

Clustering of Acoustic Environments with Variational Autoencoders for Hearing Devices

Luan Vinícius Fiorio, *Graduate Student Member, IEEE*, Ivana Nikoloska, *Member, IEEE*,
Wim van Houtum, *Member, IEEE*, and Ronald M. Aarts, *Fellow, IEEE*

Abstract—Particularly in hearing devices, the environmental context is taken into account for audio processing, often through classification. Traditional acoustic environment classification relies on classical algorithms, which are unable to extract meaningful representations of high-dimensionality data, or on supervised learning, being limited by the availability of labels. Knowing that human-imposed labels do not always reflect the true structure of acoustic scenes, we explore the (unsupervised) clustering of acoustic environments using variational autoencoders (VAEs), presenting a structured latent space suitable for the task. We propose a VAE model for categorical latent clustering employing a Gumbel-Softmax reparameterization with a time-context windowing scheme, tailored for real-world hearing device scenarios. Additionally, general adaptations on VAE architectures for audio clustering are also proposed. The approaches are validated through the clustering of spoken digits, a simpler task where labels are meaningful, and urban soundscapes, which recordings present strong overlap in time and frequency. While all variational methods succeeded when clustering spoken digits, only the proposed model achieved effective clustering performance on urban acoustic scenes, given its categorical nature.

Index Terms—Clustering algorithms, acoustic environments, variational autoencoders, hearing devices.

I. INTRODUCTION

AN acoustic environment or soundscape comprise the collection of all acoustic phenomena in a certain space perceived by a listener [1]. Noise from acoustic environments are known to not just affect one’s well-being [2], but also deteriorates one’s ability to understand speech (intelligibility), especially in low signal-to-noise ratios with non-stationary signals [3]. Hearing devices, such as hearing aids, often take the acoustic environment into account for sound processing, changing its characteristics under different listening conditions [4]. Naturally, sound classification algorithms have been developed for recognizing acoustic environments [5] and were extended to machine learning methods [6], [7].

While most neural networks (NNs) for acoustic environment classification in hearing devices are trained in a supervised manner [6], the scarcity of labels has led to the development of

semi-supervised learning algorithms [8], e.g., pseudo-labeling [9], which can generate labels for unknown data based on a few labeled samples [10]. However, labels are human-imposed categorizations that may not reflect the structure and statistical properties of the data [11], and a strong overlap in time and frequency is present for acoustic environment recordings [12]. Evidence of such behavior can also be observed by the unoptimal classification accuracy of acoustic scenes obtained through supervised learning [13]. Therefore, clustering algorithms, independent of labels, are of interest for acoustic environment categorization in hearing devices.

Clustering is the process of grouping data into clusters given similarities and characteristics of features [14]. Well-established methods like k-means and Gaussian mixture models (GMMs) are still widely applied, but unable to capture complex relations as they are limited to local dependencies in data points, and especially struggle with high-dimensionality data [15]. An alternative to classical clustering algorithms lies on deep generative models, which are capable of processing high-dimensionality data and rely on unsupervised learning, being independent of labels and capable of extracting meaningful information over the data distribution.

The two most common generative methods modified for (unsupervised) clustering are variational autoencoders (VAEs) [16] and generative adversarial networks (GANs) [17]. While both perform clustering based on latent variables, VAEs have a more structured latent space [18] which directly affects clustering [16], and often present a smaller-sized model when compared to GANs. Additionally, GAN models are known for a notorious unstable training [19]. Given the hardware-constrained application of hearing devices and the aforementioned characteristics of generative models, we focus on variational autoencoders for the clustering of acoustic environments. VAEs were considered for clustering images and text, usually employing a Gaussian mixture model as their latent space distribution [16], [20], [21], [22], [23]. A generalization was achieved with the use of variational information bottleneck [22], presenting competitive results. Nevertheless, none of these approaches were applied to the clustering of audio signals/acoustic environments.

In this work, we propose a VAE-based clustering model tailored for operation with audio signals¹, having the constraints of a hearing device in mind. Given the discrete nature of clustering, we modify the generative semi-supervised

This work was supported by the Robust AI for Safe (radar) signal processing (RAISE) collaboration framework between Eindhoven University of Technology and NXP Semiconductors, including a Privaat-Publieke Samenwerkingen-toeslag (PPS) supplement from the Dutch Ministry of Economic Affairs and Climate Policy.

L. V. Fiorio, I. Nikoloska, W. van Houtum, and R. M. Aarts are with the Eindhoven University of Technology, Eindhoven, 5612 AP, The Netherlands (e-mails: l.v.fiorio@tue.nl, i.nikoloska@tue.nl, w.j.v.houtum@tue.nl, r.m.aarts@tue.nl). L. V. Fiorio and W. van Houtum are also with NXP Semiconductors, High Tech Campus 60, Eindhoven, 5656 AG, The Netherlands (e-mail: luanvinicius.fiorio@nxp.com, wim.van.houtum@nxp.com).

¹Parts of the content in this manuscript are present in our previous preprints [24], [25]. This article is a combination, extension, and thorough revision of those earlier documents, which were not published in a conference or journal.

model (M2 model) from [26] for unsupervised clustering through the categorical latent space. For reparameterization and efficient training, the model employs a Gumbel-Softmax [27] function, which also allows for further robustness improvement. Furthermore, in view of the limitations of hearing devices, we devise a sliding-window scheme that leverages the Gumbel-Softmax operation for increasing the robustness of the proposed clustering model. Finally, the proposed methods are validated with two tasks: the clustering of spoken digits, representing a scenario where labels are meaningful; and the clustering of urban acoustic environments, where labels are merely an abstraction and the audio content strongly overlaps in time and frequency.

II. VARIATIONAL CLUSTERING

The problem we tackle is the clustering of acoustic environments. More specifically, we consider a dataset $\mathbf{X} = \{\mathbf{x}\}_{i=1}^N$ with N independent and identically distributed samples. We want cluster $\mathbf{X}$ in K different clusters, using a function h with parameters ν , without relying on labels (unsupervised learning). As a constraint, the solution h should be sufficiently lightweight to be applied in a hardware-constrained hearing device. Additionally, it should allow for an imperfect availability of environmental recordings, e.g., cases when part of the audio files may be contaminated with foreground-sound speech and cannot be used for soundscape clustering. In the following, we start from the evidence lower bound of variational inference and expand it towards the clustering of audio signals.

A. Continuous variational Inference

We consider a model with parameters θ and a continuous latent variable $\mathbf{z}$ representing the underlying data distribution. The model's ability to represent data can be measured by its likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, since it tells us how probable it is to obtain $\mathbf{x}$ given parameters θ . If $\mathbf{z}$ was known, we could find the model by directly maximizing $p_\theta(\mathbf{x}|\mathbf{z})$ with respect to θ . As $\mathbf{z}$ is an unobserved (latent) variable, we instead aim to maximize the marginal likelihood, $p_\theta(\mathbf{x})$. However, $p_\theta(\mathbf{x})$ is intractable for the majority of real-world problems given its intractable posterior $p_\theta(\mathbf{z}|\mathbf{x})$ – which relies on knowledge over $\mathbf{z}$. Therefore, we introduce a *variational* distribution with parameters ϕ to approximate the posterior $q_\phi(\mathbf{z}|\mathbf{x}) \approx p_\theta(\mathbf{z}|\mathbf{x})$. With $q_\phi(\mathbf{z}|\mathbf{x})$, we can find the *variational lower bound* $\mathcal{L}$, also called the evidence lower bound (ELBO) [28]:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} \right] \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\theta(\mathbf{z})) \\ &= \mathcal{L}(\theta, \phi). \end{aligned} \quad (1)$$

In (1), $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\theta(\mathbf{x}|\mathbf{z})]$ is the expectation over the distribution $q_\phi(\mathbf{z}|\mathbf{x})$ of $\log p_\theta(\mathbf{x}|\mathbf{z})$, and represents the reconstruction error of $\mathbf{x}$ from $\mathbf{z}$, i.e., how well the latent variables $\mathbf{z}$ explain the data $\mathbf{x}$. The D_{KL} term stands for the Kullback-Leibler (KL) divergence, which quantifies the difference between the prior $p_\theta(\mathbf{z})$ and the variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$.

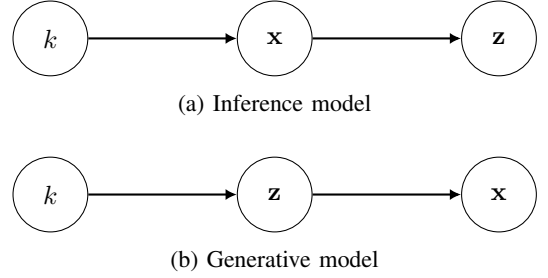


Fig. 1: Inference and generative models for clustering with a continuous latent variable $\mathbf{z}$.

