N-CORE: N-View Consistency Regularization for Disentangled Representation Learning in Nonverbal Vocalizations

Anonymous ACL submission

Abstract

Nonverbal vocalizations are an essential component of human communication, conveying rich information without linguistic content. However, the computational analysis of nonverbal vocalization faces significant challenges due to a lack of lexical anchors in the data, compounded by biased distributions of imbalanced multi-label data. While disentangled representation learning has shown promise in isolating specific speech features, its application to nonverbal speech remains unexplored. In this paper, we introduce N-CORE, a novel supervised framework designed to disentangle representations in nonverbal vocalizations by leveraging N views of the audio sample to learn invariance to specific perturbed features. We find that N-CORE achieves competitive performance compared to the baseline methods when tested for emotion and speaker classification tasks on the VIVAE, ReCANVo, and ReCANVo-Balanced datasets. We further propose an emotion perturbation function for audio signals that preserves speaker information, and validate speech transformation functions on nonverbal vocalizations. Our work informs research directions on the application of paralinguistic speech, including privacy-preserving encoding, clinical diagnoses of atypical speech, and longitudinal analysis of communicative development.

1 Introduction

Nonverbal vocalizations (NVVs) are a fundamental component of human communication, encompassing a diverse range of non-speech sounds such as laughter, sighs, cries, and other sounds that convey rich affective information without relying on linguistic content (Cowen et al., 2019). Interpreting these paralinguistic signals is vital for comprehensive modeling of human communication and the development of emotionally intelligent AI systems (Tzirakis et al., 2023). However, the computational analysis of NVVs presents unique challenges that differentiate them from conventional speech processing tasks.

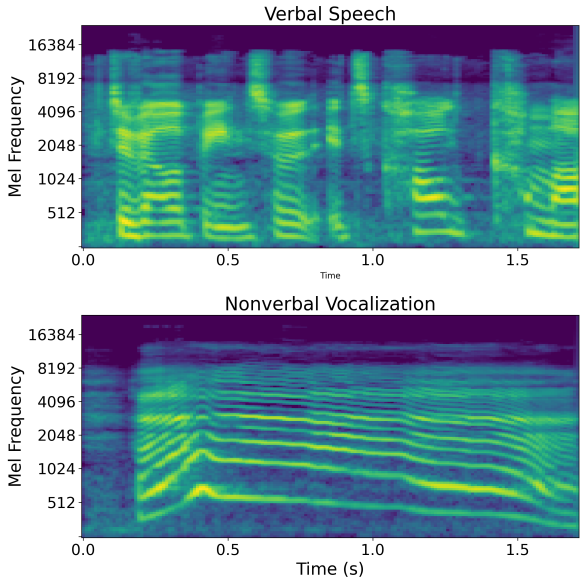


Figure 1: Comparison of mel-spectrograms from verbal (top) and nonverbal vocalizations (NVVs; bottom). The syllabic structure of word-based speech results in specific temporal variations that are less common in NVVs.

A primary challenge in NVV analysis is the scarcity of annotated data. Unlike speech corpora that can leverage millions of hours of recorded content, NVV datasets are typically limited to hundreds of hours of data (Baird et al., 2022; Koudounas et al., 2025), which leads to the suboptimal performance of modern data-hungry machine learning (ML) and deep learning (DL) methods. This limitation is exacerbated by the substantial bias caused by the low diversity in emotion and speaker labels in these datasets. For example, models trained on these limited datasets often encode speaker characteristics like pitch range, vocal timbre, and articulation patterns that confound affective computing, and affective features such as dynamic intensity, prosodic contours, and fundamental frequency (F0) variations that con-

found speaker identification (Pei et al., 2024). This leads to poor generalization across demographic groups and emotional categories, especially in low-resource datasets.

Several ML methods have attempted to address these challenges through various audio representation learning approaches. Foundation models such as HuBERT (Hsu et al., 2021b) and Wav2Vec2 (Baevski et al., 2020) have demonstrated remarkable success in learning generalized speech representations that can be fine-tuned for downstream tasks. These models are predominantly trained on verbal corpora, where canonical phoneme structures and linguistic content serve as strong structural priors. In contrast, NVVs lack the phoneme-based priors these models exploit, causing them to struggle when encoding paralinguistic sounds (Lane et al., 2015; Tzirakis et al., 2023). Figure 1 illustrates these differences by comparing mel-spectrograms of verbal speech and NVVs, highlighting how verbal speech has more complex spectral variability and clear transitions in temporal segmentation as compared to NVVs, which may assist representation learning (Nagamine et al., 2015).

Disentangled representation learning (DRL), the process of separating different informational factors in data, has been extensively explored in the speech domain for tasks including emotion recognition (Yuan et al., 2024; Xi et al., 2022), depression detection (Ravi et al., 2022), and voice conversion (Zuo et al., 2024; Wang et al., 2021a). However, applying DRL methods to NVVs presents unique challenges due to the absence of lexical anchors and how their prosodic characteristics simultaneously encode both speaker and emotion information. Conventional DRL methods on speech data often depend on augmentation strategies that preserve lexical content while altering specific features (Tu et al., 2024; Hsu et al., 2019); however, in the absence of lexical content invariant to perturbations, a single transformation may either disrupt useful information or allow uninformative artifacts to persist in the signal.

In this paper, we investigate DRL in NVVs. Our contributions are summarized as follows:

- We propose N-CORE (N-View COnsistency REgularization), a novel framework for supervised DRL of NVVs by using N perturbed views of an audio signal.
- We propose a novel transformation method to perturb emotion components in NVV while retaining speaker characteristics. We further examine the validity of an existing speaker perturbation method on NVVs.
- We comprehensively benchmark audio foundation models, DRL methods, and state-of-the-art representation learning methods on emotion and speaker classification tasks across three NVV datasets. To the best of our knowledge, we are the first to study DRL in NVVs.

2 Related Work

2.1 Machine Learning for Nonverbal Vocalizations and Paralinguistic Speech

Early work using ML to process NVVs relied predominantly on hand-engineered feature sets, with Schuller et al. (2013) establishing the ComParE acoustic feature set that captured spectral, prosodic, and voice quality parameters for paralinguistic analysis of social signals, conflict, and emotion, with application to autism diagnosis. This was further refined by the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and extended GeMAPS (eGEMAPS) frameworks Eyben et al. (2015), providing a standardized feature extraction framework optimized for affective computing applications. These approaches have been successfully employed for classifying NVVs (Lefter and Jonker, 2017; Narain et al., 2020), typically using traditional machine learning classifiers such as Support Vector Machines (Cortes and Vapnik, 1995) and Random Forests (Breiman, 2001).

The advent of deep learning (DL) has significantly advanced the processing of NVVs, learning representations directly from raw waveforms and bypassing manual feature engineering. Convolutional neural networks (CNNs) like ResNet-50 (He et al., 2016) have been used to process paralinguistic speech for understanding nonverbal emotion (Hsu et al., 2021a), speaker classification (Xu et al., 2024), and judging singing voice quality (Xu et al., 2022). The emergence of self-supervised learning has revolutionized ML for speech, with models like Wav2Vec2 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021b) achieving state-of-the-art performance on various speech processing benchmarks. While these models were primarily trained on linguistic content, they have strong transfer learning capabilities in paralinguistic tasks (Tzirakis et al., 2023; Shah and Johnson, 2025; Phukan

et al., 2025), although with performance limitations due to domain mismatch in their training and evaluation datasets. Specifically tailored to NVVs, Koudounas et al. (2025) developed Voc2Vec, which implements a self-supervised learning objective to pre-train the Wav2Vec2 architecture over multiple NVV datasets.

