvements in outcomes like **INT** and **DIS** further support this interpretation since these constructs often rely on subtle vocal cues, such as pause distribution, pitch variability, and vocal tone as established by prior works (Kotov et al., 2024). WhiSPA-394 is better equipped to model due to its expanded embedding space. By injecting dimensions with psychological relevance into the alignment process, the model bridges the gap between the prosodic information in speech and the textual semantics used to train baseline models like WhiSA. This dual alignment likely enhances the model's ability to capture both the what (semantic content) and the how (affective delivery) of speech, enabling more accurate predictions of psychological scales.

Contrastive loss criteria led to richer representations of audio. Investigation of the choice of alignment objective towards performance (Table 4) revealed that Noise Contrastive Estimation (NCE) consistently produced a better-aligned model than cosine similarity (CS). NCE outperforms CS Equation 1 across all models, likely because NCE optimizes for discriminative learning, encouraging more separation between positive and negative samples in the embedding space. This enhances the model's ability to encode nuanced semantic and psychological cues, as reflected in WhiSPA-394's better alignment and task performance. When comparing WhiSPA-394 and WhiSPA-384, we notice the recurring trend with NCE granting a greater optima during alignment than CS as exemplified in Table 4. However, WhiSPA-384 holds its ground in HiTOP, achieving comparable correlations. This suggests that WhiSPA-394's architecture may generalize well to diverse datasets but thrives in highly semantic and affective audio contexts like WTC. All results noted in Table 2 and Table 3 were from models trained with NCE.

Model	PCL	HiTOP			VAL
		INT	DIS	THD	*(segment)*
SBERT-384	**.36**	.54	.55	**.40**	.47
Whisper-384	.23	.39	.33	.28	.38
WhiSA-384	.29	.55	.53	.37	.50*
WhiSPA-384$_r$	.34	.56*	.53	.39	**.53***
WhiSPA-394	.35	.57*	.54	.37	.51*
WhiSPA-394 & SBERT-384	**.36**	**.58***	**.56**	.39	.52*

Table 5: Pearson r correlations for predicting psychological measures (10-fold cross-validation with ridge regression). Acoustic **VAL** was regressed on the segment level for 300 human-annotated segments from WTC and HiTOP. *Higher is Better.* **Bold** indicates the highest correlation for each measure. * indicates statistically significant ($p < .05$) predictions compared to SBERT-384.

WhiSPA captures semantic features of its text-based teacher without needing additional SBERT representations. Adding SBERT representations to WhiSPA offers little benefit, as it already captures nearly all the information from its teacher model. The results in Table 5 underscore the slight increase in correlations after adding text-based embeddings from SBERT-384 with WhiSPA, as seen in the last row. WhiSPA, trained through a student-teacher alignment paradigm, appears to achieve a semantic and psychological optimum during convergence. This is evident in its substantial performance gains over Whisper, which lacks the semantic and psychological depth provided by language models.

Notably, WhiSPA-394 demonstrates clear improvements in specific measures such as **INT** and **VAL**, with gains of $+3$ and $+6$ Pearson points, respectively, when compared to SBERT-384. While these increments may seem marginal, they substantiate our claim that WhiSPA-394 effectively captures nearly all the information encoded by its text-based teacher model, SBERT-384. Further, when WhiSPA-394 representations are augmented with SBERT embeddings, the downstream task performance exhibits statistically insignificant gains, suggesting minimal benefit.

Ultimately, these findings highlight an important observation: The potential of cross-modal alignment may be constrained by the representational efficacy of the teacher model. The marginal returns from text-based augmentation indicate that WhiSPA has already learned to encode the critical semantic and psychological cues provided by its teacher, reflecting the success of the alignment process.

6 Conclusion

Our findings demonstrate that WhiSPA effectively integrates semantic and psychological information from speech, enhancing state-of-the-art audio representations for psychological and mental health assessments. By aligning Whisper's audio embeddings with SBERT's text embeddings enriched with psychological features, we found not only consistent improvement for ten self-supervised tasks but also significantly greater accuracies over 11 downstream person-level psychological tasks as compared to modern audio models. Finally, we found little benefit in adding SBERT to WhiSPA embeddings. This suggests that the cross-modal alignment objective we employed affords audio-encoder-decoder models like Whisper the advantages of text-only transformer-LMs that have been trained on much larger datasets than audio foundation models. We see this as part of a growing body of work to create holistic multi-modal foundation models that result in a richer and more authentic representation of human communication.