This can be seen as a regularization to make the variational distribution closer to the desired/known prior distribution of the latent variables. The objective of variational inference is to maximize the ELBO (denoted as $\mathcal{L}$) in terms of θ and ϕ . This corresponds to learning a good approximation $q_\phi(\mathbf{z}|\mathbf{x})$ to the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$, while also optimizing the generative model $p_\theta(\mathbf{z}|\mathbf{x})$.

B. Extension to clustering

Given the complexity of the problem, we consider the VAE [28] framework because of its capacity to learn the underlying distribution of data through a structured latent [16]. Such characteristics positively affects clustering performance [18]. The VAE is composed by a neural network encoder, which takes the data $\mathbf{x}$ to a latent representation $\mathbf{z}$, and a NN decoder that generates data based on the latent variables.

To allow clustering behavior, we choose a multivariate Gaussian mixture model (GMM) prior $p_\varphi(\mathbf{z})$, replacing $p_\theta(\mathbf{z})$ in (1), given by

$$p_\varphi(\mathbf{z}) = \sum_k \pi_k \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (2)$$

where k is one component in the mixture, i.e., a cluster. The multivariate GMM defined in (2) is based on multivariate Gaussian $\mathcal{N}$ with means $\boldsymbol{\mu}_k$ and variances $\boldsymbol{\Sigma}_k$, whose dimension is $d_{\boldsymbol{\mu}_k} = d_{\boldsymbol{\Sigma}_k} = d_{\mathbf{z}}$, linearly combined by a weighting vector π_k . The GMM parameters are condensed in a variable $\varphi := \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, for K components.

As shown in Fig. 1, for inference, the data are generated from a component k of an unknown GMM $q_\phi(\mathbf{x}|k)$. In addition, the latent variables have a GMM prior $p_\varphi(\mathbf{z})$. All parameters of the model (θ , ϕ , and φ) are optimized simultaneously to maximize the ELBO, modified from (1) to meet the GMM prior:

$$\begin{aligned} \mathcal{L}(\theta, \phi, \varphi) &= \\ &\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\mathbf{a}_{\mathbf{x}} \log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\varphi(\mathbf{z})). \end{aligned} \quad (3)$$

Notice that we explicitly include the term $\mathbf{a}_{\mathbf{x}}$, which is sound activity mask for $\mathbf{x}$ over time. Specifically for audio processing, $a_{\mathbf{x},t} = 1$ when time step t contains sound activity, and $a_{\mathbf{x},t} = 0$ for a silent time step. This is necessary to avoid the clustering of silence/zero padding patterns. The ELBO in (3), except from the extension of the sound activity mask, is equivalent to the variational information bottleneck with Gaussian mixture model (VIB-GMM) proposed by [22].

Importantly, when using backpropagation, the cost function based on (8) can be written as

$$\min_{\theta, \phi, \varphi} \left(-\frac{1}{N} \sum_{i=1}^N \mathcal{L}^{(i)}(\theta, \phi, \varphi) \right), \quad (4)$$

where $\mathcal{L}^{(i)}$ denotes the ELBO for the i -th data sample, and the negative sign reflects that we minimize the negative ELBO during training.

Nevertheless, in practice, we cannot directly sample $\mathbf{z} \sim p_\phi(\mathbf{z})$ to compute the reconstruction error in (8), since sampling is a non-differentiable operation. To enable gradient-based optimization, we apply the reparameterization trick [28] combined with Monte Carlo sampling [22],

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \approx \frac{1}{M} \sum_{m=1}^M \log p_\theta(\mathbf{x}|\mathbf{z}^{(m)}), \quad (5)$$

where each sample $\mathbf{z}^{(m)}$ is obtained via

$$\mathbf{z}^{(m)} = \boldsymbol{\mu}_\phi + \boldsymbol{\Sigma}_\phi^{1/2} \cdot \boldsymbol{\epsilon}^{(m)}, \quad \boldsymbol{\epsilon}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (6)$$

The KL divergence term in (8), which measures the divergence between a single-component multivariate Gaussian and a Gaussian mixture model (GMM) with C components, does not have a closed-form solution. Nonetheless, it can be approximated under the assumption that both covariance matrices are diagonal,

$$\boldsymbol{\Sigma}_\phi = \text{diag}(\{\Sigma_{\phi,j}\}_{j=1}^{d_z}), \quad \boldsymbol{\Sigma}_k = \text{diag}(\{\Sigma_{k,j}\}_{j=1}^{d_z}), \quad (7)$$

resulting in

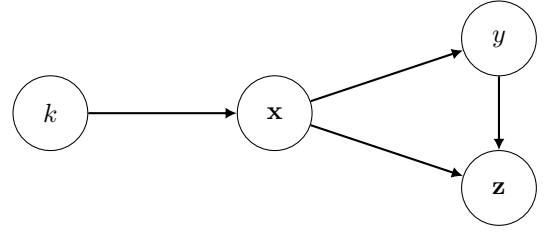
$$\begin{aligned} D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) || p_\varphi(\mathbf{z})) \approx & -\log \sum_c \pi_k \exp \left(-\frac{1}{2} \sum_j \left[\frac{(\mu_{\phi,j} - \mu_{k,j})^2}{\Sigma_{k,j}} \right. \right. \\ & \left. \left. + \log \left(\frac{\Sigma_{\phi,j}}{\Sigma_{k,j}} \right) - 1 + \frac{\Sigma_{\phi,j}}{\Sigma_{k,j}} \right] \right). \end{aligned} \quad (8)$$

Moreover, the cluster assignment k is determined by the component with the highest posterior probability:

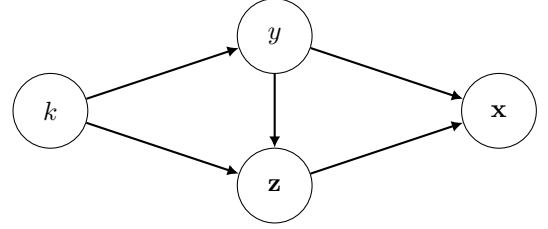
$$q_\phi(k|\mathbf{x}) = p_\phi(k|\mathbf{z}) = \frac{\pi_k \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})}, \quad (9)$$

$$k = \arg \max_k p_\phi(k|\mathbf{z}). \quad (10)$$

As of hearing devices, the data $\mathbf{X}$ in question is composed of audio recordings, which strongly overlap in time and frequency. A continuous latent variable might not be sufficient for enforcing separate and compact clusters. Categorical latent variables, on another hand, enforce the posterior for each data point to be concentrated on one of K bins, which is, naturally, close to the task of clustering. Next, we extend variational clustering to include a categorical latent variable.



(a) Inference model



(b) Generative model

Fig. 2: Inference and generative models for clustering with a categorical (y) and a continuous ($\mathbf{z}$) latent variable.

III. CONTINUOUS-CATEGORICAL VARIATIONAL CLUSTERING

The inclusion of a categorical latent variable in the inference and generative models, as presented in Fig. 2, is similar to the M2 architecture² [26] proposed for semi-supervised learning. Here, data $\mathbf{x}$ is assumed to come from an unknown discrete cluster k . In the continuous-categorical case, each observation $\mathbf{x}$ is considered to be generated by a class k of a categorical latent variable y and a continuous latent variable $\mathbf{z}$. The joint distribution is given by

$$p_\theta(\mathbf{x}, y, \mathbf{z}) = p_\theta(\mathbf{x}|y, \mathbf{z})p_\theta(y)p_\theta(\mathbf{z}). \quad (11)$$

We are interested in the true posterior $p_\theta(y, \mathbf{z}|\mathbf{x})$, as it tells us how likely each latent configuration is given an observed data point. However, given the intractability of $p_\theta(y, \mathbf{z}|\mathbf{x})$, we define a *variational* distribution with parameters v and ϕ ,

$$q_{v,\phi}(y, \mathbf{z}|\mathbf{x}) = q_\phi(\mathbf{z}|\mathbf{x}, y)q_v(y|\mathbf{x}), \quad (12)$$

such that $q_{v,\phi}(y, \mathbf{z}|\mathbf{x}) \approx p_\theta(y, \mathbf{z}|\mathbf{x})$. Therefore, we can find the ELBO for the continuous-categorical case as

$$\begin{aligned} \mathcal{L}(\theta, v, \phi) = & \mathbb{E}_{q_{v,\phi}(y, \mathbf{z}|\mathbf{x})}[\mathbf{a}_\mathbf{x} \log p_\theta(\mathbf{x}|y, \mathbf{z})] \\ & - \lambda(D_{KL}(q_\phi(\mathbf{z}|y, \mathbf{x}) || p_\theta(\mathbf{z})) + D_{KL}(q_v(y|\mathbf{x}) || p_\theta(y))). \end{aligned} \quad (13)$$

In (13), $\mathbb{E}_{q_{v,\phi}(y, \mathbf{z}|\mathbf{x})}[\mathbf{a}_\mathbf{x} \log p_\theta(\mathbf{x}|y, \mathbf{z})]$ is the expectation of $\mathbf{a}_\mathbf{x} \log p_\theta(\mathbf{x}|y, \mathbf{z})$ over $q_{v,\phi}(y, \mathbf{z}|\mathbf{x})$, representing the reconstruction error. $D_{KL}(q_\phi(\mathbf{z}|y, \mathbf{x}) || p_\theta(\mathbf{z}))$ and $D_{KL}(q_v(y|\mathbf{x}) || p_\theta(y))$ are, respectively, the KL divergence terms for the continuous and the categorical encoders, serving as regularization by approximating their output distribution to the chosen priors. As previously done for (3), a sound activity mask of $\mathbf{x}$ over time, $\mathbf{a}_\mathbf{x}$, is added to the reconstruction error, such that the clustering is focused on actual sound

²Notice that Rui Shu hinted, in his blog, that the M2 model could be a natural choice for clustering [29].

parts, not silence or zero padding. Additionally, a non-trainable Lagrangian variable λ is included for smoother training and should be tuned based on data.