In clinical applications, ML approaches have been instrumental in analyzing atypical vocalizations. Bone et al. (2017) developed a classification framework to identify distinctive acoustic signatures in the vocalizations of children with autism spectrum disorder (ASD). Similarly, Narain et al. (2022) demonstrated that ML methods could effectively classify affective and communicative functions in NVVs from individuals with ASD. Further, these techniques have been applied to speech therapy (Mulfari et al., 2021), automatic speech recognition (Mulfari et al., 2023), and speech conversion (Doshi et al., 2021) for individuals with atypical speech.

2.2 Disentangled Representation Learning in Speech

DRL aims to separate different informational factors in data, enabling models to extract and manipulate independent semantic dimensions (Wang et al., 2024b). In speech processing, DRL typically focuses on separating speaker characteristics, linguistic content, emotion, or other features from each other (Williams, 2022). This separation is valuable for tasks such as voice conversion (Luong and Tran, 2021), speech recognition (Trinh and Braun, 2022), and emotion recognition (Yuan et al., 2024), where isolating specific attributes leads to improved performance. Many DRL methods for speech leverage lexical content and phoneme sequences in speech (Hsu et al., 2019), which act as a stable anchor against the disentanglement of various attributes like emotion or speaker identity, which are conveyed through prosodic modulations (Chu et al., 2006). The application of these techniques to NVVs, which lack explicit lexical anchors and have entangled speaker and emotion information in their prosodic features, remains an unexplored domain, motivating us to investigate DRL in NVVs.

A prominent approach for DRL involves a gradient reversal layer (GRL) (Ganin and Lempitsky, 2015), enabling end-to-end training of classifiers invariant to characteristics like domain (Lu et al., 2022) and speaker identity (Oneaţă et al., 2021). Autoencoder-based methods are also widely used to learn disentangled latent spaces by imposing specific constraints on the latent distribution (Yingzhen and Mandt, 2018; Nam et al., 2024). Subsequent frameworks like NANSY (Choi et al., 2021) and ContentVec (Qian et al., 2022) learn speaker-invariant speech representations by encouraging models to learn similar representations for two audio samples with different speaker information; however, this sole perturbed sample may not expose the model to the spectrum of varied features that may exist in a dataset. Further, these methods are limited to DRL for speaker-invariant representations, as they rely solely on speaker perturbation. To address these gaps, we propose N-CORE, which uses N views of perturbed samples from an audio signal for increased sample diversity, and an emotion perturbation method that preserves valuable speaker information.

3 Methodology

In this section, we describe N-CORE, our proposed supervised DRL framework to encode NVVs by isolating emotion- and speaker-specific information. Our method uses HuBERT as a backbone encoder and applies audio perturbations to suppress either emotion or speaker information while preserving the inverse features. We generate N perturbed views per audio sample to encourage invariance across a broader distribution of irrelevant variations, regulated by a pairwise distance loss for consistency regularization. Finally, we use two classification heads, one with a GRL mechanism, to simultaneously learn required features while performing supervised disentanglement of emotion and speaker information in the representations. We train the model via a composite objective that balances regularization, cross-entropy, and gradient reversal losses. Figure 2 presents the architecture of the N-CORE framework.

3.1 Problem Formulation

Let X represent an acoustic signal encompassing an NVV with a positive label y^+ and a negative label y^-. We aim to learn a representation model $R = f(X)$ that maps X to a learned embedding $x \in \mathbb{R}^D$, encapsulating the core components of the NVV represented by y^+ while discarding information that describes y^-. Specifically, if the learning objective is to classify for the emotion label y_e, x must retain information pertinent to the

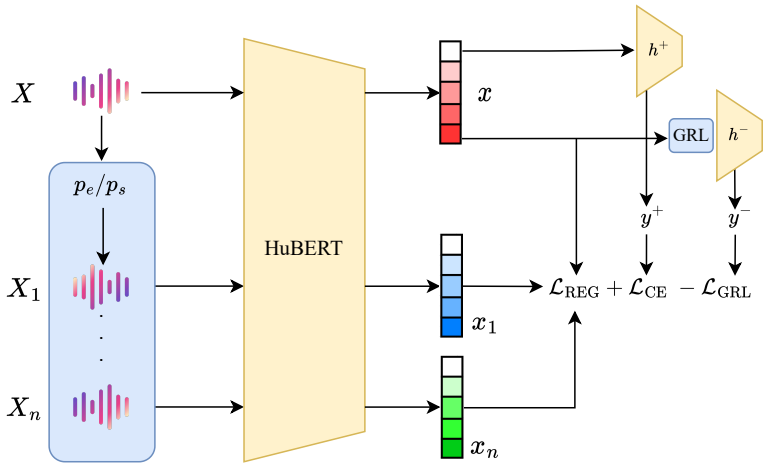


Figure 2: Our proposed framework, N-CORE, to classify for label y^+ and disentangle features that inform the label y^-. Perturbation functions p_e or p_s are used to create N views of X for consistency regularization. Cross-entropy loss is used for classifying y^+ with classification head h^+, and a GRL is used for adversarial disentanglement with respect to y^- using classification head h^-.

underlying emotion expressed in X while remaining uninformative with respect to speaker label y_s. Conversely, when classifying for y_s, x should encapsulate speaker-specific traits from X while discarding affective content descriptive of y_e. Achieving such disentanglement is challenging given the inherent entanglement of emotion- and speaker-related information in the acoustic signal.

3.2 Representation Learner

We use the HuBERT-Base model (Hsu et al., 2021b) pre-trained on 960 hours of speech data from the LibriSpeech dataset (Panayotov et al., 2015) as our feature encoder for its representation learning capabilities in both emotion and speaker recognition tasks (Wang et al., 2021b). HuBERT learns a neural embedding x from the raw audio signal X by encoding essential phonetic, prosodic, and stylistic information (Kharitonov et al., 2021), as $x = HuBERT(X)$.

3.3 Feature-Invariant Audio Perturbation

Emotion Perturbation. We aim to disrupt affective information in the audio signal while preserving speaker characteristics. The emotion perturbation function p_e comprises three transformations: 1) We compute the Short-Time Fourier Transform (STFT) of X, resulting in a spectrogram $S(X)$ with n_{spec} non-overlapping frequency bands. We randomly permute η_1 of these bands, retaining the rhythm and energy essential for speaker identification (Quatieri et al., 1994), while distorting content information (Davis and Johnsrude, 2003). 2) We normalize intensity by adjusting the waveform's RMS to a fixed target η_2 in order to suppress dynamic intensity correlated with emotion features (Koolagudi and Rao, 2012). 3) We flatten the pitch of the speaker to the average in their pitch contour f_0, effectively flattening prosodic variance and the affective content it withholds (Mozziconacci, 2002).

Speaker Perturbation. We adopt the audio transformation pipeline designed by Choi et al. (2021) for the NANSY framework to perturb speaker information while preserving the underlying content information. Similar to ContentVec (Qian et al., 2022), the speaker perturbation function p_s comprises three transformations: 1) scaling formant frequencies by a factor of ρ_1; 2) scaling F0 in every frame by ρ_2, and 3) applying a random equalizer to account for channel variations.

3.4 N Perturbed Views

Prior work on feature-invariant representation learning (Qian et al., 2022; Tu et al., 2024; Wang et al., 2024a) typically generates only a single perturbed version of each input and then enforces invariance between them. This one-shot strategy inherently constrains the diversity of transformations exposed to the model, making it less robust to unseen distortions.

