


EMPIRICAL RESEARCH

Open Access



# Multi-rate modulation encoding via unsupervised learning for audio event detection

Sandeep Reddy Kothinti<sup>1</sup> and Mounya Elhilali<sup>1\*</sup> 

## Abstract

Technologies in healthcare, smart homes, security, ecology, and entertainment all deploy audio event detection (AED) in order to detect sound events in an audio recording. Effective AED techniques rely heavily on supervised or semi-supervised models to capture the wide range of dynamics spanned by sound events in order to achieve temporally precise boundaries and accurate event classification. These methods require extensive collections of labeled or weakly labeled in-domain data, which is costly and labor-intensive. Importantly, these approaches do not fully leverage the inherent variability and range of dynamics across sound events, aspects that can be effectively identified through unsupervised methods. The present work proposes an approach based on multi-rate autoencoders that are pretrained in an unsupervised way to leverage unlabeled audio data and ultimately learn the rich *temporal* dynamics inherent in natural sound events. This approach utilizes parallel autoencoders that achieve decompositions of the modulation spectrum along different bands. In addition, we introduce a rate-selective temporal contrastive loss to align the training objective with event detection metrics. Optimizing the configuration of multi-rate encoders and the temporal contrastive loss leads to notable improvements in domestic sound event detection in the context of the DCASE challenge.

**Keywords** Audio event detection, Multi-rate processing, Temporal contrastive loss, Unsupervised learning, Variational autoencoder

## 1 Introduction

Audio analytics has gained a surge in significance in recent years, especially in analyzing everyday complex audio recordings. Audio event detection (AED), which seeks to identify and temporally locate audio events simultaneously, forms one of the core tasks of audio analytics. AED has implications across a broad spectrum of areas spanning audio content retrieval, healthcare, voice assistants, security monitoring, and audio captioning [1–3]. As the technology gets deployed more broadly, it is

becoming increasingly important to develop robust systems that can tackle the complexity of real-world audio events.

The emergence of deep learning has enabled substantial growth in the AED domain. Machine learning methods are tremendously effective when high-quality manually annotated data is available at scale, and systems have been developed to leverage a variety of training label granularities. Annotations for AED can be *weak*, where clip-level annotations indicate different audio events present in the clip, or *strong*, where each audio event is annotated with its precise temporal boundaries in the clip. Generally, the commonly adopted metrics for AED require precise temporal boundaries during inference, which makes strongly labeled data favored for training supervised models, though weakly labeled data can be

\*Correspondence:

Mounya Elhilali  
mounya@jhu.edu

<sup>1</sup> Laboratory for Computational Auditory Perception, Johns Hopkins University, Baltimore 21218, MD, USA

leveraged in training using semi-supervised techniques. Still, acquiring strongly labeled data at scale is generally impractical due to the time and cost of annotation, which explodes combinatorially with the variety of co-occurring sounds [4, 5]. In addressing this challenge, most AED research has focused on developing robust modeling methods using large-scale unlabeled datasets with limited labeled datasets [6], popularly known as semi-supervised audio event detection.

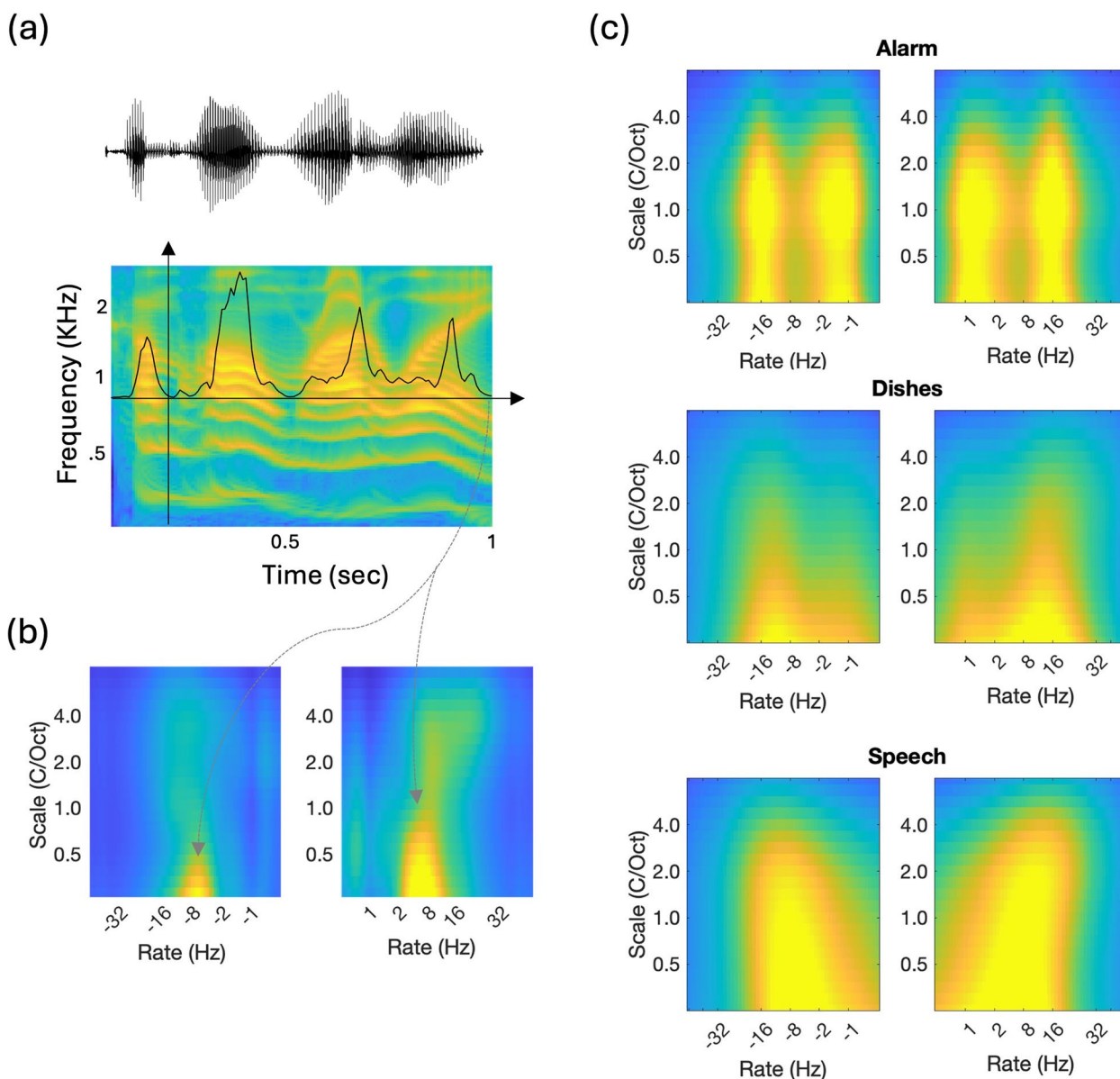
Semi-supervised techniques are well suited for the AED task, given the abundance of multimedia content and audio recordings across settings and environments. A spectrum of semi-supervised methods exists that leverage both limited labeled data with abundant unlabeled data [7]. One such approach in AED models is self-training, where iteratively more accurate models are trained by assigning pseudo labels to the unlabeled data [8, 9]. Conversely, the mean-teacher method [10], which has been widely adopted in the AED domain, combines pseudo-labeling with model averaging to train models without explicitly applying pseudo-labels. Despite notable successes, self-training methods require in-domain or compatible datasets to ensure alignment between the training and test data categories, signal distributions, and characteristics. These limitations often lead to biases in the models and lack of generalizability [11].