For the continuous-categorical case, the clusters can be chosen based on the maximum probability of data point $\mathbf{x}$ belonging to the k th cluster $q_v(k|\mathbf{x}) = p_\theta(y|k)$, obtained as

$$k = \arg \max_n \{y_n\}_{n=1}^{d_y}, \quad (14)$$

where $d_y = K$ is the dimension of y .

While the continuous latent variable $\mathbf{z}$ can be sampled by the reparameterization trick, as mentioned in Section II, the same trick cannot be applied to a categorical distribution. We, thus, consider a Gumbel-Softmax function [27] to sample from a categorical distribution and have a smooth training, also allowing for the application of the time-context windowing described in Section IV. We describe the Gumbel-Softmax distribution in the following.

A. Sampling from a categorical distribution

The foundation of this approach is the Gumbel-Max trick [30], which allows sampling from a categorical distribution using a deterministic transformation combined with noise. Consider a categorical distribution over K classes with probabilities $[\pi_1, \dots, \pi_K]$. The Gumbel-Max trick samples from this distribution by computing

$$y_{\mathbb{1}} = \text{one_hot} \left(\arg \max_k [\log \pi_k + g_k] \right), \quad (15)$$

where each $g_k \sim \text{Gumbel}(0, 1)$ is generated via

$$g_k = -\log(-\log(u_k)), \quad u_k \sim \text{Uniform}(0, 1), \quad (16)$$

transforming the sampling process into a maximization over noisy logits. Nevertheless, the $\arg \max$ operation is not differentiable, posing a problem for gradient-based optimization.

To address it, we replace the $\arg \max$ with a differentiable approximation: the softmax function, which leads to the Gumbel-Softmax (GS) distribution [27]:

$$y = \frac{\exp((\log \pi_k + g_k)/\tau)}{\sum_{j=1}^K \exp((\log \pi_j + g_j)/\tau)}, \quad \text{for } k = 1, \dots, K. \quad (17)$$

Here, τ is a temperature parameter that controls the smoothness of the distribution. As $\tau \rightarrow 0$, the GS approaches a categorical distribution (i.e., becomes more discrete). For higher τ , the distribution becomes smoother and more continuous.

To illustrate, Fig. 3 shows the GS distribution plot for 10 classes, with different values of τ . The softmax temperature τ is specially interesting when the GS distribution is used for clustering, as a smaller τ results in more distinct and dense clusters. Therefore, as a natural choice, we choose to cluster over y with a monotonic reduction of τ over training. Next, we propose a time-context windowing scheme for the continuous-categorical model, such that its operation in inference mode can be applied for real-world hearing devices.

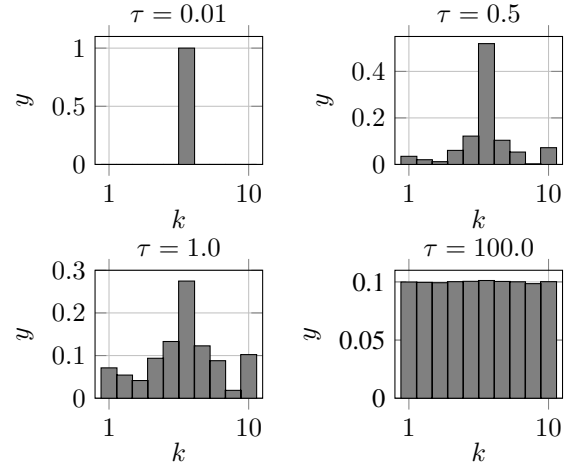


Fig. 3: Gumbel-Softmax distribution example plot for 10 classes with different values of τ .

IV. TIME-CONTEXT WINDOWING

In practical hearing devices, a voice activity detection mask [31] is often considered for detecting traces of speech in the captured audio. To allow for a clear clustering of acoustic environment, we assume a device where a voice-activity mask is used for separating parts of the signal where only background sound is present, as done in [32]. Therefore, for a complete audio recording, only a subset of its length can be used for clustering the soundscape.

We propose a windowing scheme, described in the following, that allows a categorical-based clustering model to cluster the environment without access to a complete background-only audio trace. The sliding window operates in inference mode and consists of sliding through the time dimension of the input, combining the logits π over the whole sequence for a more reliable estimation of cluster probabilities through the GS function. This extension is based on a sliding window approach previously applied in speech enhancement [33] and communications [34].

Given the spectrogram dataset $\mathbf{X}$, each sample $\mathbf{x}$ consists of a sequence of length T (representing the time dimension), with F frequency bins per time step. Our method applies a sliding window of length w across the i -th sample vector $\mathbf{x}^{(i)} = [x(1), x(2), \dots, x(T)]$. For each window position j , we extract a segment

$$\mathbf{x}_j = [\mathbf{x}(j), \mathbf{x}(j+1), \dots, \mathbf{x}(j+w-1)]. \quad (18)$$

This windowed segment is fed into a neural network model h_v , which outputs the corresponding logits $h_v(\mathbf{x}_j) = \pi_j$. All logits from the N_w windows are then averaged as

$$\bar{\pi} = \frac{1}{N_w} \sum_{j=1}^{N_w} \pi_j. \quad (19)$$

Finally, the averaged logits $\bar{\pi}$ are passed through the GS function to produce the output probabilities y . The sliding window approach is exemplified in Fig. 4.

It is worth mentioning that the inference overlap between windows in inference can be chosen based on a trade-off of computational complexity and performance. A smaller

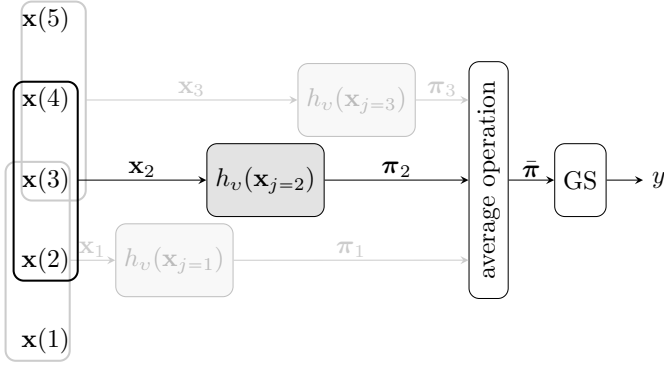


Fig. 4: Example of sliding window clustering inference for $w = 3$, hop of 1 sample, with $N = 5$ samples. $\mathbf{x}(n)$ is the n -th time bin of an input sample $\mathbf{x}^{(i)}$, $\mathbf{x}_j$ is the input window j with estimated logits π_j , $\bar{\pi}$ are the average logits, GS is the Gumbel-Softmax function, and y are the cluster probabilities.

hop results in better performance with the cost of elevated computational complexity. For this work, we utilize a hop of 50% the size of the window, which we consider to be a fair trade-off. On the other hand, the window size must also be taken into account during design stage. The length of the window should be enough to capture relevant changes in sound.

During training, the recurrent layer's final hidden state at window $j - 1$ is carried as initial state at window j , but the gradients of the carried state are detached for training efficiency. For the first window of each data sample, the initial state is initialized as zero. The training losses, calculated with (13), are obtained per-window and combined at the end of each batch of samples, making sure that the contextual relation between windows in a data sample is taken into account. Importantly, no overlap between windows is considered in training to avoid overfitting. Furthermore, training is executed with full knowledge of the acoustic environment recording, taking entire files composed of background noise only. We also consider pure-noise files for evaluation. Notice that, in practice, a voice activity detector is often used to filter-out audio traces where speech dominates the signal [32], being out of scope for this work. Next, we propose variational autoencoder architectures for the context-aware clustering of audio.

V. MODEL ARCHITECTURES

We consider two different variational autoencoder architectures for the application of unsupervised audio clustering. Each architecture is aimed at the optimization of the ELBOs derived in Sections II and III, and is described as follows.

