On Learning Useful Variational Autoencoder Representations

Applications in Audio Modelling and Hearing Loss Treatment

Høegh, Rasmus Malik Thaarup

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

PhD thesis

On Learning Useful Variational Autoencoder Representations

Applications in Audio Modelling and Hearing Loss Treatment

Rasmus M. Th. Høegh

WSAudiology

DTU

PhD thesis title: On Learning Useful Variational Autoencoder Representations: Applications in Audio Modelling and Hearing Loss Treatment
Location, year: Kongens Lyngby, 2022
Author: Rasmus Malik Thaarup Høegh
Supervisors: Professor Morten Mørup and Jens Brehm Bagger Nielsen, PhD
Co-supervisors: Assistant Professor Abigail Anne Kressner, Professor Lars Kai Hansen, and Adam Westermann, PhD

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Ørsteds Plads
Building 340
2800 Kongens Lyngby, Denmark

Summary

Representation learning systems utilise large amounts of unlabelled data to learn how to extract useful information, often to facilitate learning in some other task, such as improving learning on another, smaller supervised data set. One approach to representation learning is variational autoencoders (VAEs), a type of deep latent variable model. Such models learn to infer representations by modelling the distributions of data using variational inference. A simple VAE consists of an inference network producing distributions over a latent variable given an input datum and a generative network that produces a probabilistic estimate of the original input. The model is optimised to ensure that the reconstructed input is as close as possible to the original input or, equivalently, that as few as possible distortions are introduced in the reconstruction. The model also optimises the amount, or rate, of information used in producing the reconstructions, measured as a divergence from a set prior distribution.

This thesis explores how to learn VAE representations that are *useful* in that they both generalise well and code for information relevant to a given task wherein one should desire to use the model. The first two contributions explore, respectively, how learnt representations can be used to do model-based active learning for efficient measurement of a person's hearing loss and how a VAE perspective can improve speaker separation models. The contributions show that rate-distortions trade-offs affect the learnt representation. The first contribution shows how rate-distortion trade-offs affect the learnt representation's ability to inform active learning sequential acquisition. Furthermore, the second contribution shows how the generalisation of a speaker separation representation is improved by explicitly optimising for low rates, which existing models are not doing Notably, the probabilistic framework allows the models to quantify uncertainty. The utility of such quantifications is highlighted based on results showing how they allow for estimating a model's performance without knowledge of the ground truth reference.

The final contribution considers an extension of deep hierarchical VAEs that uses differential equations as expressive modelling components. Instead of using discrete Gaussian latent variables, the model relies on neural stochastic differential equations to construct a hierarchy of continuously deep latent processes. Furthermore, it is argued that the model displays continuity properties—based on experiments that vary the number of numerical integration steps used in approximating the latent processes—thus allowing, e.g., trading off computational complexity and performance within a single, trained model.

Based on the combined contributions, a discussion is provided of representation learning from a VAE perspective, the benefits of rate-distortion analysis in the context of generalisation, the use of quantified uncertainty in real-world problem settings, and the ability to incorporate inductive biases in probabilistic frameworks.

Resumé (Danish)

Systemer til repræsentationslæring gør brug af store mængder ikke-superviseret (en: unlabelled) data til at lære at ekstrahere brugbar information, ofte for at kunne facilitere læring i en anden sammenhæng, såsom at forbedre superviseret modellering af et mindre datasæt. En tilgang til repræsentationslæring er variationelle autoindkodere (en: variational autoencoders, VAEs), som er en type af dybe modeller med skjulte (en: latent) variable. Sådanne modeller lærer at inferere repræsentationer ved at modellere datafordelinger gennem brug af variationel inferens. En simpel VAE består af et inferensnetværk, som producerer fordelinger af en skjult variabel, og et generativt netværk, som producerer probabilistiske estimater af det oprindelige input. Modellen optimeres til at sikre, at det rekonstruerede input er så tæt som muligt på det oprindelige input, eller, tilsvarende, at få forvrængninger (en: distortions) introduceres i rekonstruktionen. Modellen optimerer også mængden, eller raten, af information som bliver brugt til at lave rekonstruktionen, målt som en divergens fra en given a priori fordeling.

Denne afhandling undersøger, hvordan man kan lære VAE repræsentationer, som er *brugbare*, i den forstand, at de generaliserer godt og koder for information, som er relevant for den sammenhæng, hvori modellen skal bruges. De første to bidrag undersøger, henholdsvis, hvordan lærte repræsentationer kan bruges i modelbaseret aktiv læring til effektivt at opsamle en persons høretab, og hvordan et VAE-perspektiv kan forbedre modeller til taleseparation. Afvejelser af rate-forvrængning (en: rate-distortion tradeoffs) påvirker de lærte repræsentationer. Det første bidrag viser, hvordan afvejningerne af rate-forvrængning påvirker den lærte repræsentations evne til at informere sekvensiel opsamling i aktiv læring. Det andet bidrag viser, hvordan generalisation af talesepara-tionsrepræsentationer forbedres ved at optimere mod lave rater, hvilket eksisterende modeller ikke gør. Den probabilistiske tilgang muliggører også, at modellerne kan kvantificere usikkerhed. Der argumenteres for nytten af sådanne kvantificeringer baseret på resulter som viser, at de tillader estimering af en models ydeevne uden viden om det rigtige output.

Det sidste bidrag undersøger udvidelsen af dybe hierakiske VAEs til at gøre brug af differentialligninger som ekspressive komponenter til brug i modelleringen. I stedet for at bruge diskrete Gaussiske skjulte variable, gør modellen brug af neurale, stokastiske differentialligninger til at konstruere et hieraki af kontinuert dybe skjulte processer. Det argumenteres at modellen udviser kontinuitet—baseret på eksperimenter som varierer antallet af skridt brugt i den numeriske integrations tilnærmelse af de skjulte processer—og dermed tillader, for eksempel, at afveje beregningskompleksitet og ydeevne, efter træning, med kun én model.

Baseret på de samlede bidrag gives der en diskussion af repræsentationslæring fra et VAE-perspektiv. Særligt diskuteres nytten af analyse af rate-forvrængning i fortolkningen af generalisation, brugen af kvantificeret usikkerhed i virkelige problemstillinger, og muligheden for indarbejdningen af induktive bias i probabilistiske tilgange.

Preface

The following PhD thesis was submitted as partial fulfilment of the requirements of the PhD programme at the Technical University of Denmark (DTU), Department of Computer Science and Applied Mathematics (DTU Compute). The PhD project was done in partnership with WS Audiology (WSA, former Widex) in the period from May 2019 to October 2022. The project was partially funded by Innovation Fund Denmark. The work done during the project was conducted in equal parts at the Section for Cognitive Systems (CogSys) at DTU Compute in Kongens Lyngby, Denmark, and the Artificial Intelligence Accelerator (AIA) at WSA in Lynge, Denmark.

The primary supervisors were Professor Morten Mørup (DTU Compute, Section for Cognitive Systems) and Jens Brehm Bagger Nielsen, PhD (Head of AIA, WSA). Co-supervisors were Assistant Professor Abigail Anne Kressner (DTU Health Tech, Hearing Systems Group), Professor Lars Kai Hansen (DTU Compute, Section for Cognitive Systems), and Adam Westermann, PhD (Vice President Innovation, WSA). The PhD project included an external stay at the International Computer Science Institute (ICSI) and the University of California, Berkeley (UCB) with host supervisor Professor Michael W. Mahoney (ICSI, UCB, Lawrence Berkeley National Laboratory) in Berkeley, California from February 2022 to September 2022.

The PhD project included two periods of leave; the first from March 2020 to May 2020 to work as a research assistant at the Section for Cognitive Systems with Professor L. K. Hansen on applying audio deep learning models for COVID-19 cough characterisation, and another period from June 2021 to September 2021 working as a machine learning engineer with J. B. B. Nielsen at the WSA AIA, developing audio deep learning models and building upon work from the PhD project.

Acknowledgements

I have benefitted from many fantastic supervisors, all of whom I would like to thank for their help and guidance. I wish to thank Adam Westermann for his help in ensuring the project became a reality and to thank Lars Kai Hansen for the opportunity to work alongside him on the COVID-19 project. I would especially like to thank Abbie Kressner for all she has taught me over the past seven years since she first supervised my Bachelor's project—and for always looking out for opportunities and helping me seize them. I would also like to extend a special thanks to Jens Nielsen for unfailingly taking the time to discuss ideas, both big and small, and for teaching me how to navigate the world of industrial research; few, I'm sure, navigate the world of WSA better than Jens, both figuratively and circular-literally. I am incredibly grateful to Morten Mørup; a special thanks for having taught me more than anyone about the fantastic world of machine learning and for always patiently and enthusiastically discussing ideas—even when "just a couple of minutes" yet again became an hour.

Beyond my supervisors, I would like to thank my colleagues at the Hearing Systems group, at the Section for Cognitive Systems, and at WS Audiology for having an ever-curious outlook and creating a highly motivating environment. I would like to thank the AIA at WSA and my office mates over the years at CogSys for always being up for a chat about my projects, their projects, confusing results, or the most recent bug. I would also like to thank Michael Mahoney, his group, and ICSI for welcoming me to Berkeley.

I would like to thank my family, too. I am thankful for my talented and supportive siblings, whom I have had the pleasure to look up to and learn from all my life; for my father, who would have been boundlessly proud of this work; and for my mother, for always believing in me and supporting me. Lastly, I would like to thank Trine; her loving support has been the unshakeable foundation without which this work would not have been possible.

Aalborg, Denmark, 18th October 2022

Rasmus M. Th. Høegh

List of contributions

Contributions included in this thesis:

- **Høegh, R. M. Th.**, Jespersen, C. B., Mølgaard, L. L., Kressner, A. A., Mørup, M., Nielsen, J. B. B. (2022). Rate-Distortion Trade-offs in Variational Autoencoder Representations for Sequential Acquisition Active Learning. Pre-print. (See Appendix A)
- **Høegh, R. M. Th.**, Nielsen, J. B. B., Kressner, A. A., Mørup, M. (2022). Improving Speaker Separation Generalization with Variational Inference. Pre-print. (See Appendix B)
- **Høegh, R. M. Th.**, Krishnapriyan, A. S., Hodgkinson, L., Mahoney, M. W. (2022). Hierarchical Variational Auto-Encoders using Latent Neural Stochastic Differential Equations. Pre-print. (See Appendix C)

Contributions completed during the PhD not included in this thesis (*shared first authorship):

- Olsen, K., **Høegh, R. M. Th.**, Mørup, M. (2022). Think Global Adapt Local: Learning Locally Adaptive Kernel Density Estimators. Submitted, under review.
- Olsen*, A. S., **Høegh*, R. M. Th.**, Hinrich, J. L., Madsen, K. H., Mørup, M. (2022). Combining Electro and Magnetoencephalography Data using Directional Archetypal Analysis. Front. Neurosci. 16:911034. doi: 10.3389/fnins.2022.911034
- **Høegh, R. M. Th.**, Nielsen, J. B. B., Mørup, M. (2019). Latent representation linear speaker recognition using deep transfer learning. In International Symposium on Auditory and Audiological Research (ISAAR) 2019, Nyborg, Denmark, August 2019.

Contents

CHAPTER 1

Introduction

1.1 Representations and usefulness

▲ *What is a representation, what makes a representation useful, and how do we discover useful representations?*

We make sense of the world through models, representing complex aspects with simple ones. With these models in hand, we might gain deeper insights into our world or make valuable predictions. While, famously, all models are wrong [Box, 1976], their usefulness stems from whether the models *represent* relevant aspects of the world. Representation learning aims to discover which aspects to represent and how to extract information about them [Goodfellow et al., 2016, Chap. 15]. The resultant knowledge we might refer to as representations or models; a loose distinction between the two, if in truth there is any difference, could be that representations are components of a model, and the synthesis of many representations alongside their interplay makes up a model. We can learn representations by analysing large amounts of data, distilling core aspects of the measured phenomena while doing away with unimportant ones. For instance, a dataset consisting of many hours of audiobooks would allow a representation learning system to determine core elements of speech, such as types of harmonic structures, the presence of silent gaps, and components of speech like phonemes [Chorowski et al., 2019; van den Oord et al., 2017]. The system's design, however, determines what kind of representation we learn. Unsurprisingly, the *usefulness* of the representation depends on its intended application. A simple, tangible representation of spoken language is the corresponding written text. If we wish to use such a representation to add subtitles to a movie, it is a useful representation. If we were interested in keeping track of the current speaker in an online meeting, the written text alone would have done away with much of the information in the spoken language that easily allows for discerning speakers. While an accurate corresponding written text is a simple representation of spoken language, it is not invariably useful.

Feature engineering One way to obtain useful representations is through feature engineering. Feature-engineered systems extract information from signals relevant to the "downstream task" . This extraction is guided by knowledge about the domain of the task, such as knowledge of essential characteristics of the data or signal as well as the specific signal processing needed for efficient extraction of said characteristics [Bishop and Nasrabadi, 2006, Chap. 1]. Above two types of downstream tasks are mentioned, transcription and speaker identification. The domain knowledge used in building feature-engineered systems might be that particular spectral representations of speech, such as mel-frequency cepstrum coefficients, are useful representations of the raw audio [Morgan et al.,

2004]. Consider a yellow system that takes as input a black audio signal and produces a grey representation (i.e., set of features):

If the representation captures the characteristics of a speaker, a simple classification model—the orange component—can be trained to analyse the representation and estimate the speaker's identity. If we have a good feature extraction system, we can generally build high-performing systems even with very little data. However, engineering the appropriate representations can be research intensive and costly, especially once existing domain knowledge—possibly built over decades or centuries of research—is already exploited. Additionally, the system might not extract all the relevant aspects of the signal, and the assumptions that went into the design of the yellow feature extractor might limit the system's overall performance.

Representations in supervised deep learning Representations learnt from data can often be potent additions to systems that integrate domain knowledge. Simple deep learning systems that do supervised learning will have the learnt representation more implicitly instead of a single separate feature extraction component feeding into a classification component, as above. The representations of deep learning systems will have learnt to combine low-level representations into increasingly abstract, high-level concepts as activations from earlier levels propagate through the network. In simple image classification tasks, low-level features of early levels are edges or textures. These are combined to produce higher-level concepts, like various types of animal ears, snouts, and so on. These higher-level objects are, in turn, further aggregated to produce the model's representation of different types of animals [Olah et al., 2018]. In modelling speech, low-level representations can be very similar to standard filterbanks that are aggregated into, for example, phonemes [Ghiasi et al., 2021], words, and so on. These representations are learnt by training models to map input data to a target output, such as the type of animal in a picture, a known transcript of some input audio, or the identity of speaker in a recording. When we have sufficient pairs of such inputs and outputs, supervised deep learning produces high-performing systems [LeCun et al., 2015]. In image classification, the output is often a (low-dimensional) class label. The term "labelled" data is often used in a broader context to refer to data where a target supervisory output is available. The supervisory signal need not be low dimensional. For example, in audio processing, this output can be a new time

series with the same or larger dimensionality. In speaker separation—estimating the component speakers in a recording of multiple, overlapping speakers—we wish to map to multiple time series of the same length as the original input. Constructing labelled datasets is costly, just like the feature engineering approach—albeit in different ways. Constructing labelled data sets can be costly in terms of monetary costs and time, for example due to payments for slow, manual, human labelling, or slow, computationally intensive simulations, and so on. Even if cost might not be an issue, some labels, or supervisory signals, are inherently scarce, simply by virtue of being rare occurences—such as rare weather events or rare diseases.

Learning without labels In the absence of sufficient amounts of labelled data, we can instead rely on learning representations from unlabelled datasets. Such datasets are often larger and more readily available. We might construct a supervisory signal from knowledge of the domain. For instance, we can a construct tasks by extracting pathces from images and predict their relative position [Doersch et al., 2015], which is a simple, early version of self-supervised learning. The learnt representation generalises if the constructed supervisory task is suitably designed, allowing us to "transfer" (parts of) the representations [Goodfellow et al., 2016, Chap. 15.2]. We might, for instance, replace the final layers that are specialised for the patch-related task task task and train new final layers on a small labelled dataset. The pre-trained layers take the place of the yellow feature extractor in the example above. We train the replaced layers to analyse the transferred representation in a new context, playing the part of the orange component. This allows us to solve problems on smaller data sets where we would otherwise be unable to train a full model—if the representation learnt from the unlabelled data is useful in this downstream task. While these are important approaches toward learning representations, we focus on another type of model, generative models. As we will discuss later, these models aim to learn the data distribution and, through this generative task, seek to learn a representation of data.

Generalization Modern representation learning systems are incredibly expressive and arguably capable of learning representations more complex than classic feature-engineered systems. As a consequence, representation learning can significantly improve how well a given task can be solved. Yet, these capabilities are a double-edged sword. The flexibility allowing for expressive learnt representation also allows for learning characteristics that do not generalise if the pattern that the system has learnt does not apply in new situations. Useful representations will generalise such that the represented aspects of the data are also present and informative when various aspects of the environment change. Returning to the online meeting speaker identification example, a sufficiently flexible system might

learn to associate a particular speaker in the training data with their office's background noise or with characteristics of the recording hardware. The system will work only as long as the speaker is in the same place with the same microphone. These example aspects of the recording are relatively high-level, but the issue of generalisation applies to low-level, less tangible facets, too. One approach to learning robust representations is to learn under uncertainty with probabilistic models. In particular, generative models aim to learn a representation of the data by learning the distribution of data. These models aim to assign a high probability (probability density or mass) to their training data. Ensuring that the model assigns a high probability to the training data also ensures that the model assigns a low probability to things not represented in the data; since probability distributions are normalised, increasing the probability of something under the model decreases the probability of others.

Autoencoders and uncertainty We will consider one approach to modelling with uncertainty, namely variational inference [Bishop and Nasrabadi, 2006, Chap. 10], especially from the perspective of auto-encoding and a type of model called variational autoencoders (VAEs) [Kingma and Welling, 2014, 2019; Rezende et al., 2014]. Simplified, key aspects in their construction can be summarised as:

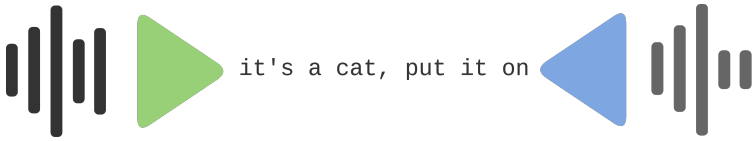
1. compression is comprehension[1],
2. construction guides compression[2],
3. certainty fools construction.