In contrast, our approach samples N distinct perturbations drawn independently from the original audio signal X. By exposing the model to a spectrum of variations, we increase the range of uninformative factors the encoder is encouraged to

ignore, reduce reliance on any single perturbation pattern, and promote the learned feature space to consistently encode all N views of X into a tight cluster in the representation space. Multiple perturbations are especially crucial in NVVs, in the absence of lexical anchors that could be preserved after perturbation (Ko et al., 2015). We regularize the pairwise distances among all views of x by measuring the average squared distance across all unique pairs as a loss function:

$$\mathcal{L}_{\text{REG}} = \frac{\sum_{i=1}^{M} \sum_{j=i+1}^{N} ||x_i - x_j||_2^2}{\frac{N(N+1)}{2}} \quad (1)$$

where the denominator term denotes the number of unique pairs among x's $N+1$ views, including the unperturbed representation. This loss encourages the model to create the same representation for all views of x, disentangling label-relevant information from uninformative features.

3.5 Classification

We project x to two separate classification heads h^+ and h^- that use cross-entropy to classify for labels y^+ and y^-, respectively. This step operates solely on the unperturbed x, and none of its augmented views. Both heads share the same underlying structure: a two-layer multilayer perceptron (MLP) with a ReLU activation and dropout in between. To enforce invariance to y^-, we precede h^- with a GRL that scales embeddings by $-\alpha$, encouraging the model to disentangle and suppress features corresponding to y^- in its learned representations. We obtain losses $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{GRL}}$ as follows:

$$\mathcal{L}_{\text{CE}} = -\sum_{k^+=1}^{K^+} y_{k^+}^{+} \log\left[h^+(x)\right]_{k^+} \quad (2)$$

$$\mathcal{L}_{\text{GRL}} = -\sum_{k^-=1}^{K^-} y_{k^-}^{-} \log\left[h^-\left(\text{GRL}_\alpha(x)\right)\right]_{k^-} \quad (3)$$

3.6 Training Objective

Our model is trained by optimizing a composite objective function comprising the three losses obtained from equations 1, 2, and 3, calculated for each input audio embedding x and its N perturbed views.

$$\mathcal{L}_{\text{total}} = \lambda_{\text{REG}} \cdot \mathcal{L}_{\text{REG}} + \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}} - \lambda_{\text{GRL}} \cdot \mathcal{L}_{\text{GRL}} \quad (4)$$

where λ_{reg}, λ_{CE}, and λ_{GRL} represent scaling factors that regulate the contribution of each loss towards $\mathcal{L}_{\text{total}}$. The optimizer minimizes $\mathcal{L}_{\text{total}}$ by maximizing the negative term $\mathcal{L}_{\text{GRL}}$, designed to learn representations that are invariant to the secondary label y_2.

4 Experimental Settings

4.1 Datasets

We evaluate our methods on three NVV datasets: Variably Intense Vocalizations of Affect and Emotion (VIVAE) (Holz et al., 2022), Real-World Communicative and Affective Nonverbal Vocalizations (ReCANVo) (Johnson et al., 2023), and ReCANVo-Balanced. For each dataset, we evaluate performance on emotion and speaker recognition tasks. We use a train/test split of 80/20 for all datasets. Detailed dataset statistics are presented in A.2.

VIVAE. The VIVAE corpus comprises 1,085 non-speech emotion vocalizations produced by 11 non-professional female actors, 20-39 years old, who were instructed to express six affective states: achievement/triumph, sexual pleasure, surprise, anger, fear, and physical pain across multiple intensity levels.

ReCANVo. The ReCANVo dataset contains 7,077 NVVs collected from eight non- and minimally-speaking individuals, ranging in age from 6-23 years old and diagnosed with various neurodevelopmental disorders, including ASD, cerebral palsy, and genetic neurodevelopmental disorders. Classes with sample counts across all participants reaching n$\geq$100 were taken from this dataset, yielding a derived dataset of 6,551 utterances distributed among seven functions: delighted, dysregulated, frustrated, laughter, request, self-talk, and social. This derived dataset is highly imbalanced with an imbalance factor of 18.66.

ReCANVo-Balanced. We use a multi-stage sampling procedure to create a balanced subset from ReCANVo by extracting 100 samples for each emotion class. Within each emotion category, participant diversity was maximized by systematically distributing the sample selection, with the constraint that no single participant would contribute a majority of samples for any given emotion class.

4.2 Baselines

We conduct a comprehensive benchmark of established audio ML methods on NVVs. Specifically, we evaluate HuBERT (Hsu et al., 2021b), Wav2Vec2 (Baevski et al., 2020), Voc2Vec (Koudounas et al., 2025), HuBERT-ER and HuBERT-SID (Yang et al., 2021), HuBERT-GRL

and Wav2Vec2-GRL (Ganin and Lempitsky, 2015), SACE (Dutta and Ganapathy, 2024), ContentVec (Qian et al., 2022), and our proposed method, N-CORE. Detailed implementation details are given in Appendix A.1.

5 Experimental Results

5.1 Emotion Classification with Speaker Disentanglement

Table 1 presents the results for emotion recognition.

Foundation Models. In line with previous research on emotion and speaker classification (Wang et al., 2021b), HuBERT consistently achieves the highest performance across all metrics in all datasets compared to the Wav2Vec2 family of models. The Voc2Vec model was trained exclusively on NVVs, allowing it to outperform Wav2Vec2 with the same architecture, demonstrating the advantage of domain-specific pre-training. Further, its self-supervised training objective may enable it to avoid overfitting and classification unfairness (Liu et al., 2021), as demonstrated by the differential in F1-Score and UAR compared to Wav2Vec2. However, despite being specifically designed for NVVs, Voc2Vec underperforms HuBERT on ReCANVo and ReCANVo-Balanced while matching its performance on VIVAE, suggesting that domain-specific pre-training may not solely surpass the representation learning power of a more suitable model.

Domain-Specific Models. Notably, neither HuBERT-ER nor HuBERT-SID outperforms the baseline HuBERT model, which may be attributed to the domain shift between the spoken word datasets used during finetuning and the NVV datasets used for this evaluation. Further, fine-tuning on a smaller corpus limits these models' generalizability to out-of-distribution data.

Gradient Reversal-based Models. The addition of GRL improves performance for both HuBERT and Wav2Vec2 models across all datasets. These results support our hypothesis that using adversarial training to explicitly disentangle speaker information leads to more robust representations less influenced by speaker-specific characteristics and biases.

DRL Frameworks. ContentVec outperforms SACE across all datasets, which can be attributed to its superior HuBERT backbone compared to SACE's Wav2Vec2 backbone. N-CORE outperforms all methods on VIVAE and ReCANVo-Balanced but falls short for ReCANVo in terms of F1 and UAR, which may be due to the dataset's interweaved speaker and emotion distributions, where models could be relying on speaker characteristics to classify for emotions due to a biased sample distribution (see 6), and N-CORE's superior DRL capabilities ended up penalizing its performance. ReCANVo-Balanced mitigates this imbalance, and N-CORE outperforms all methods here.

5.2 Speaker Classification with Affect Disentanglement

Table 2 presents the results for speaker recognition.

Foundation Models. For the ReCANVo and ReCANVo-Balanced datasets, Voc2Vec performs notably worse than HuBERT and Wav2Vec2, despite ReCANVo being a part of its pre-training corpus. Voc2Vec also uses the VIVAE dataset for pre-training, on which it performs the best, followed by HuBERT and Wav2Vec2, respectively.

Domain-Specific Models. HuBERT-ER shows competitive performance in speaker identification capabilities compared to the baseline model and even the specialized HuBERT-SID model on ReCANVo, but exhibits a substantial drop on VIVAE, highlighting the importance of task-specific pre-training. However, the model performs poorly on ReCANVo-Balanced, suggesting that it could successfully be using affective information to classify speakers on ReCANVo.

Gradient Reversal-based Methods. On the VIVAE dataset, both models demonstrate substantial performance gains after disentanglement. On ReCANVo and ReCANVo-Balanced, the HuBERT model shows a slight improvement in performance after emotion disentanglement, whereas Wav2Vec2 experiences a minor decline relative to its baseline.