A. VAE with continuous prior

Fig. 5 illustrates the neural network-based variational autoencoder model used for clustering, which assumes a continuous prior over the latent space – in this case, a GMM. The *inference model*, shown in Fig. 1a, is represented by the encoder $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi)$. We implement the encoder as a neural network f with parameters ϕ and input $\mathbf{x}$, producing the multivariate output $f_\phi(\mathbf{x}) = [\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi]$. Similarly,

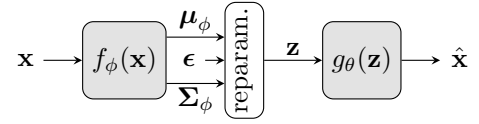


Fig. 5: Diagram of the neural network-based VAE clustering model with a (continuous) Gaussian mixture model prior. $q_\phi(\mathbf{z}|\mathbf{x})$ is represented by a NN $f_\phi(\mathbf{x}) = [\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi]$. The decoder $p_\theta(\mathbf{x}|\mathbf{z})$ is a NN $g_\theta(\mathbf{z}) = [\hat{\mathbf{x}}]$.

the *generative model*, shown in Fig. 1b, is represented by the decoder $p_\theta(\mathbf{x}|\mathbf{z}) = \hat{\mathbf{x}} = g_\theta(\mathbf{z})$. Here, g is a neural network with parameters θ that takes the latent variable $\mathbf{z}$ as input and generates the reconstructed data $\hat{\mathbf{x}}$. For inference mode, however, the generative part of the model is not needed, as we are only interested in clustering in the latent space. Therefore, only the encoder $f_\phi(\mathbf{x})$ is necessary in inference time, with clustering decision taken with (9) and (10).

Moreover, generative models tend to require a massive number of parameters to achieve desirable performance [35]. Thus, we define a reduced-size convolutional-recurrent variational autoencoder composed of a convolutional-recurrent encoder $q_\phi(\mathbf{z}|\mathbf{x})$, a latent space with prior $p_\varphi(\mathbf{z})$, and a (mirrored) convolutional decoder $p_\theta(\mathbf{x}|\mathbf{z})$. We define the VAE with continuous prior as the VIB-GMM model, the same used in [21], [22], and its audio clustering (AC) version with a convolutional-recurrent structure, the VIB-GMM-AC, which detailed schematic can be seen in Fig. 7a.

We set the latent space dimension with $d_z = 10$, which has shown to be sufficient in preliminary tests. Also, we opt for the lowest complexity operation with $M = 1$ in the Monte Carlo sampling from (5). As previously described, the prior distribution over the latent space $p_\varphi(\mathbf{z})$ is modeled as a GMM. In our experiments, we use $K = 10$ components in the mixture, matching the number of classes in the dataset. Furthermore, to adapt the model architecture to audio clustering, linear layers are used in the latent space to map the number of channels to the chosen latent dimension. This is a necessary modification compared to the fully connected VAE. The first linear layer transforms the encoder output channels to $2 \cdot d_z$, where the first d_z elements represent the mean of the latent distribution, and the last d_z the variance. The second linear layer maps the latent representation to the decoder input dimension.

For the GMM prior $p_\phi(\mathbf{z})$, defined in (2), the initialization is as follows: i) the weighting vector π_k is drawn from a uniform distribution with lower and upper bounds of 0.0 and 1.0, respectively; ii) the mean vectors $\boldsymbol{\mu}_k$ are initialized using Xavier uniform initialization; and iii) the covariance matrices $\boldsymbol{\Sigma}_k$ are initialized to zero. Note that the zero initialization of the covariance matrices implies initially deterministic components, which are updated during training. Neural network layers are initialized with the Kaiming uniform initialization for the weights and zero for the biases.

B. VAE with continuous-categorical prior

The neural network-based application of the continuous-categorical clustering model is shown in Fig. 6 and follows the inference and generative models from Fig. 2. For the

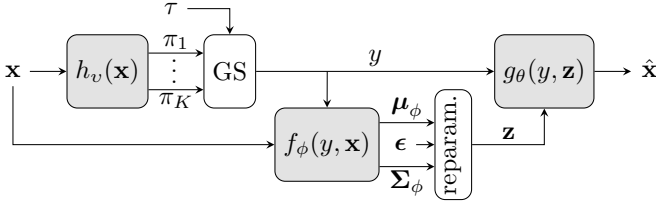


Fig. 6: Diagram of the continuous-categorical neural network-based VAE model for clustering considering both continuous ($\mathbf{z}$) and a categorical (y) latent variable. $q_v(y|\mathbf{x})$ is represented by a NN $h_v(\mathbf{x}) = [\pi_k]$ and τ is the Gumbel-Softmax temperature. $f_\phi(y, \mathbf{x}) = [\mu_\phi, \Sigma_\phi]$ is a NN equivalent to $q_\phi(\mathbf{z}|y, \mathbf{x})$. The decoder $p_\theta(\mathbf{x}|y, \mathbf{z})$ is a NN $g_\theta(y, \mathbf{z}) = [\hat{\mathbf{x}}]$.

discriminative inference encoder $q_v(y|\mathbf{x})$, we use a neural network denoted by h with parameters v , which outputs K (soft) probabilities $h_v(\mathbf{x}) = [\pi_1, \dots, \pi_K]$. The inference encoder $q_\phi(\mathbf{z}|y, \mathbf{x})$ is represented by another NN, denoted by f with parameters ϕ , such that $f_\phi(y, \mathbf{x}) = [\mu_\phi, \Sigma_\phi]$, where μ_ϕ and Σ_ϕ are used to sample $\mathbf{z}$ via the re-parameterization trick [28] $\mathbf{z} = \mu_\phi + \Sigma_\phi^{1/2} \cdot \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. Furthermore, the generative decoder is modeled by a NN g with parameters θ , which reconstructs the data $\mathbf{x}$ as $g_\theta(y, \mathbf{z}) = [\hat{\mathbf{x}}]$. Importantly, during inference time for clustering, only $h_v(\mathbf{x})$ and the GS distribution (17) with hard outputs are required.

The architecture considered for the audio clustering M2 model follows similar characteristics to the VIB-GMM-AC, nonetheless, including one additional encoder and minor changes, which can be seen in Fig. 7b. Both inference networks $h_v = q_v(y|\mathbf{x})$ and $f_\phi(y, \mathbf{x}) = q_\phi(\mathbf{z}|y, \mathbf{x})$ are convolutional-recurrent layers. A linear layer is used in model h to output K values π_k , dimension d_y , which are used as input to the GS distribution. Model f also uses a linear layer for latent dimension conversion, as explained for the VIB-GMM-AC. The outputs of both GS and continuous reparameterization are concatenated and input to the decoder, $g_\theta(y, \mathbf{z}) = p_\theta(\mathbf{x}|y, \mathbf{z})$, which is a mirrored version of the encoder. Hereafter, we refer to this model as M2-AC: M2 model extended for audio clustering. The detailed schematics of the M2-AC model is shown in Fig. 7b. Similarly as before, all layers are initialized using Kaiming uniform initialization for the weights, and all biases are set to zero.

C. Practical Implementation to Audio

We identified several key implementation aspects in the continuous-categorical audio clustering settings that significantly impact performance, supported by ablation studies presented in Appendix A, and described in the following.

- 1) Audio signals exhibit strong time-context dependencies, especially in longer sequences. Incorporating recurrent layers in the encoder architecture substantially improves performance. In contrast, omitting them can lead to unstable training behavior, including loss collapse;
- 2) Both the encoder and decoder need two-dimensional convolutions. We found that only applying convolutions along the frequency dimension – combined with global time pooling – resulted in poor clustering performance.

This highlights the importance of preserving both time and frequency structure;

- 3) When input signals are zero-padded to meet a fixed duration or contain irrelevant silence, a time mask should be applied during reconstruction error computation. This ensures that non-informative regions do not negatively affect the model’s learning. Refer to the time mask $\mathbf{a}_x$ in the ELBO formulations (3) and (13).

While the use of sound activity mask has been previously noted in sequence modeling literature [36], it has not been explicitly addressed in the context of generative audio clustering.

VI. EXPERIMENTS

In this section we describe the experiment configuration for the different datasets, what metrics we use to evaluate clustering performance, the baselines, and results.

A. Configuration

We separate our results in two main parts. In part 1, we apply the clustering methods to a dataset where the labels are not a complete abstraction, but an important feature of the data. For this case, we look into accuracy metrics as well as clustering metrics, as they should be synergetic and a good clustering should also mean a good classification. Still, time-frequency overlap is present to some extent, which should early-saturate the accuracy metrics. On the other hand, part 2 relates to datasets where the labels are merely a human abstraction with low correlation with the actual time-frequency data features. This has been previously observed in literature [11], [37], and is tied to the societal, cultural, and institutional contexts or biases in which labels are constructed. This is common in background noise-like datasets, where the labels stand for where the recordings were made (airport, metro station, etc), and the time-frequency content often strongly overlaps between classes which share common acoustic properties [38]. Therefore, knowing that clustering quality is untied from the classification task, we focus on clustering metrics. The data considered for both tasks are described next.