If we produce something that is a compressed yet adequate representation of a signal, we have relied on some understanding of the signal—maybe in the sense that we have realised specific patterns can be expressed efficiently or that certain aspects are uninformative noise. A green system that takes an audio waveform and produces a transcript is a compressed version of the input signal:

[1] Zenil [2019] attributes the phrase to Chaitin; in his "Meta Math!", Chaitin summaries the scientific method as "Understanding is compression! To comprehend is to compress!" [Chaitin, 2004, p. 54].

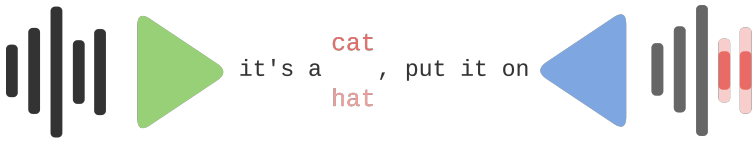
[2] Hawking [2001, p. 83] shows a picture of R. Feynman's blackboard at the time of his death with the phrase "what I cannot create, I do not understand" [Granger, 2017].

If compression is comprehension, we can now learn representations simply by learning to do compression instead. One approach is to learn how to invert the compression process. We might have a learnable version of the green component compress—or encode—our input signal and then have a similarly learnable blue component reconstruct—or decode—the representation. This concept, auto-encoding, allows us to learn representations by minimising some notion of distance between the input and the output (reconstructed input):



A perfect system, solely in terms of the input to output distance, can be arbitrarily achieved by letting both the encoder and the decoder be identity functions, both simply outputing their input unchanged. As far as "distilling" the input information into a simpler representation, identity functions do poorly, and the representation is not any more, or any less, useful than the input signal. We generally reduce the representation capacity so that the model learns to choose what aspects of the input it represents by introducing a bottleneck, constraining the amount of information the system can send to reconstruct its input. The chosen bottleneck imposes characteristics on the representation. Simply reducing the amount of "numbers" sent through by some factor compared to the input signal allows for learning a compressed representation. Other constraints are useful, too, and, for instance, we might be interested in whether the overall magnitude of the representation remains low, that as few dimensions as possible activates for any given input (sparsity), or that dimensions are informative on their own (disentangled) [Goodfellow et al., 2016, Chap. 14]. We might think of these choices as various types of simple inductive biases.

If we endow the model with the capability to express and characterise uncertainty, the model will be more robust to these learnt variations. Instead of relying on deterministic values to represent the representation, we allow the model to describe the distribution of possible values. We might allow the green encoder to express uncertainty about the encoded words and allow the decoder to express uncertainty about how to reconstruct the representation back into a waveform (here darker red indicates higher certainty):



The model might be entirely certain how to represent all aspects of the input except for whether the input audio says cat or hat. Considering only the reconstruction ability, a model without uncertainty might only express one waveform and outputs the more certain, darker red one. The waveform matches the input waveform poorly. Conversely, suppose the model now expresses some uncertainty and is allowed to produce distributions. Then, the model reconstructs better—in expectation—since the light red waveform that matches the input better is part of the model's estimated distribution, even if it has a lower probability. Allowing the model to express uncertainty also provides a tool for quantifying how sure the model might be about a prediction, possibly indicating the quality of the prediction. Accurate uncertainty quantification allows us to choose when to trust the model predictions or, for instance, fall back on a less performant but more robust system. The fallback could be the feature-engineered simple system we considered earlier.

Variational inference Inference using probabilistic models allows us to learn under uncertainty, and we will consider variational inference in particular. Just like models generally represent complex aspects of the world with simple aspects, variational inference relies on representing complex distributions with simpler ones. Generally, we cannot directly infer anything about the complex distributions that would completely describe our phenomena of interest. However, we can often characterise useful aspects of these complex distributions if we use simplified distributions instead. In these models, the representation we are trying to learn might be considered a latent variable, an unobserved factor we are trying to infer based on the observed data. Depending on various choices in the design of our system, there will be a "true" distribution of the latent, unobserved representation for some input—a true posterior distribution of the representation. The word posterior indicates the notion that this distribution is the distribution obtained after observing data. Generally, we cannot feasibly infer the true posterior distributions. We make the problem of learning these models feasible by introducing simplified, guiding distributions and optimising them such that they are close to the true posterior distributions. The simplified distributions are, for this reason, called approximate posterior distributions. Usually, quantifying this closeness relies on a particular notion of a difference between two distributions called the Kullback-Leibler divergence; this divergence will increase if the distributions assign different probabilities to the same outcomes. Modelling with these distributions allows us to impose characteristics on the learnt distributions, for instance, by choosing a particular family of approximate posterior distributions. The true and approximate posterior distributions are affected by various design choices, such as the distribution of the representation before observing data (i.e., the prior distribution) and the architecture we use to construct or parameterise

the distributions.

Rate-distortion trade-offs We can control the expressivity of modern representation learning systems with the variational autoencoder framework, by considering rate-distortion analysis [Shannon et al., 1959]—where, the notion of controlling the information sent through the VAE bottleneck is achievable by controlling a "rate" [Alemi et al., 2018]. If the system changes the distribution a lot when seeing data compared to our choice of prior, we might say that the system encodes a large amount of information in the distribution. This (informational) rate is quantifiable as the Kullback-Leibler divergence between the approximate posterior and the prior. Just like we can quantify the rate, we can also quantify how the system changes or distorts the input signal. This we do as the input signal's probability under the distributions for the reconstructed data—if a high probability is assigned, the system did not distort a lot. Sending through large quantities of information helps the system introduce fewer distortions. Conversely, if we wish to limit the amount of information we send, we might accept a higher level of distortion. In the cat/hat example, the distortion from the "cat" representation is higher than the "hat". The encoding of "cat" (i.e., the approximate posterior) might be closer to the prior, so it incurs a lower rate—it may be that animals were more often the topic than clothing in the training data. This trade-off between good reconstruction and divergence from a prior is called the rate-distortion trade-off. By doing rate-distortion analysis, we can characterise the system's behaviour as a function of these trade-offs.

1.2 Thesis overview

Before delving into a more technical presentation of the above topics in Chapter 2, we present a brief overview of the contributions presented in this thesis and how the topics introduced above tie into the contributions, both individually and as a whole. A summary of the contributions are provided in Chapter 3 and they can be found in their entirety in the Appendix. We return to the questions of this section in Chapter 4 based on the findings of the contributions.

At the beginning of this introduction, we asked ▲ *what is a representation, what makes a representation useful, and how do we discover useful representations?* Representations are simple representations of complex phenomena, and using the representations allows us to model the real world. A representation's usefulness partly comes from an ability to generalise (i.e., being robust to unimportant changes) and partly from its relevance (i.e., whether the represented information is pertinent to the task at hand). One way to learn representations is through variational autoencoders, and the usefulness of these representations can be achieved, for instance, by imposing desired characteristics on the representation or making use of the quantified uncertainty of the probabilistic models. Our focus will be variational autoencoders, and the primary ideas we wish to explore in this thesis, as depicted in Figure 1, revolve around learning useful representations through variational inference, uncertainty quantification, and rate-distortion analysis.

In particular, we are interested in addressing two central questions:

▲ **Variational Inference and Uncertainty Quantification** *How can variational inference and uncertainty quantification in representation learning be used to improve the usefulness of a learnt representation?*

▲ **Rate-Distortion Analysis** *What insights does characterising a representation's usefulness from the perspective of rate-distortion trade-offs analysis provide?*

We will address these questions in three separate, but related, contributions. The first contribution considers active learning of audiograms with a variational autoencoder (VAE) representation, or variational autoencoder acquisition (VAEcquisition)—we will denote this contribution with ⋈ **VAEcquisition**. The second contribution considers audio modelling, or more specifically speaker separation, using variational inference (VI) for a particular class of models called encoder-masker-decoder (EMD) models, or variational inference encoder-masker-decoders (VI-EMDs)—this contribution will be referred to as ⋈ **VI-EMD**. The

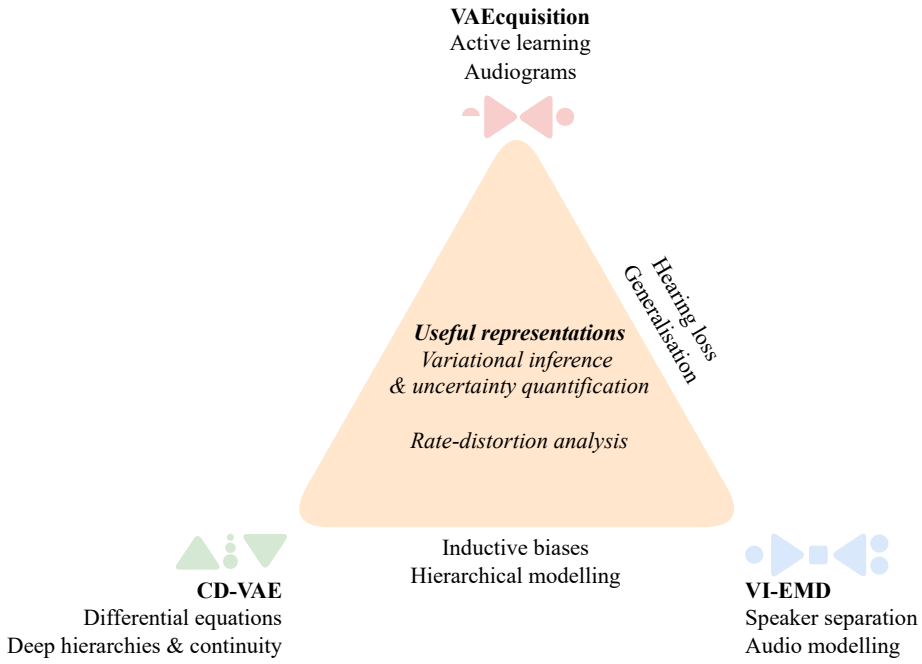


Figure 1. An overview of the contributions and topics in this thesis.

final contribution considers an extension of deep hierarchies in variational autoencoders using differential equations, and introduces a model called continuously deep variational autoencoder (CD-VAE)—and this contribution will be referred to as **CD-VAE**.

VAEcquisition The first contribution (titled "Rate-Distortion Trade-offs in Variational Autoencoder Representations for Sequential Acquisition Active Learning") considers the problem of actively learning about data. In basic active learning, we use a model to decide which unlabelled data are the most cost-effective ones to label. We, instead, consider a sequential acquisition problem of partial data, i.e. data which has unobserved, or missing, dimensions. The model chooses which dimension of a partially observed datum should be observed to learn the most about the datum. We utilise a variational autoencoder that has learnt how to encode partial data and produce distributions over all observed and unobserved dimensions. Beyond the general question of **Useful representations**, this study specifically investigates:

VAEcquisition: Audiogram Acquisition *How can we utilise a learnt variational autoencoder representation to efficiently acquire audiograms?*

Audiograms are characterisations of a person's hearing ability as a function of frequency, and they are used in hearing loss diagnosis and treatment. Measuring audiograms can be time-consuming, and making the process more efficient or accurate improves diagnostics and treatment. The work will, for instance, analyse how well an audiogram representation trained on an American (United States) data set generalises to a German population. This study also allows us to explore the uncertainty quantification capabilities of variational autoencoder representations; we will use the model's uncertainty to estimate whether a sequential acquisition process can be stopped or should carry on. We will consider how the learnt representations' ability to support acquisition and estimate when to stop is affected by rate-distortion trade-offs.

Improving Speaker Separation Generalisation with Variational Inference Following this, yet remaining in the realm of hearing-related topics, we will consider representation learning in audio modelling and, more specifically, the problem of speaker separation. A series of deep learning-based speaker separation systems follow an architectural pattern called encoder-masker-decoder systems. The model makes use of an architectural inductive bias in the form of separation-by-masking, similar to how earlier speaker separation systems relied on masking feature-engineered spectrotemporal representations; these systems encode a mixture of speakers, produce masks for each estimated component and decode the masked representation to produce a single estimate speaker While their performance exceeds that of earlier methods, the performance degrades in new environments. Again, beyond exploring **Useful representations**, this study particular investigates variational inference (VI) encoder-masker-decoders (EMDs):

VI-EMD: Speaker separation *How does recasting encoder-masker-decoder speaker separation models in a variational autoencoder framework affect generalisation?*

The variational version of these models is a simple hierarchical model, producing distributions of encodings and, subsequently, producing masks based on the encodings. Additionally, this study will explore inductive biases, such as how different priors affect the learnt representation. Similar to the stopping estimation in **VAEcquisition**, we will investigate whether uncertainty quantification

enables us to determine the quality of separation, so that we can estimate the quality of the separation without knowledge of ground truth reference signals.

Hierarchical Variational Auto-Encoders using Latent Neural Stochastic Differential Equations Differential equations are a well-studied mathematical object. Incorporating them in representation learning systems allows us to induce characteristics of systems typically described well by differential equations, such as dynamical systems with a state that changes over time. Neural differential equations are differential equations where neural networks parametrise the rate of change of the state. In variational autoencoders, we incorporate these as a modelling component by defining the latent object as solutions to neural differential equations. Similar to the simple hierarchy in **VI-EMD**, the model considered in this contribution uses hierarchies of latent variables. Deep hierarchical variational autoencoders are strong generative models, and increasing the depth of the hierarchical model improves the performance. In some sense, these deep hierarchical models produce a series of latent states that evolve down through the hierarchy dependent on previous states and a controlling signal from an encoder. In this view, they resemble a dynamical system where the hierarchical latent states are the state of a dynamical system that evolves not over time but the depth of the hierarchy. Because of this, we ask:

CD-VAE: **Hierarchies as Dynamical Systems** *How can we interpret the latent representations in deep hierarchical variational autoencoders as stochastically evolving dynamical systems?*

CD-VAE: **Continuity** *What are the benefits, if any, of a continuous-depth formulation of hierarchical variational auto-encoders using neural stochastic differential equations?*

For dynamical systems, a central characteristic of many systems is continuity—for instance, whether the state or the rate of change is continuous in time. We also explore the usefulness of an inductive bias towards a continuous depth (CD) hierarchical variational autoencoder as a particular aspect of exploring **Useful representations**.

CHAPTER 2

Background

In the following, an introduction to some central background studies is provided. The presentation focuses on aspects that are at the core of the contributions and on aspects that facilitate an understanding of the contribution summaries provided in Chapter 3. Less central but still relevant background material is briefly introduced, but for these aspects, the contributions in the appendix provide further details.

2.1 Deep generative models

Deep generative models use deep learning architectures to construct generative models of data [Goodfellow et al., 2016, Chap. 20]. Generative models form representations of data by learning how to model the data distribution. Generally, such models are trained on large, unstructured (or unsupervised, unlabelled) but readily available data sets. The resulting representation can then be used as a starting point for training other models where only smaller supervised data sets are available, or yet other tasks, such as increasing the resolution of an input, contructing continuations of the input, producing other views of it, determining whether the input is out-of-distrubution for the learnt distibution, procucing a compression scheme, and so on [Bond-Taylor et al., 2021; Townsend et al., 2019].

One type of deep generative models is autoregressive (AR) models, such as MADE [Germain et al., 2015], PixelCNNs [Salimans et al., 2017; van den Oord et al., 2016b], and WaveNets [van den Oord et al., 2016a]. Generally, AR models were until recently the stronger of deep generative models, but models achieving better performances in terms of (bounds on) the likelihood have now been introduced. Recent developments include the introduction of (normalising) flows and their extensions [Dinh et al., 2015, 2017], energy-based-models [Goodfellow et al., 2016, Chap.16] and extensions [Che et al., 2020], and diffusion models and their extensions [Ho et al., 2020; Kingma et al., 2021; Rombach et al., 2022; Sohl-Dickstein et al., 2015; Song et al., 2021]. In this thesis, the contributions study variational autoencoders. Unifying frameworks exists, such as SurVAE flows, combining VAEs with flows [Nielsen et al., 2020] and Autoregressive Diffusion Models combining ARs models with diffusion models [Hoogeboom et al., 2022]. Various connections exists between all the framework, and for example Child [2021] discusses how deep VAEs generalise AR models, Che et al. [2020] discuss how generative adversarial networks (GANs) are energy-based models, Kingma et al. [2021] discuss how variational diffusion models are an infinitely deep limit of VAEs, and Song et al. [2021] discuss how diffusion models can be transformed into flows.

2.2 Variational autoencoders

VAEs are latent variable models, and they learn to model the distribution of data through amortised approximate inference [Kingma and Welling, 2014; Rezende et al., 2014]. In particular, VAEs use deep learning models to parametrise the variational distributions. A VAE optimises an evidence lower bound (ELBO) for a data point $\boldsymbol{x}$:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &\geq \mathcal{L}(\boldsymbol{x};\varphi,\theta)\\ &= \mathbb{E}_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log p_\theta\left(\boldsymbol{x}|\mathbf{z}\right)\right] - D_{\mathrm{KL}}(q_\varphi(\mathbf{z}|\boldsymbol{x})||p(\mathbf{z})) = -(D+R). \quad (2.1)\end{aligned}$$

Here, the parameters for an encoder, or inference network, are denoted by φ, and parameters for the decoder, or generative network, θ. The approximate posterior is $q_\varphi(\mathbf{z}|\boldsymbol{x})$ (approximating a true, but generally intractable posterior, $p_\theta(\mathbf{z}|\boldsymbol{x})$, over the latent variable $\mathbf{z}$), and $p(\mathbf{z})$ denotes the used prior. The bound can be expressed as the negation of the summed distortion D—i.e, the negative log-likelihood (NLL) of the input—and the divergence between the approximate posterior and the prior, also called the rate, R. The Kullback-Leibler divergence (KL) divergence, D_{KL}, for some distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ is defined as:

$$D_{KL}(p(\mathbf{x})||q(\mathbf{x})) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} = \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \geq 0 \quad (2.2)$$

In simple VAEs, the approximate posterior and prior used are a Gaussian distribution, allowing for an analytical expression in evaluating the KL.