Another widely used semi-supervised method is transfer learning, where the parameters of a *pretrained* model are retrained for a target task. When large-scale labeled data exists for a different but related task, the pretraining can be supervised, leveraging task similarities. For example, the PANN models [12] trained on audio tagging often benefit AED [13]. When no such large-scale labeled data is available, the pretraining task is unsupervised or self-supervised, where the goal is to learn representations that would be useful in the target task. In the AED domain, generative [14] and discriminative [15] unsupervised learning methods are employed to pretrain on unlabeled data. Recently, self-supervised methods [16] and contrastive methods with data augmentation [17] are shown to help when finetuned on AED datasets. These methods inject prior knowledge through unsupervised training that can benefit the downstream task.

A defining characteristic of audio events is that the dynamics of natural audio events span a wide range of modulations [18]. Figure 1a depicts a time waveform of a speech utterance along with a time-frequency spectrogram. A cross section across a frequency channel highlights the temporal dynamics of syllabic rate of speech revealing a clear 4 Hz pattern with 4 clear peaks over the 1-s signal. Figure 1b reflects the spectrotemporal modulation derived from this specific utterance. The  $x$ -axis reflects temporal modulations across the entire signal,

while the  $y$ -axis highlights the range of spectral modulations across frequency channels. Since the representation captures changes along both time and frequency, positive and negative modulations are reflected in terms of positive joint spectrotemporal changes (positive rates shown in right panel as well as negative rates shown in the left panel). The figure shows a clear peak near 4 Hz along both positive and negative rates that are commensurate with the 4 Hz syllabic rate highlighted in Fig. 1a. To contrast with the specific speech utterance, Fig. 1c shows spectrotemporal modulations for three distinct classes, part of the DESED dataset [6] averaged across a large number of signals to reflect common modulation patterns within each class. This illustrates how three commonly occurring sounds have broad variations in their signal modulations. The  $x$ -axis emphasizes the changes in temporal dynamics or the rate of an energy change over time for each sound event, while the  $y$ -axis highlights the energy change along the spectral dimension. Notable in the figure is how events like the clinking of dishes unfold over fast dynamics in the 4 Hz–16 Hz range with little energy in slower modulations; meanwhile, speech energy is expected to be concentrated around the typical syllabic rate of speaking around 6–10 Hz [19, 20]. The human auditory system has been shown to leverage these dynamics to facilitate the parsing of complex scenes into individual sound sources and events [21]. This decomposition relies on rate-selective mappings that project the one-dimensional signal along parallel representations that offer a multi-resolution analysis that is both flexible and robust in dealing with the complex dynamics of behaviorally relevant sounds [22]. Leveraging these rate-selective decompositions can lead to more refined tracking of rates at which the event of interest lies [23]. Signal pre-processing and normalization have also benefited from multi-rate decompositions with notable benefits for AED in urban audio settings using techniques like multi-rate per-channel energy normalization (PCEN) [24].

In the present study, we embrace the diverse dynamics of sound events in everyday environments and explore learning unsupervised mappings of audio signals constrained by these dynamics. Previous works on multi-rate modeling have utilized structural constraints to enforce dynamics. For example, Chakrabarty et al. [25] used a set of conditional restricted Boltzmann machines with varying conditioning lengths to fix the dynamics, which were beneficial in tracking feature spaces to detect events in AED tasks [26]. Alternatively, modeling the dynamics of the latent space using stochastic variables has been explored in recent years, such as variational recurrent neural networks (VRNNs) [27] and Kalman VAE [28] models. These approaches impose a temporal dependency across the latent representations using a variational



**Fig. 1** **a** Waveform and time-frequency spectrogram of example speech signal. A cross section at  $F = 850$  Hz shows the changes in spectral energy near that frequency band and highlights the syllabic rate of the speech. **b** Spectrotemporal modulation profile of the speech utterance highlights temporal peaks near 4 Hz, which are commensurate with the peaks observed in the cross-section. **c** Spectrotemporal modulation profiles for three different audio classes from the DESED dataset. The x-axis represents temporal modulations that reflect how fast sound dynamics unfold over time (in units of Hz). The y-axis represents spectral modulations that indicate the spectral spread of energy of the frequency profile of the sound event (in units of cycles/octave). Note that the speech profile shown in **c** reflects an average over a large number of speech utterances, while the profile shown in **b** is an example derived from one signal

approximation. In the present work, we follow the variational formulation used by these dynamical models to constrain the latent space. Importantly, we map the signal onto parameterized modulation decompositions where different bisections of the signal dynamics are tracked separately in the latent space by focusing the pretraining on modulation tracking of audio signals. The pretraining

is performed in a completely unsupervised fashion before refining the representation for the AED task.

The latent space of the modulation-constrained encoders from the VAE is subjected to changes commensurate with detecting events over time in the AED training stage. The existing supervised training objectives often consider the latent representations over time

to be conditionally independent and consequently apply an objective that does not truly reflect the objectives of identifying boundaries accurately. In other words, typical loss function based cross-entropy measure treat all time points as equally important or informative. Recently proposed temporal contrastive loss and its variants [29, 30] use objective functions that treat event boundaries and steady event regions differently and impose a temporal dependence on the latent space. These temporal-based objectives emphasize frames at the boundaries of sound events by regularizing consecutive samples of the latent space. Specifically, these functions assume that frames within an event are expected to have more stationary behavior since the event is continuously unfolding, while frames near event boundaries (near onsets or offsets) should reveal bigger changes in feature representation owing to transition near an onset or offset of the sound event. Building on this principle, the current work extends the concept of temporal loss to combine two principles: (i) function differently near event boundaries versus within event; and (ii) operate over time scales commensurate with the constrained dynamics of the embeddings. This later principle allows the changes of model embeddings over time to operate at the same timescale as the dynamics of the multirate encoders themselves.

Overall, the objective of this paper is to demonstrate how multi-rate dynamic structure naturally occurring in audio events can be leveraged for audio event detection problems. Section 2 details the proposed multi-rate latent space models. Details of the model training and their usage in the downstream task are provided in Section 3. Results on the downstream task and some breakdown of the model components are discussed in Section 4, followed by our summary of the observation and future directions in Section 5.

## 2 Methods

The proposed method is structured around a two-stage training process. In the first stage, a set of variational autoencoders (VAE) called modulation VAEs (ModVAE) are trained with modulation constraints on the latent space. In the second stage, the encoders of the ModVAE autoencoders are refined for event detection using a controlled training procedure commensurate with the modulation constraints imposed on the initial mapping. These procedures are described in detail below.

### 2.1 Modulation variational autoencoders

The proposed ModVAE uses an encoder-decoder architecture as shown in Fig. 2a. The main objective of the training ModVAE is to enforce a low-pass modulation rate on half of the latent embeddings and a high-pass

modulation rate on the remaining embeddings. Using a spectrogram representation of the audio signal  $\mathbf{X} \in \mathbb{R}^{T,F}$  as input, the encoder  $\Phi$  produces two approximate posterior distributions  $q(\mathbf{Z}_{low}|\mathbf{X})$  and  $q(\mathbf{Z}_{high}|\mathbf{X})$ . The approximate posteriors are parameterized by means  $\boldsymbol{\mu}_{low}, \boldsymbol{\mu}_{high} \in \mathbb{R}^{T',D}$  and variances  $\sigma_{low}^2, \sigma_{high}^2 \in \mathbb{R}^{T',D}$ . Here,  $T' \leq T$  is the number of samples after any down-sampling in the encoder, and  $D$  is the size of the latent embedding. The latent space essentially consists of two parts, as shown in Fig. 2a. The means and variances are used with the reparameterization trick [31] to generate two separate stochastic latent embeddings  $\mathbf{Z}_{low}$  and  $\mathbf{Z}_{high} \in \mathbb{R}^{T',D}$ .

$$\mathbf{Z}_{low} = \boldsymbol{\mu}_{low} + \boldsymbol{\sigma}_{low} \odot \boldsymbol{\epsilon}_1, \quad \boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \mathbf{I}) \quad (1)$$

$$\mathbf{Z}_{high} = \boldsymbol{\mu}_{high} + \boldsymbol{\sigma}_{high} \odot \boldsymbol{\epsilon}_2, \quad \boldsymbol{\epsilon}_2 \sim N(\mathbf{0}, \mathbf{I}) \quad (2)$$

The latent space of a VAE is constrained with a prior distribution, typically an isotropic Gaussian distribution with zero mean and unit variance, which does not impose any explicit dynamics on the latent. In this work, by adjusting the prior mean to be within the band of modulation (low-pass or high-pass), we enforce a modulation rate on the latent space. To generate the prior, we use a second encoder model  $\Psi$ , which consists of a slow exponential moving average of the encoder  $\Phi$ , to generate two latent vectors using the same spectrogram input passed to the encoder  $\Phi$ . This slowly moving average is observed to empirically stabilize the convergence of the model relative to instantaneous embeddings. Two infinite impulse response (IIR) filters with low-pass and high-pass characteristics with the same cutoff frequency  $f_c$  are applied on  $\boldsymbol{\mu}'_{low}, \boldsymbol{\mu}'_{high}$  to generate prior means  $\hat{\boldsymbol{\mu}}_{low}$  and  $\hat{\boldsymbol{\mu}}_{high}$  respectively. The prior distributions are given below.