1) *Spoken digits*: spoken digit recognition consists of identifying which digit was spoken in an audio utterance. The main acoustic feature is the digit itself, while other sound characteristics are minor features. For this task, we consider the AudioMNIST dataset [39], with 30000 audio samples – of which 24000 are randomly selected for training, 3000 for validation, and 3000 for testing – where each file contains the audio recording, resampled at 16 kHz, of a spoken digit. The speakers are of different gender and age. The raw audio data are preprocessed as follows. First, we pad zeros to each audio sample until the desired duration of 1 second is achieved. The padded audio is then applied to a short-term Fourier transform (STFT), with length of 960 samples, Hann window of the same size, and a hop of 480 samples. Moreover, we take the module of the output of the STFT and limit the frequency range to 128 frequency bins – approximately 6 kHz. Such a frequency range has showed to be sufficient for classification tasks on AudioMNIST in previous tests [40]. Finally, the spectrograms

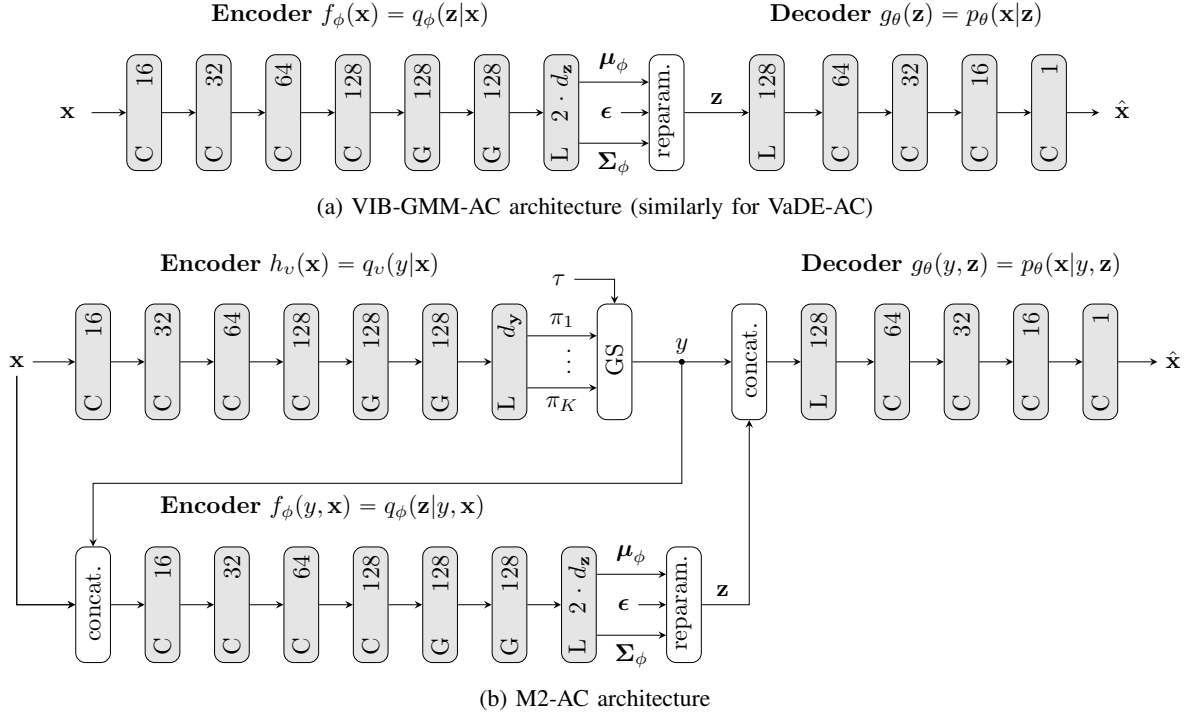


Fig. 7: Proposed VAE architectures for continuous prior (a) and continuous-categorical latent spaces (b). The number in each layer indicates output channels. The C encoder layers consist of Conv2D, while G layers are GRUs. The decoder layers are transposed convolutions Conv2D.T. All C kernels are (8,8) with stride (2,2) and padding (3,3). Every C layer is followed by 2D batch normalization and a ReLU activation, except the last convolution before a GRU layer, which employs a tanh, and the last convolution of the decoder, followed by a sigmoid function. Linear layers (L) have no activation. “concat” is a concatenation in the channel (latent) dimension.

are normalized by their mean and variance, with ranges limited from 0 to 1 by a min-max adjustment.

The input to the neural network is the padded and preprocessed full second of audio, with 128 frequency bins and 99 time bins, which zero-frequency bin is removed. Additionally, we leverage time dependencies by feeding the entire duration of a file to the NN [33]. Importantly, we generate an activity detection vector $\mathbf{a}_x$ containing zeros in the zero-padded time-indexes, and ones otherwise. This vector multiplies both target and prediction during the calculation of the reconstruction loss, as in (3) and (13).

2) *Acoustic environments*: we devise the classification of urban acoustic scenes using real-world datasets. Such a task is much more challenging as the (background) sound features resemble noise and greatly overlap in time and frequency. We consider two different datasets: TAU2019 [41], with 1200 audio recording files of 10 seconds from different acoustic scenes; and the UrbanSound8K (US8K) [42] dataset, with 8732 sound excerpts of 10 different acoustic scenes, mostly 4-seconds long. Differently from AudioMNIST, we now use a mel-frequency cepstrum to reduce the input dimension without limiting frequency range, as acoustic scene classification can benefit from the broader range.

We resample data to 16 kHz. The US8K audio files are zero-padded to four seconds, where we also employ a sound activity mask for the calculation of the reconstruction error. The mask $\mathbf{a}_x$ is always one for the TAU2019 dataset since all files are 10-seconds long, not requiring zero-padding. We obtain an STFT

of 960 samples with a Hann window the same size and 50% overlap. Lastly, we apply a mel-frequency scaling with 128 bins. The cepstrum is normalized by its mean and variance, and a min-max normalization to limit values from 0 to 1. We feed the NNs with a time-context window [33] of 4 seconds for the US8K dataset and 10 seconds for the TAU2019 dataset, which are the maximum duration of the dataset’s files.

Differently from the AudioMNIST case, urban scene classification has direct application for hearing devices [43], [44], where different acoustic scenes result in different processing, which is proportional to the constraints of the device. We take two cases into account: the same number of clusters as labels in the dataset; and a reduced number of clusters. Specifically for the considered urban acoustic scene datasets, we consider 10 clusters for the higher end, as it matches the number of labels in the dataset. For the lower end, we take 5 clusters into account, as it is a significant reduction from 10, merging similar clusters, but still sufficient for effectively calculating clustering metrics. In practice, a higher number of clusters results in a more complex and “specialized” processing, representing a higher-end version of, e.g., a hearing aid device. On the other hand, the lower cluster number could represent a more affordable version of the same device.

B. Metrics

The considered accuracy metrics, used for analysis in part 1, and the clustering metrics, taken into account for both parts (1 and 2) of the results, are described in the following.

TABLE I: AudioMNIST dataset results, either by considering labels as clusters or by applying K-means, GMM-EM, and the aforementioned variational autoencoder models. The best results are highlighted in bold.

Method	Accuracy (%) $\uparrow$	NMI $\uparrow$	Silhouette $\uparrow$	DBI $\downarrow$	CHI $\times 10^3 \uparrow$	Total param. (M)	Enc. param. (M)
None (labels)	100.00	1.00	-0.04	5.56	0.10	NA	NA
K-means	18.40 $\pm$ 1.22	0.10 $\pm$ 0.02	0.13 $\pm$ 0.01	2.04 $\pm$ 0.02	0.69 $\pm$ 0.07	NA	NA
GMM-EM	17.62 $\pm$ 0.33	0.09 $\pm$ 0.07	0.13 $\pm$ 0.01	1.95 $\pm$ 0.04	0.69 $\pm$ 0.03	NA	NA
VaDE	69.07 $\pm$ 7.90	0.76 $\pm$ 4.41	0.23 $\pm$ 0.16	1.65 $\pm$ 0.16	6.37 $\pm$ 41.78	15.26	7.63
VaDE-AC	77.98 $\pm$ 9.11	0.81 $\pm$ 0.05	0.25 $\pm$ 0.02	1.51 $\pm$ 0.10	0.61 $\pm$ 0.05	2.07	1.31
VIB-GMM	78.26 $\pm$ 5.08	0.78 $\pm$ 4.37	0.23 $\pm$ 0.02	1.56 $\pm$ 0.10	0.64 $\pm$ 0.03	15.26	7.63
VIB-GMM-AC	70.78 $\pm$ 2.80	0.71 $\pm$ 0.02	0.21 $\pm$ 0.01	1.61 $\pm$ 0.26	0.54 $\pm$ 0.03	2.00	1.25
M2-AC ($\lambda = 0.5$)	76.30 $\pm$ 7.58	0.78 $\pm$ 0.05	0.97 $\pm$ 0.01	0.07 $\pm$ 0.01	40.14 $\pm$ 10.96	3.32	1.24

1) *Unsupervised accuracy*: in (unsupervised) clustering tasks, the numeric labels may not correspond directly to the ground truth labels. We then consider an unsupervised approach for calculating accuracy, which consists of finding the matching truth labels for the clusters via the Hungarian algorithm [45]. Unsupervised accuracy ranges from 0 to 100%.