2.3 Rate-distortion analysis

Optimising the unmodified ELBO produces a model that will, generally, attain the best likelihoods—even if optimisation procedures that, e.g., anneal the rate term in or introduce some level of freedom with "free bits" might improve learning. However, optimising a different balance between the rate and distortion by re-weighting the rate with some factor β allows models to explore other rate-distortion (RD) trade-offs [Higgins et al., 2017]. For instance, increasing the weight of the rate-term can produce, in some sense, more disentangled latents [Burgess et al., 2018]—if an isotropic Gaussian prior is used, this effectively penalises any co-variance between dimensions more strictly, thus producing more independent latent dimensions. Such trade-off considerations, and RD analysis [Alemi et al., 2018, 2017; Poole et al., 2019], provides a framework for analysing the learnt representations of VAEs. In Alemi et al. [2018], it is shown how models

reside on an RD plane and that this plane is divided into different regions. One region is "infeasible" because no models can attain rate-distortion trade-offs that sum to the values in this region. This region is bounded by a line between the level corresponding to the entropy of the data on each of the axes (when considering discrete data), since a model cannot, on average, attain distortions lower than the entropy of the data. Beyond the infeasible region is a feasible region wherein the rate-distortion trade-offs could be achieved, in principle. However, the finite capacity of models results in one last region of "realisable" models, wherein all models will be for a given capacity. Depending on the specific model's encoding and decoding capacities, the model will fall along a particular RD curve within this realisable region. The model would move along this curve when varying a weight β on the rate during optimisation.

Kingma and Welling [2019] provide a more general introduction to VAEs and their various extensions. Such extensions include, for example, using the unsupervised objective alongside a supervised one in semi-supervised learning [Kingma et al., 2014], modelling with more expressive priors [Chen et al., 2017; Kingma et al., 2016; Tomczak and Welling, 2018], optimising a tighter bound on the likelihood [Burda et al., 2015], and using consistency regularisation [Sinha and Dieng, 2021]. In particular, the contributions considered in this thesis will build on three specific VAE constructions: a very deep variational autoencoder (VD-VAE), partial variational autoencoder (P-VAE) and a latent neural stochastic differential equations (SDEs).

2.4 Hierarchical variational autoencoders

Various hierarchical VAEs use a hierarchy of latent variables to produce more expressive models, alongside other improvements [Child, 2021; Hazami et al., 2022; Maaløe et al., 2019; Salimans, 2016; Sønderby et al., 2016; Vahdat and Kautz, 2020]. Specifically, in VD-VAEs, a hierarchy of N latent variables is designed, $\mathbf{z} = \{\mathbf{z}_0, \ldots, \mathbf{z}_N\}$. The VD-VAE uses a particular bottom-up and top-down architecture and residual blocks to parametrise the approximate posterior and prior distributions. The latent variables are structured such that a top-most variable, $\mathbf{z}_0$, is independent of the other latent variables. All other latent variables in the hierarchy depend on the latents above them. The prior and approximate posterior take the following forms, respectively (using $\mathbf{z}_{<N}$ to denote latents higher up in the hierarchy than N):

$$p_\theta\left(\mathbf{z}_0\right) p_\theta\left(\mathbf{z}_1\right)|\mathbf{z}_0), \ldots, p_\theta\left(\mathbf{z}_N|\mathbf{z}_{<N}\right) \tag{2.3}$$

$$q_\varphi\left(\mathbf{z}_0|\boldsymbol{x}\right) q_\varphi\left(\mathbf{z}_1|\mathbf{z}_0, \boldsymbol{x}\right) \ldots q_\varphi\left(\mathbf{z}_N|\mathbf{z}_{<N}, \boldsymbol{x}\right). \tag{2.4}$$

The VD-VAE ELBO, $\mathcal{L}_{\text{VD}}$, becomes similar to that in (2.1), but now includes a sum of KL divergences:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &\geq \mathcal{L}_{\text{VD}}(\theta, \varphi; \boldsymbol{x}) \\ &= \mathbb{E}_{q_\varphi(\mathbf{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\mathbf{z})] - \sum_{n=1}^{N} D_{\text{KL}}(q_\varphi(\mathbf{z}_n|\mathbf{z}_{<n}, \boldsymbol{x})||p_\theta(\mathbf{z}_n|\mathbf{z}_{<n})) \\ &\qquad - D_{\text{KL}}(q_\varphi(\mathbf{z}_0|\boldsymbol{x})||p_\theta(\mathbf{z}_0)). \quad (2.5)\end{aligned}$$

2.5 Missingness

Another extension that will be considered in this thesis is VAEs that can handle missing, or partially observed, data [Ipsen et al., 2021; Ma et al., 2018, 2019; Mattei and Frellsen, 2019; Nazábal et al., 2020]—specifically, the P-VAE introduced by Ma et al. [2018] and an incorporation of the improved way of embedding an identity of dimensions (introduced in Ma et al. [2019]). In these models, the input datum is partially observed, $\boldsymbol{x}_{\mathcal{O}}$. The dimensions are split into a set of observed, $x_o, o \in \mathcal{O}$, and unobserved, $\mathcal{U}$ dimensions. For such models, a partial ELBO, $\mathcal{L}_p$, is optimised, which, again, resembles the standard ELBO in (2.1), but now with the replacement of the "fully observed" distortion with a partial distortion over only observed dimensions and, similarly, replacing the fully observed approximate posterior with an encoding conditioned only on the partially observed datum:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}_{\mathcal{O}}) &\geq \mathcal{L}_p(\theta, \varphi; \boldsymbol{x}) \\ &= \mathbb{E}_{q_\varphi(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}})}\left[\sum_{o \in \mathcal{O}} \log p_\theta(x_o|\mathbf{z}) - D_{KL}(q_\varphi(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}})||p(\mathbf{z}))\right]. \quad (2.6)\end{aligned}$$

The encoding handles partial data by using a summing operation to be invariant to the number of observed dimensions and uses an embedding procedure to learn "identifiers" of the individual data dimensions—this could be one-hot encodings of the data dimension index, but learning them improves performance. When trained, this model can encode a partial datum, and the generative network can produce distribution over all dimensions, including the unobserved ones, $\mathcal{U}$. The uncertainty in these predictive distributions can be used as an acquisition function in active learning.

2.6 Dynamical systems in deep learning

The last type of VAE extension relies on a dynamical systems view of latent objects. Making use of a dynamic systems perspective in deep learning has proven valuable. For example, the approach of interpreting existing systems as a discretisation of a dynamical system enables residual networks to be understood as Euler discretisations of an ordinary differential equation (ODE) [Haber and Ruthotto, 2017; He et al., 2016]. The view can be used to understand existing systems, to improve upon them, or to create new discrete systems by discretising continuous systems with desired properties, both in supervised learning approaches [Chang et al., 2018; Chen et al., 2018; Erichson et al., 2021] and in generative modelling [Grathwohl et al., 2018; Hodgkinson et al., 2021]. In particular, the view can also be used to learnt model representations that display continuity, both in a modelling temporal phenomena [Krishnapriyan et al., 2022] and continuity over the layers, or "depth", of a model [Queiruga et al., 2021, 2020; Xu et al., 2022].

Differential equations that have, e.g., their rate of change (and initial values) defined by learnt neural networks are called neural (ordinary) differential equations [Kidger, 2021]. Neural differential equations constitute expressive modelling components that also build dynamical systems properties into the learnt repsentations—for example, enabling that a state can be evolved from any given point using the learnt dynamics (i.e., the dynamics parametrised by a neural network given the differential equation's rate of change). Neural ODEs have been extended to also use SDEs [Øksendal, 2013], thus producing neural SDEs [Jia and Benson, 2019; Kidger et al., 2021a; Tzen and Raginsky, 2019] [Kidger et al., 2021b; Look et al., 2022], and even neural stochastic partial differential equations [Salvi and Lemercier, 2021].

The VAE perspective on neural SDEs comes from using the neural SDEs as latent objects [Li et al., 2020], for instance replacing discrete Gaussian distributions with stochastic processes defined as solutions to SDEs that, in turn, have their evolution defined by neural networks. These neural networks parametrise how the system drifts (a component similar to the rate of change in an ODE) and how the system is affected by stochasticity from a random Brownian motion, called a diffusion. The following aligns with the presentation of latent SDEs given by Li et al. [2020]. Processes are considered over some time horizon, $\mathbb{T} = [0, T]$. In order to do variational inference (VI), two processes are defined to learn a neural latent SDE. The processes share a diffusion term controlled by a k-dimensional Brownian motion, $\{W_t\}_{t\in\mathbb{T}}$:

$$\mathrm{d}\tilde{Z}_t = h_\xi(\tilde{Z}_t, t)\mathrm{d}t + \sigma_\psi(\tilde{Z}_t, t)\mathrm{d}W_t, \tag{2.7}$$

$$\mathrm{d}Z_t = h_\zeta(Z_t, t, c)\mathrm{d}t + \sigma_\psi(Z_t, t)\mathrm{d}W_t. \tag{2.8}$$

The first differential equation defines the prior process, and the second defines the approximate posterior process. That is, $\{\tilde{Z}_t\}_{t\in\mathbb{T}}$ is the prior process and $\{Z_t\}_{t\in\mathbb{T}}$ is the approximate posterior process. Here, the processes are of dimensionality k. The prior drift is parametrised by a function $h_\xi : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}^k$, which takes as input the state and the current time and produces the drift (a deterministic rate of change of the state). The approximate posterior drift $h_\zeta : \mathbb{R}^k \times \mathbb{R} \times \mathbb{R}^{k_c} \to \mathbb{R}^k$ is similar but takes an extra data-dependent input, $c \in \mathbb{R}^{k_c}$. Finally, the two processes share the same diffusion, $\sigma_\psi : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}^k$. Because the processes share the same diffusion (and when some other regularity conditions are met, see Li et al. [2020] for a derivation of the ELBO relying on Girsanov's theorem), the training of latent neural SDEs amounts to optimising a latent neural SDE ELBO, $\mathcal{L}_{\text{SDE}}$:

$$\begin{aligned}\log p_{\xi,\psi}\left(\boldsymbol{x}\right) &\geq \mathcal{L}_{\text{SDE}}\left(\varphi, \xi, \zeta, \psi; \boldsymbol{x}\right) \\ &= \mathbb{E}\left[\log p_{\xi,\psi}\left(\boldsymbol{x}|\boldsymbol{z}\right) - \int_0^T \frac{1}{2}|u\left(\boldsymbol{z}_t, t\right)|^2 \mathrm{d}t - \mathrm{KL}\left(q_\varphi(\mathbf{z}_0|\boldsymbol{x})||p(\mathbf{z}_0)\right)\right]. \quad (2.9)\end{aligned}$$

Here, the ELBO retains a "regular" KL term resulting from the parametrisation of an initial value distribution using φ as parameters for a network parametrising the posterior initial value. The expectation is under the approximate posterior process distribution as well as under the approximate posterior initial value distribution. The integrand, u, is defined through $\sigma_\psi\left(\boldsymbol{z}_t, t\right) u\left(\boldsymbol{z}_t, t\right) = h_\zeta\left(\boldsymbol{z}_t, t, c\right) - h_\xi\left(\boldsymbol{z}_t, t\right)$. This component in the loss measures how different the two processes' drifts are (under a scale defined by the diffusion term). This integral is the KL divergence between the approximate posterior process and the prior process and can be referred to as a path rate. The more information from the data-dependent signal, c, is used to affect the approximate posterior process, the more the drifts will differ and thus incur a path rate.

2.7 Applications

In this section, two applications that the contributions will explore are presented, namely speaker separation and active learning of audiograms.

2.7.1 Speaker separation

In speaker separation, the task is to estimate component speech from a (noisy) mixture of multiple speakers. Systems solving this task, also referred to as the cocktail party problem [Cherry, 1953], has seen improvements by deep learning

approaches [Hershey et al., 2016; Nachmani et al., 2020; Wang and Chen, 2018; Zeghidour and Grangier, 2021] and by posing the problem of separation as a deep supervised learning problem.

A series of modern deep learning speaker separation networks build on time-domain audio separation networks (TasNets) [Luo and Mesgarani, 2018, 2019]. These models follow an encoder-masker-decoder structure: an encoder network takes as input a mixture and produces an encoding, which is parsed to a masker. The masker, based on the encodings, produces a set of masks matching the number of estimated speakers the model is trained to produce. The maskings are then applied to the original mixture encoding by elementwise multiplication, and the resulting masked encodings are decoded to produce estimates of the component speech. The original TasNet introduced in Luo and Mesgarani [2018] has been extended in various ways [Chen et al., 2020; Luo et al., 2020; Subakan et al., 2021; Tzinis et al., 2020], notably by a variant replacing a recurrent neural network in the original model with a convolutional network [Luo and Mesgarani, 2019]. While these models perform well in domains seen during training, poorer generalisation to new, unseen domains remains a barrier to their real-world adoption [Cosentino et al., 2020; Kadioglu et al., 2020].

These speaker separation models are trained to optimize a scale-invariant signal distortion ratio (SI-SDR) [Le Roux et al., 2019] (comparing a ground truth source, s, and an estimated source $\hat{s}$):

$$\text{SI-SDR}(s, \hat{s}) = 10 \log_{10} \left(||\alpha s||^2 / ||\alpha s - \hat{s}||^2 \right), \quad \alpha = \hat{s}^\top s / ||s||^2. \tag{2.10}$$

Training with this objective makes the loss invariant to differences in the overall power of the estimate time series, and using this loss, as opposed to, e.g., a root-mean-square error (RMSE) loss, is central for training the models well [Heitkaemper et al., 2020]. Other improvements to the training procedures for these models include, e.g., permutation invariant training (PIT) [Yu et al., 2017], which matches estimated sources to ground truth references, and mixture invariant training (MixIT) and extensions [Tzinis et al., 2022; Wisdom et al., 2020], which uses a similar approach to enable learning from mixtures by constructing mixtures-of-mixtures.

2.7.2 Audiograms and active learning

Audiograms consist of measurements of a person's audible thresholds ("how well they hear") at a given set of frequencies. An audiogram can be used diagnostically since it describes a person's hearing and shows, for example, the type and extent of potential hearing loss [Moore, 2012, Chap 2]. Furthermore, using the measurements of lost audibility, an audiogram can be used prescriptively in hearing loss treatment with hearing aids to set the levels of needed hearing

aid compensation (frequency-specific gains) [Kates, 2008]. Improving efficiency (through partial automation and measurement assistance, or even full automation) of the acquisition of audiograms would help increase the amount people with access to hearing-related health care [Mahomed et al., 2013; Margolis and Morgan, 2008].

A common active learning problem called a pool-based active learning cycle [Settles, 2009] determines which data from a pool of unlabelled data should be labelled so as to improve a supervised learning problem the most with the new labels. A similar problem aims to determine which dimensions of a partially observed datum—in other words: from a "pool of unobserved dimensions"—should be observed so as to improve some prediction based on the partially observed datum. In Ma et al. [2019], such variable-wise active learning per data instance is considered, showing, for example, how a P-VAE can be used to do active learning in a risk assessment task on medical data. Acquisition of audiograms using active learning approaches has previously relied on Gaussian processes [Gardner et al., 2015; Schlittenlacher et al., 2018; Song et al., 2015]. These approaches use previously acquired dimensions to inform the acquisition of further dimensions. In these, a Gaussian process regressor, for instance, estimates a function characterising a person's audibility as a function of frequency. Using this estimated function, the model then uses its point of maximal entropy to guide acquisition.

CHAPTER 3

Contribution summaries

3.1 VAEcquisition

The following is a summary of the first contribution, **VAEcquisition**, titled "Rate-Distortion Trade-offs in Variational Autoencoder Representations for Sequential Acquisition Active Learning". The contribution, in its full extent, is provided in Appendix A.

The contribution studies model-based active learning using P-VAEs [Ma et al., 2019] to reason about partially observed data in sequential acquisition of unobserved dimensions. For the active learning process, the study investigates how acquisition estimation performance (estimating full data from partial observations) and uncertainty quantification (early termination of the acquisition process) is affected by trade-offs in rate-distortion of the used representation. The experiments consider two types of data. The first is synthetic data sets constructed using archetypal analysis [Cutler and Breiman, 1994] to mimic core aspects of audiogram data in a controlled manner. The second type is real audiogram data sets from both the United States and Germany, the NHANES [CDC, 1999–2022] and the HÖRSTAT [von Gablenz and Holube, 2015] data sets, respectively. Given these data sets consisting of fully observed data, missing-at-random missingness is imposed on the data during training. In this way, a partial representation can be learnt using the P-VAE. Experimental results presented are for models that optimise the partial ELBO using an adaptive re-weighting of the rate-term, such that the trade-off between the rate and the partial distortion achieves a range of pre-specified target rates.

Given partial observations, the trained models can produce estimates of distribution over the full observations. This is done by estimating an encoding of the partial datum using a partial encoder, and subsequently using the generative network to produce distribution over a fully observed datum from the encoding. The model evaluation includes the performance of the acquisition estimation, which measures how well estimates of full data match a ground truth reference. When these error metrics are summed over all steps in an acquisition process (i.e., going from a fully unobserved datum to a fully observed datum in a sequential acquisition process), a single scalar value is produced. This scalar, or "the area under the information curve", summarises the acquisition process's performance.

While the model quantifies its uncertainty in reconstruction with the partial distortions and the rate of the partial encoding (producing the partial ELBO), the uncertainty in the predictive distribution of unseen dimensions allows the model to inform an acquisition process. By acquiring the dimension with the maximal variance in the predictive distribution, the experimental results show that the acquisition process outperforms (in attained areas under the information curves) both a random acquisition and a "single best" greedily optimised ordering.

The study shows how uncertainty quantification enables an estimation of the

model's error without access to the ground truth reference. Specifically, the combined uncertainty quantification of the partial ELBO components (i.e., the rate and the partial distortion) and an estimate of the predictive distribution entropy in unseen dimensions form the basis for a linear estimate of the VAEs error in estimating the full audiogram. Given a predefined error threshold, such an estimate can be used to terminate the acquisition process. The uncertainty quantification performance of the linear model is compared to a simple model that stops at a single best index during acquisition across all sequences. The experimental results show that a model that uses uncertainty quantifications can estimate the model's error better than the baseline model. Since the number of measurements needed to achieve this set level of accuracy changes with the specifics of an audiogram, the procedure thus allows for datum-specific adaptation. This procedure is evaluated using both the absolute offset (i.e., the offset to a model which used a stopping index based on ground truth estimation errors) and the calibration of the average stopping index error (whether the model, on average, stops at approximately the desired level of accuracy). In these evaluations, using the datum-specific uncertainty is an improvement over the baseline single best stopping index.

