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

**Rajath Rao[1], Adithya V Ganesan[1], Oscar Kjell[1], Jonah Luby[1], Akshay Raghavan[1], Scott Feltman[1] Whitney Ringwald[2], Ryan L. Boyd[3], Benjamin Luft[1], Camilo Ruggero[3], Neville Ryant[4], Roman Kotov[1], H. Andrew Schwartz[1]**

[1]Stony Brook University, [2]University of Minnesota

[3]University of Texas at Dallas, [4]University of Pennsylvania

**Correspondence:** rajath.rao@stonybrook.edu, has@cs.stonybrook.edu

## Abstract

Current speech encoding pipelines often rely on an additional text-based LM to get robust representations of human communication, even though SotA speech-to-text models often have a LM within. This work proposes an approach to improve the LM within an audio model such that the subsequent text-LM is unnecessary. We introduce **WhiSPA** (**Whi**sper with **S**emantic and **P**sychological **A**lignment), which leverages a novel audio training objective: contrastive loss with a language model embedding as a teacher. Using over 500k speech segments from mental health audio interviews, we evaluate the utility of aligning Whisper's latent space with semantic representations from a text autoencoder (SBERT) and lexically derived embeddings of basic psychological dimensions: emotion and personality. Over self-supervised affective tasks and downstream psychological tasks, WhiSPA surpasses current speech encoders, achieving an average error reduction of 73.4% and 83.8%, respectively. WhiSPA demonstrates that it is not always necessary to run a subsequent text LM on speech-to-text output in order to get a rich psychological representation of human communication.

## 1 Introduction

Human communication is inherently multimodal, but AI integration of modalities is often fragmented (Lazaro et al., 2021; Gu et al., 2017). Speech to text models, like Whisper (Radford et al., 2022), are often pipelined into text-based language models (LMs) (Chuang et al., 2020) in order to get the most accurate speech-based representations (see Figure 1). This often results in redundant computational costs from having two LMs in the pipeline (one within the audio model and one for the text LM) and representations remain incomplete of the full spectrum of human expressions (Zhang et al., 2023; Lian et al., 2023). This is especially important for psychological and social

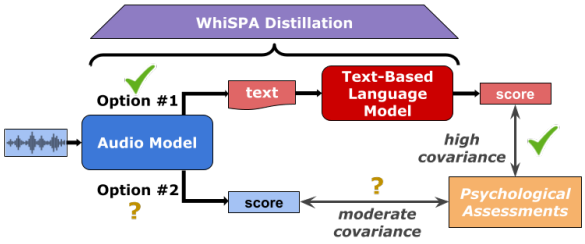



Figure 1: Speech processing pipelines that are further processed by language models (option #1 above) often yield higher accuracies than those produced solely by SotA audio models (option #2). Our distillation introduces a speech encoder which streamlines the pipeline and exhibits similar performance to a text-based LM.

scientific applications where representations from text-based LMs demonstrate superior performance than direct speech representations (Lukac, 2024; Chen et al., 2024).

Here, we seek to bridge the *semantic and psychological representation gap* between speech-based LMs present in audio models and text-based LMs. We introduce a speech encoding model, **WhiSPA**[1] (**Whi**sper with **S**emantic and **P**sychological **A**lignment), which aligns a pre-trained speech recognition model, Whisper (Radford et al., 2022), with the latent dimensions from SBERT (Reimers and Gurevych, 2019), intended to better capture semantics and deeper psychological information (V Ganesan et al., 2022; Park et al., 2014). Such alignment reduces computational and memory inefficiencies, circumventing the need for a second text encoder, as it enables a unified cross-modal representation between speech and language models. Still, since text is derivable from speech, speech should intrinsically be mappable to the same rich semantic features from the text.

Our focus on psychological or human-level tasks reflects a growing demand for foundation models to better understand the intrinsic qualities of hu-

[1]https://github.com/humanlab/WhiSPA



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22529–22544

man communication (Soni et al., 2024). As Clark and Schober (1992) put it, *"The common misconception is that language has to do with words and what they mean. It does not. It has to do with people and what they mean."* and specifically how well the representations capture information *about the people* communicating (Hovy and Yang, 2021; Soni et al., 2022). More specifically, psychological studies have suggested mental health attributes are highly multimodal as they are influenced by subtle nuances in voice and content (Sartori and Orrù, 2023; Chen et al., 2024).

Our **main contributions** include: (1) The development of **WhiSPA** (**Whi**sper with **S**emantic and **P**sychological **A**lignment), with a novel alignment objective, (2) Evaluation of the hypothesis that aligning text and audio latent spaces can significantly enhance audio-based representations for a deeper semantic and psychological understanding of human communication, (3) Demonstration of significant accuracy improvements in self-supervised tasks and downstream psychological tasks over systematically tested variants of WhiSPA. We find that: (a) aligning with text-based semantic and psychological representations drastically improves audio representations, including SotA person-level psychological assessments; (b) a Noise Contrastive Estimation loss yielded a more optimal convergence in aligning Whisper's latent space with semantic and psychological dimensions. and (c) for downstream psychological tasks, there was almost no benefit in utilizing SBERT representations on top of WhiSPA's, suggesting the same information from a text LM can be captured with the LM of the audio model and thus it is not necessary to pipeline another text LM after the audio model.