7 Limitations

Currently, WhiSPA's training scope idealizes lexically derived psychological features, which can negatively impact its ability to discern acoustic information. Although we have shown drastic improvements in yielding semantically contextual audio embeddings, it begs the question of how much more *affectively contextual* the embeddings can be for psychological predictions and emotion recognition tasks. This would require injecting psychological features that are not only lexically derived but also acoustically–since acoustics and vocal prosody are known to convey more affect (Low et al., 2020).

Subsequently, since the alignment procedure places equal weight on all dimensions of embedding, the semantic dimensions outnumber the psychological dimensions. This is likely resulting in the convergence of WhiSPA on a semantic similarity optima. Potentially, we would like to explore a multi-objective weighted loss criteria where parameters can be tuned for placing higher priority on affect over semantics in some cases.

8 Ethical Implications

The multimodal WhiSPA model holds significant potential for improving mental healthcare assessments by providing rich insights into individuals' states of mind through speech analysis. However, multimodal approaches increase ethical considerations due to the richer and more diverse forms of personally identifiable information (PII) they capture compared to unimodal models. In addition to text content, the WhiSPA model processes acoustic and prosodic features — including tone of voice, speech patterns, and emotional expressions — which can inadvertently reveal sensitive details like gender, ethnicity, emotional state, and health conditions. This expanded data scope raises the risk of re-identification, making it essential to implement stringent data security and handling, including compliance with privacy regulations such as GDPR and HIPAA.

Security & Privacy. Moreover, the potential for misuse or unauthorized exploitation of such detailed multimodal data necessitates robust ethical guidelines for its storage, processing, and application. Transparency in how these models are trained and used is critical to building trust among clinicians and patients. Finally, ongoing efforts to mitigate algorithmic biases and ensure fairness are important, as errors in multimodal assessments could disproportionately impact vulnerable populations or lead to incorrect diagnoses if not carefully managed.

The WTC and HiTOP recordings took place in a clinical setting at the Stony Brook WTC Health and Wellness Program. Each participant gave informed consent and was fully informed about the study, that it was voluntary to take part, and that they had the right to withdraw at any time without giving a reason or that it would affect their treatment. After the interview, participants were debriefed (for more details about the WTC data collection, see (Kjell et al., 2024); for more details about the HiTOP data, see (Kotov et al., 2022, 2024). The studies and data uses were approved by the Institutional Review Board at an undisclosed university for privacy reasons.

Software. Adhering to the ideals of open and reproducible science, we will make the WhiSPA software code base, along with the trained models and secure dimensional representations of the data, openly available. These representations strictly comply with established security protocols, ensuring that no individual can be identified nor any anonymity safeguard compromised. Nevertheless, direct access to the underlying data remains restricted in accordance with privacy and security measures.

Additionally, AI-based tools were employed throughout the project to assist in code development and report formulation, including the use of ChatGPT and other similar consumer generative AI. Such integration aligns with established best practices and guidelines, ensuring that the technical accuracy, integrity, and scientific rigour of the work remain uncompromised while benefiting from enhanced efficiency and streamlined workflows.

Acknowledgments

The work presented in this paper stems from the immense hours of audio recordings from the **WTC** and **HiTOP** datasets. We greatly thank the participants, creators, maintainers, and interviewers from the studies for enabling this research.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *Preprint*, arXiv:2202.01374.

E B Blanchard, J Jones-Alexander, T C Buckley, and C A Forneris. 1996. Psychometric properties of the PTSD checklist (PCL). *Behav. Res. Ther.*, 34(8):669–673.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Yung-Sung Chuang, Chi-Liang Liu, Hung-Yi Lee, and Lin shan Lee. 2020. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *Preprint*, arXiv:1910.11559.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Yue Gu, Xinyu Li, Shuhong Chen, Jianyu Zhang, and Ivan Marsic. 2017. Speech intention classification with multimodal deep learning. *Adv. Artif. Intell.*, 10233:260–271.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning. *Preprint*, arXiv:2004.11362.

Oscar Kjell, Adithya V Ganesan, Ryan Boyd, Joshua Oltmanns, Alfredo Rivero, Scott Feltman, Melissa Carr, Benjamin Luft, Roman Kotov, and H. Schwartz. 2024. Demonstrating high validity of a new ai-language assessment of ptsd: A sequential evaluation with model pre-registration.