Finally, the learnt representations' abilities to inform full audiogram estimation and to inform the stopping procedure are compared to the representations' rate-distortion trade-off. In synthetic data, small or large corruptions can be introduced to the archetypal components that define the training data sets underlying generation mechanism. When considering the area under the information curve, evaluation on a data set generated from small corruptions resulted in optimal rates that were around, or slightly higher than, the optimal rate for an uncorrupted data set. In contrast, large corruptions resulted in optimal rates that were much lower than the optimal rate in the uncorrupted setting. On the real data sets, the generalisation from NHANES to HÖRSTAT displayed behaviour more aligned with the results for the synthetic results with smaller corruptions. Importantly, the RD analyses showed how the optimal rates were not the same across evaluation metrics and data sets, thus showing how a choice must be made in favour of either, e.g., high performance in estimating the full observations or having high quality stopping estimation. Notably, the optimal rates did not invariably align with the optimal ELBO.

3.2 VI-EMD

The following is a summary of the second contribution **VI-EMD**, titled "Improving Speaker Separation Generalization with Variational Inference". The contribution, in its full extent, is provided in Appendix B.

The primary contribution of this study is to re-cast these encoder-masker-decoder networks in a VAE-framework. Learning encodings and masks in the encoder-masker-decoder networks are formulated as a hierarchical latent variable model and optimised using VI. The study provides a detailed construction of speaker separation in the context of VI and introduces an ELBO for modelling component speech.

The deterministic versions of the models rely on a SI-SDR-based objective. In order to allow the probabilistic framework to learn under the same scale-invariance, the work introduces a scale-invariant observation model using Bayesian linear regression (BLR). In this, the differences in scale between a target speech time series and an estimated time series is seen as a regression coefficient which can be analytically inferred and corrected for, thus allowing for efficient training of VI-EMDs. The study provides a comparison to the standard SI-SDR objective, showing that the SI-SDR objective and its probabilistic counterpart optimise comparable quantities.

The experiments show results for both synthetic data sets and for real data sets. As a controlled setting, the study introduces a Gaussian pulse data set in which simulated speech utterances are constructed as Gaussian pulses with "speaker"-specific frequency ranges. Furthermore, two data sets were considered: LibriMix [Cosentino et al., 2020], which is used as the seen, training domain, and VCTK [Yamagishi et al., 2019], which is used as the unseen domain. A detailed discussion of the data foundation is provided, discussing, e.g., the consequences of realistic speech material and noise models.

The VAE framework enables the analysis of RD trade-offs, where the separation performance can be interpreted as an auto-encoding distortion of the component speech in the input mixture. The study presents experiments that show how optimising the rate alongside the component speech distortion produces models that generalise better to unseen domains than their deterministic counterparts. This is seen the original convolutional time-domain audio separation network (Conv-TasNet) [Luo and Mesgarani, 2019] and its VI counterpart, and it, similarly, seen in the improved, successor successive downsampling and resampling of multi-resolution feature (SuDoRMRF) [Tzinis et al., 2020] and its VI counterpart. The results indicate that the deterministic models, which solely optimise the speaker separation objective—i.e., without regard to the resultant rate—generalise poorly. This can be interpreted as a consequence, in part, of a sub-optimal RD trade-off in the learnt representation.

The study compares training VI-EMDs under various priors, including standard Gaussian priors, log-normal priors, Gamma priors, and learnt autoregressive flow priors. Furthermore, a multitasking VI-EMD is presented, which not only optimises the speaker separation but also optimises an auto-encoder task of the input mixture in a semi-supervised learning setup. The multitasking framework

allows for the specification of a prior inducing that mixture encodings resemble known (during training) component speech encodings. Notably, the multitasking model is also a generative model of the input mixture, such that an auto-encoding ELBO can be determined even without access to ground truth component speech signals. Experiments show how this input mixture density is informative of the separation performance, thus allowing for knowing when the model is capable of separating speakers and when the model's uncertainty about its input is too high to produce good separation.

3.3 CD-VAE

The following is a summary of the third contribution **CD-VAE**, titled "Hierarchical Variational Auto-Encoders using Latent Neural Stochastic Differential Equations". The contribution, in its full extent, is provided in Appendix C.

Deep hierarchical VAEs are competitive density estimation models [Child, 2021; Sønderby et al., 2016]. In particular, Child [2021] showed how increasing the stochastic depth improves the performance of VD-VAEs, i.e. using a deeper hierarchy while otherwise keeping the size of the model constant in terms of parameter count. This contribution constructs a continuously deep VAE, building on the VD-VAEs. The model replaces discrete Gaussian variables in the hierarchy formulation with latent stochastic processes defined by neural SDEs. In so doing, the model relies on numerical SDE solvers to estimate the solutions to the neural SDEs that define the latent processes.

Experimental results are provided for training the model on a simple synthetic Poisson equation-based data set and on binarized MNIST. The experimental evaluation of the CD-VAE focuses on investigating its continuity properties. For both data sets, the experiments investigate the ELBO achieved when varying the integration step sizes after training. Results show that the CD-VAE learns a depth-continuous representation in that it generalises well to step sizes different from the constant step size used in the SDE solvers during training. The CD-VAE is compared to a version of the model that does not optimise how well-behaved the latent stochastic processes are with respect to a prior process (an "ODE-like" CD-VAE); such a model does not display the same properties of depth-continuity and instead displays an optimum in ELBO around the step size used during training.

This exploration of continuity contributes to an building body of works that explore how depth-continuity in model representations can be achieved by using dynamical system components in deep learning [Queiruga et al., 2021, 2020; Xu et al., 2022] and continuity in representations, in general [Krishnapriyan et al., 2022]. Furthermore, the model adds to ongoing efforts that can be said to lie

in the intersection of various modern deep generative frameworks, incorporating elements (SDEs) that are central to diffusion models [Kingma et al., 2021] and continuous normalizing flows [Chen et al., 2018].

CHAPTER 4

Discussion & Conclusion

Before considering the overarching questions underlying all three contributions that were presented in Chapter 1, the contribution-specific questions are addressed based on the summaries in the previous chapter.

Partial VAEs allow efficient audiogram acquisition In the first contribution, the question of audiogram acquisition was considered, asking:

> **VAEcquisition**: **Audiogram Acquisition** *How can we utilise a learnt variational autoencoder representation to efficiently acquire audiograms?*

By using an architecture that can encode partial data, a P-VAE, this contribution shows how a learnt representation of audiograms can be used to acquire audiograms efficiently by:

1. making use of the predictive distribution uncertainty resulting from partial encodings to inform acquisition, and

2. utilising the probabilistic framework's uncertainty quantification in estimating the model's prediction quality, allowing the model to terminate acquisition when it is estimated that prediction quality is sufficiently good.

Both of these points allow for a model-based active learning procedure more efficient than baseline models that do not use datum-specific uncertainty in the acquisition (such as single best ordering acquisition approaches or single best stopping indices).

Variational inference improves speaker separation generalisation The second contribution addressed the generalisation of speaker separation models, asking:

> **VI-EMD**: **Speaker separation** *How does recasting encoder-masker-decoder speaker separation models in a variational autoencoder framework affect generalisation?*

In order to answer this question, the contribution constructs VI-EMDs and introduces the components needed for a VI-based encoder-masker-decoder architecture similar to the existing models (like TasNets), including introducing a scale-invariant observation model, a latent hierarchy formulation of the encoding and masking outputs, and a speaker separation ELBO.

The contribution discusses how the probabilistic approach learns to characterise, and be robust to, small perturbations in the representation, which provides a perspective on how the VAE framework improves generalisation. More importantly, the study shows how optimisation of encoding and mask rates improves generalisation to new domains. The VAE perspective thus provides a rate-distortion trade-off explanation for—some aspects of—the generalisation gap in deterministic speaker separation models that do not explicitly consider the information rates of the encodings and the applied masks.

Furthermore, the VAE-formulation of the speaker separation problem readily enables a semi-supervised learning approach. The contribution shows how to train a multitasking system that jointly learns to model input mixtures and target component speech. This system provides a basis for systems that can harness the learning signal from generative modelling of arbitrary audio input in improving speaker separation performance. Even the largest speaker separation data sets are smaller than unstructured data sets of audio. Notably, speaker separation datasets often rely on some simulated approach in order to have access to ground truth component signals (i.e., simulated mixtures from isolated speech recordings mixed with noise). However, with the multitasking VI-EMD, the model can let separation in such data sets be informed by learning from realistic audio mixtures (i.e., real recordings of multiple speakers in noisy environments) that would otherwise not be included given that lack of the component speech recordings. This aspect of the VI-EMDs addresses a central limitation (only learning from data sets where component speech is available) of existing frameworks and is thus a promising avenue for improving generalisation by learning from more realistic audio.

Hierarchical VAEs as continuous dynamical systems The last contribution centred around two questions. The first question considered a dynamical systems formulation of hierarchies, asking:

> **CD-VAE**: **Hierarchies as Dynamical Systems** *How can we interpret the latent representations in deep hierarchical variational autoencoders as stochastically evolving dynamical systems?*

One way to interpret latent representations in deep hierarchical VAEs as stochastically evolving dynamical systems is to replace the—usually—discrete Gaussian distributions with stochastic processes. This can be done by making use of neural stochastic differential equations to define the latent processes and using differentiable numerical integration frameworks to produce solutions to the differential equations and to provide the gradients needed during training. In this, the contribution builds on existing work on latent neural SDEs. This study,

however, contributes the application of this dynamical systems perspective in re-interpreting a specific, successful deep hierarchical VAE architecture called VD-VAEs. To interpret a VD-VAE's latents as a stochastically evolving dynamical system, the existing neural latent SDE formulation is expanded in two main ways: (i) by introducing a hierarchy of SDEs that operate on different spatial resolutions in the latent processes and (ii) by introducing convolutional architectures for the drift functions in the latent processes. The latter is needed in order to match the VD-VAE, but, more importantly, it is also required to efficiently scale learning of the processes to the pre-requisite high-dimensional latent processes. The introduced model, CD-VAE, constitutes a deep hierarchical VAE wherein the latent representations are a stochastically evolving dynamical system.

The second question was concerned with continuity, asking:

> **CD-VAE**: **Continuity** *What are the benefits, if any, of a continuous-depth formulation of hierarchical variational auto-encoders using neural stochastic differential equations?*

In modelling simple image problems, such as a binarised version of MNIST, the experimental results in this contribution show how the representation learnt by a CD-VAE can be evaluated with different granularities of step sizes in the numerical integration used in determining the latent process evolutions. Specifically, the representation generalises in such a way as to benefit from a smaller step size than seen during training, thus achieving a lower ELBO when increasing the computational budget. Furthermore, the contribution discusses how inducing continuity in a model representation, for instance: (i) allows the use of adaptive step size solvers during training and post-training application, (ii) enables the use of the stochastic adjoint method to increase the depth of the hierarchy at constant memory complexity, and (iii) allows for an efficient parametrisation of functions by sharing weights across the depth of the model. Ensuring continuity of the learnt representation might also be a useful inductive bias for improving generalisation in the same way that more simple forms of regularisation improve generalisation.

Variational inference and uncertainty quantification in representation learning Central to all contributions was the use of variational inference, and each contribution explores quantified uncertainty in some manner. In Chapter 1, the concept of usefulness in representations was discussed, where it was argued that a useful representation is characterised by both generalising well and providing relevant information for a considered task. The question of usefulness in the context of VI and uncertainty quantification (UQ) was posed as:

Variational Inference and Uncertainty Quantification *How can variational inference and uncertainty quantification in representation learning be used to improve the usefulness of a learnt representation?*

Jointly, the studies show, in significantly different applications, how VAE representations can ensure that more useful representations are learnt. For example, in **VI-EMD**, it was shown how generalisation to the unseen dataset VCTK of the speaker separation performance improved when using a VI-based approach, thus improving the generalisation aspect of the representations' usefulness. In both **VI-EMD** and **VAEcquisition**, it was shown how UQ enables estimation of a model's performance on a considered task (speaker separation and audiogram acquisition, respectively) without access to ground truth references. That is, the learnt representations had learnt to characterise uncertainty in a manner relevant to the considered downstream task—in these examples, these aspects of UQ thus contribute to the relevance aspect of usefulness. Furthermore, the uncertainty quantification of the P-VAE's predictive distributions based on the partial encodings is the main component allowing for efficient audiogram acquisition in **VAEcquisition**; similarly, the variational inference is the main building block that allows the **VI-EMD** to learn representations that are more robust than its deterministic counterparts; and, VI is the framework that directly enables the CD-VAEs to learn well-behaved stochastic processes with meaningful (non-degenerate) diffusion behaviours in **CD-VAE**. The UQ applications estimating the quality of a given model's downstream task performance are a vital component in making highly expressive models like deep learning systems viable for use in real-world settings; even if they "do not fail gracefully", the example UQ aspects presented in **VI-EMD** and **VAEcquisition** show the feasibility of knowing when the system cannot be trusted such that a user might fall back on a more reliable, but less expressive, model (in **VI-EMD**) or such that an acquisition process should continue for a longer time (in **VAEcquisition**).

Rate-distortion analysis Each contribution trained models that optimised an ELBO, thus considering some trade-off of rate and distortion. Regarding the usefulness of a learnt representation, it was asked:

Rate-Distortion Analysis *What insights arise by characterising a representation's usefulness from the perspective of rate-distortion trade-offs analysis?*

Through rate-distortion analysis of learnt representations, and especially how

these trade-offs affected downstream tasks, the contributions provide experimental results that showed, firstly, how optimising distortion alone does not provide useful representations—even if the representation generalises, the information coded for is not relevant to the downstream tasks. Examples of this include the high-rate, low-distortion models with no train-test gap shown in **VAEcquisition**'s Figure 9. The information that produces the low distortions generalises to the test set, yet the information is not just irrelevant but detrimental to the model's ability to aid acquisition.

Secondly, while optimising an unmodified ELBO, in some sense equally balancing the rate and distortion of the learnt representation, produces the best likelihood models, such a trade-off is not necessarily conducive to improving downstream task performance. For example, in **VI-EMD**, models that balanced the rate and distortion equally could not, to any discernable degree, learn how to do speaker separation but required much higher rates to achieve competitive performance. Similarly, the examples in **VAEcquisition** show how the optimal rates for various metrics of downstream performance did not in all cases align with the optimal ELBO.

Thirdly, the results in **VI-EMD** on various priors and **CD-VAE** on the latent process path rate show how the rate-distortion perspective is intrinsically linked to the choices made for priors. For example, rate-distortion trade-offs in a VI-EMD that uses a Gaussian approximate posterior and another that uses a log-normal, or a Gamma, approximate posterior will behave in drastically different ways, despite potentially incurring an equal number of nats in the loss from the resulting rate-term. Similarly, while an "ODE-like" **CD-VAE** might incur overall rates similar to a corresponding **CD-VAE**, their properties in terms of continuity, and consequent usefulness, are very different. The rate-distortion trade-offs are not directly comparable between different approximate posterior and prior distributions (or processes).

The rate-distortion perspective, alongside the probabilistic framework, provides a principled way to impose characteristics on representations through the loss during optimisation or through construction/architectural biases, both of which can be thought of as inductive biases allowing improved learning. For example, the loss incurred in a CD-VAE by behaving dissimilar from a simple stochastic process uninformed by data ensures that the model learns a well-behaved continuous representation through the loss. Similarly, the choice of a log-normal approximate posterior for the latent masking variable in a VI-EMD is a principled way to—directly through the model's construction—ensure that the masks are non-negative. Another example of an architectural, inductive bias explored in this work is the separation-by-masking used in VI-EMDs, guiding the models to learn to separate by encoding the mixture in such a way that simple element-wise masking is sufficient for separation.

A view of the models considered in the contributions is that they attempt to both (i) find optimal rate-distortion trade-offs given some set expressivity of the model (this is especially the case for **VAEcquisition** and **VI-EMD**) and (ii) increase the expressivity of the model (this is especially the case for **CD-VAE**). In the RD analysis view presented in Alemi et al. [2018], the former amounts to moving along a set given RD curve to improve the usefulness of the learnt representation and the latter amounts to trying to construct models that push the encoding and decoding capabilities, thus expanding the region of "realisable models".

Underlying the approaches in this work is some notion of efficiency in learning, such as forcing a model to prioritise what information to code for given some set model expressivity (and computational budget) within a given informational rate budget. This approach is a different approach to representation learning than that of recent large representation learning models, or foundation models [Bommasani et al., 2021], such as large language models like GPT-3 [Brown et al., 2020], large audio models like wav2vec and GLSM [Baevski et al., 2020; Lakhotia et al., 2021], and large language/image-models like DALL-E [Ramesh et al., 2022]. Simplistically, large foundation models increase model expressivity greatly and with the scaling comes emergent properties allowing the learnt representations to solve tasks with little to no supervision (for instance, through prompts) [Bommasani et al., 2021]. The models considered in this thesis—which are "small" by comparison—approach the problem of representation learning under a much more restrictive model capacity, instead relying on inductive biases (like that of separation-by-masking in VI-EMDs and the use of latent neural SDEs in CD-VAEs). Note that the foundation models do benefit from such inductive biases, too, but rely more prominently on large capacity and large amounts of computational power instead. In the RD perspective, especially relevant since many of the large models share encoder-decoder structures with VAEs, these models might be said to rely on large computational power and model capacity to push the boundary of realisable models; a conjecture is that these models also learn representations with very high rates and select the information after training (e.g., with finetuning, prompt engineering) instead of prioritising during training like the models considered in this work.

As discussed in **CD-VAE**, modern deep generative models like VAEs, flows, diffusion models, and autoregressive models all share common ground, especially when considered from the perspective of deep hierarchies and from an SDEs perspective. Understanding these commonalities and unifying the frameworks have been beneficial and will likely continue to be so. In the context of RD analysis, another central approach within the broader context of representation learning can be incorporated, namely GANs. In RD analysis, the distortion can be thought of as a datum-specific distance between an input and an output

introduced by the composition of an encoding and a decoding function—and the contributions provided in this thesis extensively discuss trade-offs between this distortion and the rate. However, the encoder and decoder also induce some overall (i.e. not datum-specific) distribution of reconstructed signals, which can be different from the actual data distribution (the distribution of data on the input side). Because of this, the rate-distortion trade-offs can be extended to consider a further component which Blau and Michaeli [2019] call perception, resulting in rate-distortion-perception trade-offs. The perception measures a divergence between the input data distribution and the distribution of data from the decoder (independent of a specific datum) [Blau and Michaeli, 2019; Zhang et al., 2021]. The use of "perception" is based on the fact that a model might produce an overall distribution of outputs from the decoder that generally matches the distribution of general types of input, thus having "good perceptual quality". Notably, a model might produce a good perception metric even with very poor distortions; if the model decoder ignores the encoded information yet still produces outputs that "look" like real data (in some sense ignoring an encoder and just having a GAN decoder), the model has high distortion but low perception. In GANs, the divergence between the true data distribution (approximated by an empirical data distribution) and the learnt model's distribution is minimised, e.g. by minimising a Wasserstein distance [Arjovsky et al., 2017]. Blau and Michaeli [2019] shows how, at a given rate, a model can only improve distortion in a trade-off with perception and vice versa. Originally, VAEs produced significantly lower-quality samples (low-quality, blurry images, for instance) in comparison to early GANs. The fact that recent VAEs, such as the VD-VAEs, produce higher quality samples might be seen as a consequence of having increased the models' capabilities in utilising higher rates—pushing the boundary of realisable models in the RD planes, or, equivalently, that early model, having had only low rates available and actively optimised distortions, did so at the expense of poor perceptual quality. Since the contributions in this thesis have shown various views of the utility of RD analysis, the extra aspect of perception trade-offs—and thus incorporating adversarial approaches like GANs into the analysis framework—might prove very useful in producing models with both high perceptual quality and low distortions. In audio modelling and speaker separation, poor perceptual quality is a central aspect of the model evaluations, and improvements might be brought about with rate-distortion-perceptions perspectives in learning representations of speech.