## 2 Background

This work builds on top of Whisper (Radford et al., 2022), OpenAI's SotA automatic speech recognition (ASR) foundation model. We chose Whisper over other alternatives such as HuBERT and Wav2Vec2-BERT, since previous works (Kyung et al., 2024; Yang et al., 2023) have shown that Whisper has a stronger language encoding module at capturing speaker attributes due to its pretraining objective of transcription/translation.

Recent advances in foundational speech technologies, like Whisper and HuBERT, have vastly improved the performances on speech recognition tasks (Radford et al., 2022; Hsu et al., 2021). However, they have limited ability to capture deeper semantics and speaker attributes compared to a text-based language model (Chen et al., 2024; Dong et al., 2022). Prior works that have addressed this have targeted a very narrow scope of psychological attributes (Busso et al., 2008). These gaps underscore the need for methodologies that bridge speech encoders' acoustic robustness with the psychological depth of text-based language models—a challenge we address by embedding fundamental psychological dimensions present in one's speech.

Multi-level fusion architectures leveraging both acoustic and lexical features have shown to improve performance in emotion recognition, speaker identification, and other downstream tasks. For instance, (Zhao et al., 2022) demonstrates that coattention-based early fusion and late fusion using Wav2Vec2.0 (Baevski et al., 2020; Schneider et al., 2019) and BERT (Devlin et al., 2019) outperform SotA emotion recognition benchmarks. Other recent works inject acoustic nuances into language models using textual descriptions of speech characteristics (Wu et al., 2024) or common-sense reasoning through historical utterances from the speaker (Fu, 2024). However, this approach does not fully leverage the cross-modal dependencies between text and audio, as it remains unimodal, relying solely on textual inputs rather than raw acoustic representations.

Prior works in cross-modal alignment provide foundational insights for this integration. Compositional Contrastive Learning (Chen et al., 2021) distilled audio-visual knowledge into video representations by aligning teacher-student embeddings across modalities, embedding rich semantics from teacher-audio and image models into the student-video model. In another work, Dong et al. (2022) improved the accuracy of intent classification of spoken language by employing a contrastive loss using both speech and language features. These works highlight that the cross-modal alignment objective embeds information from different modalities into shared spaces to capture their relationships, while contrastive learning aids in grouping related inputs across different modalities (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 (Devlin et al., 2019) to paired speech-text data, and SLAM (Speech-Language Aligned Models) (Bapna et al., 2022), which op-

timized joint embedding spaces to improve downstream tasks like speech recognition and audio-text retrieval. To the best of our knowledge, this is the first work to perform cross-modal learning to endow the foundational speech model with richer semantic and psychological representations.

## 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 World Trade Center (9/11) 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 `whisper-large-v2`.

| Dataset | WTC | HiTOP |
|---|---|---|
| Total Segment Duration (*hr*) | ∼252 | ∼474 |
| 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).

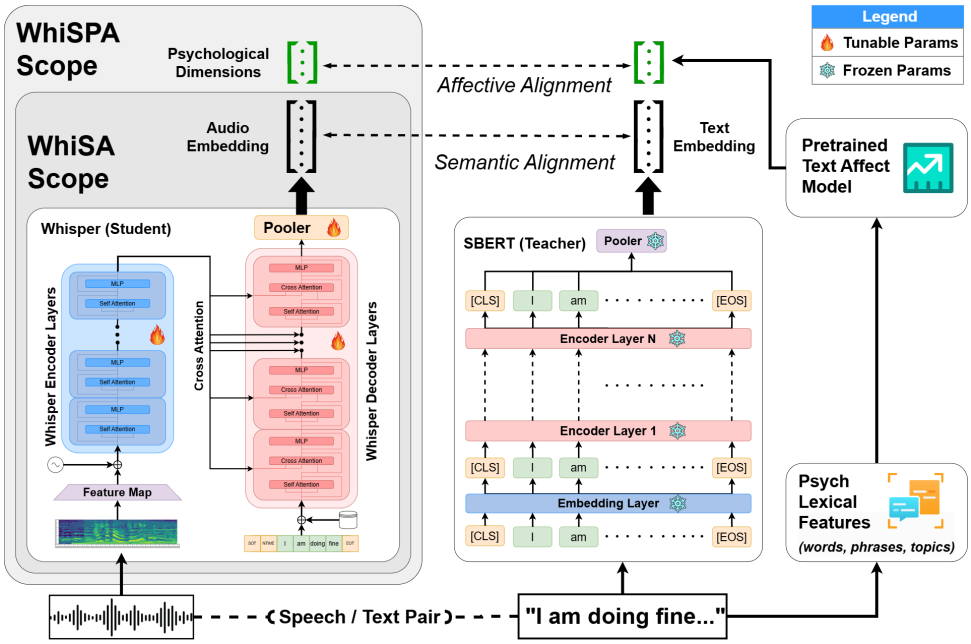



Figure 2: Diagram of WhiSA and WhiSPA training procedure involving a student-teacher model paradigm. Whisper (left) is semantically aligned to the ground truth embeddings encoded by SBERT (right). When PsychEmb features are included in the alignment function, the WhiSPA framework semantically and psychologically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

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 (Figure 8). 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 characteristics (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. After extracting self-supervised PsychEmb dimensions for each segment across both datasets, we perform a 80:10:10 (train/val/test) split.