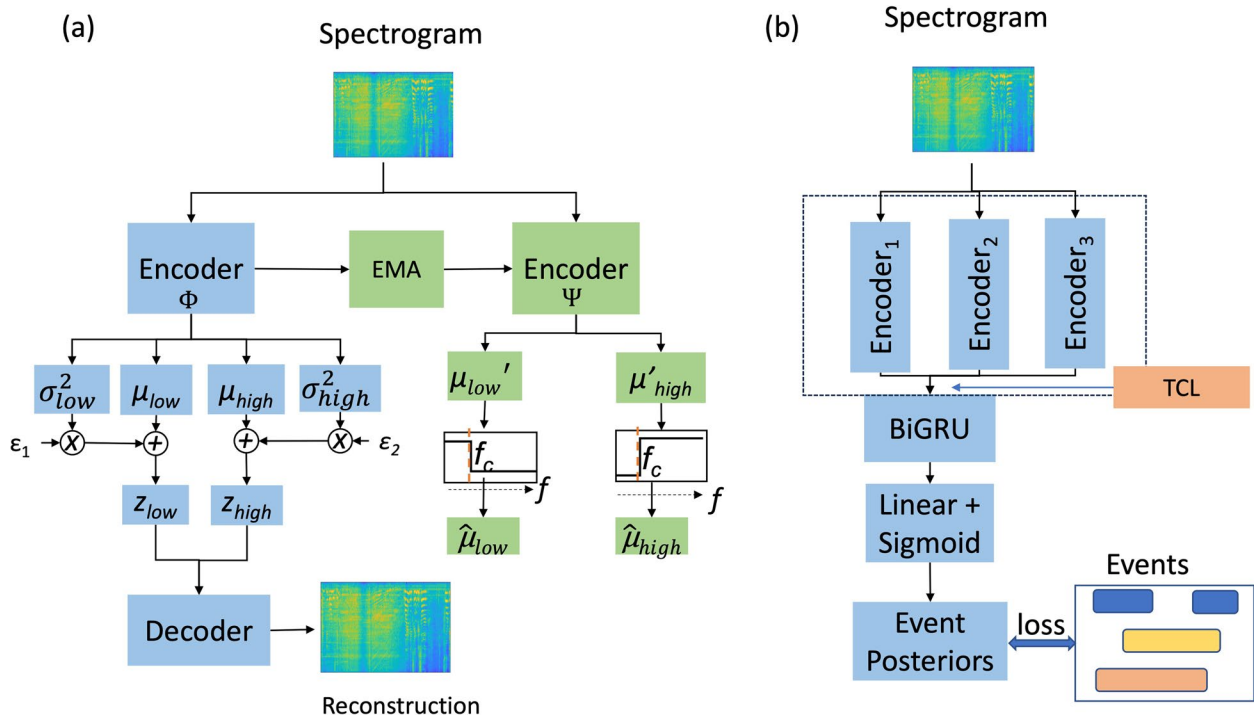
$$p_{\mathbf{Z}_{low}} = N(\hat{\boldsymbol{\mu}}_{low}, \mathbf{I}) \quad (3)$$

$$p_{\mathbf{Z}_{high}} = N(\hat{\boldsymbol{\mu}}_{high}, \mathbf{I}) \quad (4)$$

The decoder reconstructs the spectrogram  $\mathbf{X}'$  by processing concatenated latent embeddings,  $\mathbf{Z}_{low}, \mathbf{Z}_{high}$  from their respective approximate posterior distributions. The encoder-decoder combination is trained by minimizing the loss function, which is the negative of the empirical lower bound on the likelihood, given by

$$\mathcal{L}(X) = \|\mathbf{X} - \mathbf{X}'\|_F^2 + D_{KL}(q_{\mathbf{Z}_{low}} || p_{\mathbf{Z}_{low}}) + D_{KL}(q_{\mathbf{Z}_{high}} || p_{\mathbf{Z}_{high}}) \quad (5)$$

Here,  $D_{KL}$  represents KL-divergence. The loss function,  $\mathcal{L}$ , balances the reconstruction of the input



**Fig. 2** Block diagram detailing the proposed method. **a** ModVAE pipeline. Parameters and models indicated with blue boxes are trainable via backpropagation. Models in green boxes are computed as an exponential moving average of the corresponding learned parameters (blue boxes). **b** AED system pipeline. Outputs of 3 different encoders are concatenated and passed through a BiGRU and a fully connected layer to produce event posteriors

spectrogram and the dynamic constraints on the embeddings. By applying backpropagation on the loss function, the encoder  $\Phi$  is updated with a minibatch gradient descent training. After updating with each batch, the second encoder model  $\Psi$  is updated as

$$\Psi \leftarrow \gamma \Psi + (1 - \gamma) \Phi \quad (6)$$

## 2.2 AED with ModVAE encoders

The encoders from the ModVAEs,  $\Psi$ , are finetuned using the DESED database [6] by adopting the baseline from the DCASE2023 challenge task 4a. The baseline uses a teacher-student framework to train the AED system using a mix of strongly labeled (events with time stamps), weakly labeled (event tags at segment level), and unlabeled data (no event labels), with a convolutional RNN (CRNN) model as backbone model. The semi-supervised training objective consists of a frame-level binary cross entropy (BCE) for strongly labeled data, clip-level BCE for weakly labeled, and a mean-squared error (MSE) between the teacher and student predictions for the whole data.

$$\mathcal{L}_{ssup} = \mathcal{L}_{strong}^{BCE} + \mathcal{L}_{weak}^{BCE} + \beta L_{total}^{MSE} \quad (7)$$

Here,  $\beta$  is the scaling factor to balance the supervised and self-supervised loss functions. In this work, we modify the model architecture used in the baseline system to replace CRNN with multiple CRNN encoders working in tandem that are trained as part of ModVAE training. The model generates event posteriors over time by concatenating the CRNN encoder outputs and passing them through a bidirectional gated recurrent unit (BiGRU) and a fully connected layer (Fig. 2b). The parameters of all the components are trained using the same training procedure as the DCASE baseline system.

## 2.3 Temporal contrastive loss

We adopt the temporal contrastive loss (TCL) defined in Kothinti et al. [29] as a regularization along with the semi-supervised loss used in the DCASE baseline during training. The original premise of the temporal contrastive loss concept is to impose temporal smoothness on the embedding space during events while enhancing the edges, forcing all the dimensions of the latent space to change coherently. In the present work, we build on this concept in conjunction with the controlled rates from the ModVAE and introduce a lag term in the contrastive loss, which is selectively chosen based on the operating rates

of the specific ModVAE encoder. In doing so, the latent representations are regulated at a rate of change that is commensurate with the variational priors imposed in the encoder. In addition, we modified the existing definition of temporal contrastive loss to suit the soft mix-up data augmentation method used in the baseline training approach.