On learning useful variational autoencoder representations

> *What is a representation, what makes a representation useful, and how do we discover useful representations?*

In this thesis, learning useful representations using variational autoencoders has been explored. For one, the usefulness of the learnt representations was demonstrated by their ability to *generalise*—for instance, by improving audio modelling speaker separation performance in new, unseen conditions in **VI-EMD** and in allowing a single model to generalise to new hierarchical depths in **CD-VAE**. For another, the usefulness of the learnt representations was demonstrated by their ability to provide *relevant* information facilitating downstream tasks—for instance, by providing estimates of the model performance using the representations' uncertainty quantification in both **VAEcquisition** and **VI-EMD** and in enabling efficient acquisition of audiograms through predictive distributions from partial data in **VAEcquisition**. The presented studies show how an approach to learning useful representations is variational autoencoders, especially when coupled with consideration of rate-distortion trade-offs, incorporation of inductive biases through the variational inference model construction, and when the representations' uncertainty quantifications are utilised in the considered application.

Appendices

Overview of appendices

In the appendices, the following contributions are provided:

- Appendix A, **VAEcquisition**: "Rate-Distortion Trade-offs in Variational Autoencoder Representations for Sequential Acquisition Active Learning"
- Appendix B, **VI-EMD**: "Improving Speaker Separation Generalization with Variational Inference"
- Appendix C, **CD-VAE**: "Hierarchical Variational Auto-Encoders using Latent Neural Stochastic Differential Equations"

APPENDIX A

VAEcquisition: Rate-Distortion Trade-offs in Variational Autoencoder Representations for Sequential Acquisition Active Learning

Rate-Distortion Trade-offs in Variational Autoencoder Representations for Sequential Acquisition Active Learning

Rasmus M. Th. Høegh[1,2,3,*] *rmth@dtu.dk*

Caspar B. Jespersen[3] *caspar.jespersen@wsa.com*

Lasse L. Mølgaard[3] *lasse.moelgaard@wsa.com*

Abigail A. Kressner[2,4] *aakress@dtu.dk*

Morten Mørup[1] *mmor@dtu.dk*

Jens B. B. Nielsen[3] *jens.nielsen@wsa.com*

[*] *Corresponding author,* [1] *Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Denmark,* [2] *Hearing Systems, Department of Health Technologies, Technical University of Denmark (DTU), Denmark,* [3] *WS Audiology A/S, Denmark,* [4] *Copenhagen Hearing and Balance Center, Ear, Nose, and Throat (ENT) and Audiology Clinic, Rigshospitalet, Copenhagen University Hospital*

Abstract

In this work, we consider the active learning problem of sequentially acquiring unobserved dimensions from a partially observed datum. We inform the acquisition process by a variational autoencoder (VAE) representation learnt on large, readily available data. As in traditional active learning settings, the aim is to learn while balancing, in an informed manner, costly measurements (in time, compute, money, etc.) against their utility. In particular, we consider the application of VAE representations in characterizing a person's hearing loss as a function of frequency. We show that a representation can improve the acquisition process by providing distributions over the full datum given partial observations. VAE representations balance good reconstruction (low distortion) against divergence from a prior (rate), and we study rate-distortion trade-offs for the downstream acquisition performance. We show that the downstream utility of a learnt representation is rate-dependent and that these choices need not coincide with the rate providing the optimal evidence lower bound (ELBO); emphasis on different aspects of the downstream acquisition lead to different choices for the target rate. Furthermore, we show how the model's uncertainty in characterizing observed dimensions (the partial ELBO) and unseen dimensions (the predictive variance) provides information that can enable early termination of the acquisition process given some desired accuracy. Our results point to the importance of tuning the VAE rate with respect to downstream tasks and highlights the utility of the associated VAE uncertainty quantification for decision-making.

1 Introduction

Representation learning aims to learn useful, well-behaved representations. The intent is to harness large amounts of readily available data in learning a representation, such that these learnt representations can, e.g., be repurposed to facilitate downstream tasks or used concurrently to improve learning from scarce or costly labelled data (Kingma & Welling, 2019). This work explores how to use a learnt representation to efficiently learn about partially observed data in an active learning setting.

Variational auto-encoders (VAEs) learn representations by jointly optimizing an encoder and a decoder network (Kingma & Welling, 2014; Rezende et al., 2014). The encoder maps data to a latent space, whereas the

decoder learns to map from the latent space back to the original data space. A VAE is trained by optimizing the evidence lower bound (ELBO). One view is that this amounts to minimizing the *distortions* introduced by the composition of the encoder and decoder function under a constraint on the *rate* of information passed through the latent space. In particular, we are interested in how a variational autoencoder representation is affected by the trade-off between characterizing the observed data well (quantified as a negative log-likelihood, or distortion) while keeping the representation *well-behaved* against some predefined prior—quantified as a Kullback-Leibler divergence (KL), or rate. There is a tension between a "well-behaved" representation with a low rate and a model that captures as much as possible about the data with low distortion. The standard ELBO will strike one balance that directly optimizes the marginal likelihood, but optimizing modified objectives that penalize the rate to a lesser or greater extent—i.e., other rate-distortion (RD) trade-offs—can improve the properties of the learnt representation (Higgins et al., 2017; Alemi et al., 2018).

While models can achieve low distortions at a high informational rate, we explore whether such learnt representations are also useful for acquisition—or whether, conversely, models favouring lower rates at the expense of increased distortions learn representations better suited for the downstream task. In a standard active learning problem, a system determines which datum from a pool of unlabelled data should be labelled next, or a "pool-based active learning cycle" (Settles, 2009). We instead consider the process of sequentially acquiring dimensions of a datum that is initially partially or wholly unobserved. How to sequentially acquire information, variables, or dimensions about a datum in an efficient, informed manner is a problem of general relevance, including medical applications. An exampling of this is determining which tests to use in determining a diagnosis of a patient (Ma et al., 2019). Sequential feature acquisition in this manner has been studied using, for example, Gaussian processes (Song et al., 2015), reinforcement learning procedures (Contardo et al., 2016), and variational autoencoders (Ma et al., 2019).

In this work, we focus specifically on the problem of measuring audiograms in a time-efficient manner. An audiogram is a graphical representation of an individual's audible thresholds as a function of frequency. Traditionally, an audiogram is measured at a given set of frequencies, and taken together, these thresholds jointly characterize the hearing loss, or lack thereof, of a person—see, for instance, Moore (2012, Chap 2, Sec. 3, Fig. 2.3). Beyond its diagnostic purpose, the audiogram is used prescriptively to treat an individual's hearing loss—for instance, by defining frequency-specific gains in a hearing aid to compensate for the loss of audibility (Kates, 2008). Measuring a complete audiogram is a time-consuming process; with a large and increasing need for hearing healthcare services, improving the accuracy and efficiency of how we acquire audiograms becomes an integral part of increasing capacity (Margolis & Morgan, 2008; Mahomed et al., 2013). In practice, experienced clinicians tend to rely on their domain knowledge to hasten the process to determine when it is acceptable and appropriate to measure, for example, only a specific subset of frequencies. We are interested in doing the same in an automated way, such that audiogram acquisition becomes equally as fast, without cost to accuracy, with less experienced clinicians or in fully automated systems.

We are interested in estimating a complete audiogram from as few frequencies as possible. To do this, we define a model that determines which frequencies to measure in an informed, sequential manner to arrive at a sufficiently accurate estimate with as few measurements as possible. We will refer to a model that achieves this a having good estimation performance. The number of measurements needed for a model with good estimation performance will, however, be directly dependent on the desired accuracy. Furthermore, the number of measurements and at which frequencies will vary from individual to individual. Therefore, we are interested in a model capable of quantifying the uncertainty of its estimate for a specific individual. As the acquisition is in progress, we can determine at which point the process can be stopped. We will refer to a model that achieves this as one that has good uncertainty quantification.

The main contribution of this paper is to evaluate acquisition estimation performance and uncertainty quantification as a function of rate-distortion trade-offs in variational autoencoder representations of audiogram data. In particular, we analyze the problem of efficiently acquiring audiograms and explore how rate-dependant qualities of the representation affect (i) the efficiency and accuracy of estimating complete audiograms from partially observed audiograms and (ii) the ability to quantify the uncertainty of the estimation accurately. This we do by considering both synthetic data with ground truth and two openly available audiograms datasets, NHANES (CDC, 1999–2022) and HÖRSTAT (von Gablenz & Holube, 2015).

2 Related work

We build on the partial VAE (Ma et al., 2018), and specifically the PointNet Plus version presented with the "Efficient Dynamic Discovery of High-value Information" (EDDI) framework (Ma et al., 2019). The partial VAE enables learning approximate posteriors for partial data. Additionally, Ma et al. (2019) consider how to efficiently acquire dimensions sequentially to minimize an "area under the information curve" (the accumulated error over all measurements) introducing an information reward acquisition function—and how to estimate this reward efficiently. While Ma et al. (2019) focus on predictions of a target dimension for tabular data, we focus on autoencoding improvements of partially observed data. Ma et al. (2019) consider the target variable separate from the set of possible acquisition dimensions, but in this work we consider an auto-encoding acquisition task, aiming to improve the representation of all unobserved dimensions without an "unobservable" target dimension. We further focus on the sequential process of acquiring audiograms. In particular, we extend the analysis to consider rate-distortion trade-offs on the estimation performance (estimating the full audiogram) and the uncertainty quantification (ability to terminate early accurately).

Extension of the ability of variational autoencoders to handle missing data include the heterogeneous-incomplete VAE (Nazábal et al., 2020), which introduced a model for handling both missing data and data of heterogeneous types (continuous, discrete, etc.). Mattei & Frellsen (2019) introduce the missing data importance weighted bound, and this work was extended to consider more realistic models of the missingness, beyond a missing-at-random process (Ipsen et al., 2021).

Given a data set of fully observed and labelled data and some partially observed data, active feature-value acquisition (AFA), see, e.g., Saar-Tsechansky et al. (2009). determines which partially observed data should be (wholly) acquired given some cost. AFA is related to, but not the same as, the sequential process we consider. We distinguish between acquiring fully observed data points given a pool of partially observed data points (i.e., AFA) and the presently considered problem, which Ma et al. (2019) denote "active variable selection", that sequentially (one at a time) determines more dimensions of a given datum.

We attempt to quantify the uncertainty of full audiogram estimates given partially observed data. Uncertainty in imputation estimates giving missing data can help guide acquisition, for instance, by improving the effectiveness of acquiring labels by characterizing the uncertainty of multiple imputations and using this in an acquisition function (Zheng & Padmanabhan, 2002; Han & Kang, 2021).

Acquiring audiograms efficiently is an active area of study—for a review of automated audiometry, see Mahomed et al. (2013). One set of studies has used Bayesian active learning using Gaussian Process (GP) regression to measure audiograms efficiently (Gardner et al., 2015; Song et al., 2015; Schlittenlacher et al., 2018). Using a GP to model a specific subject's hearing threshold as a (continuous) function of frequency, these approaches use the estimated function's point of maximal variance to guide the acquisition process ("which point to acquire next"). Active learning with GPs has also been used to efficiently learn an end user's individual preference for the setting of hearing-aid parameters used directly to adjust the hearing aids automatically (Nielsen et al., 2014). Such models improve acquisition by letting previously acquired dimensions guide the acquisition of further dimensions. In addition to previously acquired dimensions for a given acquisition, the presently considered approach integrates a learnt representation of real-world audiograms to inform acquisition and produce estimates of the full audiograms, thus integrating information learnt from large data sets of audiograms.

3 Methods

3.1 Data

NHANES and HÖRSTAT data We model audiograms from both NHANES ($N = 12813$) and HÖRSTAT ($N = 2986$). The United States' Centers for Disease Control and Prevention publishes the National Health and Nutrition Examination Survey (NHANES). NHANES includes pure-tone air-conduction audiometry data of participants in the United States (CDC, 1999–2022), and here we use data from 1999-2004, 2011-12, and 2015-16, excluding years that, e.g., considered only specific demographics. The HÖRSTAT study includes similar measurements for a survey done in Germany (von Gablenz & Holube, 2015). While a

larger set of frequencies are occasionally measured, we only consider results for 0.5, 1, 2, 3, 4, 6 and 8 kHz (for both ears) due to consistency across years and data sets. In addition to the measured frequencies, we include the subject's age. Datasets like NHANES include a lot of information about other aspects that might inform the acquisition process (beyond the age, as considered here), including information pertinent to the hearing thresholds (such as information on recent colds or earaches, audiometric tests like tympanometry), general demographics data (level of education, country of birth), and questionnaire data (occupation, physical activity). The VAE framework readily enables the integration of this knowledge in a scalable manner.

Synthetic data For controlled comparisons, we study a generated data set of synthetic audiograms. These audiograms are generated by randomly sampling "archetypal" audiograms. We sample five examples from a 15-dimensional ($7 \cdot 2 + 1$, seven frequencies for each ear, and the age) multivariate normal distribution defined by a mean roughly matching a sloping hearing loss on each ear (poorer hearing at higher frequencies). We construct a covariance matrix which induces relationships between: the age dimension (higher age, higher hearing loss), neighbouring frequencies (similar hearing loss for, for example, 1 kHz and 2 kHz), across ears at the same frequency (1 kHz on the right and left ear co-vary), and overall on the ear (all frequencies on one ear co-vary). Overall this produces archetypal audiograms with statistics roughly matching some essential characteristics of true audiograms in a controlled manner. The archetypes are the basis for generating a dataset of $N = 2^{12} = 4096$ synthetic audiograms using an archetypal analysis generation mechanism (Cutler & Breiman, 1994) resulting in data points that resides within the convex hull of the generated archetypal examples. We generate convex combinations of the archetypes by sampling weights of archetypes from a uniform Dirichlet distribution, $\mathrm{Dir}(\mathbf{1})$. Finally, the synthetically generated audiograms from the archetypal analysis process are corrupted with a zero-mean Gaussian observational noise with a standard deviation of 5.0 dB (or 5.0 years for the age dimension). The examples are split into training, validation and test partitions. The validation partition is used to, e.g., monitor learning and calibrate the uncertainty quantification model, whereas the test partition is for the final evaluation.

Missingness Given data sets with fully observed data, we mask data completely at random to impose missingness/partiality. A percentage of missing dimensions, ρ, is sampled uniformly from $[0, 1]$ for each datum. Following an initial random draw of ρ, we further randomly change the missingness level for 10% of data to be completely missing and 10% of data to be fully observed. This process is either done once during dataset generation (static missingness) or repeatedly at each batch generation (dynamic missingness). The latter acts as a regularization/augmentation scheme similar to random input dropout. To show the effect of augmentation, we compare dynamic and static missingness. We note that the mechanism behind dimensions being missing influences how we should treat the "missingness". If there is some structure to the missing dimensions, for instance that they are missing not at random as opposed to missing (completely) at random, this will influence the learnt representation (Mattei & Frellsen, 2019; Ipsen et al., 2021). We note that the simple missingness we use presently does not reflect the structure that comes with the acquisition process and the missingness patterns it induces. Learning the representation under the actual acquisition process' missingness might prove beneficial but beyond the scope of this paper.

3.2 Model

Partial data representation learning We assume that a partially observed datum has a set of observed dimensions $\mathcal{O}$ and unobserved dimensions $\mathcal{U}$ that jointly correspond to a fully observed datum, $\boldsymbol{x} \in \mathbb{R}^D$. Given an observed partial datum, $\boldsymbol{x}_{\mathcal{O}}$, the inference network initially embeds each observed dimension, x_d, $d \in \mathcal{O}$, producing embedding vectors of fixed sizes (here 16). Then, the embeddings are aggregated across the observed dimensions. The aggregated embeddings are the input to the network that parametrizes an approximate posterior distribution over a latent variable, $\mathbf{z}$, conditioned on the partially observed datum, $q_{\boldsymbol{\varphi}}(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}})$, where $\boldsymbol{\varphi}$ are the collected parameters of the inference network. We train these partial VAEs using the PointNet Plus variation introduced in Ma et al. (2019).

The generative network produces distributions of the fully observed data, $\mathbf{x}$, given the latent variable, $p_\theta(\mathbf{x}|\mathbf{z})$. We assume the observed data dimensions are conditionally independent given the latent variables, such that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = \prod_{u \in \mathcal{U}} p_{\boldsymbol{\theta}}(\mathrm{x}_u|\mathbf{z}) \prod_{o \in \mathcal{O}} p_{\boldsymbol{\theta}}(\mathrm{x}_o|\mathbf{z})$, where $\boldsymbol{\theta}$ are the parameters of the generative network. We follow common practice and use Gaussians for the observation distribution.

We optimize the inference and generative networks jointly by optimizing a lower bound on the marginal likelihood of the observed data, the partial ELBO, $\mathcal{L}^p$:

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_{\mathcal{O}}) \geq \mathcal{L}_p = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\varphi}}(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}})}\left[\sum_{o\in\mathcal{O}} \log p_{\boldsymbol{\theta}}\left(x_o|\mathbf{z}\right) - D_{KL}\left(q_{\boldsymbol{\varphi}}\left(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}}\right)||p\left(\mathbf{z}\right)\right)\right] = -(D^p + R), \tag{1}$$

where D^p, the partial distortion, denotes the expectation of the negative log-likelihood of the observed data given the approximate posterior, and R, the rate, similarly denotes the expected KL-divergence from the prior, $p\left(\mathbf{z}\right)$, where we choose a standard isotropic Gaussian.