## 4 Methodology

Aligning audio representations directly with a text-based language model allows us to infuse the audio model's latent space with the rich semantic and affective details typically provided by text representations, thereby eliminating the need for a separate text LM. While this approach does not explicitly leverage the unique acoustic features of speech, it prioritizes efficiency by avoiding redundant processing and consistently delivers a semantically enriched representation—an advantage that is particularly critical for psychological and social scientific applications (Lukac, 2024; Chen et al., 2024).

**Model Architecture.** We begin with the Whisper[2] encoder-decoder backbone (Radford et al., 2022), which does not run autoregressively. During training, audio segments are previously transcribed with `whisper-large-v2`, making it entirely self-supervised. Likewise, SBERT and PsychEmb representations were encoded using these transcriptions. As seen in the Whisper (Student) portion of Figure 2, 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. Throughout this paper, we denote the pre-trained Whisper model as **Whisper-384**, where the numeric suffix refers to the embedding dimensionality.

### 4.1 Alignment Objective

While fusion architectures focus on merging acoustic-textual features throughout layers, *we contrast this paradigm* by directly aligning cross-modal latent spaces for deeper semantic and psychological representations from audio, bypassing the need for task-specific fusion architectures. Our 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[3] 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, given by the following equation:

$$\mathcal{L}^{CS} = \sum_{i \in \mathcal{I}} \mathcal{L}_i^{CS} \tag{1}$$

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

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

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

[2]Whisper-384 version: `whisper-tiny`

[3]SBERT-384 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 embeddings 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 $\tau$, temperature, represents a tunable scalar parameter which is set to 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 $384 \times 10$ learnable projection matrix, $P$, to transform Whisper embeddings of dimensionality 384 to 394, which is then passed through a $TanH$ activation. This model goes by the name **WhiSPA-394**. To address the numerical instability issues from modeling the PsychEmb dimensions in its absolute range, we standardize and scale them to match SBERT's distribution of embedding values. Refer to Appendix subsection A.2 for more information on training.

## 5 Results & Discussion

We consider three popular, robust speech encoders as baselines: Wav2Vec2-BERT[4] (Communication et al., 2023; Chung et al., 2021), HuBERT[5] (Hsu et al., 2021), and Whisper (Radford et al., 2022), which are referred to as **W2V2B**, **HuBERT**, 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, we encode audio segments for each participant and aggregate them with a statistical mean to represent person-level embeddings for the tasks in Table 2 and Table 3.

[4]W2V2B version: `wav2vec2-bert-CV16-en`
[5]HuBERT version: `hubert-large-ls960-ft`

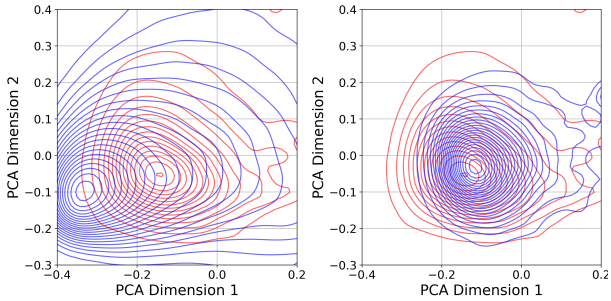
| Dataset | Model | Traits | | | | | | | | | | States | | | | Dispositions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | OPE | | CON | | EXT | | AGR | | NEU | | VAL | | ARO | | ANG | | ANX | | DEP | |
| | | $r(\uparrow)$ | $mse(\downarrow)$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ |
| HiTOP | *SBERT-384* | .73 | **.11** | **.83** | **.07** | .68 | .11 | .75 | .06 | .77 | .09 | .69 | .001 | .81 | **.000** | .59 | **.03** | .60 | **.01** | .61 | .04 |
| | W2V2B | .63 | .14 | .69 | .12 | .75 | .09 | .60 | .09 | .72 | .10 | .65 | .001 | .73 | **.000** | .45 | .04 | .51 | .02 | .64 | **.03** |
| | HuBERT | .67 | .13 | .71 | .11 | **.77** | **.08** | .57 | .10 | .70 | .11 | .66 | .001 | .73 | **.000** | .48 | .04 | .48 | .02 | .58 | .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 | *SBERT-384* | .65 | .35 | .78 | .29 | .73 | .33 | .73 | .25 | .73 | .34 | .62 | .003 | **.86** | **.002** | .62 | .14 | .56 | .11 | .59 | .11 |
| | W2V2B | .33 | .54 | .51 | .55 | .34 | .64 | .37 | .46 | .34 | .64 | .32 | .004 | .51 | .005 | .31 | .21 | .14 | .16 | .22 | .15 |
| | HuBERT | .35 | .54 | .57 | .50 | .39 | .61 | .44 | .43 | .42 | .60 | .38 | .003 | .53 | .005 | .36 | .20 | .15 | .16 | .22 | .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: **Self-Supervised Prediction Accuracies for Psychological Traits, States, and Dispositions.** Averaged person-level embeddings were fit to a ridge regression with 10-fold cross validation. **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 W2V2B. † indicates *statistically significant* ($p < .05$) predictions compared to Whisper-384.