2) *Normalized mutual information*: the normalized mutual information (NMI) is an information theoretic approach that evaluates the clustering quality by measuring the amount of shared information between clustering assignments and truth labels [46]. Its range is from 0 to 1.

3) *Silhouette score*: the Silhouette score [47] measures how similar a data point is to its own cluster in comparison to other clusters. It combines cohesion (how close data points within a cluster are) and separation (how distinct is a cluster from another). The range is from -1 to +1: -1 indicates misclassification; 0 tells us that clusters overlap; and +1 indicate optimal clustering.

4) *Davies-Bouldin index*: the Davies-Bouldin index (DBI) [48] is defined as the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DBI indicates better clustering – in terms of compactness and separation. Its range can vary from 0 to infinity.

5) *Caliński-Harabasz index*: the Caliński-Harabasz index (CHI) [49] measures the ratio of the sum of between-cluster to within-cluster dispersion – distinctiveness. Better-defined clusters are indicated by a higher value, and the range is from 0 to infinity.

C. Baselines

For the sake of comparison, we perform clustering using two traditional methods, named K-means [50] and the optimization of a Gaussian mixture model using the expectation-maximization (EM) algorithm (GMM-EM) [51]. Other classical approaches and derivations are assumed to achieve similar performance to K-means and GMM-EM.

We also compare our proposed M2-AC model directly to the VIB-GMM [22] with its lower-sized alternative VIB-GMM-AC. Additionally, as noticed in [22], the Variational Deep Embedding (VaDE) [21] is a subset of the VIB-GMM model. Therefore, for comparison, we also consider a model trained with the VaDE ELBO [21], also including a sound activity mask for audio clustering:

$$\mathcal{L}_{\text{VaDE}}(\theta, \phi) = \mathbb{E}_{q_{\phi}(\mathbf{z}, c | \mathbf{x})} [\mathbf{a}_{\mathbf{x}} \log p_{\theta}(\mathbf{x} | \mathbf{z})] - D_{KL}(q_{\phi}(\mathbf{z}, c | \mathbf{x}) || p_{\theta}(\mathbf{z} | c)). \quad (20)$$

Equations (8), (9), and (10) can be adapted to the VaDE ELBO in (9) by replacing $p_{\phi}(\mathbf{z})$ with $p_{\theta}(\mathbf{z} | c)$ and $q_{\phi}(\mathbf{z} | \mathbf{x})$ with $q_{\phi}(\mathbf{z}, c | \mathbf{x})$. The VaDE model uses the same architecture of the VIB-GMM model. Thus, we also consider its audio clustering version (VaDE-AC) as the one from Fig. 7a.

Other variations in literature are often a derivation of the VaDE and the VIB-GMM, and they often are unable to outperform these methods consistently [23], so we let them out of the scope of this paper. For all considered models, we use the same encoder and decoder architectures, detailed in Section V, which include the modifications proposed in Section V-C. All models are trained 10 times, independently, for 500 epochs, with a learning rate decreasing exponentially from $5e-4$ to $5e-5$, using Adam optimizer. In the following, we present the results for the spoken digits clustering task.

D. Performance

1) *Spoken digits*: for spoken digit clustering, K-means and GMM-EM achieve insufficient accuracy and NMI metrics, only enhancing clustering scores when compared to the truth labels. We also tried to apply dimensionality reduction techniques on the pre-processed data with the traditional methods, but no performance improvement was obtained.

The variational autoencoder approaches are successful in achieving a sufficiently high accuracy (average of 74.5%) and NMI (average of 0.77). The method achieving the highest accuracy is the VIB-GMM with 78.26%, with its low-complexity counterpart, the VIB-GMM-AC, arriving at 70.78%. The difference in parameters is, however, dramatic, with a reduction of 13.26 M parameters from the original VIB-GMM model. On another hand, the VaDE of higher complexity (15.26 M parameters) achieves lower accuracy, of 69.07%, when compared to the hardware-friendly version (2.07 M parameters), with 77.98% accuracy. It appears that the VaDE’s less general loss function was more efficiently optimized by the audio clustering variant, with convolutions and recurrent layers instead of fully connected layers. The different behavior from VIB-GMM and VaDE versions seems counter intuitive, showing a strong dependence of loss function and neural network architecture. The silhouette, DBI, and CHI scores, nonetheless, are not very satisfactory for the VaDE and VIB-GMM variations. The consistent failure of the Gaussian mixture-like latent space in achieving high clustering metrics gives us a hint that the choice might not be appropriate for the task when more challenging scenarios are considered.

The M2-AC, on the other hand, achieves high accuracy and NMI (76.30% and 0.78, respectively), and is capable of

TABLE II: Clustering metrics on the test set of the mentioned datasets, by applying K-means and the aforementioned variational autoencoder models, averaged over 10 independent runs. The best results are highlighted in bold.

Dataset	Clusters	Method	Silhouette $\uparrow$	DBI $\downarrow$	CHI $\times 10^3 \uparrow$	Total param. (M)	Enc. param. (M)
TAU2019	10	K-means	0.03 $\pm$ 0.01	3.48 $\pm$ 0.03	0.00 $\pm$ 0.00	NA	NA
		VIB-GMM-AC	-0.02 $\pm$ 0.02	5.07 $\pm$ 0.27	0.25 $\pm$ 0.01	2.20	1.28
		M2-AC ($\lambda = 2.0$)	0.77 $\pm$ 0.01	0.33 $\pm$ 0.02	6.97 $\pm$ 0.38	3.67	1.26
		M2-AC-w ($\lambda = 2.0$)	0.53 $\pm$ 0.01	0.61 $\pm$ 0.01	2.03 $\pm$ 0.05	3.27	1.24
	5	K-means	0.06 $\pm$ 0.01	3.12 $\pm$ 0.06	0.53 $\pm$ 0.00	NA	NA
		VIB-GMM-AC	0.02 $\pm$ 0.00	3.41 $\pm$ 0.14	0.34 $\pm$ 0.03	2.20	1.28
		M2-AC ($\lambda = 2.0$)	0.79 $\pm$ 0.06	0.30 $\pm$ 0.07	14.76 $\pm$ 4.01	3.55	1.24
		M2-AC-w ($\lambda = 2.0$)	0.59 $\pm$ 0.01	0.54 $\pm$ 0.01	4.98 $\pm$ 0.10	3.24	1.23
UrbanSound8K	10	K-means	0.14 $\pm$ 0.02	1.90 $\pm$ 0.14	0.20 $\pm$ 0.02	NA	NA
		VIB-GMM-AC	0.09 $\pm$ 0.01	2.35 $\pm$ 0.21	0.11 $\pm$ 0.02	2.03	1.25
		M2-AC ($\lambda = 2.0$)	0.73 $\pm$ 0.11	0.39 $\pm$ 0.14	1.17 $\pm$ 0.48	3.37	1.24
		M2-AC-w ($\lambda = 2.0$)	0.55 $\pm$ 0.01	0.61 $\pm$ 0.02	0.44 $\pm$ 0.03	3.27	1.24
	5	K-means	0.18 $\pm$ 0.01	1.63 $\pm$ 0.10	0.32 $\pm$ 0.03	NA	NA
		VIB-GMM-AC	0.10 $\pm$ 0.02	2.33 $\pm$ 0.34	0.14 $\pm$ 0.03	2.03	1.25
		M2-AC ($\lambda = 2.0$)	0.78 $\pm$ 0.09	0.32 $\pm$ 0.10	3.19 $\pm$ 1.47	3.32	1.24
		M2-AC-w ($\lambda = 2.0$)	0.61 $\pm$ 0.01	0.52 $\pm$ 0.01	1.10 $\pm$ 0.07	3.24	1.23

maintaining very high clustering metrics, with the silhouette metric almost at its maximum and lowering the DBI to almost zero, as well as achieving a very high CHI score. This indicates that the choice of a categorical latent space for clustering better aligns to the nature of the task, which is inherent categorical. Interestingly, no method was able to achieve an accuracy higher than 80.00%. This result reflects on the time and frequency overlap of the spoken digits: even though the label of the digit is the principal feature, their physical overlap (of sound) provides a natural limit to perfect unsupervised classification. With those insights in mind, we analyze next the more challenging task of environment sound clustering.

2) *Acoustic environments*: for this challenging task, GMM-EM did not converge when the TAU2019 or UrbanSound8K data were considered. The VaDE and its lower-complexity variation, the VaDE-AC, consistently presented divergence or latent space collapse³, even for different random seeds.