Rate-distortion trade-offs In addition to optimizing Eq. 1, we explore models that are optimized towards a modified objective using a re-weighting factor, β (Higgins et al., 2017):

$$\mathcal{L}_{p,\beta} = -(D^p + \beta R), \tag{2}$$

We can achieve a model more penalized by divergences from the prior by increasing β, resulting in an approximate posterior distribution more closely resembling the isotropic Gaussian implicitly (enforcing a decrease in the rate, R). If we decrease β, we can achieve representations that better fit the data due to a decreased distortion, D^p. By exploring various weights, we can characterize the RD trade-off, which we can visualize as a curve on an RD-plane. Specifically, we choose to use an adaptive schema that changes the weight β such that the rate matches a set target rate using a schema matching the formulation used in Dieleman et al. (2021, Eq. 7).

Predicting full observations from partial data We use a trained partial VAE to acquire an audiogram sequentially by using the approximate posterior given the partial observation at any given time in the process to estimate the unobserved dimension. We will denote the number of measurements by m. For an audiogram of 14 total frequencies, a full sequence of estimated audiograms will be $\hat{\boldsymbol{x}}^{m=1}, \ldots, \hat{\boldsymbol{x}}^{m=14}$, where we have $D = M + 1$; the extra dimension is age, which we always observe). Here each $\hat{\boldsymbol{x}}^m$ is the estimate of the full audiogram at the given iteration of acquisition/number of measurements. The distribution over the unobserved dimensions, $\mathbf{x}_{\mathcal{U}}$, given a partial datum, $\boldsymbol{x}_{\mathcal{O}}$, are determined as $p_{\boldsymbol{\theta}}\left(\mathbf{x}_{\mathcal{U}}|\boldsymbol{x}_{\mathcal{O}}\right) = \int p_{\boldsymbol{\theta}}\left(\mathbf{x}_{\mathcal{U}}|\mathbf{z}\right) p_{\boldsymbol{\theta}}\left(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}}\right) d\mathbf{z} \approx \int p_{\boldsymbol{\theta}}\left(\mathbf{x}_{\mathcal{U}}|\mathbf{z}\right) q_{\boldsymbol{\varphi}}\left(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}}\right) d\mathbf{z}$. In producing an estimated $\hat{\boldsymbol{x}}$, we can combine the known $\boldsymbol{x}_{\mathcal{O}}$ with the mean of this predictive distribution for each unobserved dimension, x_u. In order to determine the next dimension, $i \in \mathcal{U}$, to be acquired, we use an acquisition function, $R(u, \boldsymbol{x}_{\mathcal{O}})$, $u \in \mathcal{U}$ given by the predictive distribution variance, using the superscript m to denote the current sets of unobserved and observed dimensions:

$$i^m = \underset{u\in\mathcal{U}^m}{\arg\max}\, R(u, \boldsymbol{x}_{\mathcal{O}^m}) = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\varphi}}(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}})}\left[\mathrm{Var}\left(p_{\boldsymbol{\theta}}\left(\mathrm{x}_u|\mathbf{z}\right)\right)\right]. \tag{3}$$

This we approximate using the sample variance of samples from the posterior predictive of the unobserved dimensions given multiple samples from the approximate posterior. As a baseline, we compare the performance of the acquisition process informed by the learnt representation with a random acquisition process.

3.3 Evaluation

Estimation of full audiogram Given ground truth knowledge of the fully observed audiogram, $\boldsymbol{x}$, we can evaluate how well the estimate at any given point in the process approximates the full audiogram. We can consider the root mean square error, e^m_{RMSE}:

$$e^m_{\mathrm{RMSE}} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left(x_d - \hat{x}_d^m\right)^2}. \tag{4}$$

Had we modelled with a fixed observation distribution variance of 1.0, evaluating this measure would correspond to evaluating the (negative log-)likelihood under the observation model distribution if we normalized

by the number of dimensions. Without such a normalization by the number of dimensions, we get the negative log-likelihood (NLL) error:

$$e^m_{\text{NLL}} = \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\varphi}}(\mathbf{z}|\boldsymbol{x}_{\mathcal{O}})}\left[-\sum_{d=1}^{D}\log p_{\boldsymbol{\theta}}\left(\mathrm{x}_d|\mathbf{z}\right)\right]. \tag{5}$$

Since we model variance in the observation distribution, the RMSE error and the NLL need not coincide. Following (Ma et al., 2019), we will refer to the error of the estimate as a function of the number of measurements, m, as an information curve. The area under the information curve is a single scalar value that then describes the quality of the acquisition process estimation performance:

$$A_{\text{RMSE}} = \sum_{d=1}^{D} e^m_{\text{RMSE}}. \tag{6}$$

Uncertainty quantification for stopping the acquisition process Given some uncertainty quantification metrics from the model, we are interested in evaluating how well these metrics can inform whether or not we can stop the acquisition process—that is, whether the current estimate is sufficiently good, assuming some predefined level of accuracy. We train a simple, linear ridge regression model to produce a prediction of the error, $\hat{e}^m_{\text{RMSE}}$, on a validation set. We estimate the regularization strength on the validation set using leave-one-out cross-validation across a range of possible values. The linear model input features are the acquisition function, the partial ELBO, and the current number of measurements, $m = |\mathcal{O}^m|$. The acquisition function, $R(u, \boldsymbol{x}_{\mathcal{O}^m})$ for all u, informs the error estimation about uncertainty in predictive distribution of unseen data. The partial ELBO, $\mathcal{L}^m_p$, informs the error estimation about uncertainty in seen data. All values can be evaluated without knowledge of the full audiogram. In this way, we can use it in a scenario where we do not know the full audiogram and estimate whether the models' estimates are sufficiently good.

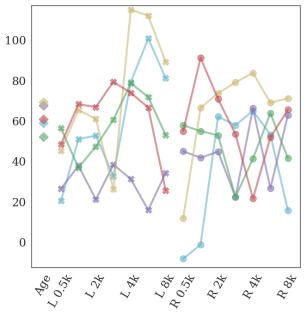
Absolute offset error A simple way of evaluating the acquisition performance would be to consider each measurement in all sequences as independent points and compare the predicted error with the true error across all such points. For this, any correlation will be partially driven by the sequential acquisition process since observing more and more dimensions reduces the error. Instead, we evaluate the uncertainty quantification by determining an offset between (a) when the model estimated it was done and (b) when the model was done according to ground truth knowledge. Given some threshold on the error, τ_e, each sequence of acquisitions has a specific measurement index, $s \in [1, \ldots, M]$, where the true at first falls below the criterion, $s = \arg\min_m e^m$ s.t. $e^m < \tau_e$ (dropping the subscript RMSE for clarity). During testing, we apply the same threshold to the linear model's prediction of the error. That is, we determine the estimated $\hat{s} = \arg\min_m \hat{e}^m$ s.t. $\hat{e}^m < \tau_e$. If not completely correct, the estimated stopping index can then be either too early or too late, which we define as the absolute offset:

$$O = |s - \hat{s}|. \tag{7}$$

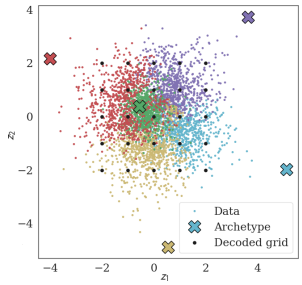
We report the absolute offsets as the average across all considered sequences of acquisitions (all audiograms in the test set). A similar approach to the evaluation of stopping an active learning process is considered in Ishibashi & Hino (2020). We compare the stopping performance of the linear model using uncertainty quantification to a fixed single best stopping index trained on the validation data. This baseline is a threshold on m without any possibility of datum-specific adaptation and amounts to fitting a linear model of the error with the measurement number as the only input feature.

4 Results and discussion

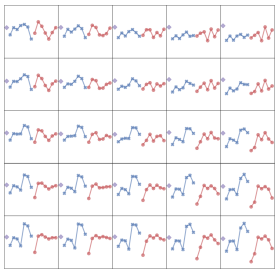
In Section 4.1, we analyze a learnt representation of the synthetic audiogram data. On this same representation, we study a single acquisition sequence in Section 4.2 and consider the corresponding information curves in Section 4.3. We show results for the uncertainty quantification for early stopping in Section 4.4 and results for comparing the use of the acquisition function to baselines in Section 4.5. Finally, in Section 4.6, we will present an analysis of rate-distortion trade-offs and their downstream effects for both synthetic and real data.



(a) Five archetypes (different colours), each of fifteen dimensions: age (diamond), and seven frequencies on each ear, left (crosses, L) and right (circles, R).



(b) Encoded data and archetypes in a two-dimensional representation. Placement of decoded grid points shown in x-space (c).



(c) Decoded grid of latent points shown in (b). Age (magenta diamond), hearing thresholds for left (blue, cross) and right (red, circle) across learnt representation.

Figure 1: Example of simple synthetic audiogram representation.

4.1 Learnt representation

As a simple example, we consider a simple two-dimensional learnt representation for a synthetic dataset consisting of five archetypal components. Figure 1a shows the archetypes generated through the process described in Section 3. These form the convex hull of the generated data in the data space, and all data points are combinations of these archetypes, "forming a continuum between the data archetypes". Encodings of the archetypes and all data are shown on Figure 1b. The encodings of the archetypes reside in the extreme parts of the representation. For all dimensions in the data space, the green component has values similar to or less extreme than other components. The VAE has learnt to express the green component efficiently using characteristics of the other components and has placed the component in the highest density region under the prior. The other components that describe more unique characteristics of the generated data each occupy a quadrant of latent space. We can decode a grid in the simple two-dimensional latent space and visualize the learnt representation as a two-dimensional grid of 15-dimensional data points in data space. We show such a visualization in Figure 1c, and we see how, for example, the first dimension (z_1) interpolates between the red and magenta archetype.

4.2 Acquisition sequence

Given a learnt representation, like the one visualized in Figure 1b and Figure 1c, the problem of estimating a full audiogram given partial observations reduces to searching in the two-dimensional representation for parts of the space that correspond well with the observed dimensions. The example sequence in Figure 2 visualizes the full audiogram alongside the current estimate (1$^{\text{st}}$ column), the acquisition function (2$^{\text{nd}}$ column), and the approximate posteriors of the full and partial posteriors (3$^{\text{rd}}$ column). Ideally, the estimates (mean in green and distribution in grey) match the full audiogram (magenta, blue and red) in the 1$^{\text{st}}$ column, and the partial approximate posterior (green) overlaps completely with the full approximate posterior (black) in the 3$^{\text{rd}}$. The first row in Figure 2 shows the model's a priori distribution over the data, i.e., with no observed dimensions. We assume that the age is always known, and the starting point is thus the second row. Knowing the age shrinks the green partial posterior closer to the black full posterior and improves the green estimate in the data space.

We observe data with a Gaussian observation noise with a standard deviation of 5 dB. The noise is a simple approximation of the inherent noise arising from measuring an audiogram, also approximately corresponding to the standard quantization of the hearing threshold levels into 5 dB steps. We retain this observation noise in the observation distribution, resulting in the estimated variance always being at 5 dB for the observed

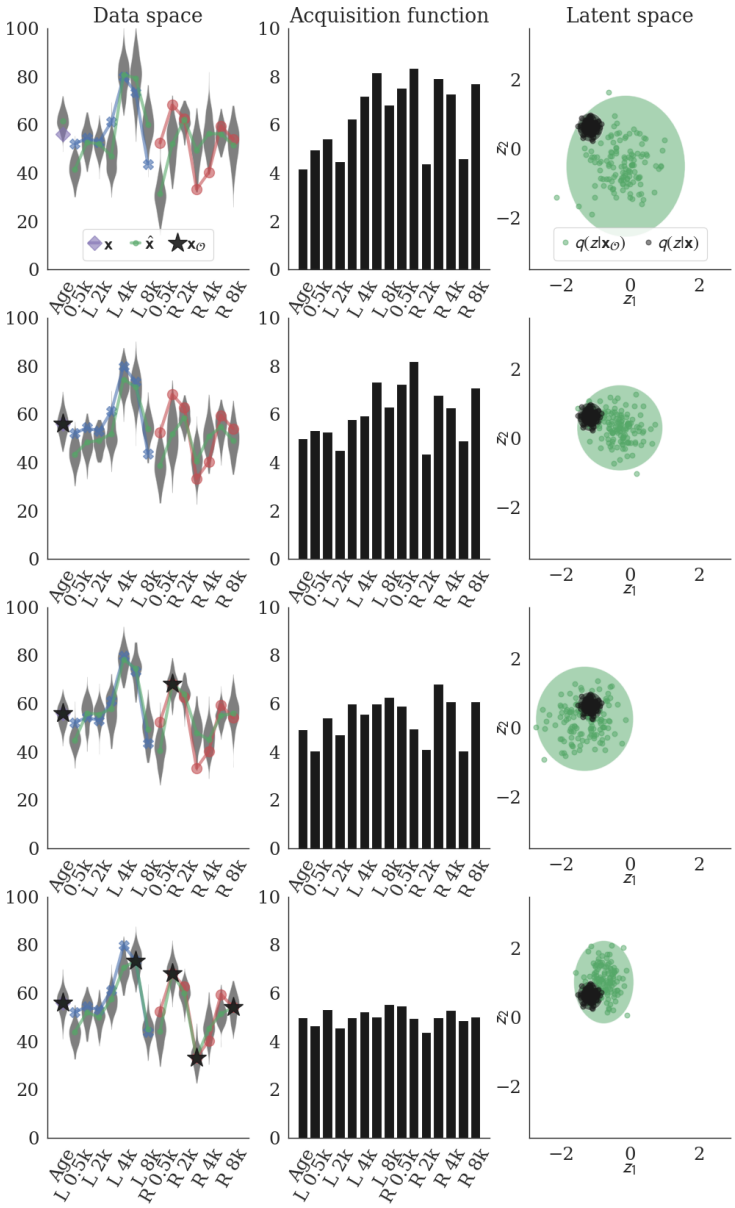
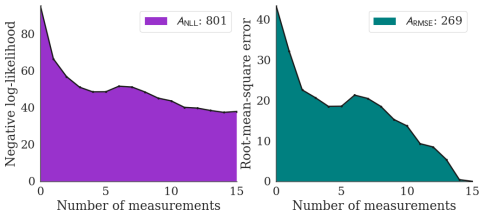
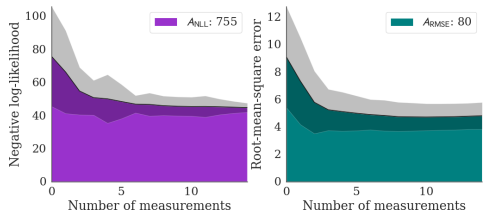


Figure 2: Example sequence of acquisition (each row down showing more further acquired dimensions). 1st column: estimated audiogram (green: mean, gray: distribution), ground truth fully observed (magenta star: age, blue cross: left ear, red circle: right ear), and observed dimensions (star). 2nd column: acquisition function. 3rd column: latent space distributions (black: ground truth, fully observed data approximate posterior, green: current partial posterior).



(a) Single example matching example in Figure 2.



(b) Across 128 test data acquisitions, shading corresponds to the estimated mean ± the estimated standard deviation.

Figure 3: Information curves, and the area under the information curves, for the negative expected log-likelihood (left, indigo) and the root-mean-square (right, teal) versions.

dimensions in the acquisition function plots. In the second row in Figure 2, we see that the dimension with the highest variance (here shown as the estimated standard deviation) is the 1 kHz dimension on the right ear. In the third row in Figure 2, we subsequently see how observing this dimension results in updated posteriors and estimates for the full audiogram. Note how the general audiogram structure captured in the learned representation informs the estimates for all dimensions. Observing this dimension not only improves the estimates for the surrounding frequencies on the right ear but also on the left ear. The last row in Figure 2 then shows the result of having repeated this procedure three additional times, i.e., by observing the dimension with the highest variance, which—here—improves the estimates and tightens the partial posterior around the full approximate posterior.

4.3 Information curves

Information curves summarize the performance of the acquisition sequence. Figure 3a shows the information curves for the full number of measurements of the sequence shown in Figure 2. As a function of measurements, we see that the errors tend to decrease. For the sequence considered, the model's estimate is temporarily, slightly degraded by the measurements made after five measurements. With no observation noise, the estimate would be perfect at the final point on the curve. In that scenario, the errors would be either the density at the mean of a multivariate Gaussian with a scalar standard deviation of 5 dB (for the negative log-likelihood information curve) or zero for the root-mean-square error information curve. However, since we have observation noise, the average final estimates are, at best, matching this noise. In expectation, the root mean square error would, for instance, end up being 5 dB (even if it, here, is lower due to the specific sample of the observation noise). Note that the learnt VAE also parametrizes the variances of the observation distribution; if this was not the case and the VAE had instead used a fixed variance of 1^2 dB^2, the two information curves would have been the same. We can summarize the performance of the acquisition estimate across all stages of the process by determining the area under the information curves (the coloured areas), which shows that this specific sequence of acquisition attained performance of $A_{\text{RMSE}} = 267\,\text{measurement} \cdot \text{dB}$ (the unit here indicating the metric as being the result of integrating the error metric measured in dB over the first axis of the information curve which has units of number of measurements). This scalar is, of course, a simplification of the process. An asymmetric cost might be associated with observing different dimensions, or early estimation errors might be critical in the interest of early termination. In such cases, the area under the information curve metric would be too simplistic. We visualize, in Figure 3b, the general performance of the acquisition process across many sequences (as opposed to just one, as above) as the mean and standard deviation at each measurement across the sequences. The shaded area indicates a standard deviation above and below the mean across 128 test acquisitions.

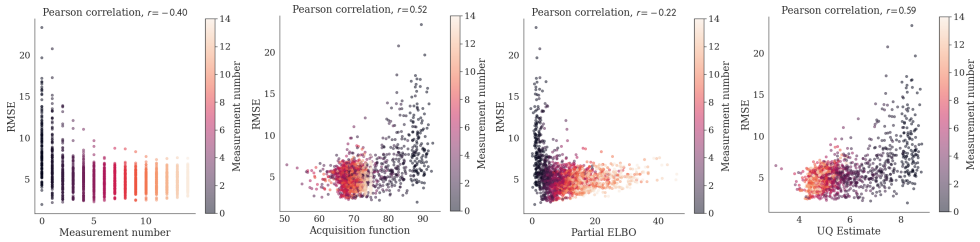


Figure 4: Root-mean-square error as a function of (from left to right) the measurement number, the mean acquisition function (mean variance), the partial ELBO, and the linear UQ estimate. Each dot is the estimate error and predictor at a given point in a sequence for 128 test sequences coloured by the measurement number.