**Alignment improved the models' ability to capture psychological dimensions from language.** We evaluated the speech-based models' ability to capture the psychological dimensions of language 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 (W2V2B and Whisper) across all ten dimensions on both metrics. Compared to Whisper, which was evidently a stronger baseline than W2V2B ($Avg\Delta = 36$ 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 W2V2B 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).

Interestingly, deriving 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




(a) Before Alignment (b) After Alignment

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

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 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 speech-based 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

| Model | HiTOP | | | | | | | | | | | | WTC | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | INT | | DIS | | ANT | | SOM | | THD | | DET | | PCL | | REX | | AVO | | NAM | | HYP | |
| | $r(\uparrow)$ | $mse(\downarrow)$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ |
| W2V2B | .50 | .17 | .46 | .21 | .35 | .11 | -.00 | .24 | .27 | .11 | .32 | .20 | .14 | 133.19 | .14 | 12.08 | .07 | 3.99 | .13 | 10.98 | .10 | 17.17 |
| HuBERT | .50 | .17 | .53 | .19 | .36 | .11 | .07 | .23 | .28 | .11 | .31 | .20 | .21 | 129.86 | .22 | 11.72 | .07 | 3.99 | .19 | 10.80 | .15 | 16.99 |
| Whisper-384 | .39 | .19 | .33 | .24 | .33 | .11 | .07 | .23 | .28 | .11 | .29 | .20 | .23 | 128.85 | .21 | 11.77 | .06 | 4.00 | .19 | 10.87 | .23 | 16.41 |
| WhiSA-384 | .55† | .16 | .53† | **.19** | **.43**† | **.10** | .22† | .23 | .37† | **.10** | .33† | **.18** | .29† | 119.68 | .27† | 11.26 | .19† | 3.90 | .26† | 10.12 | .28† | 15.56 |
| WhiSPA-384$_r$ | .56† | **.15** | .53† | **.19** | .42† | **.10** | **.23**$*$ | **.22** | **.39**† | **.10** | **.39**† | .19 | .34† | 119.24 | **.30**† | 11.23 | .17 | 3.88 | .31† | **10.08** | **.32**† | 15.54 |
| WhiSPA-394 | **.57**† | **.15** | **.54**† | **.19** | **.43**† | **.10** | .22† | **.22** | .37† | **.10** | .38† | .19 | **.35**† | **118.91** | **.30**† | **11.18** | **.20** | **3.85** | **.32**† | 10.09 | **.32**† | **15.48** |

Table 3: **Self-Reported/Annotated Prediction Accuracies for Psychological Scales.** Averaged person-level embeddings were fit to a ridge regression with 10-fold cross validation. **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 W2V2B. † indicates *statistically significant* ($p < .05$) predictions compared to Whisper-384.

in capturing deeper psychological conditions such as **INT** (internalizing) ($\geq$ 16 Pearson points) and **DIS** (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). 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). This is likely because NCE optimizes for discriminative learning, encouraging more separation between positive and negative samples in the embedding space (Ye et al., 2022), enhancing the model's ability to encode nuanced semantic and psychological cues. 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

| 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*.

| Model | HiTOP | | | | | | | | | | | | WTC | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | INT | | DIS | | ANT | | SOM | | THD | | DET | | PCL | | REX | | AVO | | NAM | | HYP | |
| | $r(\uparrow)$ | $mse(\downarrow)$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ | $r$ | $mse$ |
| **Transcribe → Text Model** (*Upperbound*) | | | | | | | | | | | | | | | | | | | | | | |
| *SBERT-384* | .54* | .16 | **.55*** | **.19** | **.43*** | **.10** | .16* | .23 | **.40*** | **.10** | **.40*** | **.18** | **.36*** | **118.21** | **.32** | **11.08** | **.24*** | **3.80** | **.32*** | **10.08** | **.33*** | **15.43** |
| *SBERT-1024* | .65* | **.13** | .59* | .17 | .51* | **.09** | .20 | .23 | **.43*** | **.09** | **.44*** | **.18** | .37* | 118.67 | .32* | 10.99 | **.27*** | **3.68** | .31* | 10.26 | .31* | 15.66 |
| **Audio Model** | | | | | | | | | | | | | | | | | | | | | | |
| Whisper-384 | .39 | .19 | .33 | .24 | .33 | .11 | .07 | .23 | .28 | .11 | .29 | .20 | .23 | 128.85 | .21 | 11.77 | .06 | 4.00 | .19 | 10.87 | .23 | 16.41 |
| WhiSPA-394 | **.57*** | **.15** | .54* | **.19** | **.43*** | **.10** | **.22*** | **.22** | .37* | **.10** | .38* | .19 | .35* | 118.91 | .30* | 11.18 | .20 | 3.85 | **.32*** | 10.09 | .32* | 15.48 |
| **Audio Model (Scaled Up)** | | | | | | | | | | | | | | | | | | | | | | |
| Whisper-1024 | .57 | .15 | .52 | .19 | .44 | .10 | **.23** | **.22** | .34 | .10 | .37 | .19 | .28 | 126.03 | .23 | 11.81 | .11 | 3.96 | .27 | 10.54 | .26 | 16.29 |
| WhiSPA-1034 | **.67*** | **.13** | **.62*** | **.16** | **.53*** | .10 | .21 | **.22** | .40* | .10 | **.44*** | **.18** | **.41*** | **114.44** | **.34*** | **10.89** | **.27*** | 3.70 | **.37*** | **9.84** | **.34*** | **15.42** |

Table 5: **Performance of WhiSPA Distilled Across Larger Dimensionalities.** Averaged person-level embeddings were fit to a ridge regression with 10-fold cross validation. **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 Whisper-(384/1024).