Differently, as we can see in Table II, K-means was able to converge but achieved very poor clustering metrics. Moreover, the VIB-GMM-AC could process the data without any divergence or latent collapse, however, its higher-complexity format, the VIB-GMM, often would result in a collapse of the clusters. This could probably be remedied by a tailored initialization of weights, which is out of scope for this comparison. Indifferently, we can suppose that the problem for the VaDE probably stems from the loss function, given that VIB-GMM-AC and VaDE-AC have a very similar neural network architecture, but different losses, with the additional considerations in the VaDE ELBO (20) becoming too strict of a constraint when applied to high-dimensional highly-overlapping data. Furthermore, by removing the VaDE loss function constraint, the general form considered for the VIB-GMM-AC was able to hold a stable training. Nevertheless, the metrics show unsuccessful clustering. Such scores are in line with expectations from the previous experiment, where we pointed out that the Gaussian mixture latent space is not the most interesting choice for clustering given its continuous nature, while clustering is, in essence, a categorical problem.

By considering a categorical latent space for clustering,

the M2-AC model was able to achieve satisfactory silhouette, DBI, and CHI scores for both TAU2019 and UrbanSound8K datasets. Considering windows of 1 second of audio, we also applied the time-context windowing approach proposed in Section IV resulting in the M2-AC-w variation. The windowing approach reduced even further the complexity of the model with an obvious penalization in all clustering metrics, which is expected given that the new estimation approximates the entire signal by separate windows, averaging each output. The windowing approach is still, however, able to perform successful clustering for both considered datasets.

For a closer connection to practical applications, we freely modify the number of clusters to 5 and replicate the results for both datasets. This could represent a more affordable system, for example. From the results, we can notice that silhouette and DBI metrics are increased by small amounts, indicating that clusters are almost as compact and separate as before. A greater increase is perceived for the CHI score, which tells us that the lower number of clusters results in less disperse clusters, probably due to the reduction of outliers.

For a comprehensive visualization of the clustering process, we present an example of the raw and clustered data points of both datasets through the M2-AC-w approach in Fig. 8, obtained through t-distributed stochastic neighbor embedding (t-SNE). Note how the raw data have strongly-overlapped data points, which are nicely split apart in different clusters after the method is applied. Notice, also, how outliers are more present for 10 clusters than for 5 clusters, which explains the bigger leap in the CHI metric on the quantitative results. Importantly, there is always overlap, which happens given the time and frequency overlap of sound data. To achieve close-to-perfect classification given its abstract labels, one would have to rely on supervised learning, which would focus solely on that task and would less reflect the physical behavior of audio signals.

VII. CONCLUSION

Based on an existing continuous-categorical variational inference process, we proposed a clustering model tailored for audio applications, the M2-AC, which follows a convolutional-recurrent architecture and allows for substantial reduction in parameters. The proposed model uses a Gumbel-Softmax for

³We refer to latent space collapse when all data is classified to a single cluster.

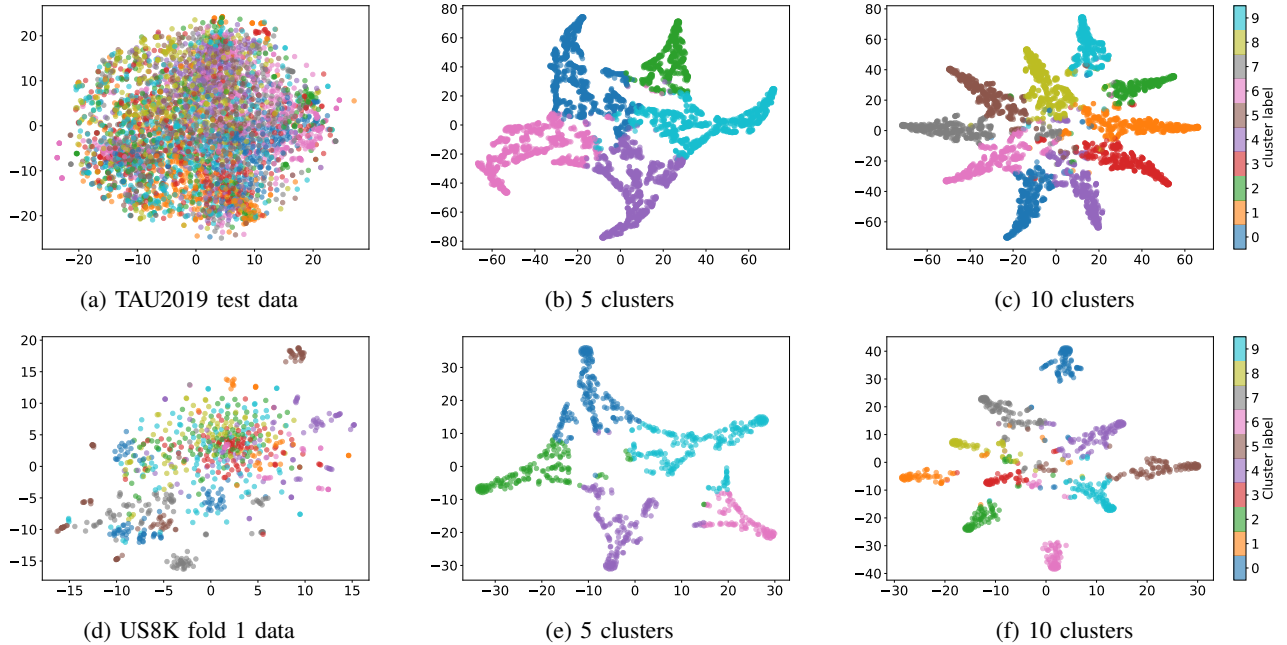


Fig. 8: TAU2019 test set and US8K fold 1 raw and clustered data points by the M2-AC-w approach. Data size reduced for plotting with t-SNE. Each color represents a cluster (labels shown in bar plot), and each circle in the plot is a data point.

categorical sampling during backpropagation-based training. To facilitate practical application and further reduce complexity, we also proposed a windowing scheme for clustering that is applied to the Gumbel-Softmax function logits, averaging encoder outputs for time-windows that are substantially smaller than the total duration of the audio signal. For a spoken digits clustering task, all considered variational approaches achieved high accuracy metrics, however, the proposed model achieve much higher clustering metrics, given its categorical nature, aligned with the clustering problem. Differently, the only method to succeed for clustering acoustic environment recordings was the M2-AC and its windowed variation M2-AC-w. While the spoken digits problem indicated a possible failure of the Gaussian mixture latent space, this was confirmed through the more challenging task, resulting in the divergence or latent collapse of most methods that would cluster through a Gaussian representation.

For future works, the application of the M2-AC variants in a mixture-of-experts system could be considered, detaching the router from the experts, allowing for a Pareto-optimal estimation of each expert through neural architecture search. On another hand, the unsupervised nature of the proposed models could be taken into account for source localization problems, transforming the clusters into a grid of positions, possibly reducing the necessary prior knowledge on the task.

APPENDIX A ABLATION STUDIES

For the ablation study on the practical implementation to audio clustering (Section V-C), we consider the problem of clustering of spoken digits with the AudioMNIST dataset, as described in detail in Section VI-A1. We utilize the M2-AC model architecture mentioned in Section V. The models are trained for 500 epochs each, 10 times (otherwise specified),

independently. The Adam optimizer is considered, with a decreasing learning rate from $5e-4$ to $5e-5$. The model without recurrent layers is defined by removing all GRU layers from the architecture shown in Fig. 7b. In the case of 1D convolutions, all Conv2D and BatchNorm2D are modified to Conv1D and BatchNorm1D operating only in the frequency axis, with kernel 8, stride 2, and padding 3, maintaining the same output channel values as indicated in the schematic.

The ablation study results in Table III, with metrics described in Section VI-B, show that the utilization of 2D convolutions and recurrent layers in synergy achieves the best clustering results, as well as high unsupervised accuracy for the spoken digits when compared to the dataset labels. Notice that removing recurrent layers results in a much higher standard deviation in terms of accuracy and NMI. That reflects how a different initialization can strongly affect the results of the model, showing that recurrent layers add substantial robustness to the process. Additionally, the inclusion of a sound activity mask in the ELBO (13) is crucial for the correct clustering, otherwise, the model is unable to focus solely on the sound-rich part of the audio file, sometimes even clustering zero-padding patterns as we observed by looking at individual results, i.e., the clustering metrics without activity mask cannot be trusted for zero-padded data. Importantly, three independent runs had their latent space collapsed when the mask was removed, and all data was clustered into a single cluster.

ACKNOWLEDGMENT

The authors would like to thank Alex Young, Bruno De-fraene, Frans Widdershoven, Johan David, and Yan Wu for the insightful comments during the development of this work. Additional thanks go to José Núñez Kasaneva for the discussion that planted the initial seed for the use of variational autoencoders.

TABLE III: Clustering and accuracy metrics on the test set of AudioMNIST by applying the M2-AC model with different configurations related to the practical implementation to audio. Average over 10 independent runs* with standard deviation.