4.4 Uncertainty quantification and stopping criterion

Given an estimate of the root-mean-square error, we can allow the acquisition process to terminate early. We visualize the evaluation procedure for the same, simple two-dimension latent model as also considered in Figure 1, Figure 2 and Figure 3. The uncertainty quantification (UQ) relies on a simple linear model of three predictors: the measurement number (which, when used alone, is our baseline), the mean acquisition function, and the partial ELBO. Figure 4 shows how the ground truth root-mean-square-error estimate varies as a function of each predictor (first three plots from left) and as a function of the linear model prediction (last plot on the right). The Pearson correlation is the highest for the combined model, but the correlation is driven, in part, by sequentially having acquired more observations. Instead, as described earlier, we evaluate the model's ability to estimate a stopping index based on the error estimate of the linear model.

Figure 5 shows a comparison between two approaches and the ground truth. The first row shows our baseline model, which estimates the current root-mean-square error based on the current number of measurements alone. The result, here, is a single stopping index at nine measurements across all sequences. The second row constructs the error estimate by including the acquisition function and the partial evidence lower bound as features to allow stopping based on sequence-specific uncertainty quantification. Finally, we compare these models to an oracular model that perfectly knows when to stop (last row). The false negatives (purple outcomes) are sequences where the estimated stopping index is later than needed, i.e., the true error fell below the threshold before the model stopped. Similarly, false positives (red) are situations where the estimated error is too low compared to the true error, resulting in early termination of the acquisition—before the estimate was sufficiently close to the ground truth.

We see that the sequence-specific uncertainty quantification allows for a reduction in absolute offset errors, in expectation, of about two measurements for this example. The UQ allows the model to stop two measurements closer to the correct stopping index on average. This improvement is, for this model, primarily a consequence of reducing the amount of much too late stopping estimations over the baseline at the cost of an increase in the number of sequences prematurely stopped.

Depending on the use case of the model, the cost of early/late errors might be asymmetric. In this model, we can account for such asymmetries by calibrating the threshold (τ_e) used to define stopping; to match a desired cost-sensitive balance, we could use a threshold that has been shifted either negatively or positively from τ_e when applying the threshold to the estimated error. These linear models could likely be improved, for instance, by including acquisition sequence history and interactions between the features; our purpose here is to show the benefit of UQ with a simple model, not to produce a highly performant estimator with a more elaborate model.

4.5 Acquisition functions

In this section, the synthetic data remains unchanged. However, compared to the simpler models in previous sections, we now train models with an increased capacity in terms of latent dimensions by increasing the

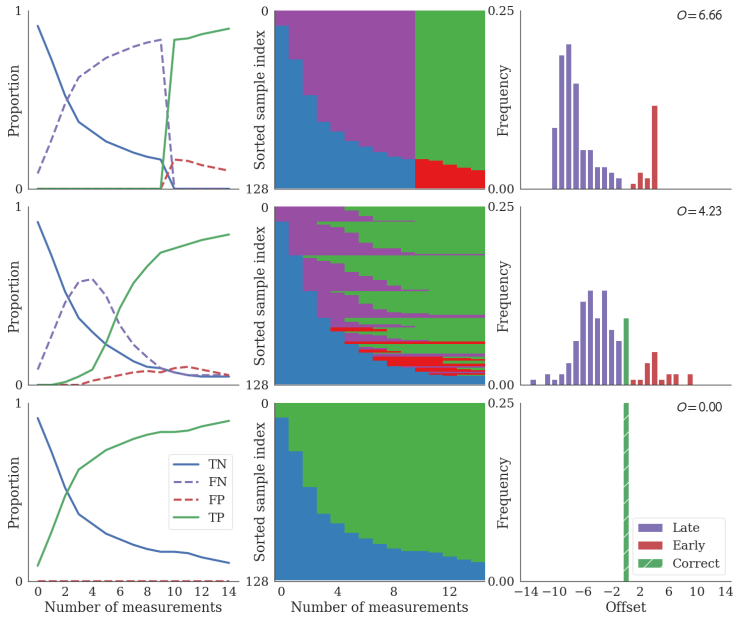


Figure 5: Visualization of stopping criterion evaluation. 1st column: proportion of outcomes across all test sequences in a binary classification of whether the process is done or not (true/false negative/positive: T/F N/P). 2nd column: the 128 test sequences and the same outcomes (same colors, legend in 3rd row). 3rd column: a histogram of offsets (too early, correct, or too late), and the top right corner shows the mean absolute offset. 1st row: single best stopping index. 2nd: linear model using uncertainty quantification features. 3rd: ground truth (dashed histogram bar indicates it has been cut off).

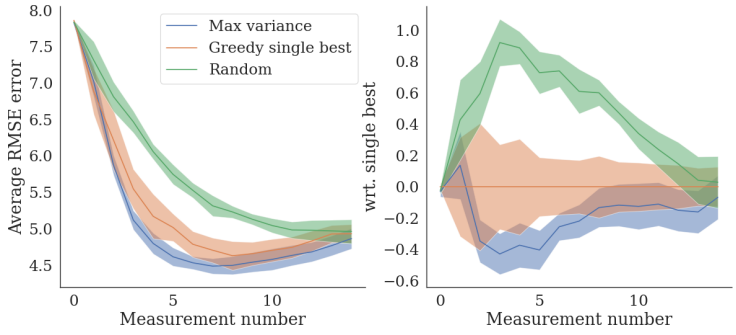


Figure 6: Left: RMSE information curve for three acquisition functions. Right: normalized to the greedy single best curve.

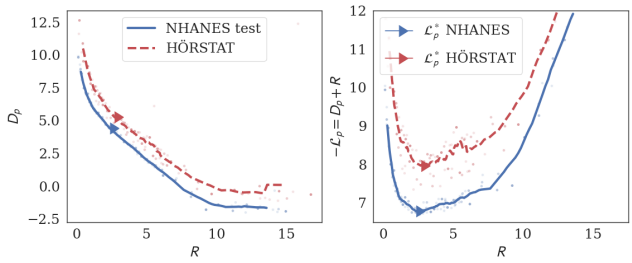


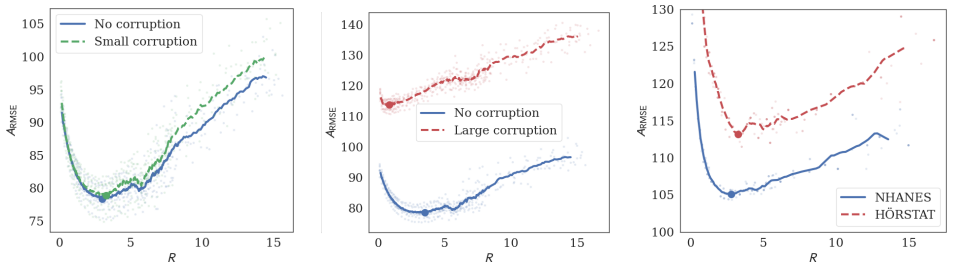
Figure 7: Results for NHANES-trained model evaluated on NHANES and HÖRSTAT test data. Left: rate-distortion curves. Right: Partial (negative) ELBO as a function of rate. Each trained model gives rise to two dots, on when evaluated on the training domain test set (blue) and one when evaluated on unseen test data (red). Left: the partial distortion. Right: the partial ELBO. The marker shows the optimal ELBO in each domain.

number of latents to four, with the same encoder and decoder capacity. The models optimize an unmodified partial ELBO on the same dataset. We report the standard deviation across ten replicates of the model (ten different model seeds on the same fixed synthetic data set) for each acquisition function.

We consider the information curve for different acquisition functions. Figure 6 shows the root-mean-square-error information curves on synthetic data when using three different acquisition strategies. As comparisons to the adaptive strategy, we in green show a process randomly selecting the next dimension, and, in orange, a process with one fixed ordering for all sequences. The single best ordering is found by greedily optimizing the ordering at each step across a validation set to select the dimension that would minimize the expected error. Finally, in blue, we show the acquisition process that acquires the dimensions based on the max variance acquisition function. We see that both the single best ordering and max variance acquisition outperform the random acquisition. The max variance acquisition function process is an improvement over a single best ordering. Except for the first measurement, the max variance acquisition has a lower error as a function of the number of measurements and a lower variance over the sequences. The differences between the models disappear as we observe more dimensions. For this data, the estimations at fully observed audiograms can, in expectation, be poorer than estimates with fewer measurements because the observation model variance can be lower than 5 dB for unobserved dimensions but is fixed to 5 dB when a dimension is observed.

4.6 Rate-distortion trade-offs

Figure 7 shows how increased rates allow for increasingly lower distortions and vice versa. The left figure shows the partial distortion as a function of rate. The right figure shows the same information but as the partial ELBO as a function of rate. For this figure, and the following similar to it (Figure 8, Figure 9), the curves are moving averages when sorting the models based on the rate. Each dot is a model trained towards a given target rate. We train models with target rates from 0.125 to 15.0, with increasingly poorer resolution for increased target rates—i.e., we train models with target rates from 0.125 to 1.5 in steps of 0.125 (0.125, 0.250, ...), from 1.5 to 8.0 in steps of 0.25 (1.5, 1.75, ...), and from 8.0 to 15.0 in steps of 1.0 (8.0, 9.0, ...). For all remaining experiments exploring rate-dependant qualities, we use this range of targets rates. For each target rate, we trained either three models (three different seeds) for the larger real data sets, or ten different models for the smaller synthetic datasets. The minimal partial ELBO is marked for both considered data sets, NHANES (blue) and HÖRSTAT) (red). The models shown here were all trained on NHANES, and the blue curve is unseen test data from NHANES. The red curve, instead, shows the model RD trade-off on an unseen data set (HÖRSTAT)). We see a general increase in partial distortion across all rates—i.e., the information coded for does not result in a higher quality of reconstruction in the unseen dataset. We anticipated that the optimal rate on the inter-dataset evaluation would be lower—i.e., that the red marker would be to the left of the blue marker—but the optimal rate is slightly higher in the unseen domain than



(a) Synthetic data, unseen domain with small corruption.

(b) Synthetic data, unseen domain with large corruption.

(c) Real data, training domain is the NHANES data set (blue), the unseen domain is the HÖRSTAT data (red).

Figure 8: Area under the curve as a function of rate for test data from training domain (blue) and an unseen domain (green or red). For the synthetic datasets, the unseen domain are defined by a corruption of underlying archetypal components in generating the dataset. Large circular markers indicate the optimal rate within a domain. The smaller transparent dots indicate individual models trained towards different target rates.

in the training domain. Simply penalizing the KL-term more harshly is not inducing better generalization based on this evaluation of the NHANES-trained model on HÖRSTAT).

In Figure 8, we show how the area under the information curve is affected by the representation rate and how this affects generalization. As a simple example, we construct a corrupted version of the synthetic dataset that the model saw during training. The underlying archetypal components are corrupted with Gaussian noise with a standard deviation of either 0.5 (large corruption, Figure 8b) or 0.05 (small corruption, Figure 8a) times the overall standard deviation of the training dataset across all dimensions. In the small corruption setting (left, green), the performance is only slightly decreased, with worse performance for higher rates. When the corruption is small, the optimal rate remains about the same or slightly higher. When the corruption is large, the area under the information curves are much worse (red curve, center plot). In the large corruption setting, the best generalization is attained with rates lower than the rates that optimized the performance on the original model dataset. We show a similar analysis for the real data in Figure 8c. As for the rate-distortion trade-off in Figure 7, the optimal rate for the area under the information curve in the unseen domain is similar to, or slightly higher than, the optimal rate test data in the seen data set.

Figure 9 shows four metrics for the partial VAE acquisition process as a function of the model rate on synthetic and real data. The first row shows the results for models trained on a dataset with static missingness—a scenario where the dataset we learn from *is* partially observed. The second row shows the result of dynamically inducing missingness in a dataset, reflecting the situation where we attempt to learn representations of partial data by imposing some missingness pattern on a dataset with fully observed data. The dynamic missingness works as augmentation, where the missingness pattern changes each time a datum is part of a batch. We can apply dequantization with new dequantization noise sampled at each batch generation. Dequantization accounts for the fixed 5 dB intervals in which real audiogram thresholds are, typically, measured. The second-row synthetic data includes this dequantization as well as the dynamic missingness, too. The last row shows the results for the models trained on the NHANES dataset using the dynamic missingness and dequantization.

For all rows, the first column shows the partial distortion as a function of rate for both the training data (dashed, green) and test data (full, blue). For the first row, the distortion continues to fall as the rate increases for the training data. In contrast, the test data distortion reaches a minimum at a rate of about 5 nats (indicated by the round, blue marker). Balancing the rate and the distortion equally results in the partial ELBO (marked by a red triangle) at a lower rate of about 3 nats. At higher rates, the models overfit, producing poorer generalization to the test set. This gap disappears when considering the models trained

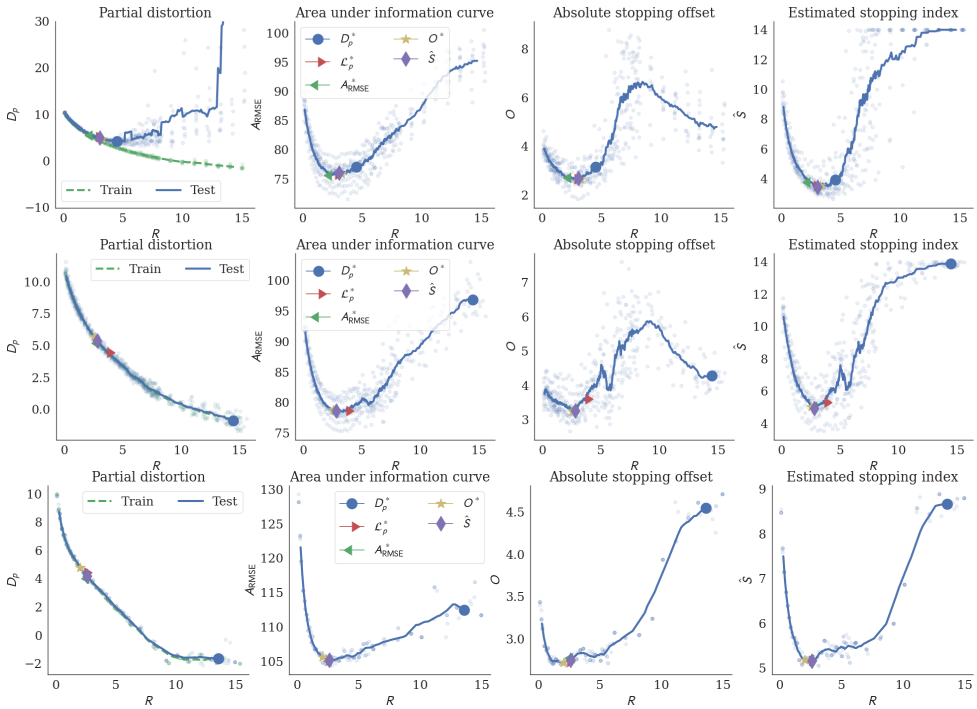


Figure 9: Partial distortion (1st column), acquisition prediction error (2nd column), acquisition uncertainty quantification performance as stopping offset (3rd column) and estimated stopping index (4th column) as a function of rate. 1st row: synthetic data with fixed missingness. 2nd row: synthetic data with dynamic missingness and dequantization. 3rd row: real data, NHANES. For the synthetic data, the stopping threshold on estimated error was set at 5 dB RMSE, and for the real, more complicated data, a threshold of 7 dB RMSE was used. D_p: partial distortion (blue circle, superscript asterisk indicating optimal value). O: absolute offset (yellow star). $\mathcal{L}_p$: partial ELBO (red right caret/triangle). $\hat{S}$: estimated stopping index (magenta diamond). A_{RMSE} (green left caret): area under the information curve for root-mean-square error. The superscript start indicates optimal value)

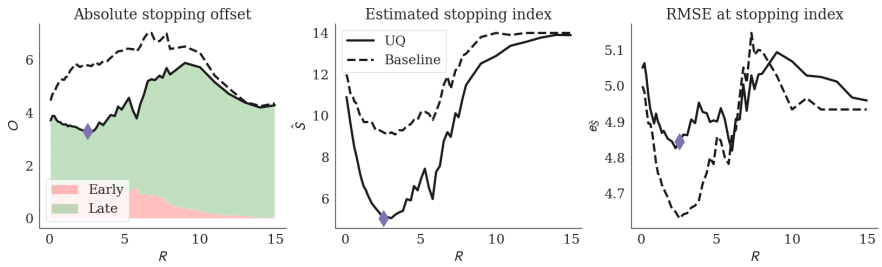
(synthetic and real) with dynamic missingness and dequantization on the second and third row. Here, the lowest test distortions are attained at much higher rates for these models (the blue, circular marker is far to the right). The second plot shows the area under the information curve (A_{RMSE}, for the RMSE error) as a function of rate (we only show the test data curve here). The models that produce low partial distortions, even on the test data, do not produce good estimates of the full audiogram. This is true in the setting where the model overfits to the training data for the higher rates (first row). However, the same behavior is seen for the scenario where the models do not display a gap between the RD curves for training and test data (second and third row). That is, poorer acquisition performance at higher rates is not a result of overfitting to the training data in a straightforward sense but rather a consequence of too high informational rates. The extra information produces lower distortions even on the test set, but the information is not valuable for the downstream acquisition task. The optimal rate for the area under the information curve metric (marked by a green triangle) is considerably below the optimal distortion (blue circle) and nearer the optimal ELBO (marked by red triangular marker). The acquisition uncertainty estimation performance, as quantified using the absolute stopping offset (O), is shown in the third column. Again, the models with the lowest distortions produce poor uncertainty quantification, but the optimal rate (marked by a yellow star) is closer to the rate of the partial ELBO. The same is true for the last column, showing the optimal stopping index (lowest amount of measurements until the model estimates it is done, marked by a magenta diamond). In the synthetic data with dynamic missingness (second row), both the optimal area under the information curve, absolute offset and lowest estimated stopping index are at rates lower than the optimal partial ELBO. For the static missingness, only the area under the information curve appears to have a lower optimal rate of the considered compared to the optimal rate for the partial ELBO. For the real data, only the absolute stopping offset appears to have a lower optimal rate than the ELBO.