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.

**WhiSPA effectively scales to larger dimensionalities.** To investigate the effects of utilizing a larger teacher LM, we conducted experiments with `all-roberta-large-v1` (∼330M) paired with `whisper-medium` (∼796M) as the audio backbone, each with 1024 embedding dimensions. Table 5 shows that the distillation process remains effective for aligning larger student-teacher model configurations, further validating its scalability and generalizability for the downstream task. When comparing Whisper-384 to WhiSPA-394, we observe an **average error reduction of 83.38%**, while the 1024-sized models show an **even greater reduction of 86.61%.** A larger audio backbone also improves the consistency with which the student model outperforms its language-based teacher, likely due to enhanced context retention afforded by a larger parameter space. For example, WhiSPA-394 surpasses its teacher in only 2 of 11 outcomes, whereas WhiSPA-1034 does so in 7 of 11 psychological assessments. These findings underscore the effectiveness of our distillation strategy, particularly for larger models offering greater embedding dimensionality.

**WhiSPA captures semantics without the need for appending SBERT representations.** The last row in Table 6 underscores the marginal increase in correlations after appending SBERT embeddings to WhiSPA. WhiSPA, trained through a student-teacher alignment paradigm, appears to reach 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. However, the potential of cross-modal alignment may be constrained by the representational efficacy of the teacher model(s). On human-annotated audio segments, all of the WhiSPA variants achieve substantial improvements in capturing acoustic valence. In comparison with Whisper-384, WhiSPA-384$_r$ exhibits a gain of $+15$ Pearson points in **VAL** (acoustic valence) which exemplifies the reduction in the semantic/psychological gap between audio models and text-based models. Notably in Figure 4, WhiSPA-1034 demonstrates clear improvements in the majority of outcomes, with an average gain of $+2$ Pearson points, when compared to its teacher, SBERT-1024.

Ultimately, these findings highlight **two important observations:** (1) WhiSPA effectively captures nearly all the information encoded by its text-

| 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 6: **Comparison of Audio and Text Models for Predicting Psychological Scales.** Acoustic valence (**VAL**) was regressed on 300 human-annotated audio segments. SBERT-384 utilizes a cascaded pipeline (Whisper transcript → SBERT encoding). *Higher* is *Better*. * indicates statistically significant ($p < .05$) predictions compared to SBERT-384.

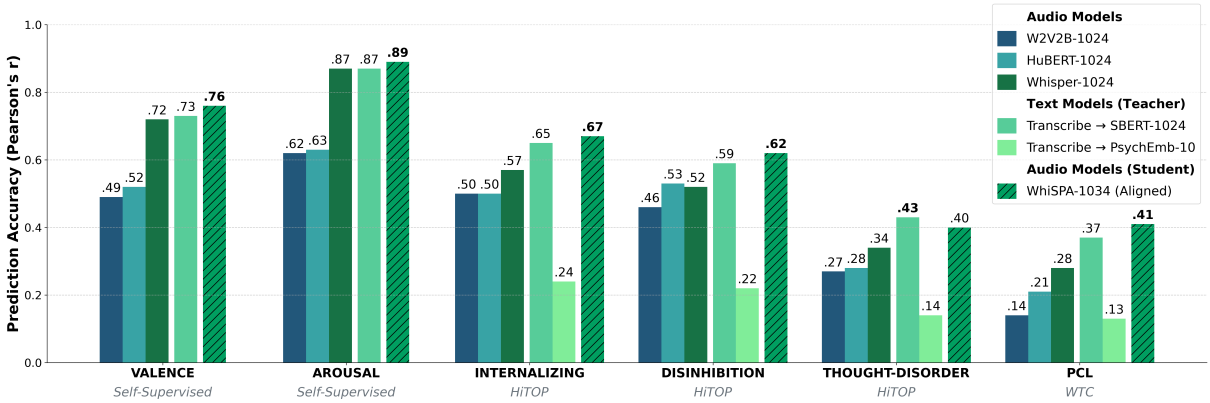



Figure 4: **WhiSPA Bridges the Semantic/Psychological Representation Gap.** WhiSPA consistently outperforms every baseline audio model and, in most cases, matches or exceeds the performance of the text-based language model teacher.

based teacher model, SBERT. (2) The marginal returns from appending text-based representations indicate that WhiSPA successfully learns to encode the critical semantic and psychological cues provided by its teachers, reflecting the success of the distillation.