Roman Kotov, David C Cicero, Christopher C Conway, Colin G DeYoung, Alexandre Dombrovski, Nicholas R Eaton, Michael B First, Miriam K Forbes, Steven E Hyman, Katherine G Jonas, Robert F Krueger, Robert D Latzman, James J Li, Brady D Nelson, Darrel A Regier, Craig Rodriguez-Seijas, Camilo J Ruggero, Leonard J Simms, Andrew E Skodol, Irwin D Waldman, Monika A Waszczuk, David Watson, Thomas A Widiger, Sylia Wilson, and Aidan G C Wright. 2022. The hierarchical taxonomy of psychopathology (HiTOP) in psychiatric practice and research. *Psychol. Med.*, 52(9):1666–1678.

Roman Kotov, Holly Frances Levin-Aspenson, Camilo Ruggero, Holly Levin-Aspenson, and Katherine Jonas. 2024. Interview for the hierarchical taxonomy of psychopathology (iHiTOP).

May Jorella Lazaro, Sungho Kim, Jaeyong Lee, Jaemin Chun, Gyungbhin Kim, EunJeong Yang, Aigerim Bilyalova, and Myung Yun. 2021. A review of multimodal interaction in intelligent systems.

Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10).

Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig. Otolaryngol.*, 5(1):96–116.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilevar. 2021. Exploring the role of bert token representations to explain sentence probing results.

Gregory Park, H. Schwartz, Johannes Eichstaedt, Margaret Kern, Michal Kosinski, David Stillwell, Lyle Ungar, and Martin Seligman. 2014. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Claire Roman and Philippe Meyer. 2024. Analysis of glyph and writing system similarities using Siamese neural networks. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 98–104, Torino, Italia. ELRA and ICCL.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *Preprint*, arXiv:1904.05862.

Adil Soubki, John Murzaku, and Owen Rambow. 2024. Multimodal belief prediction. In *Proc. INTERSPEECH 2024*. ISCA.

Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates. 2017. Cold fusion: Training seq2seq models together with language models. In *Interspeech*.

Adithya V Ganesan, Vasudha Varadarajan, Juhi Mittal, Shashanka Subrahmanya, Matthew Matero, Nikita Soni, Sharath Chandra Guntuku, Johannes Eichstaedt, and H. Andrew Schwartz. 2022. WWBP-SQT-lite: Multi-level models and difference embeddings for moments of change identification in mental health forums. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 251–258, Seattle, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. *Preprint*, arXiv:2205.02444.

Chuan Zhang, Daoxin Zhang, Ruixiu Zhang, Jiawei Li, and Jianke Zhu. 2023. Bridging the emotional semantic gap via multimodal relevance estimation. *Preprint*, arXiv:2302.01555.

A Appendix

A.1 Data Description

A.1.1 HiTOP.

The HiTOP dataset consists of video-recorded interviews conducted between World Trade Centre

responder participants and clinicians. Each recording is annotated with the outcomes derived from the HiTOP structured interview, which includes a standardized set of questions designed to assess a comprehensive set of mental health dimensions, including aspects of internalizing (e.g., questions about distress and fear), dis-inhibited externalizing (e.g., questions about substance abuse and antisocial behaviours) and more.

Outcomes in HiTOP The HiTOP outcomes were derived from the structured clinical interview (Roman and Meyer, 2024), where we used the total score of the six dimensions including: i) internalizing (INT; e.g., dysphoria, lassitude), ii) disinhibited externalizing (DIS; e.g., alcohol use, drug use), iii) antagonistic externalizing (ANT; e.g., attention seeking, callousness), iv) somatoform (SOM; e.g., conversion, somatization), v) thought disorder (THD; e.g., psychotic and disorganized thought patterns), vi) detachment (DET; e.g., intimacy avoidance, suspiciousness)