Let  $\mathbf{h}_{i,t}$  be an intermediate representation of an event detection model at time  $t$  for an input  $\mathbf{X}_i$  and  $\mathbf{y}_{i,t}$  be a row vector of the label matrix  $\mathbf{Y}_i$  at time  $t$ . The proposed temporal contrastive loss function is defined as:

$$\mathbf{y}_{diff} = \|\mathbf{y}_{i,t} - \mathbf{y}_{i,t-\tau}\|_1 \quad (8)$$

$$\mathbf{z}_{diff} = \|\mathbf{z}_{i,t} - \mathbf{z}_{i,t-\tau}\|_1 \quad (9)$$

$$\begin{aligned} \mathcal{L}_{TCL,\tau} = & -\alpha_1 \sum_{i=1}^N \sum_{t=\tau+1}^{T'} \mathbf{y}_{diff} \mathbf{1}_{>0}(\mathbf{y}_{diff}) \mathbf{z}_{diff} \\ & + \alpha_2 \sum_{i=1}^N \sum_{t=\tau+1}^{T'} \mathbf{1}_{=0}(\mathbf{y}_{diff}) \mathbf{z}_{diff} \end{aligned} \quad (10)$$

Compared to the previous definitions of TCL, the two modifications to note here: (i) in the boundary condition, the loss objective is scaled in proportion to the change in the labels provided by the mix-up procedure, and (ii) the difference operations of the label and the latent space have a delay factor  $\tau$ , which will allow a rate-selective contrastive loss. The delay factor, when applied to the label vector ( $\mathbf{y}_{diff}$ ), smears the boundary location, thereby allowing for a different rate of change controlled by the lag factor  $\tau$ . The TCL objective is applied to the strongly labeled samples from the training dataset and combined with the overall semi-supervised loss.

$$\mathcal{L}_{total} = \mathcal{L}_{ssup} + \mathcal{L}_{TCL,\tau_1}^{enc_1} + \mathcal{L}_{TCL,\tau_2}^{enc_2} + \mathcal{L}_{TCL,\tau_3}^{enc_3} \quad (11)$$

Here  $\mathcal{L}_{TCL,\tau_1}^{enc_1}$  is the TCL objective for the ModVAE encoder  $enc_1$  with TCL lag  $\tau_1$ . The TCL objective in this work is applied to the BiGRU output from each ModVAE encoder as shown in Fig. 2b, and the hyperparameters  $\alpha_1$  and  $\alpha_2$  are adjusted accordingly.

### 3 Experiment setup

#### 3.1 Datasets

The primary dataset for our experiments is the DESED dataset [6], featured in the DCASE challenge task 4. This dataset includes real and synthetic audio clips, each around 10 seconds long. The real clips include 1578 weakly labeled, 14,412 unlabeled in-domain, and 1168 strongly labeled validation segments. The synthetic clips consist of 10,000 strongly labeled training and 2500 validation segments. The DESED dataset encompasses

ten domestic sound classes, abbreviated as Alarm (A), Blender (B), Cat (C), Dishes (Di), Dog (D), Shaver (Sh), Frying (F), Water (W), Speech (S), and Vacuum (V). These events have a variety of duration profiles and can be categorized into long (Blender, Shaver, Frying, Water, Vacuum) and short (Alarm, Cat, Dishes, Dog, Speech) duration events [32], based on the mode of their duration distribution.

For the unsupervised training of ModVAEs, we use balanced segments from the AudioSet [33] in combination with the DESED dataset to augment the training data. AudioSet comprises 10-s audio clips with sound events, including those beyond domestic environments. For finetuning the ModVAE encoders on the AED task, we incorporate the DESED dataset and 3,387 strongly labeled AudioSet segments [34] specific to domestic audio events.

All samples are resampled to 16 kHz with single-channel audio. For all the experiments, 128-dimensional Log-Mel spectrograms are used as the input features, with a window length of 1024 samples and a hop size of 256 samples. The features are normalized using a mean-variance normalization with statistics computed across the dataset. During the training, *mix-up* [35] is applied on weakly and strongly labeled datasets for half of the randomly chosen batches. For the strongly labeled segments, the labels are mixed at frame level, and for weakly labeled, the labels are mixed at clip level.

#### 3.2 Modulation VAE: models and training

The encoder of the ModVAE uses CRNN with seven layers of CNN and one BiGRU layer, followed by two fully connected projection layers. The CNN layers have 2D convolutions followed by average pooling in each layer. Convolution kernels are of size 3 with stride length 1. The pooling layers have a downsampling factor of 2 along the time dimension in the first two layers and along frequency for all seven layers. The BiGRU layer has 128 cells in each direction. The fully connected layers transform the CRNN output into the mean and variance vectors of the approximate posterior of the VAE. The encoder output dimensionality is 200, with 100 embeddings for each band (low-pass and high-pass). We use two 8th-order IIR elliptical filters with 60dB attenuation in stop bands to filter the embeddings. Each embedding dimension is passed through the filters in forward and backward directions to eliminate the delay caused by the filters [36]. Multiple ModVAEs with cutoffs in the range [0.8 Hz, 7.2 Hz] (with 0.8 Hz increments) are trained to span the modulation spectrum.

ModVAEs are trained using the DESED training set (which includes synthetic, weak, and unlabeled data) and the AudioSet balanced train segments. We divide the

total training dataset into a 9:1 split for training and validation. Each ModVAE undergoes 50 training epochs with the Adam optimizer [37], set at a learning rate of 0.0002. We choose a  $\gamma$  value of 0.999 for the moving average, consistent with other similar methodologies [10, 38]. The KL divergence in the ModVAE loss is annealed over the first ten epochs, with linear increments at each epoch. The learning rate is set to halve every ten epochs or if there is no improvement in the validation loss. The ModVAE version with the minimum validation loss is chosen as the final model.

### 3.3 Event detection models

We adopt the DCASE2023 challenge baseline training methodology for the event detection models. The DESED dataset detailed above is used for the training and validation of the AED models. We use the same training conditions as the baseline system except for the training epochs, which are increased to 250 (from 200) since we notice a reduced variance in the performance with more epochs. The model performance on the synthetic validation is used to choose the best model in the training process.

The DCASE baseline system uses a CRNN model with 7 CNN and 2 BiGRU layers and a fully connected classification layer. This baseline model consists of about 1M parameters. Building on the DCASE baseline architecture, the proposed pipeline replaces the CRNN layers with 3 CRNNs combined using a single layer of BiGRU followed by a fully connected layer. This proposed model consists of about 3M parameters by tripling the front embeddings. Each of these CRNNs is initialized with encoders of ModVAEs with different cutoff frequencies. We refer to this model architecture as a 3×CRNN to denote the number of CRNN components in the system. To compare the benefits of the ModVAE initialization, we train two variations on the 3×CRNN structure: (i) the model parameters are initialized randomly, and (ii) the model parameters are initialized with VAE encoders. For the VAE-based initialization, we train three different VAEs with standard Gaussian prior constraints with a similar training procedure as the ModVAEs and use the VAE encoders as initialization. The coherence-based TCL objective is also tested by adding contrastive loss specific to encoder embeddings with a specific lag. For 3 CRNNs, three different lags are applied to test for any additional benefits of tuning the contrastive loss for each ModVAE encoder.