DRL Frameworks. ReCANVo's data imbalance proves to be challenging for N-CORE across both datasets. However, the model outperforms all other methods on the uniformly distributed VIVAE and ReCANVo-Balanced datasets. Notably, ContentVec outperforms all methods on ReCANVo despite being trained to be invariant to speakers, indicating that speaker perturbation may not transform all speaker features, and that models can still benefit from it.

5.3 Data Analysis

ReCANVo's data imbalance reflects real-life data distributions, where multi-label data often exhibit inherent biases (Schultheis et al., 2022). In this

Model	VIVAE			ReCANVo			ReCANVo-Balanced		
	Acc	F1	UAR	Acc	F1	UAR	Acc	F1	UAR
HuBERT	58.06	56.51	56.81	66.97	55.07	54.53	32.86	28.89	32.86
Wav2Vec2	51.15	50.69	50.49	61.94	50.45	49.61	24.29	15.89	24.29
Voc2Vec	57.60	57.07	57.03	61.17	53.21	52.24	29.29	27.79	29.29
HuBERT-ER	53.00	47.60	51.19	65.60	52.87	51.84	32.86	29.42	32.86
HuBERT-SID	57.14	56.26	56.51	63.54	54.39	54.54	32.14	30.59	32.14
HuBERT + GRL	59.91	59.16	59.20	67.73	**57.66**	**57.28**	35.71	33.59	35.71
Wav2Vec2 + GRL	53.00	51.69	51.85	63.46	52.99	53.31	32.14	30.19	32.14
SACE	53.00	51.47	52.14	64.00	52.83	53.61	27.86	23.97	27.86
ContentVec	59.91	59.41	59.27	65.06	56.97	55.94	31.43	27.74	31.43
N-CORE	**64.06**	**63.01**	**63.52**	**67.96**	55.74	54.90	**42.86**	**39.53**	**42.86**

Table 1: Comparison of model performance on the emotion classification task for VIVAE, ReCANVo, and ReCANVo-Balanced. The best results are highlighted in **bold** and the second-best results are underlined.

Model	VIVAE			ReCANVo			ReCANVo-Balanced		
	Acc	F1	UAR	Acc	F1	UAR	Acc	F1	UAR
HuBERT	60.83	56.90	58.72	93.97	92.38	92.91	77.86	77.93	77.59
Wav2Vec2	56.68	54.58	55.27	94.20	92.40	92.31	75.00	73.86	75.31
Voc2Vec	65.90	65.07	65.13	90.92	89.84	89.32	72.14	71.12	72.19
HuBERT-ER	51.61	46.05	48.69	93.82	92.30	92.51	59.29	55.37	56.62
HuBERT-SID	59.91	58.09	58.26	93.36	91.62	92.21	76.43	76.05	75.55
HuBERT + GRL	71.43	67.65	69.12	94.51	93.27	93.63	77.14	77.78	77.84
Wav2Vec2 + GRL	64.06	59.90	61.08	93.44	91.72	91.59	72.86	72.77	72.64
SACE	50.69	46.79	48.16	92.75	90.78	91.05	60.71	58.68	58.49
ContentVec	65.90	64.59	64.85	**95.96**	**94.95**	**95.13**	76.43	76.16	76.32
N-CORE	**75.12**	**74.25**	**74.54**	94.97	93.61	93.49	**80.71**	**80.01**	**80.43**

Table 2: Comparison of model performance on the speaker classification task for VIVAE, ReCANVo, and ReCANVo-Balanced. The best results are highlighted in **bold** and the second-best results are underlined.

context, affective vocalizations reflect the idiosyncratic behaviors of individual autistic speakers, and since the vocalizations are not acted, some samples may naturally lie between two emotional categories. These facets limit model performance for emotion classification despite the dataset's relatively large number of samples, with performance deteriorating significantly on ReCANVo-Balanced.

Universally, speaker identification proves more challenging on the VIVAE dataset across all models, with significantly lower performance compared to ReCANVo. This dataset contains acted vocalizations from adults, where emotional expressiveness tends to converge on shared cultural templates for what each affective vocalization is expected to sound like. This reduces inter-speaker variability by masking natural speaker-specific cues, making it more difficult for models to distinguish between speakers, especially compared to spontaneous, real-world vocalization datasets like ReCANVo. Disentanglement was particularly effective for speaker classification on VIVAE, suggesting that DRL excels for datasets containing relatively homogeneous speakers.

All the models demonstrated remarkably high performance on speaker identification for ReCANVo, which may be due to the diverse age range of the dataset and the idiosyncratic forms of NVVs across individuals with autism (Pegado et al., 2020), making the accurate speaker classification a relatively easier ML task. The competitive performance of all models on the small-scale ReCANVo-Balanced dataset shows that even a relatively small corpus of NVVs can help create effective speaker recognition systems for unique populations.

5.4 Cross-Verification of Perturbation

We conducted a cross-verification experiment to validate the efficacy of our affect and speaker perturbation functions by applying each to both clas-

sification tasks in VIVAE, and our results are presented in Table 3. Applying speaker perturbation p_s to speaker classification or emotion perturbation p_e to emotion classification significantly degrades performance, indicating the successful disruption of cues that the respective perturbation function targets. Conversely, applying the inverse pairing for the tasks leads to improved performance, indicating that the model learns to become invariant to the perturbed features, and that the respective transformations do not disrupt features informative to the classification task. This experiment validates our proposed p_e, and the applicability of both these transformations to NVVs.

Task	Perturbation	Performance		
		Acc.	F1	UAR
Speaker	p_e	**75.12**	**74.25**	**74.54**
	p_s	70.97	66.16	68.21
Emotion	p_e	61.29	61.18	60.64
	p_s	**64.06**	**63.01**	**63.52**

Table 3: Cross-Verification of signal perturbation efficacy using N-CORE on VIVAE. The best results are highlighted in **bold**.

5.5 Optimal number of perturbations

To identify the optimal number of perturbations (N) for N-CORE, we evaluated the model's classification accuracy on VIVAE while varying N from 1 to 7, with results presented in 3. We find that $N = 5$ leads to the best result for this dataset; however, this may vary with dataset size and the distribution of multi-labeled samples.

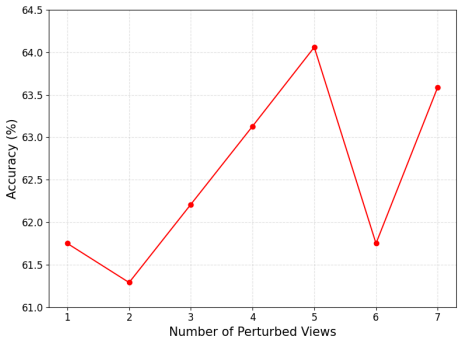


Figure 3: Accuracy vs. number of perturbed views with N-CORE for emotion classification on VIVAE. The y-axis is limited from 61.0 to 64.5 for clarity.

5.6 Ablation Study

We conduct a systematic ablation study on N-CORE to evaluate the individual contribution of its components, and the results are presented in Table 4. Our findings show a clear progression in performance across all metrics as we sequentially add GRL, regularization loss, and especially N perturbed views to the base HuBERT model.

Component				Performance		
HB	GRL	RL	NV	Acc.	F1	UAR
✓				58.06	56.51	56.81
✓	✓			59.91	59.16	59.20
✓	✓	✓		61.75	61.03	60.98
✓	✓	✓	✓	**64.06**	**63.01**	**63.52**

Table 4: Ablation studies were conducted on the N-CORE for emotion recognition in VIVAE. The abbreviations HB, GRL, RL, and NV refer to HuBERT, Gradient Reversal Layers, Regularization Loss, and N-Views, respectively. The final row corresponds to the entire framework. The best results are highlighted in **bold**.