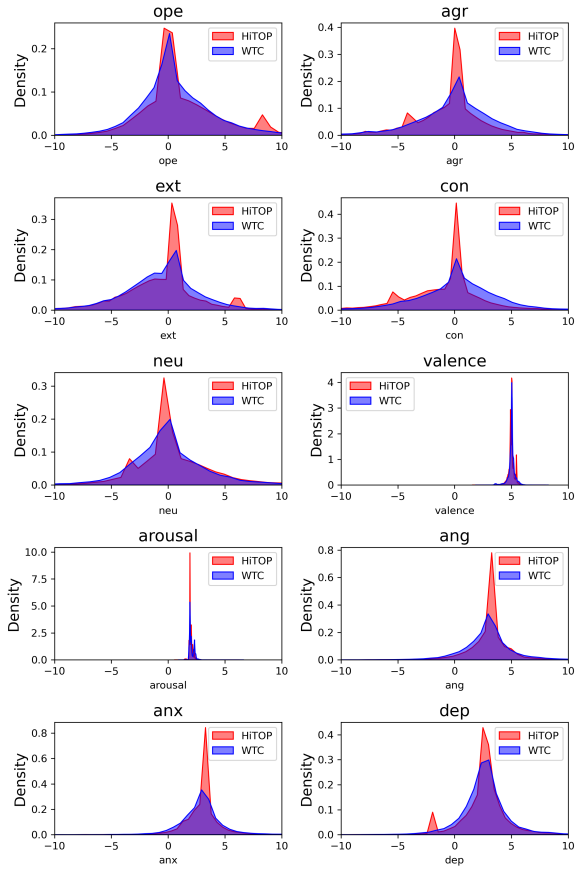


Figure 4: Standardized distributions of PsychEmb dimensions for each segment across both datasets. The distribution of WTC is shown in blue. The distribution of WTC is shown in red.

A.1.2 WTC.

In the WTC dataset, participants were recorded in a private room during their clinical visit while responding to questions displayed on a screen as part of an automated clinical interview. These questions prompted participants to reflect on both positive (e.g., What are three things you currently look forward to the most?) and negative aspects of their lives across different time frames (past, present, and future). Topics included general life experiences (e.g., the best and worst experiences, challenges, and support systems) and significant events such as COVID-19 and 9/11 (e.g., How does 9/11 affect you now?). A full list of the questions is provided in (Kjell et al., 2024).

To enhance generalizability, the questions were designed to be broad and used everyday language, avoiding clinical jargon or references to specific symptoms. Instructions on the screen advised participants not to read the questions aloud and to aim for at least 60 seconds of response time per question. Throughout the development phase, the questions were refined over three iterations to improve engagement and elicit more detailed responses. However, for the evaluation phase, the same set of questions was used for all participants. On average, recordings for those who met a threshold of at least 150 words lasted 7.5 minutes (SD = 4.1; range = 1.1 to 43.0 minutes).

The data, from its source, totalled 1437 participants (Female = 7%, Male = 93%; Mean age = 57.9, SD = 8.0 years; 14.5%).

Outcomes in WTC The PCL score and subscales were derived from the PTSD CheckList (PCL) (Blanchard et al., 1996), which consists of 17 items designed to measure the severity of PTSD symptoms according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria. Participants rate their experiences over the past month using a scale from 1 (not at all) to 5 (extremely). We calculated both the overall score (PCL) and scores for the four subscales. These subscales are Re-experiencing (REX; e.g., intrusive thoughts related to trauma), Avoidance (AVO; e.g., evading trauma-related thoughts), Emotional Numbing (NAM; e.g., difficulty recalling aspects of the trauma), and Hyperarousal (HYP; e.g., disturbances in sleep patterns). Reliability, as measured by Cronbach's alpha, was acceptable across all scales (.70).

A.2 Training

The research done for devising WhiSPA's framework resulted from iterations of tweaking and testing architectures, loss criteria, parameters, and hyperparameters.

For the methodology presented in this paper, we provide the following configurations for reproducibility:
Pooling: $MEAN$. Learning Rate: 1×10^{-5}. Weight Decay: 1×10^{-2}. Temperature (τ): 0.1. Batch Size: 900. Number of Epochs: 50. Number of workers (CPU cores): 16. These configurations result in a total average training time of $\sim$ 20 hours.

We discovered that the efficacy of Equation 2 highly depends on the batch size. It should be stated that larger batch sizes allow for greater degrees of repulsion and attraction in the cross-modal embedding space. While training WhiSA and WhiSPA, we utilized a batch size of 900 and distributed them across 3 NVIDIA RTX A6000 devices with 48GB of VRAM each.

Additionally, we use open-source licensed pretrained models from HuggingFace. Our programmatic implementation for deep learning is done with PyTorch. When it comes to evaluation, we utilize Differential Language Analysis Tool Kit (DLATK) for correlating regression results across specified groups (i.e., $user\_id$ or $segment\_id$)

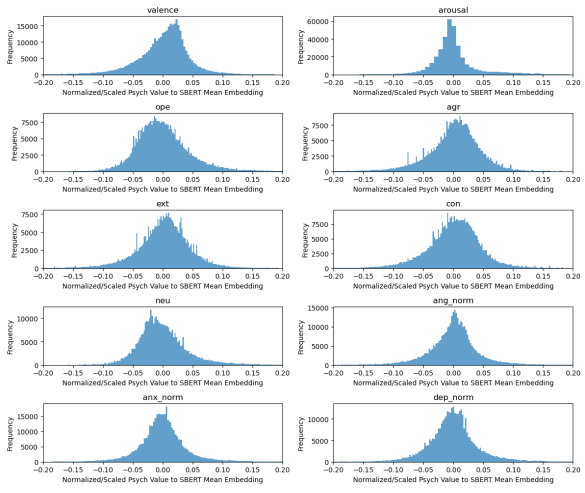


Figure 5: Distributions of psychological features standardized and scaled to the distribution of SBERT's mean embedding value before augmentation for WhiSPA alignment training.