Event detection performance is measured on the strongly labeled validation set of the DESED dataset using threshold-independent polyphonic sound event detection scores (PSDS) [39]. Two scores are computed

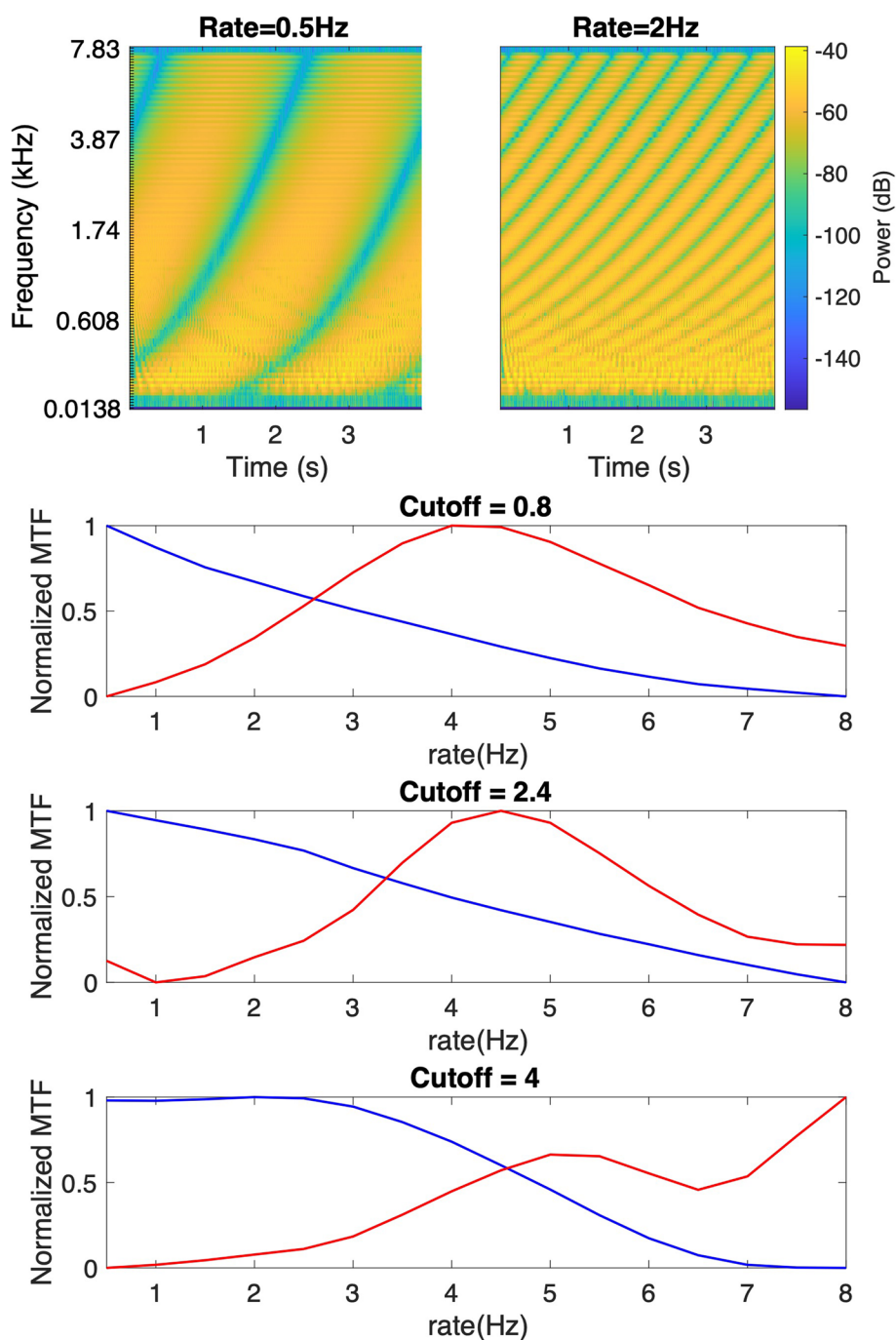
targeting accurate temporal detection (PSDS1) and minimal class confusion (PSDS2).

## 4 Results

ModVAE encoders are tested for rate-selective responses to check the efficacy of the training methodology. A range of ripple waveforms with different spectrotemporal modulations characterized by a rate and scale are generated to confirm the selectivity of the model. Figure 3 shows two such ripple patterns for rates 0.5 Hz and 2 Hz. These inputs are fed through the ModVAE encoders, and the low-pass and high-pass embeddings are analyzed for phase-locked energies for a given rate and scale. By sweeping the rates over the range of [0 Hz, 8 Hz], rate response plots are generated and shown in the bottom three panels of Fig. 3. As seen from these plots, both low-pass and high-pass embeddings of the modulation encoders exhibit different rate-selectivity. With increased cutoff frequency, the frequency at which low-pass and high-pass curves intersect increases.

The AED performance of the ModVAE encoders when retrained on the DESED dataset are shown in Table 1. The PSDS scores of the models are averaged across five training runs. The first row shows the DCASE baseline results in our experiments, which closely matches the DCASE 2023 challenge baseline. For ModVAE initialization, we find the best combination for the three encoders to be 0.8 Hz, 2.4 Hz, and 4 Hz. With initialization from ModVAE encoders, the proposed 3×CRNN model performs significantly better than the baseline model, improving PSDS1 by 3% and PSDS2 by 8% relatively, indicating better event boundaries and event classifications. The 3×CRNN model with random initialization performs better than the DCASE baseline model. This improvement could be from increased computational complexity with almost three times more parameters. Given the same number of parameters, the 3×CRNN model with VAE initialization performs better than a randomly initialized model on PSDS1 and PSDS2 evaluation, showing the benefits of unsupervised pretraining. Based on the incremental gains, we can infer that VAE initialization works better than random initialization, and ModVAE initialization offers the most gains when trained with the same data.

For testing the effectiveness of the temporal contrastive loss, we experiment with lag values of 1, 2, 4, and 8 samples with various combinations for the ModVAE encoders with three different rates. This experiment examines the link between dynamics of autoencoder embeddings which are constrained to specific temporal profiles (i.e., operating at cutoffs of 0.8, 2.4, and 4 Hz), and the constraints of the temporal loss which itself can present lags over segments



**Fig. 3** Rate selectivity of the ModVAE encoders

of time. We report PSDS scores for the best lags (8, 4, 2) for the three encoders with cutoffs (0.8 Hz, 2.4 Hz, 4 Hz). To elucidate the rate-selectivity of the contrastive loss for the three encoders, we also show the performance when the same lags are applied across the encoders. Table 2 presents the PSDS from different configurations. As can be seen from this table, different lags have varying effects on PSDS1

and PSDS2 scores. If we combine PSDS1 and PSDS2 as reported in DCASE annual challenges, we see that the lag values of (8,4,2) best suit the set of rates considered here. This result strengthens the link between the dynamics of the embeddings emerging from the ModVAE encoders and the constraints of the loss function which can operate over different temporal dynamics represented by the lags.



**Table 1** PSDS values for different systems. Average (avg) and standard deviations (std) are computed from 5 iterations of each model

Model	PSDS1	PSDS2
DCASE2023 baseline	0.365 ± 0.010	0.581 ± 0.003
3×CRNN random-init	0.363 ± 0.004	0.594 ± 0.007
3×CRNN VAE-init	0.374 ± 0.003	0.607 ± 0.009
3×CRNN ModVAE-init	<b>0.375 ± 0.006</b>	<b>0.627 ± 0.005</b>

**Table 2** PSDS values for different TCL lags showing mean and standard deviation. The standard deviations are computed from 5 iterations of each model

TCL lags	PSDS1	PSDS2
No TCL	0.375 ± 0.006	0.627 ± 0.005
1,1,1	0.379 ± 0.008	0.641 ± 0.004
2,2,2	0.376 ± 0.004	<b>0.642 ± 0.001</b>
4,4,4	0.379 ± 0.004	0.637 ± 0.007
8,8,8	0.377 ± 0.002	0.636 ± 0.006
8,4,2	<b>0.385 ± 0.001</b>	<b>0.642 ± 0.003</b>

Next, we analyze the performance of components in the proposed model to elucidate their contributions. For this purpose, we initialize the CRNN models with different encoder combinations. First, we examine the performance of a 1×CRNN system initialized with a single ModVAE encoder. As shown in Table 3, performances vary across the different encoders. The encoder with a cutoff at 2.4 Hz performed better than the other