Convolution type	Recurrent layers	Activity mask	Accuracy (%) $\uparrow$	NMI $\uparrow$	Silhouette $\uparrow$	DBI $\downarrow$	CHI $\times 10^3$ $\uparrow$
2D	Yes	Yes	76.30 $\pm$ 7.58	0.78 $\pm$ 0.05	0.97 $\pm$ 0.01	0.07 $\pm$ 0.01	40.14 $\pm$ 10.96
1D	Yes	Yes	22.32 $\pm$ 11.49	0.13 $\pm$ 0.18	0.90 $\pm$ 0.02	0.15 $\pm$ 0.03	12.42 $\pm$ 6.10
2D	No	Yes	39.20 $\pm$ 21.46	0.38 $\pm$ 0.28	0.85 $\pm$ 0.12	0.25 $\pm$ 0.15	13.74 $\pm$ 11.27
2D	Yes	No	52.69 $\pm$ 5.36	0.58 $\pm$ 0.64	0.94 $\pm$ 0.01	0.13 $\pm$ 0.06	21.13 $\pm$ 4.27

* For the case without activity mask, only 7 runs were considered as the remaining 3 had their latent space collapsed.

REFERENCES

- [1] International Organization for Standardization, "Iso 8253-1:2010 - acoustics — audiometric test methods — part 1: Pure-tone air and bone conduction audiometry," <https://www.iso.org/standard/43601.html>, ISO, Geneva, Switzerland, 2010, edition 2, Confirmed in 2021.
- [2] C. D. Francis, P. Newman, B. D. Taff, C. White, C. A. Monz, M. Levenhagen, A. R. Petrelli, L. C. Abbott, J. Newton, S. Burson, C. B. Cooper, K. M. Frstrup, C. J. McClure, D. Mennitt, M. Giamellaro, and J. R. Barber, "Acoustic environments matter: Synergistic benefits to humans and ecological communities," *Journal of Environmental Management*, vol. 203, pp. 245–254, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0301479717307193>
- [3] R. Wu, Z. Huang, J. Song, and X. Liang, "Improving speech perceptual quality and intelligibility through sub-band temporal envelope characteristics," in *Man-Machine Speech Communication*. Singapore: Springer Nature Singapore, 2024, pp. 235–247.
- [4] J. M. Kates, "Classification of background noises for hearing-aid applications," *The Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 461–470, 1995.
- [5] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 18, pp. 2991–3002, 2005.
- [6] P.-J. Ting, S.-J. Ruan, and L. P.-H. Li, "Environmental noise classification with inception-dense blocks for hearing aids," *Sensors (Basel)*, vol. 21(16):5406, 2021.
- [7] A. Yellamsetty, E. J. Ozmeral, R. A. Budinsky, and D. A. Eddins, "A comparison of environment classification among premium hearing instruments," *Trends in Hearing*, vol. 25, 2021.
- [8] W. Han, E. Coutinho, H. Ruan, H. Li, B. Schuller, X. Yu, and X. Zhu, "Semi-supervised active learning for sound classification in hybrid learning environments," *PLOS ONE*, vol. 11, no. 9, pp. 1–23, 09 2016.
- [9] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML 2013 Workshop: Challenges in Representation Learning*, 2013. [Online]. Available: <https://www.researchgate.net/publication/280581078>
- [10] L. V. Fiorio, B. Karanov, J. David, W. v. Houtum, F. Widdershoven, and R. M. Aarts, "Semi-supervised learning with per-class adaptive confidence scores for acoustic environment classification with imbalanced data," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [11] M. Shah and N. Sureja, "A comprehensive review of bias in deep learning models: Methods, impacts, and future directions," *Archives of Computational Methods in Engineering*, vol. 32, pp. 255–267, 2025. [Online]. Available: <https://doi.org/10.1007/s11831-024-10134-2>
- [12] L. Barbaro, A. Sourdril, J.-S. P. Froidevaux *et al.*, "Linking acoustic diversity to compositional and configurational heterogeneity in mosaic landscapes," *Landscape Ecology*, vol. 37, no. 5, pp. 1125–1143, 2022.
- [13] L. Pham, H. Tang, A. Jalali, A. Schindler, R. King, and I. McLoughlin, "A low-complexity deep learning framework for acoustic scene classification," in *Data Science – Analytics and Applications*. Springer, 2022, pp. 26–32.
- [14] A. E. Ezugwu, A. M. Ikonu, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022.
- [15] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.
- [16] N. Dilokthanakul, P. A. M. Mediano, M. Garnelo, M. C. H. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders," 2017.
- [17] S. Mukherjee, H. Asnani, E. Lin, and S. Kannan, "Clustergan : Latent space clustering in generative adversarial networks," 2019. [Online]. Available: <https://arxiv.org/abs/1809.03627>
- [18] J. T. Chauhan, "Comparative study of gan and vae," *International Journal of Computer Applications*, vol. 182, no. 22, pp. 1–5, Oct 2018. [Online]. Available: <https://ijcaonline.org/archives/volume182/number22/30062-2018918039/>
- [19] H. Thanh-Tung and T. Tran, "Catastrophic forgetting and mode collapse in gans," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–10.
- [20] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 31–35.
- [21] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: an unsupervised and generative approach to clustering," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, 2017, p. 1965–1972.
- [22] Y. Uğur, G. Arvanitakis, and A. Zaidi, "Variational Information Bottleneck for Unsupervised Clustering: Deep Gaussian Mixture Embedding," *Entropy*, vol. 22, no. 2, 2020.
- [23] J. Guo, W. Fan, M. Amayri, and N. Bouguila, "Deep clustering analysis via variational autoencoder with gamma mixture latent embeddings," *Neural Networks*, vol. 183, p. 106979, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0893608024009080>
- [24] L. V. Fiorio, B. Defraene, J. David, F. Widdershoven, W. van Houtum, and R. M. Aarts, "Unsupervised variational acoustic clustering," 2025. [Online]. Available: <https://arxiv.org/abs/2503.18579>
- [25] L. V. Fiorio, I. Nikoloska, and R. M. Aarts, "Categorical unsupervised variational acoustic clustering," 2025. [Online]. Available: <https://arxiv.org/abs/2504.07652>
- [26] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised Learning with Deep Generative Models," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.
- [27] E. Jang, S. Gu, and B. Poole, "Categorical Reparameterization with Gumbel-Softmax," in *International Conference on Learning Representations (ICLR)*, 2017.
- [28] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014*, 2014.
- [29] R. Shu, "Gaussian Mixture VAE," <https://ruishu.io/2016/12/25/gmvae/>, December 2016, accessed: 2025-06-26.
- [30] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," 2015. [Online]. Available: <https://arxiv.org/abs/1411.0030>
- [31] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.
- [32] J. Lee, Y. Jung, M. Jung, and H. Kim, "Dynamic noise embedding: Noise aware training and adaptation for speech enhancement," 2020. [Online]. Available: <https://arxiv.org/abs/2008.11920>
- [33] L. V. Fiorio, B. Karanov, B. Defraene, J. David, F. Widdershoven, W. Van Houtum, and R. M. Aarts, "Spectral Masking With Explicit Time-Context Windowing for Neural Network-Based Monaural Speech Enhancement," *IEEE Access*, vol. 12, pp. 154 843–154 852, 2024.
- [34] B. Karanov, D. Lavery, P. Bayvel, and L. Schmalen, "End-to-end optimized transmission over dispersive intensity-modulated channels using bidirectional recurrent neural networks," *Opt. Express*, vol. 27, no. 14, pp. 19 650–19 663, Jul 2019.
- [35] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," 2020.
- [36] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [37] A. Smart, D. Wang, E. Monk, M. Díaz, A. Kasirzadeh, E. Van Liemt, and S. Schmer-Galunder, "Discipline and label: A weird genealogy and social theory of data annotation," in *Proceedings of the EASST-4S 2024 Conference: Making and Doing Transformations*. Amsterdam, Nether-

- lands: European Association for the Study of Science and Technology (EASST) and Society for Social Studies of Science (4S), 2024.
- [38] H.-S. Heo, J. weon Jung, H. jin Shim, and H.-J. Yu, "Acoustic scene classification using teacher–student learning with soft labels," in *Proceedings of Interspeech 2019*. Graz, Austria: ISCA, September 2019, pp. 2843–2847.
 - [39] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek, "AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark," *Journal of the Franklin Institute*, 2023.
 - [40] A. Young, L. V. Fiorio, B. Yang, B. Karanov, W. van Houtum, and R. M. Aarts, "Hybrid real- and complex-valued neural network architecture," 2025. [Online]. Available: <https://arxiv.org/abs/2504.03497>
 - [41] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13.
 - [42] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2014, p. 1041–1044.
 - [43] S. Yook, K. W. Nam, H. Kim, S. H. Hong, D. P. Jang, and I. Y. Kim, "An environment-adaptive management algorithm for hearing-support devices incorporating listening situation and noise type classifiers," *Artificial Organs*, vol. 39, no. 4, pp. 361–368, April 2015.
 - [44] L. Lamarche, C. Giguère, W. Gueaieb, T. Aboulnasr, and H. Othman, "Adaptive environment classification system for hearing aids," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 3124–3135, May 2010.
 - [45] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
 - [46] N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *J. Mach. Learn. Res.*, vol. 11, p. 2837–2854, Dec. 2010.
 - [47] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
 - [48] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.
 - [49] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.
 - [50] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.
 - [51] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.