6 Conclusion

In this paper, we investigated DRL specifically on NVVs. We proposed N-CORE, a novel disentanglement method using N-views of perturbed audio signals to disentangle relevant features from uninformative ones. Our experiments demonstrate that multi-view perturbation enhances performance compared to traditional single-view approaches, with N-CORE achieving competitive performance on both emotion and speaker classification tasks for VIVAE and ReCANVo-Balanced datasets. We further propose a signal transformation pipeline that perturbs emotions in speech signals while preserving speaker information. Further, we validate previous perturbation techniques, finding that these transformations are generalizable to NVVs.

Our work further establishes that DRL is indeed achievable for NVVs and applies to both typical and atypical paralinguistic speech. This opens several promising directions for future research and applications, including privacy-preserving encoding of NVVs, disentangled voice conversion for NVVs, and the clinical analysis of vocalizations from non- and minimally-speaking individuals. N-CORE further empowers longitudinal studies of communicative development through NVVs that remain invariant to changes in speaker characteristics over time. The modular design of N-CORE allows it to scale with advances in DL, potentially benefiting from larger foundation models as they become available. Our work is an important step toward more inclusive and accurate computational models of human paralinguistic communication.

Ethical Considerations

Potential Risks. We acknowledge the privacy implications of technologies that can separate speaker characteristics from communicative content. While N-CORE demonstrates benefits for privacy-preserving representations by removing identifying speaker information from emotion-focused embeddings, this same capability could potentially be misused for unauthorized voice anonymization or modification. We emphasize that any deployment of these technologies should adhere to strict privacy protocols and informed consent requirements, particularly when processing data from vulnerable populations such as non- and minimally-speaking individuals.

Biases. Our experimental results highlight how dataset imbalances can significantly affect model performance. Demographic limitations of training data may introduce biases that could impact the equitable performance of these systems across different populations. We urge caution in applying these models to populations not well-represented in the training data.

Reproducibility Statement. We include implementation details and hyperparameter settings for all models in Appendix A.1. The source code for N-CORE has been submitted for review with this paper and will be released publicly upon acceptance.

Limitations

Our study primarily focuses on disentangling emotion and speaker features. NVVs, however, convey a rich spectrum of paralinguistic information, including varying levels of intensity, different communicative intents beyond broad affective categories, and other subtle cues, which N-CORE does not explicitly disentangle. The generalizability of our findings is also constrained by the two datasets and one derived dataset; while diverse, they do not encompass the full variability of NVVs across different cultures, broader age ranges, numerous real-world acoustic environments, or a wider array of clinical populations. The general challenge of limited annotated NVV data also impacts the scale at which models can be trained and validated.

N-CORE's performance, particularly for emotion classification, was comparatively lower on the highly imbalanced ReCANVo dataset for F1 and UAR, and it was outperformed by other methods for speaker classification on the same dataset. This suggests that in scenarios with extreme data imbalance or where speaker and affective cues are deeply convoluted, our model's strong disentanglement capabilities might not directly translate to optimal performance for classification.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alice Baird, Panagiotis Tzirakis, Jeffrey A Brooks, Chris B Gregory, Björn Schuller, Anton Batliner, Dacher Keltner, and Alan Cowen. 2022. The acii 2022 affective vocal bursts workshop & competition. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–5. IEEE.

Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. 2017. Signal processing and machine learning for mental health research and clinical applications [perspectives]. *IEEE Signal Processing Magazine*, 34(5):196–195.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265.

Min Chu, Yong Zhao, and Eric Chang. 2006. Modeling stylized invariance and local variability of prosody in text-to-speech synthesis. *Speech Communication*, 48(6):716–726.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3(4):369–382.

Matthew H Davis and Ingrid S Johnsrude. 2003. Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, 23(8):3423–3431.

Rohan Doshi, Youzheng Chen, Liyang Jiang, Xia Zhang, Fadi Biadsy, Bhuvana Ramabhadran, Fang Chu, Andrew Rosenberg, and Pedro J Moreno. 2021. Extending parrotron: An end-to-end, speech conversion and speech recognition model for atypical speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6988–6992. IEEE.

Soumya Dutta and Sriram Ganapathy. 2024. Zero shot audio to audio emotion transfer with speaker disentanglement. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10371–10375. IEEE.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, and 1 others. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2022. The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception. *Emotion*, 22(1):213.

Jia-Hao Hsu, Ming-Hsiang Su, Chung-Hsien Wu, and Yi-Hsuan Chen. 2021a. Speech emotion recognition considering nonverbal vocalization in affective conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1675–1686.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021b. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. 2019. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5901–5905. IEEE.

Kristina T Johnson, Jaya Narain, Thomas Quatieri, Pattie Maes, and Rosalind W Picard. 2023. Recanvo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1):523.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, and 1 others. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech*, volume 2015, page 3586.

Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology*, 15:99–117.

Alkis Koudounas, Moreno La Quatra, Sabato Marco Siniscalchi, and Elena Baralis. 2025. voc2vec: A foundation model for non-verbal vocalization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Nicholas D Lane, Petko Georgiev, and Lorena Qendro. 2015. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 283–294.

Iulia Lefter and Catholijn M Jonker. 2017. Aggression recognition using overlapping speech. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 299–304. IEEE.

Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. 2021. Self-supervised learning is more robust to dataset imbalance. *arXiv preprint arXiv:2110.05025*.

Cheng Lu, Yuan Zong, Wenming Zheng, Yang Li, Chuangao Tang, and Björn W Schuller. 2022. Domain invariant feature learning for speaker-independent speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2217–2230.

Manh Luong and Viet Anh Tran. 2021. Many-to-many voice conversion based feature disentanglement using variational autoencoder. *arXiv preprint arXiv:2107.06642*.

Sylvie Mozziconacci. 2002. Prosody and emotions. In *Speech prosody*, volume 2002, pages 1–9.

Davide Mulfari, Lorenzo Carnevale, and Massimo Villari. 2023. Toward a lightweight asr solution for atypical speech on the edge. *Future Generation Computer Systems*, 149:455–463.

Davide Mulfari, Gabriele Meoni, Marco Marini, and Luca Fanucci. 2021. Machine learning assistive application for users with speech disorders. *Applied Soft Computing*, 103:107147.

Tasha Nagamine, Michael L Seltzer, and Nima Mesgarani. 2015. Exploring how deep neural networks form phonemic categories. In *Interspeech*, pages 1912–1916.

KiHyun Nam, Hee-Soo Heo, Jee-weon Jung, and Joon Son Chung. 2024. Disentangled representation learning for environment-agnostic speaker recognition. *arXiv preprint arXiv:2406.14559*.

Jaya Narain, Kristina T Johnson, Craig Ferguson, Amanda O'Brien, Tanya Talkar, Yue Zhang Weninger, Peter Wofford, Thomas Quatieri, Rosalind

Picard, and Pattie Maes. 2020. Personalized modeling of real-world vocalizations from nonverbal individuals. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 665–669.

Jaya Narain, Kristina T Johnson, Thomas F Quatieri, Rosalind W Picard, and Pattie Maes. 2022. Modeling real-world affective and communicative nonverbal vocalizations from minimally speaking individuals. *IEEE Transactions on Affective Computing*, 13(4):2238–2253.

Dan Oneaţă, Adriana Stan, and Horia Cucu. 2021. Speaker disentanglement in video-to-speech conversion. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 46–50. IEEE.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Felipe Pegado, Michelle HA Hendriks, Steffie Amelynck, Nicky Daniels, Jean Steyaert, Bart Boets, and Hans Op de Beeck. 2020. Adults with high functioning autism display idiosyncratic behavioral patterns, neural representations and connectivity of the 'voice area'while judging the appropriateness of emotional vocal reactions. *Cortex*, 125:90–108.

Guanxiong Pei, Haiying Li, Yandi Lu, Yanlei Wang, Shizhen Hua, and Taihao Li. 2024. Affective computing: Recent advances, challenges, and future trends. *Intelligent Computing*, 3:0076.