**WhiSPA's representations are interpretable through language semantically associated with psychological dimensions.** Table 7a shows that n-grams known to be indicative of PTSD severity from prior studies (Kjell et al., 2024) —- including first-person pronouns, experienced symptoms, psychological distress, and negative affect – yield significantly higher correlations with WhiSPA's predictions compared to Whisper. In contrast, Table 7b reveals that language discussing relationships and positive affect is more negatively associated with WhiSPA's scores. These findings indicate that the contrastive loss training effectively aligns the latent space with rich semantic and psychological representations, capturing psychologically relevant linguistic markers more robustly. The highly semantic latent spaces of text-based LMs are reflected in WhiSPA's representations, especially for psychological nuances in spoken language. More quantitative analysis of our model can be found in Appendix subsection A.4

## 6 Conclusion

We claim that WhiSPA is a significant step toward more accurate representations of human communication by addressing the modal gap between text and audio, as language models often outperform audio models in predicting psychological attributes. By aligning WhiSPA's representations with SBERT's representations enriched with PsychEmb, we found consistent improvement for ten self-supervised tasks and significantly greater accuracies over 11 downstream psychological tasks. We observed only marginal improvements when appending SBERT representations to WhiSPA's, implying that the distillation process effectively captures the semantic features provided by the teacher language model. Our findings exemplify WhiSPA's effectiveness in extracting semantic and psychological features from speech, enhancing SotA audio representations for psychological and mental health assessments.

## 7 Limitations

While WhiSPA demonstrates significant advancements in providing semantically enriched audio embeddings, its current training paradigm predominantly aligns with psychological features derived from text, potentially limiting its capacity to capture critical acoustic information. This lexical bias, while beneficial for aligning with language-based models, raises an important question: *to what extent can WhiSPA's embeddings be further refined to incorporate affective context for psychological prediction?* Given that vocal prosody and acoustic features convey essential emotional and psychological cues beyond textual content (Low et al., 2020), incorporating these dimensions is crucial for a more comprehensive representation.

We acknowledge that this strong alignment with text-based language models may introduce an imbalance, diminishing the richness of acoustic cues that are particularly valuable for affective and psychological assessments. Despite WhiSPA's demon-

strated success-—matching its language model teacher in psychological prediction and surpassing state-of-the-art audio models—-there remains an opportunity to enhance its representational capacity by preserving acoustic features. To address this, future work will explore a multi-weighted dual loss objective, ensuring that WhiSPA retains a broader spectrum of information beyond textual representations. We suspect this refinement would not only improve its efficacy in psychological modeling but also enhance its versatility for general-purpose speech tasks like automatic speech recognition (ASR) and emotion recognition in conversation (ERC), where both linguistic and acoustic cues are essential.

## 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.

This work was part of a study approved by SBU's Institutional Review Board (IRB #1157153 World Trade Center Responder Language and Health Study) and (IRB #2022-00391: iHiTOP). The WTC and HiTOP recordings took place in a clinical setting at the Stony Brook WTC Health and Wellness Program where each participant gave 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

This work was supported in part by a grant from the CDC/NIOSH (U01 OH012476) and a grant from the NIH-NIAAA (R01 AA028032). The conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the CDC, NIH, or any other government organization. We greatly thank the participants, creators, maintainers, and interviewers from the WTC and HiTOP 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.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, 42(4):335–359.

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

Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. 2021. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7016–7025.

Yining Chen, Jianqiang Li, Changwei Song, Qing Zhao, Yongsheng Tong, and Guanghui Fu. 2024. Deep learning and large language models for audio and text analysis in predicting suicidal acts in chinese psychological support hotlines. *Preprint*, arXiv:2409.06164.

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.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

Herbert H. Clark and Michael F. Schober. 1992. Asking questions and influencing answers.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

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.

Jingjing Dong, Jiayi Fu, Peng Zhou, Hao Li, and Xiaorui Wang. 2022. Improving spoken language understanding with cross-modal contrastive learning. In *Interspeech*, pages 2693–2697.

Yumeng Fu. 2024. Ckerc : Joint large language models with commonsense knowledge for emotion recognition in conversation. *Preprint*, arXiv:2403.07260.

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.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

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).

Jehyun Kyung, Serin Heo, and Joon-Hyuk Chang. 2024. Enhancing multimodal emotion recognition through asr error compensation and llm fine-tuning. In *Proc. Interspeech 2024*, pages 4683–4687.

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.

Martin Lukac. 2024. Speech-based personality prediction using deep learning with acoustic and linguistic embeddings. *Sci. Rep.*, 14(1):30149.

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.

Giuseppe Sartori and Graziella Orrù. 2023. Language models and psychological sciences. *Frontiers in Psychology*, 14.

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

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, page 622–636. Association for Computational Linguistics.

Nikita Soni, H. Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2024. Large human language models: A need and the challenges. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8646, Mexico City, Mexico. Association for Computational Linguistics.

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.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. *Preprint*, arXiv:2407.21315.

Hao Yang, Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. 2023. Investigating pre-trained audio encoders in the low-resource condition. *Preprint*, arXiv:2305.17733.

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.

Zihan Zhao, Yanfeng Wang, and Yu Wang. 2022. Multi-level fusion of wav2vec 2.0 and bert for multimodal emotion recognition. *Preprint*, arXiv:2207.04697.