**Table 3** Event detection performance with various components of the proposed model. The standard deviations are computed from 5 iterations of each model

	Condition	PSDS1 avg ± std	PSDS2 avg ± std
1×CRNN	$f_c = 0.8$ Hz	0.361 ± 0.007	0.601 ± 0.007
	$f_c = 2.4$ Hz	0.371 ± 0.006	0.603 ± 0.002
	$f_c = 4$ Hz	0.365 ± 0.005	0.593 ± 0.008
	VAE	0.365 ± 0.003	0.605 ± 0.004
2×CRNN	$f_{c1} = 0.8$ Hz	0.368 ± 0.004	0.612 ± 0.007
	$f_{c2} = 2.4$ Hz		
	$f_{c1} = 0.8$ Hz	0.365 ± 0.003	0.594 ± 0.003
	$f_{c2} = 4$ Hz		
3×CRNN	$f_{c1} = 2.4$ Hz,	0.374 ± 0.009	0.611 ± 0.009
	$f_{c2} = 4$ Hz		
3×CRNN	Low-pass only	0.373 ± 0.007	0.607 ± 0.005
	High-pass only	0.376 ± 0.007	0.613 ± 0.008

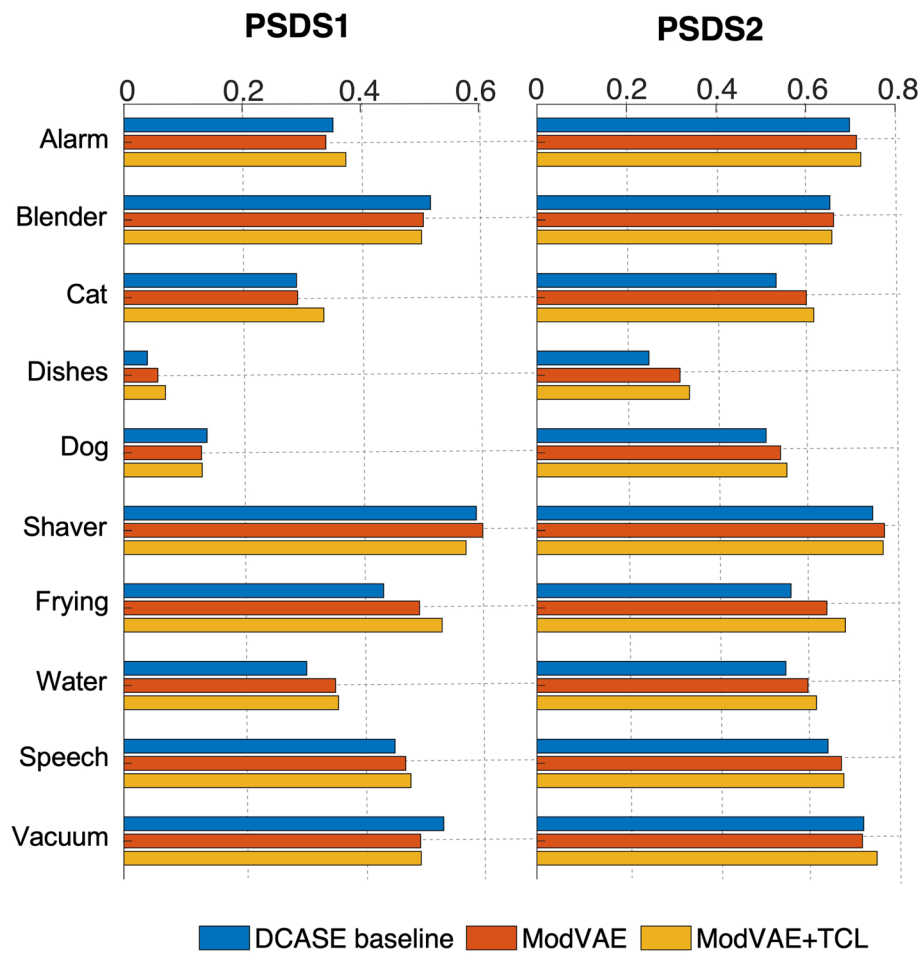
two encoders. Each of the encoders performs better than the DCASE baseline model. For a fair comparison, we train a 1×CRNN system with VAE encoder initialization. The VAE-initialized 1×CRNN performs on par with the single ModVAE encoders. Thus, a single ModVAE encoder may not offer significant advantages over a VAE encoder. The performances of the 1×CRNNs vary across the different classes of events. When grouped according to the duration profiles (Section 3.1) as shown in Table 4, the effect of rates can be observed. For the PSDS1 scores, lower cutoffs perform relatively better on long-duration events, which typically have slower modulations. The PSDS2 scores do not show clear trends as lower cutoffs are, in general, better on both short and long-duration classes.

As a follow-up, we examine 2×CRNN systems initialized with two different ModVAE encoders. In two cases, the 2×CRNN performs better than individual encoders. For example, the combination of ModVAEs with cutoffs 0.8 Hz and 4 Hz performs better than the individual models. While the combination of two ModVAE encoders performed better, there is still a gap between the 3×CRNN model and the best 2×CRNN model, indicating the necessity for one more encoder. As a final set of ablation, we test the 3×CRNN system by initializing with either low-pass or high-pass parts of the ModVAE encoders. In both cases, PSDS1 and PSDS2 drop when compared to the model with both the low-pass and high-pass components. Removing either band seems to affect both PSDS scores. This suggests that both the timing and class-specific information are being shared across both the low and high-pass bands.

To analyze the improvements in PSDS scores, we look at the classwise breakup of the PSDS scores. As shown in Fig. 4, the improvements in PSDS2 are uniformly observed across classes when initialized with ModVAE encoders. These improvements are further enhanced by the TCL objective. Since PSDS2 scores penalize cross triggers more, this improvement can be attributed to

**Table 4** PSDS values 1×CRNN models with different modulation cutoffs, analyzed for short and long sound events. Sound events are grouped based on mode of the duration, where long events are Blender, Shaver, Frying, Water, and Vacuum and short events are Alarm, Cat, Dishes, Dog, and Speech

Cutoff freqs	PSDS1		PSDS2	
	Short	Long	Short	Long
$f_c = 0.8$ Hz	0.250	0.477	0.548	0.656
$f_c = 2.4$ Hz	0.262	0.485	0.544	0.664
$f_c = 4$ Hz	0.262	0.474	0.544	0.644



**Fig. 4** Classwise performance breakup indicated by PSDS scores

the multiple encoders that might help with the interference from other classes. PSDS1 scores show mixed results with the ModVAE initialization. Adding the TCL objective seems to improve PSDS1 scores of ModVAE models for several classes except for Blender and Shaver.

## 5 Discussion

In this work, we proposed a novel unsupervised methodology to train modulation-constrained encoders using a variational framework. The objective of this scheme is to leverage the availability of large-scale audio data to learn representations that reflect the range of temporal dynamics naturally occurring in everyday sound events. By constraining the latent space along different temporal rates, the model helps identify the inherent modulation structure of different audio events and, therefore, improves detection. We test this hypothesis using a combination of encoders with varying modulation selectivity and a temporal contrastive loss function in the audio event

detection framework. The loss function is a generalized temporal contrastive constraint that emphasizes transitional frames near event onsets and offsets. Unlike typical cross-entropy losses typically embraced in the SED frameworks, the proposed temporal loss incorporates temporal sensitivity over time scales that are commensurate with the constraints imposed in the variational encoders.

In addition to the main results due to constrained latent representations (as shown in Table 1), we note that unsupervised pretraining on a large-scale dataset using different initialization methods helps the downstream audio event detection, as shown by both VAE and ModVAE initializations. The additional improvements given by ModVAE point to the benefits of the proposed methodology in addition to the unsupervised training. Since the pretraining stage is purely unsupervised, the proposed method can be extended to utilize more unlabeled data. Ablation results in Table 3 further support the role of the individual components in the proposed framework.