Orchid Chetia Phukan, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Sishir Kalita, Arun Balaji Buduru, Rajesh Sharma, SR Mahadeva Prasanna, and 1 others. 2025. Strong alone, stronger together: Synergizing modality-binding foundation models with optimal transport for non-verbal emotion recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. 2022. Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International conference on machine learning*, pages 18003–18017. PMLR.

Thomas F Quatieri, CR Jankowski, and Douglas A Reynolds. 1994. Energy onset times for speaker identification. *IEEE Signal Processing Letters*, 1(11):160–162.

Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. 2022. A step towards preserving speakers' identity while detecting depression via speaker disentanglement. In *Interspeech*, volume 2022, page 3338.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, and 1 others. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Erik Schultheis, Marek Wydmuch, Rohit Babbar, and Krzysztof Dembczynski. 2022. On missing labels, long-tails and propensities in extreme multi-label classification. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1547–1557.

Siddhant Bikram Shah and Kristina T Johnson. 2025. Multi-feature audio fusion for nonverbal vocalization classification. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Viet Anh Trinh and Sebastian Braun. 2022. Unsupervised speech enhancement with speech recognition embedding and disentanglement losses. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 391–395. IEEE.

Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien. 2024. Contrastive self-supervised speaker embedding with sequential disentanglement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Panagiotis Tzirakis, Alice Baird, Jeffrey Brooks, Christopher Gagne, Lauren Kim, Michael Opara, Christopher Gregory, Jacob Metrick, Garrett Boseck, Vineet Tiruvadi, and 1 others. 2023. Large-scale nonverbal vocalization detection using transformers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Huimeng Wang, Zengrui Jin, Mengzhe Geng, Shujie Hu, Guinan Li, Tianzi Wang, Haoning Xu, and Xunying Liu. 2024a. Enhancing pre-trained asr system fine-tuning for dysarthric speech recognition using adversarial data augmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12311–12315. IEEE.

Jie Wang, Jingbei Li, Xintao Zhao, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2021a. Adversarially learning disentangled speech representations for robust multi-factor voice conversion. *arXiv preprint arXiv:2102.00184*.

Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2024b. Disentangled representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021b. A fine-tuned wav2vec

2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*.

Jennifer Williams. 2022. *Learning disentangled speech representations*. Ph.D. thesis, University of Edinburgh.

Yu-Xuan Xi, Yan Song, Li-Rong Dai, Ian McLoughlin, and Lin Liu. 2022. Frontend attributes disentanglement for speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7712–7716. IEEE.

Anfeng Xu, Kevin Huang, Tiantian Feng, Helen Tager-Flusberg, and Shrikanth Narayanan. 2024. Audio-visual child-adult speaker classification in dyadic interactions. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8090–8094. IEEE.

Yanze Xu, Weiqing Wang, Huahua Cui, Mingyang Xu, and Ming Li. 2022. Paralinguistic singing attribute recognition using supervised machine learning for describing the classical tenor solo singing voice in vocal pedagogy. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):8.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Li Yingzhen and Stephan Mandt. 2018. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pages 5670–5679. PMLR.

Zhichen Yuan, CL Philip Chen, Shuzhen Li, and Tong Zhang. 2024. Disentanglement network: Disentangle the emotional features from acoustic features for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11686–11690. IEEE.

Lishi Zuo, Man-Wai Mak, and Youzhi Tu. 2024. Promoting independence of depression and speaker features for speaker disentanglement in speech-based depression detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10191–10195. IEEE.

A Appendix

A.1 Implementation Details

We conducted all our experiments on Python 3.9.21 and PyTorch 2.6.0 on an NVIDIA V100 GPU with 32 GB of dedicated memory. We set the batch size to 16 and trained each model for 100 epochs with an early stopping patience of 20 while monitoring validation accuracy to save the best model for each run. We used the default settings set by each tested method's authors. When unspecified, we used a learning rate of 10^{-5} with the AdamW optimizer, including for N-CORE. We use $N = 5$ for all experiments on N-CORE. We used a linear scheduler with 0.1 $\times$ the number of training steps as warmup steps. We report the maximum performance achieved for each model.

We implemented HuBERT[1], Wav2Vec2[2], and Voc2Vec2[3], HuBERT-ER[4], HuBERT-SID[5], and ContentVec[6] through the HuggingFace library. We implemented GRL from GitHub[7]. We implemented SACE[8] and from the code released by the respective.

A.2 Dataset Distribution

Detailed dataset statistics for VIVAE, ReCANVo, and ReCANVo-Balanced are presented in Tables 5, 6, and 7.

A.3 Disentanglement Training

N-CORE's DRL dynamics for emotion classification on VIVAE is illustrated through the loss and accuracy curves presented in Figure 4 and Figure 5, respectively. Figure 4 shows the emotion classification loss decreasing and stabilizing over epochs, while the adversarial speaker classification loss increases, as intended with the use of a GRL. Concurrently, Figure 5 shows that the emotion classification accuracy consistently improves until stabilization, whereas the speaker classification accuracy rapidly drops to random chance. These trends infer the model's success in learning representations that are discriminative for emotion while simultaneously becoming invariant to speaker characteristics over the training period.

A.4 TSNE Plots

We use TSNE plots to compare HuBERT and N-CORE on the testing sets of VIVAE in Figures 6 and 7, and ReCANVo in Figures 8 and 9. Representations from N-CORE were generated solely using the HuBERT backbone.

[1] https://huggingface.co/facebook/hubert-base-ls960
[2] https://huggingface.co/facebook/wav2vec2-base-960h
[3] https://huggingface.co/alkiskoudounas/voc2vec
[4] https://huggingface.co/superb/hubert-base-superb-er
[5] https://huggingface.co/superb/hubert-base-superb-sid
[6] https://huggingface.co/lengyue233/content-vec-best
[7] https://github.com/tadeephuy/GradientReversal
[8] https://github.com/iiscleap/ZEST/

Label	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	Total
achievement	16	11	12	18	20	12	17	16	18	14	7	161
anger	12	18	15	18	18	20	14	19	17	16	7	174
fear	16	17	14	18	19	19	17	18	17	13	8	176
pain	17	20	21	17	19	20	18	14	19	12	8	185
pleasure	19	19	20	17	15	19	20	20	18	18	17	202
surprise	13	16	19	20	20	21	17	21	19	14	7	187
Total	**93**	**101**	**101**	**108**	**111**	**111**	**103**	**108**	**108**	**87**	**54**	**1085**

Table 5: Data distribution of the VIVAE dataset.

Label	P01	P02	P03	P05	P06	P08	P11	P16	Total
delighted	357	43	25	235	227	39	207	139	1272
dysregulated	212	0	302	116	5	13	22	34	704
frustrated	150	56	47	283	30	781	27	162	1536
request	130	13	61	6	124	44	22	19	419
self-talk	564	34	55	286	56	503	33	354	1885
social	182	247	0	0	1	93	52	59	634
laughter	0	38	8	13	0	42	0	0	101
Total	1595	431	498	939	443	1515	363	767	**6551**

Table 6: Data distribution of the ReCANVo dataset.

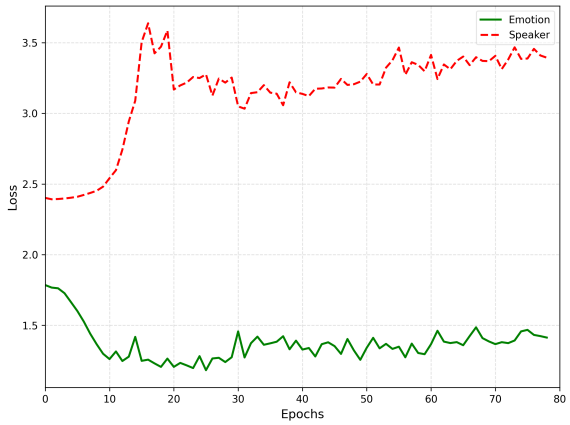


Figure 4: Loss vs. Number of Epochs for emotion classification on VIVAE.