# A Appendix

## A.1 Data Description

### A.1.1 HiTOP.

The HiTOP dataset consists of video-recorded interviews conducted between World Trade Center 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 behaviors) and more. This dataset comprises of 525 unique participants.

**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)

### 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%).

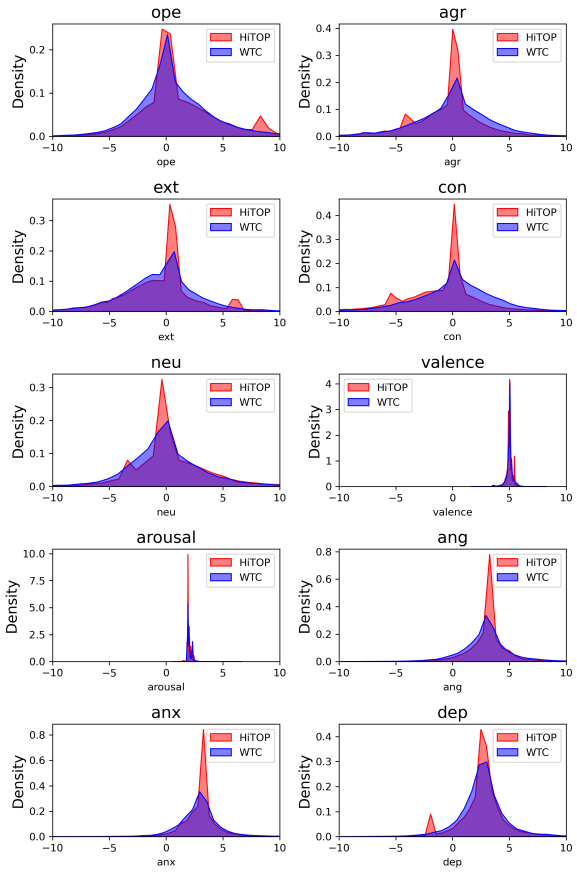



Figure 5: Standardized distributions of PsychEmb dimensions for each segment across both datasets. The distribution of WTC is shown in blue. The distribution of WTC is shown in red.

**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., intru-

sive 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 ($\geq .70$).

## 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: $1 \times 10^{-5}$. Weight Decay: $1 \times 10^{-2}$. Temperature ($\tau$): 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 $segment\_id$).

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 as visualized in Figure 6, 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.

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 which is depicted in Figure 7. 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.

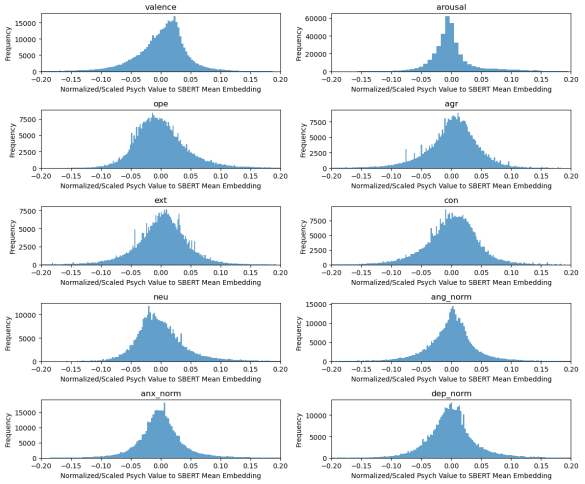



Figure 6: Distributions of psychological features standardized and scaled to the distribution of SBERT's mean embedding value before augmentation for WhiSPA alignment training.

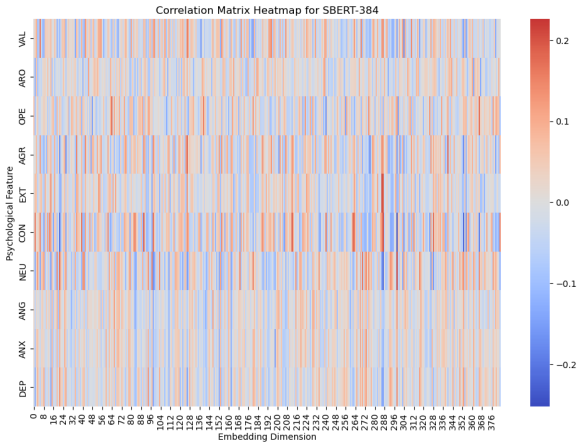



Figure 7: Pearson r correlation heatmap of SBERT-384's mean embedding. This visual displays the correlations of SBERT's 384 dimensions with each of the 10 PsychEmb dimensions.

## 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. The WTC dataset assessed PTSD symptom severity and related constructs, including anxiety and depression, using English-language data from WTC emergency responders. 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.

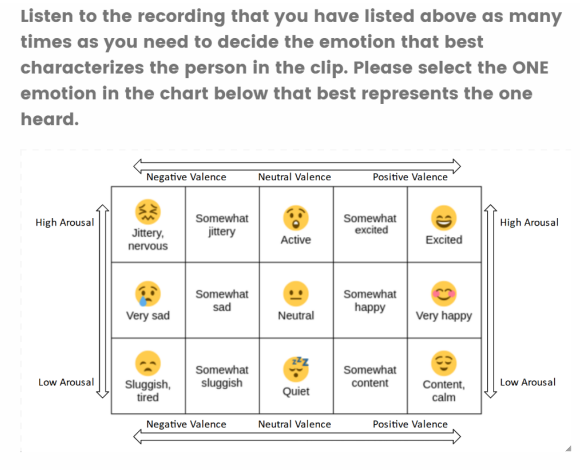



Figure 8: Annotator's affective circumplex visual grid for the task of manually annotating speech segments.