The steady increments from adding more encoders with different rate-selectivity indicate the complementary information each encoder can provide. Table 4 paints an interesting picture of the value each encoder adds, especially when the events have vastly different dynamics.

By highlighting the importance of carefully constraining the dynamics of the latent space, the present work allows us to further push forth the concept of temporal contrastive loss (TCL), whereby consecutive samples are considered differently. In general, cross-entropy loss that has been adopted in audio event detection frameworks attempts to minimize classification error averaged over all segments and time points. This classic view treats all time points as equally important and ignores the fact that event boundaries near where a sound event starts, or ends are far more information, and instances within events are expected to be stationary or reflect more coherent behavior over time. In contrast, introducing temporal loss has been shown to improve both class identification and boundary detection of sound events [29, 30]. In the present work, we further extend this notion by introducing additional constraints on the temporal coherence. As demonstrated in Table 2, imposing gradually increasing time lags commensurate with the gradually increasing modulation cutoffs results in improved event detection performance for both PSDS1 and PSDS2. This link between the two constraints (ModVAE and TCL) is important to leverage the division of the embedding space along different rates and track changes of latent representations over scales commensurate with those rates. The TCL objective is inspired by the biological phenomena of temporal coherence, which binds neurons that co-evolve together to form a single source from multiple feature dimensions [40–42]. By allowing for feature comparisons across different time scales, the TCL objective enables the adaptation of coherence phenomena at different rates. Notably, the optimal lags for individual encoders align proportionally with the time scale dictated by the ModVAE's prior constraints.

The improvements in the PSDS2 score, emphasizing classification accuracy, remain consistent across all classes. However, the PSDS1 scores, which prioritize boundary detection, exhibit variability across classes, as depicted in Fig. 4. Notably, for longer-duration classes like Shaver, Vacuum, and Blender, the ModVAE model trained with TCL displays lower PSDS1 scores than the baseline, even though its PSDS2 scores are superior. An analysis of the model's event predictions for these classes reveals it often predicts event boundaries at the clip's start and end. This behavior might stem from the scarcity of supervised training samples that feature event boundaries within the clip for these particular classes.

Among the various combinations of cutoff frequencies that are tested for the ModVAEs, we find the cutoffs of 0.8 Hz, 2.4 Hz, and 4 Hz, giving the best performance on the domestic sound classes present in the DESED dataset. These cutoff frequencies are evenly spaced out to occupy the modulation spectrum. While different sound environments might require different combinations of ModVAE encoders, the unsupervised training enables arbitrary encoders to be combined to provide the best performance on the task. Leveraging how human auditory processing projects acoustic information along parallel, multi-rate representations [22] to adapt to tackle dynamic sound environments, the current framework could be extended further and evaluated across broader types of sound classes.

Multi-view representations based on modulation rates have been studied for AED in recent studies [24, 43]. Ick et al. tested multi-rate PCEN by providing parallel views of PCEN spectrograms, where each view is normalized with a different rate as input channels of CNN. They show the training procedure can learn to take advantage of such a multi-rate presentation. Min et al. [43] incorporated biologically feasible spectrotemporal receptive fields as a learnable layer of a multi-layer CNN for the DESED task. Both these works incorporate structurally different processing steps, whereas our proposed method uses functional constraints to achieve a similar effect. From Fig. 3, it can be seen that the encoders loosely follow the priors. It is feasible to incorporate both structural constraints in the model and functional constraints to achieve better rate-selectivity and AED performance, which can be explored in future works.

Overall, the design principles used in this work build on well-known auditory processing processes to provide an unsupervised framework for audio representation geared toward audio event detection. First, the idea of multi-rate processing maps the signal along parallel, somewhat redundant, yet distributed representations of sound dynamics that highlight how sound events evolve differently over time. Second, the principle of temporal coherence further builds on this modulation representation to constrain the learning dynamics over time. Naturally, avenues being considered for future work include augmenting these principles with additional learning constraints, such as attentional mechanisms, which give rise to improved embedding representations that can further benefit the task of audio event detection.

#### Abbreviations

AED	Audio event detection
VAE	Variational autoencoder
ModVAE	Modulation VAE
PSDS	Polyphonic sound event detection score
IIR	Infinite impulse response
CNN	Convolutional neural network

RNN	Recurrent neural network
CRNN	Convolutional RNN
GRU	Gated recurrent neural network
BiGRU	Bidirectional neural network
BCE	Binary cross-entropy
MSE	Mean squared error
TCL	Temporal contrastive loss

#### Authors' contributions

Both the authors contributed to the research and the paper.

#### Funding

This work was supported by funding from ONR N00014-23-1-2050 and N00014-23-1-2086.

#### Availability of data and materials

The datasets used are open-source and can be found online following the citation. The code base will be made available on request to the corresponding author.

#### Declarations

#### Competing interests

The authors declare that they have no competing interests.