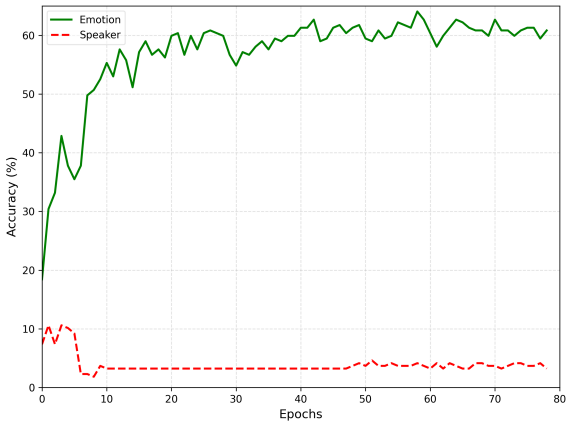
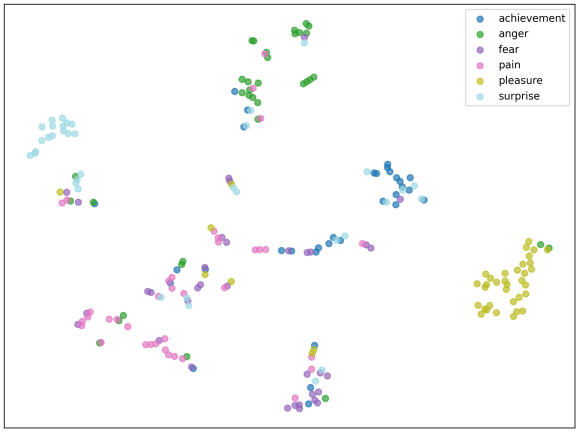


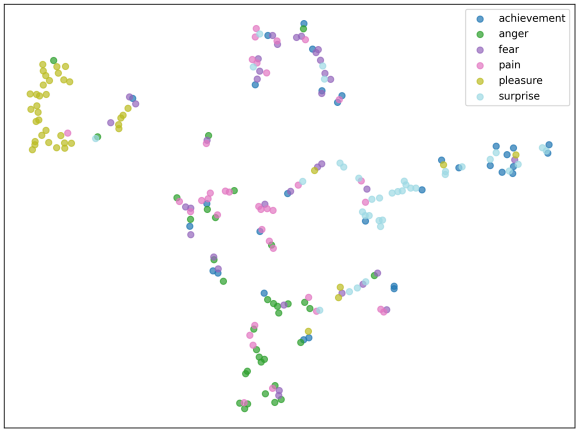
Figure 5: Accuracy vs. Number of Epochs for emotion classification on VIVAE.

Label	P01	P02	P03	P05	P06	P08	P11	P16	Total
delighted	13	13	13	13	12	12	12	12	100
dysregulated	17	0	17	16	5	13	16	16	100
frustrated	12	12	13	13	12	13	13	12	100
request	14	13	13	6	14	13	13	14	100
self-talk	13	12	13	12	12	13	13	12	100
social	20	20	0	0	1	20	20	19	100
laughter	0	38	8	13	0	41	0	0	100
Total	89	108	77	73	56	125	87	85	**700**

Table 7: Data distribution of the ReCANVo-Balanced dataset.

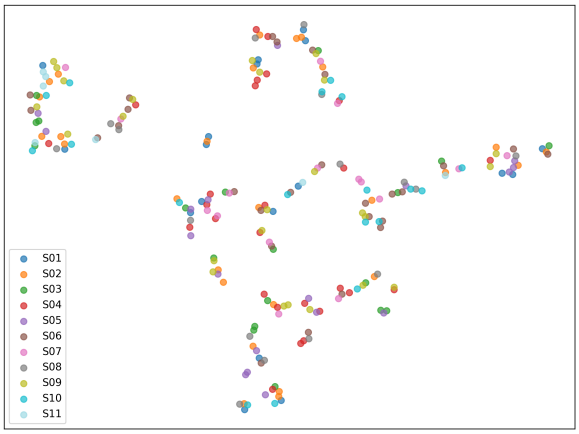


(a) N-CORE: Emotion labels highlighted.



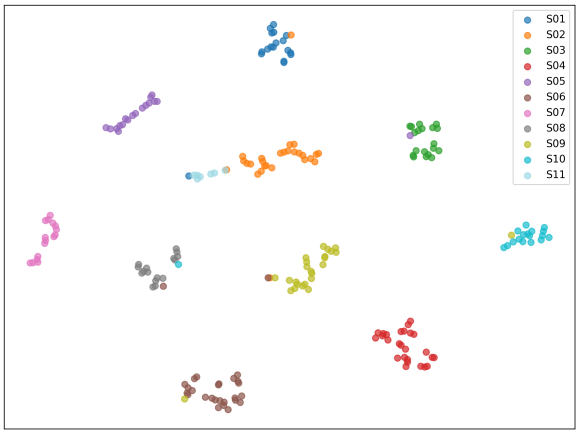
(b) HuBERT: Emotion labels highlighted.

(c) N-CORE: Speaker labels highlighted.

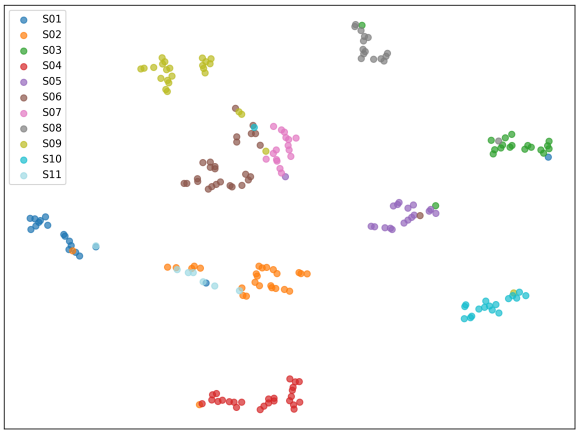


(d) HuBERT: Speaker labels highlighted.

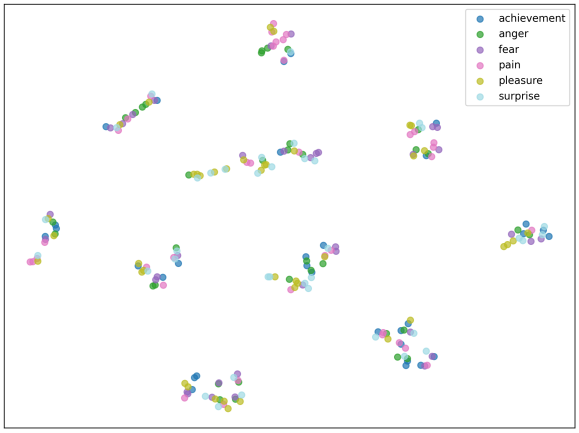
Figure 6: TSNE plots for emotion classification on VIVAE.



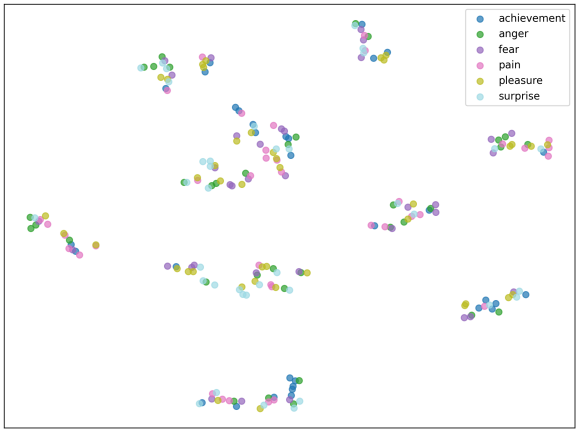
(a) N-CORE: Speaker labels highlighted.