### A.4 Quantitative Analysis

PsychEmb's lower correlations in Figure 4 should not be mistaken for poor performance. With only 10 dimensions, PsychEmb representations achieve a staggering 24 and 22 Pearson points on **INT** and **DIS** respectively, emphasizing its validity as the psychological teacher. WhiSPA's consistent improvement over the audio models is attributed to the semantic and psychological dimensions that SBERT and PsychEmb offer. Notably, WhiSPA exemplifies drastic improvements in prediction accuracy for all outcomes compared to Whisper.

While WhiSPA demonstrates substantial advancements, surpassing even its text-based LM teacher, SBERT-1024, it remains inherently constrained by the representational capacity of the teacher model. If the teacher's capabilities are limited, these deficiencies inevitably carry over to the student, even after distillation. This is evident in the **ARO** column, where arousal —- an affective dimension -— is more accurately conveyed through acoustic cues. However, WhiSPA struggles to capture and preserve the acoustic information, instead predominantly aligning with the semantic representations provided by SBERT, thus limiting its ability to fully represent the nuanced affective content inherent in speech.

Beyond demonstrating superior alignment with established PTSD markers, Table 7 highlights WhiSPA's enhanced sensitivity to psychologically meaningful language patterns. Table 7a shows that n-grams reflecting personal experiences, self-referential content (e.g., first-person pronouns), and negative affective states correlate more strongly with WhiSPA's predictions than with those of Whisper. WhiSPA appears better attuned to indicators of psychological distress, anxiety, and trauma symptoms—-an advantage likely stemming from the contrastive alignment objective with text-based representations. The model's capacity to detect nuanced emotional and cognitive expressions in spoken language is further supported by its higher effect sizes on known PTSD-relevant n-grams, underscoring that semantically oriented embeddings can bolster the recognition of clinically significant markers in audio data.

Meanwhile, Table 7b points to a distinctive negative association between WhiSPA's predicted severity scores and n-grams referencing positive affect or social relationships. This result suggests that the same semantically focused latent space that amplifies negative or distress-related terms also filters out language tied to more adaptive or supportive experiences. In practical terms, such an effect could be advantageous for screening or early detection: positive affect or relational talk might serve as a buffer or resilience indicator, thereby inversely correlating with predicted symptom severity. Taken together, these findings highlight the unique strength of WhiSPAs in capturing a wide spectrum of psychologically relevant linguistic markers, surpassing the granularity offered by audio models alone.

| n-gram | r (WhiSPA) | r (Whisper) |
|---|---|---|
| me | 0.261 | 0.211 |
| ptsd | 0.226 | 0.126 |
| mental | 0.200 | 0.076 |
| because | 0.195 | 0.190 |
| therapist | 0.188 | 0.088 |
| anxiety | 0.187 | 0.075 |
| my therapist | 0.175 | 0.089 |
| my mental health | 0.167 | 0.072 |
| stress | 0.165 | 0.055 |
| want | 0.161 | 0.098 |
| through this | 0.160 | 0.082 |
| pain | 0.158 | 0.171 |
| body | 0.156 | 0.105 |
| this | 0.155 | 0.113 |
| mental health , | 0.152 | 0.051 |
| i had no | 0.151 | 0.135 |
| depression | 0.148 | 0.101 |
| shit | 0.147 | 0.148 |
| but i can't | 0.145 | 0.041 |
| flashbacks | 0.144 | 0.113 |

(a)

| n-gram | r (WhiSPA) | r (Whisper) |
|---|---|---|
| family | -0.264 | -0.200 |
| will be | -0.201 | -0.108 |
| college | -0.199 | -0.099 |
| we've | -0.190 | -0.155 |
| will | -0.182 | -0.065 |
| wife | -0.180 | -0.068 |
| pretty | -0.176 | -0.161 |
| as | -0.172 | -0.127 |
| good | -0.170 | -0.170 |
| hopefully | -0.167 | -0.155 |
| my wife | -0.165 | -0.070 |
| graduated from | -0.163 | -0.028 |
| would | -0.159 | -0.117 |
| able | -0.154 | -0.051 |
| i would say | -0.153 | -0.098 |
| able to | -0.153 | -0.054 |
| kids will | -0.153 | -0.102 |
| would say | -0.152 | -0.100 |
| vacations | -0.151 | -0.152 |
| lucky | -0.150 | -0.101 |

(b)

Table 7: (a) Top positively correlated N-grams with WhiSPA prediction for PCL scores on the WTC dataset and the corresponding correlations with Whisper predictions. (b) Top negatively correlated N-grams with WhiSPA prediction for PCL scores on the WTC dataset and the corresponding correlations with Whisper predictions. All correlations are statistically significant ($p<.05$; Benjamini Hochberg corrected).