Received: 15 October 2023 Accepted: 14 March 2024

Published online: 01 April 2024

#### References

1. Y. Zigel, D. Litvak, I. Gannot, A method for automatic fall detection of elderly people using floor vibrations and sound - Proof of concept on human mimicking doll falls. *IEEE Trans. Biomed. Eng.* **56**(12), 2858–2867 (2009). <https://doi.org/10.1109/TBME.2009.2030171>
2. Q. Jin, P.F. Schulam, S. Rawat, S. Burger, D. Ding, F. Metzger, Event-based video retrieval using audio, in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. (International Speech Communication Association, Lyon, 2012). <https://doi.org/10.21437/Interspeech.2012-556>
3. A.O. Eren, M. Sert, in *2020 IEEE International Symposium on Multimedia (ISM)*. Audio captioning based on combined audio and semantic embeddings. (2020), pp. 41–48. <https://doi.org/10.1109/ISM.2020.00014>
4. H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, I. McLoughlin, A. Mertins, *25th European Signal Processing Conference, EUSIPCO 2017 2017-January*. What makes audio event detection harder than classification? (2017), pp. 2739–2743. <https://doi.org/10.23919/EUSIPCO.2017.8081709>
5. H. Phan, T.N.T. Nguyen, P. Koch, A. Mertins, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Polyphonic audio event detection: Multi-label or multi-class multi-task classification problem? (IEEE, 2022), pp. 8877–8881. <https://doi.org/10.1109/ICASSP43922.2022.9746402>
6. N. Turpault, R. Serizel, J. Salamon, A.P. Shah, in *DCASE Workshop*. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. (2019), pp. 253–257. <https://doi.org/10.33682/006b-jx26>
7. J.E. van Engelen, H.H. Hoos, A survey on semi-supervised learning. *Mach. Learn.* **109**(2), 373–440 (2020). <https://doi.org/10.1007/s10994-019-05855-6>
8. B. Shi, M. Sun, C.c. Kao, V. Rozgic, S. Matsoukas, C. Wang, Semi-supervised acoustic event detection based on tri-training. *IEEE Int. Conf. Acoust. Speech Signal Process.* 750–754 (2019). <https://doi.org/10.1109/ICASSP.2019.8683710>
9. S. Park, A. Bellur, D.K. Han, M. Elhilali, in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Self-Training for Sound Event Detection in Audio Mixtures. (2021), pp. 341–345. <https://doi.org/10.1109/ICASSP39728.2021.9414450>
10. A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process. Syst.* **2017-Decem**(Nips), 1196–1205 (2017). <http://arxiv.org/abs/1703.01780>. Accessed 26 Mar 2024
11. E. Arazo, D. Ortego, P. Albert, N.E. O'Connor, K. McGuinness, in *Proceedings of the International Joint Conference on Neural Networks*. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. (2020). <https://doi.org/10.1109/IJCNN48605.2020.9207304>
12. Q. Kong, Y. Xu, W. Wang, M.D. Plumbley, Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **28**, 2450–2460 (2020). <https://doi.org/10.1109/TASLP.2020.3014737>
13. L. Xu, L. Wang, S. Bi, H. Liu, J. Wang, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Semi-supervised sound event detection with pre-trained model. (2023), pp. 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10095687>
14. M. Meyer, J. Beutel, L. Thiele, in *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*. Unsupervised feature learning for audio analysis. (2017). <https://arxiv.org/abs/1712.03835v1>. Accessed 26 Mar 2024
15. S. Hornauer, K. Li, S.X. Yu, S. Ghaffarzadegan, L. Ren, Unsupervised discriminative learning of sounds for audio event classification. *ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc.* **2021-June**, 3035–3039 (2021). <https://doi.org/10.1109/ICASSP39728.2021.9413482>
16. S. Deshmukh, B. Raj, R. Singh, Improving weakly supervised sound event detection with self-supervised auxiliary tasks. *Proc. Ann. Conf. Int. Speech Commun. Assoc. INTERSPEECH* **1**, 36–40 (2021). <http://arxiv.org/abs/2106.06858>. Accessed 26 Mar 2024
17. E. Fonseca, D. Ortego, K. McGuinness, N.E. O'Connor, X. Serra, Unsupervised contrastive learning of sound event representations. *ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc.* **2021-June**, 371–375 (2020). <https://doi.org/10.1109/ICASSP39728.2021.9415009>
18. N. Singh, F. Theunissen, Modulation spectra of natural sounds and ethological theories of auditory processing. *J. Acoust. Soc. Am.* **106**, 3394–3411 (2003)
19. T.M. Elliott, F.E. Theunissen, The modulation transfer function for speech intelligibility. *PLoS Comput. Biol.* **5**(3), e1000302 (2009)
20. M. Elhilali, in *Timbre: Acoustics, Perception, and Cognition*, ed. by K. Siedenburg, S. Charalampos, S. McAdams, chap. 12 (Springer, 2019), pp. 335–359. [https://doi.org/10.1007/978-3-030-14832-4\\_12](https://doi.org/10.1007/978-3-030-14832-4_12)
21. T. Chi, Y. Gao, M.C. Guyton, P. Ru, S. Shamma, Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* **106**(5), 2719–2732 (1999)
22. R. Santoro, M. Moerel, F. De Martino, G. Valente, K. Ugurbil, E. Yacoub, E. Formisano, Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci. USA* **114**(18), 4799–4804 (2017). <https://doi.org/10.1073/pnas.1617622114>
23. A. Bellur, M. Elhilali, Audio object classification using distributed beliefs and attention. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 729–739 (2020). <https://doi.org/10.1109/TASLP.2020.2966867>
24. C. Ick, B. McFee, Sound event detection in urban audio with single and multi-rate PCEN. *ICASSP IEEE Int. Conf. Acoust. Speech Signal Process. Proc.* **2021-June**, 880–884 (2021). <https://doi.org/10.1109/ICASSP39728.2021.9414697>
25. D. Chakrabarty, M. Elhilali, A Gestalt inference model for auditory scene segregation. *PLoS Comput. Biol.* **15**(1), e1006711 (2019). <https://doi.org/10.1371/journal.pcbi.1006711>
26. S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe, M. Elhilali, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Joint acoustic and class inference for weakly supervised sound event detection. (2019), pp. 36–40. <https://doi.org/10.1109/ICASSP2019.8682772>
27. J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, Y. Bengio, A recurrent latent variable model for sequential data. *Adv. Neural Inf. Process. Syst.* **2015-January**, 2980–2988 (2015). <https://arxiv.org/abs/1506.02216v6>. Accessed 26 Mar 2024
28. M. Fraccaro, S. Kamronn, U. Paquet, O. Winther, *A Disentangled Recognition and Nonlinear Dynamics Model for Unsupervised Learning*, in *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates Inc., Red Hook, 2017)
29. S. Kothinti, M. Elhilali, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Temporal contrastive-loss

- for audio event detection. (IEEE, 2022), pp. 326–330. <https://doi.org/10.1109/ICASSP43922.2022.9747468>
30. S. Park, S. Kothinti, M. Elhilali, Temporal coding with magnitude-phase regularization for sound event detection. *Proc. Ann. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2022-September*, 1536–1540 (2022). [https://doi.org/10.21437/INTER\\_SPEECH.2022-950](https://doi.org/10.21437/INTER_SPEECH.2022-950)
  31. D.P. Kingma, M. Welling, in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. Auto-encoding variational Bayes. (International Conference on Learning Representations, ICLR, 2013). <http://arxiv.org/abs/1312.6114>. Accessed 26 Mar 2024
  32. R. Serizel, N. Turpault, H. Eghbal-Zadeh, A.P. Shah, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Large-scale weakly labeled semi-supervised sound event detection in domestic environments. (2018), pp. 19–23. <https://hal.inria.fr/hal-01850270>. Accessed 26 Mar 2024
  33. J.F. Gemmeke, D.P.W. Ellis, F. Freedman, A. Jansen, W. Lawrence, C. Moore, M. Plakal, M. Ritter, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, in *Proceedings of ICASSP. Audio Set: An ontology and human-labeled dataset for audio events*. (IEEE, 2017), pp. 776–780. <https://doi.org/10.1109/ICASSP.2017.7952261>
  34. S. Hershey, D.P. Ellis, E. Fonseca, A. Jansen, C. Liu, R.C. Moore, M. Plakal, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. The benefit of temporally-strong labels in audio event classification, vol. 2021-June (Institute of Electrical and Electronics Engineers Inc., 2021), pp. 366–370. <https://doi.org/10.48550/arxiv.2105.07031>
  35. H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, in *ICLR, International Conference on Learning Representations - Proceedings*. Mixup: Beyond Empirical Risk Minimization. (2018), pp. 1–8. <https://arxiv.org/abs/1710.09412>. Accessed 26 Mar 2024
  36. F. Gustafsson, Determining the initial states in forward-backward filtering. *IEEE Trans. Signal Process.* **44**(4), 988–992 (1996). <https://doi.org/10.1109/78.492552>
  37. D.P. Kingma, J.L. Ba, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. Adam: A method for stochastic optimization. (International Conference on Learning Representations, ICLR, 2014). <https://arxiv.org/abs/1412.6980v9>. Accessed 26 Mar 2024
  38. J.B. Grill, F. Strub, F. Altché, C. Tallec, P.H. Richemond, E. Buchatskaya, C. Doersch, B.A. Pires, Z.D. Guo, M.G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, in *Advances in Neural Information Processing Systems*. Bootstrap your own latent: A new approach to self-supervised Learning. vol. 2020-December (Neural information processing systems foundation, 2020). <http://arxiv.org/abs/2006.07733>. Accessed 26 Mar 2024
  39. J. Ebberts, R. Haeb-Umbach, R. Serizel, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Threshold independent evaluation of sound event detection scores. vol. 2022-May (Institute of Electrical and Electronics Engineers Inc., 2022), pp. 1021–1025. <https://doi.org/10.1109/ICASSP43922.2022.9747556>
  40. K. Lu, Y. Xu, P. Yin, A.J. Oxenham, J.B. Fritz, S.A. Shamma, Temporal coherence structure rapidly shapes neuronal interactions. *Nat. Commun.* **8**, 13900 (2017). <https://doi.org/10.1038/ncomms13900>
  41. M. Elhilali, L. Ma, C. Micheyl, A.J. Oxenham, S.A. Shamma, Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron* **61**(2), 317–329 (2009). <https://doi.org/10.1016/j.neuron.2008.12.005>
  42. J.A. O'Sullivan, S.A. Shamma, E.C. Lalor, Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. *J. Neurosci.* **35**(18), 7256–7263 (2015). <https://doi.org/10.1523/JNEUROSCI.4973-14.2015>
  43. D. Min, H. Nam, Y.H. Park, in *Detection and Classification of Acoustic Scenes and Events*. Auditory neural response inspired sound event detection based on spectro-temporal receptive field. (2023). <http://arxiv.org/abs/2306.11427>. Accessed 26 Mar 2024

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.