(b) HuBERT: Speaker labels highlighted.

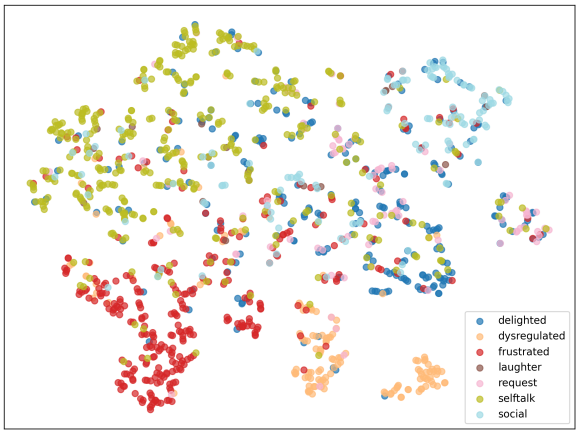


(c) N-CORE: Emotion labels highlighted.

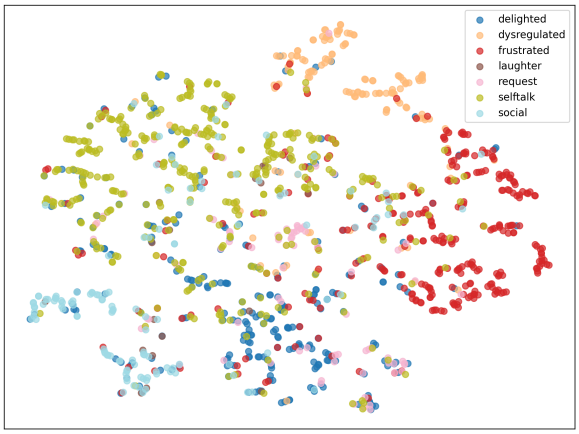


(d) HuBERT: Emotion labels highlighted.

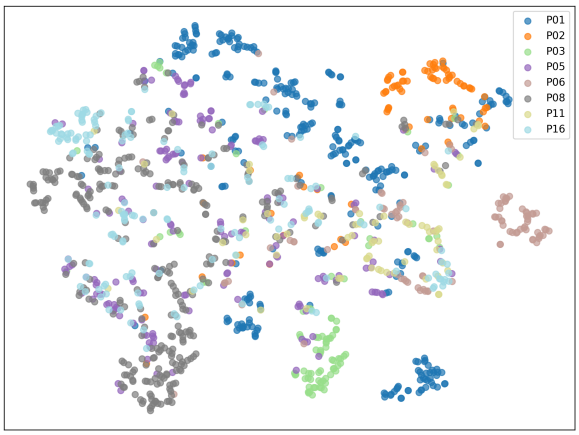
Figure 7: TSNE plots for speaker classification on VIVAE.



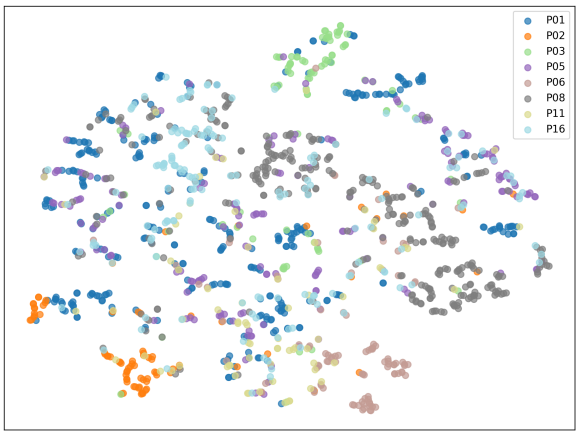
(a) N-CORE: Emotion labels highlighted.



(b) HuBERT: Emotion labels highlighted.

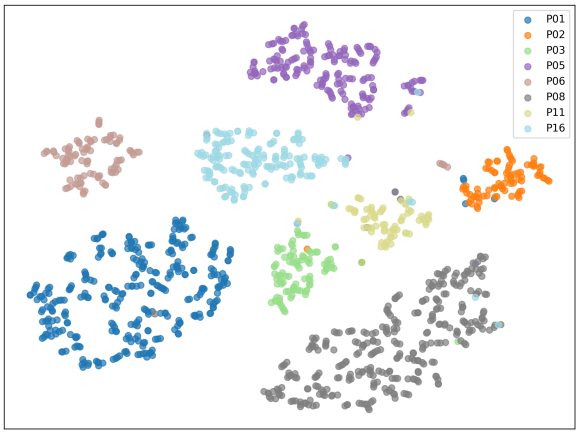


(c) N-CORE: Speaker labels highlighted.

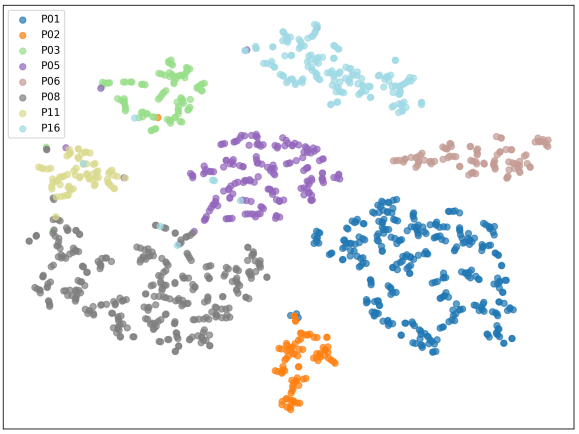


(d) HuBERT: Speaker labels highlighted.

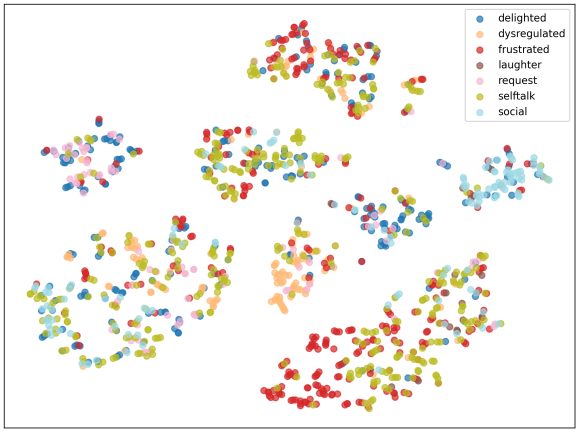
Figure 8: TSNE plots for emotion classification on ReCANVo.



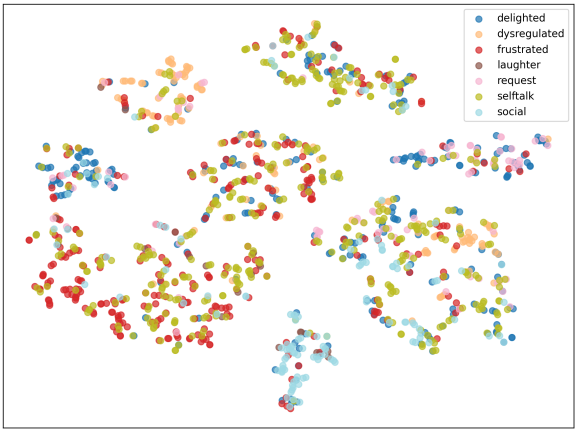
(a) N-CORE: Speaker labels highlighted.



(b) HuBERT: Speaker labels highlighted.



(c) N-CORE: Emotion labels highlighted.



(d) HuBERT: Emotion labels highlighted.

Figure 9: TSNE plots for speaker classification on ReCANVo.