Cosine similarity is sensitive to the relative magnitudes of the vectors being compared. If the added ten dimensions of psychological features have a very different scale or distribution from SBERT embeddings, they could dominate or skew the cosine similarity computation. Once either loss function is applied, (1) or (2), WhiSPA embeddings remain semantically aligned with SBERT while also encoding meaningful affective cues for downstream tasks.

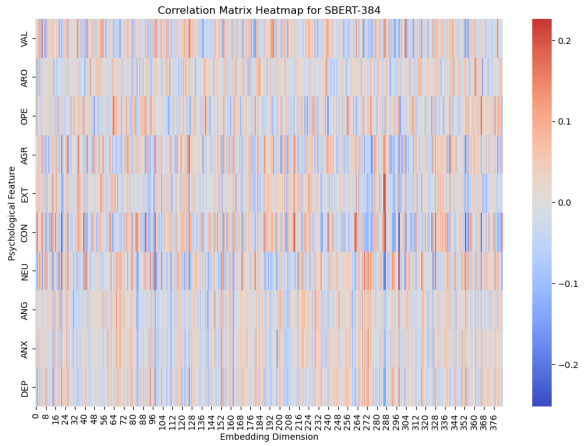


Figure 6: Pearson r correlation heatmap of SBERT-384's mean embedding. This visual displays the correlations of SBERT's 384 dimensions with each of the 10 PsychEmb dimensions.

During the training of WhiSPA, we experimented with identifying which dimensions of the teacher-model, SBERT, have the lowest correlations with PsychEmb dimensions to replace those dimensions. We decided that this approach may lead to statistical biases when training, and so we naively replaced the first 10 dimensions. One should note that the set of 10 dimensions to replace in SBERT can be chosen arbitrarily since our study experimented with this.

A.3 Annotations

Please note that the annotators were expert psychologists and co-authors.

The documentation accompanying the iHiTOP interview dataset was utilized to report the coverage of its domains, demographic information, and other relevant details. The dataset's focus on structured psychological interviews and its linguistic properties were described in the paper to contextualize its relevance to this research. This information was presented to ensure transparency and reproducibility. The WTC dataset assessed PTSD symptom severity and related constructs, including anxiety and depression, using English-language data from WTC emergency responders in the Stony Brook Health Program. Linguistic features such as

RoBERTa-large embeddings, n-grams, and LDA topics were used to analyze behavioural patterns alongside closed-vocabulary features like pronouns and death-related terms (LIWC-22). The development dataset included 1,437 participants, and the prospective dataset included 346, with a mean age of 58 years, predominantly male (93% and 91%, respectively) and white (54% and 49%). The analysis emphasized language markers of stress, anxiety, and trauma while reflecting on participants' experiences of 9/11. Ethical safeguards, including IRB approval, informed consent, and automated anonymization, ensured compliance. While comprehensive in its linguistic and demographic scope, the study was limited to English speakers and WTC responders, constraining generalizability.

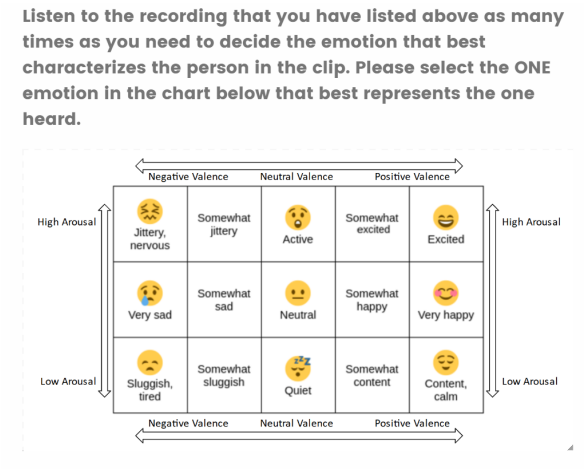


Figure 7: Annotator's affective circumplex visual grid for the task of manually annotating acoustic segments of speech from both datasets.