We compare the stopping evaluation to the baseline stopping (using only the measurement number) in Figure 10. Figure 10a show the results on the synthetic data with dynamic missingness and dequantization (the same considered on the second row of Figure 9). We see that the estimated stopping index and error in this estimation (the absolute stopping offset) are lower for all rates for the uncertainty quantification linear model compared to the baseline single index stopping method. The difference is the largest for optimal rate configurations. Another view of how the UQ model with access to the acquisition function and partial ELBO produces more well-calibrated stopping estimates than a model solely using the measurement number is shown in the right plot. This shows that the RMSE at the stopping index more closely matches the specified threshold of 5 dB RMSE in the UQ model compared to the baseline, which stops too late (the error is too low w.r.t the threshold of 5 dB). The same conclusions apply to the real data, as shown in Figure 10b—as for the results in Figure 9, the real data uses a less strict threshold (7.0 dB) for stopping to reflect the more complicated problem. Setting the threshold at 5.0 dB for the real data would result in an estimated stopping generally coinciding with the maximum number of measurements.

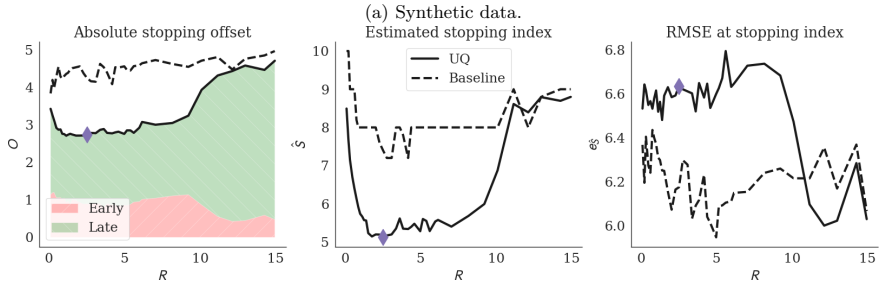
Broader Impact Statement

We provide an in-depth analysis of rate-distortion trade-offs in variational autoencoder representations. Specifically, we show how downstream performance is affected by these trade-offs. Improving our understanding of generalization in variational autoencoder representation learning frameworks through rate-distortion analysis is of general applicability; while we consider partial data and sequential acquisition, no particular aspect of this formulation hinders this type of analysis being generally useful where variational autoencoders are used.

Mahomed et al. (2013) conclude that there is a need for knowledge on difficult-to-test and different types and degrees of hearing loss. Applying a representation learning scheme with robust acquisition performance and uncertainty estimation is especially important when the diversity of the population increases. Automated audiometry can be used to provide hearing healthcare services to underserved populations and areas (Mahomed et al., 2013). When applying learning based models in such scenarios, the learnt representation might reflect the large, available data sets, whereas lack of data can result in poorer performance of the model on the underserved populations—and lack of evaluation data might blind us to deficiencies of our models. In this case, improving and understanding how representations generalize is especially relevant, and our analysis shows how we might begin to increase robustness by exploring the rate-dependency on downstream tasks.



(a) Synthetic data.



(b) Real data, NHANES

Figure 10: Stopping criterion comparison to baseline. Both representations trained with dynamic missingness and dequantization. O is the absolute stopping index, or how far away from ground truth correct stopping index the model stopped. $\hat{S}$ is the estimated stopping index, indicating at which measurement the model did stop (the estimating stopping index plus the offset would produce the true, oracular stopping index. The RMSE at the stopping index, $e_{\hat{S}}$, shows how well the model hits a designated target (of 5.0 dB for the synthetic data and 7.0 dB for the real data). The UQ-based stopping process stops earlier and more accurately than the baseline method.

5 Conclusion

We show how a partial variational autoencoder representation can be use to efficiently, sequentially acquire partially observed data, and we provide in-depth analysis of the rate-distortion trade-offs that affect the model. In particular, we explore how to efficiently measure audiograms considering both controlled, synthetic data generated using an archetypal generation mechanism and on real data from both the United States and Germany. We show how using an acquisition function based on uncertainty in the variational autoencoder representation improves acquisition over using a predetermined, globally best ordering. The acquisition function chooses the dimension of maximal variance given an encoding of the partially observed data. The same encoding provides an estimate of the full audiogram, and we show how the representation's uncertainty quantification allows more accurate early acquisition termination than single best stopping index. We use the number of measurements, the partial evidence lower bound, and the acquisition function as features in a linear model to estimate the model's error in predicting the full audiogram from the partial audiogram. We show that the uncertainty quantification termination is more accurate than a model that terminates based on a globally best stopping index alone.

Finally, we show how both the full audiogram estimation and the termination procedure performances are rate-dependent. Models that produce the lowest distortions at the expense of high rates, even on unseen test set, provide poor downstream performances. The optimal rates were not in all cases coinciding with the rate that optimized the evidence lower bound, but lower rates ("stronger regularization") did not invariably improve generalization. This points to the importance of accounting for rate-distortion trade-offs when using VAEs for data completion and uncertainty quantification as presently illustrated in the context of efficient audiogram acquisition.

References

Alex Alemi, Ben Poole, Ian Fischer, Josh Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 159–168, Stockholmsmässan, Stockholm Sweden, 2018. URL `http://proceedings.mlr.press/v80/alemi18a.html`.

Gabriella Contardo, Ludovic Denoyer, and Thierry Artières. Sequential cost-sensitive feature acquisition, 2016.

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

Sander Dieleman, Charlie Nash, Jesse Engel, and Karen Simonyan. Variable-rate discrete representation learning. *arXiv preprint arXiv:2103.06089*, 2021.

Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. cdc.gov/nchs/nhanes, 1999–2022.

Jacob R Gardner, Xinyu Song, Kilian Q Weinberger, Dennis L Barbour, and John P Cunningham. Psychophysical detection testing with bayesian active learning. In *UAI*, pp. 286–295, 2015.

Jongmin Han and Seokho Kang. Active learning with missing values considering imputation uncertainty. *Knowledge-Based Systems*, 224:107079, 2021.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*, 2017.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-MIWAE: Deep Generative Modelling with Missing not at Random Data. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=tu29GQT0JFy`.

Hideaki Ishibashi and Hideitsu Hino. Stopping criterion for active learning based on deterministic generalization bounds. In *International Conference on Artificial Intelligence and Statistics*, pp. 386–397. PMLR, 2020.

James M Kates. *Digital hearing aids*. Plural publishing, 2008.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019. doi: 10.1561/2200000056.

Chao Ma, Wenbo Gong, José Miguel Hernández-Lobato, Noam Koenigstein, Sebastian Nowozin, and Cheng Zhang. Partial VAE for hybrid recommender system. In *NIPS Workshop on Bayesian Deep Learning*, volume 2018, 2018.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *International Conference on Machine Learning*, pp. 4234–4243. PMLR, 2019.

Faheema Mahomed, De Wet Swanepoel, Robert H Eikelboom, and Maggi Soer. Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear and hearing*, 34(6):745–752, 2013.

Robert H. Margolis and Donald E. Morgan. Automated Pure-Tone Audiometry: An Analysis of Capacity, Need, and Benefit. *American Journal of Audiology*, 17(2):109–113, December 2008. ISSN 1059-0889, 1558-9137. doi: 10.1044/1059-0889(2008/07-0047).

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pp. 4413–4423. PMLR, 2019.

Brian CJ Moore. *An introduction to the psychology of hearing*. Brill, 2012.

Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107:107501, 2020. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107501. URL `https://www.sciencedirect.com/science/article/pii/S0031320320303046`.

Jens Brehm Bagger Nielsen, Jakob Nielsen, and Jan Larsen. Perception-based personalization of hearing aids using gaussian processes and active learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):162–173, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009.

Josef Schlittenlacher, Richard E Turner, and Brian CJ Moore. Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, 144(1):421–430, 2018.

Burr Settles. Active learning literature survey. 2009.

Xinyu D Song, Brittany M Wallace, Jacob R Gardner, Noah M Ledbetter, Kilian Q Weinberger, and Dennis L Barbour. Fast, continuous audiogram estimation using machine learning. *Ear and hearing*, 36(6):e326, 2015.

P. von Gablenz and I Holube. Prevalence of hearing impairment in northwestern Germany. Results of an epidemiological study on hearing status (HÖRSTAT). *HNO*, 63(3):195–214, 2015.

Zhiqiang Zheng and Balaji Padmanabhan. On active learning for data acquisition. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 562–569. IEEE, 2002.

APPENDIX B

VI-EMD: Improving Speaker Separation Generalization with Variational Inference

Improving Speaker Separation Generalization with Variational Inference

Rasmus M. Th. Høegh[1,2,*] **Jens B. B. Nielsen**[2]

Abigail A. Kressner[1,3] **Morten Mørup**[1]

Abstract

Audio processing networks with encoder-masker-decoder architectures are effective speaker separation systems, but their generalization abilities have deficiencies that limit their real-world viability. Building on existing deterministic models, we propose a variational inference approach to learn stochastic encodings as well as stochastic masks that can be applied to the encodings for separation. To facilitate this, we introduce a likelihood with scale-invariance properties similar to a commonly used separation objective. We show that the approach improves generalization to new datasets while also improving overall test performance. The probabilistic framework further enables a wide range of modeling possibilities; we consider three aspects in particular: rate-distortion analysis on speaker separation generalization, the influence of prior specification on performance, and the relation between quantified uncertainty and performance.

1 Introduction

Speaker separation has seen great advances with deep learning (DL)-based methods [1]. A series of models following an encoder-masker-decoder (EMD) architecture perform particularly well. These models, building upon time-domain audio separation networks (TasNets) [2], are separating in a learned encoding space by means of a masker (see Figure 1a). A convolutional variant [3] was shown in certain scenarios to outperform a traditionally strong oracle baseline in both objective distortion measures and perceived, subjective audio quality. Later work has improved performance by, for example, introducing improvements to the masking network such as using dual-path recurrent neural networks [4] or using attention mechanisms [5]. However, TasNet inter-dataset generalization (test data from unseen datasets) is poorer relative to the intra-dataset generalization (unseen test data from the same dataset as the training data), and similarly, when presenting mixing procedures different from the training task [6, 7]. That is, the models learn to separate speech in a manner that does not generalize well beyond the training dataset and task, limiting their utility in a real-world setting.

In standard supervised learning, a mapping is learned from a high-dimensional input to a low-dimensional supervision label. DL models with sufficient capacity will generally be able to learn near-perfect mappings for simple problems on the training data, but if these models are not adequately constrained, the implicit representation of the input will not generalize. This over-fitting is caused by the model learning to map uninformative characteristics for training data points to their labels. Models like TasNets can, for the same reasons, fail to generalize if not properly regularized, even though the supervisory signal is more high-dimensional. Generative models, on the other hand, aim to learn the distribution of data. By tasking the model with reconstructing data, representations can be learned without explicit guidance from a label or supervisory signal. The generative task provides a learning

[*]Corresponding author: rmth@dtu.dk, [1]Technical University of Denmark (DTU), [2]WS Audiology, [3]Rigshospitalet

Preprint. Under review.

signal to infer patterns that are robust (generalize well) such that the learned representations can be usefully applied in another task of interest. Such models include, for example, directly modelling the likelihood of the data [8–10] or using variational inference (VI) with variational auto-encoders (VAEs) [11, 12].

In the particular case of VAEs, the models learn an explicit representation of the input (i.e., an encoding, or latent representation), through a stochastic mapping (i.e., an inference network) from an input to a latent space. Learning this encoding is coupled with learning a similar mapping (i.e., a generative network) from the latent and back to original data. The mappings both produce distributions over either the latent space or the original data space. In learning these mappings, samples from the latent distributions are taken, and the parameterization of the distributions are updated to improve the model. Simplistically, the stochasticity of the sampling can be likened to injecting noise in the learning procedure, and it enables models to learn both to characterize the variation for a particular input and to be robust to such variations. VAEs have been successfully applied in modelling data in a wide variety of domains, such as computer vision, chemistry, natural language, and astronomy [13]. VAEs have also been applied for time-domain modelling of speech enhancement in particular [14] and of audio in general; for the latter, a high-capacity encoder-decoder can, for example, learn representations that enable voice conversion and has learned representations that correlate with high-level representations such as phonetic content [15]. VAE-based semi-supervised learning enables learning from large amounts of unlabeled data, such that the representation can be used to efficiently learn how to solve a specific task using fewer, but labelled, examples [13].

In the following, we explore whether a generative approach can characterize and improve the generalization of encoder-masker-decoders (EMDs), such as TasNets. We leverage that these networks have an encoder-decoder structure, similar to (variational) auto-encoders (AEs) (see Figure 1). Specifically, we consider speaker separation EMDs in a VI/VAE framework by recasting the learning of the EMDs as VI of latent variable encodings and masks coupled with reconstructing single source components of an input mixture. To facilitate this, we present a likelihood function with properties similar to the scale-invariant signal distortion ratio (SI-SDR) with which TasNets are mostly trained. We focus on TasNets for simplicity and introduce variational inference TasNets (VI-TasNets), but we stress that the approach can readily be used for any EMD architecture, notably also e.g. improved succesors to TasNets (of which we consider the SuDoRMRF). We investigate generalization to unseen datasets from the perspective of rate-distortion (RD) analysis, exploring the implications of jointly minimizing the separation performance (i.e., the distortion) and divergence from the prior (i.e., the rate). The prior on the latent variables further facilitate imposing desired properties, and we explore a model going beyond the standard isotropic Gaussian prior using log-normal mask priors and adaptive, additive encoding priors. Lastly, we consider a multitasking model's ability to quantify the uncertainty of its separation performance in a more realistic setting where the underlying sources are unknown.

2 Background

Speaker separation Speaker separation, and its application towards solving the cocktail party problem [16], has been studied for many decades, especially from the perspective of classical (digital) signal processing. Re-framing the problem as a DL supervised learning problem has enabled many advances in how well speaker separation can be done (see, e.g., Wang and Chen [1] for an overview). Speaker separation is the task of recovering a set of (clean) single sources, $\mathcal{S} = \{s_0, \ldots, s_N\}$, from a (potentially noisy) mixture of the components, x, under some mixture generating function, Mix. A simple expression with a set of speakers and single noise source of this could be e.g. $x = \text{Mix}(s_1, \ldots, s_N, n) = n + \sum_{i=1}^{N} s_i$, where x, s_i, n are time-series (e.g. $x = [x_0, \ldots, x_t, \ldots, x_T]$) of length T, and n is some interference/noise time-series. In this work, for Mix, we consider a simple additive mixture process for a mono-channel audio signal.

Deep learning speech separation A wide variety of approaches for DL speaker separation systems exist, but broadly speaking, an overarching difference between the approaches lie in whether the models are relying on a spectrogram or frequency-based representation of the audio or are directly operating on a time-domain representation. Similarly, a distinction can be made between high-performance, large, complex (in parameter counter, operations per second to process, etc.), high-latency systems targeted to function offline versus smaller, efficient, and real-time capable models, of which TasNets are the latter. TasNets (see Figure 1a) utilize an EMD structure directly operating on a

time-domain representation of the audio [2]. Separation is done by estimating masks that are applied to the mixture encoding such that decoding the resulting masked encodings provides estimates of the component sources in isolation. That is, an encoder ($f_\varphi(x) = z$) provides an encoding (z) for an input mixture (x). The encoding is fed to a masker ($h_\psi(z) = \mathcal{M}$) which provides a set of masks ($\mathcal{M} = \{m_0, \ldots, m_i\}$) that will be applied to the encoding. Applying a given mask (m_i) to the encoding provides a masked encoding ($\hat{z}_i = m_i \odot z$). Finally, the decoder ($\tilde{s}_i = g_\theta(\hat{z}_i)$) outputs estimates of the single sources ($\tilde{s}_i$) based on the masked encodings. While the original TasNet relied on recurrent neural network (RNN) architectures to estimate masks, a variant using dilated temporal convolutions (Conv-TasNet) was shown to learn a more efficient and better-performing separation system [3], even outperforming spectral oracular performance methods (ideal ratio masks) in some scenarios. In training these models, assigning which estimated single source corresponds to a target source is a permutation problem, and the use of (utterance level) permutation invariant training (PIT) enables the model to learn by using only the best permutation of the known target sources to estimated sources (on an utterance level) to determine the loss [17]. A limitation of these models is their requirement of separation into a fixed number of speakers; this is a central problem addressed in other work [18, 19].

Scale-invariance TasNets are usually trained towards maximizing the SI-SDR [20], which for a known source s and estimated source $\hat{s}$ is defined as:

$$\text{SI-SDR}(s, \hat{s}) = 10 \log_{10} \left(||\alpha s||^2 / ||\alpha s - \hat{s}||^2 \right), \quad \alpha = \hat{s}^\top s / ||s||^2, \tag{1}$$

where $||\circ||$ designates the 2-norm. The α coefficient rescales the target such that the error in the denominator between the estimated source and target source is measured in a way that is invariant to the overall scale (power) of the time-series. While the scale-invariance is a central part of the formulation of SI-SDR, it notably measures the error in a logarithmically scaled manner. Often a scale-invariant signal-to-distortion-ratio improvement (SI-SDRi) is reported, giving the SI-SDR increase in using the processing over using the input mixture as the estimate of the single source. A similar objective function to SI-SDR that does not re-scale the target but retains the logarithmic scaling of the errors is the logarithm of the mean-squared error (MSE). A log-MSE has been shown to enable training of TasNets in a manner similar to the SI-SDR [21], while a standard MSE does not achieve comparable results. That is, training a TasNet with an MSE loss (as opposed to using a SI-SDR or log-MSE) does not provide performant speaker separation in TasNets.

Generalization of TasNets TasNets show drops in performance when evaluated on datasets unseen during training. Kadioglu et al. [6] found significant drops in performance for a model trained on LibriTTS when evaluating it on test sets from VCTK [22] and WSJ0-2 [18] (about 6 dB poorer relative to a LibriTTS test set). However, Cosentino et al. [7] argued that these effects are partially due to differences in SNR-calculation and characteristics of the utterances in the corpora. They show that training on a different dataset, LibriMix, actually shows higher performance on the WHAM! test set (a noisy extension of WSJ0-2mix), while less significant drops (on the order of 1-2 dB) are still found when evaluating on VCTK. Their findings indicate that training on a larger, more diverse dataset reduces the inter-dataset generalization error, but that a generalization gap persists in going from e.g. LibriMix to VCTK.

Realistic data We can consider speaker separation in quiet, in simplistic noise (additive Gaussian noise), and more realistic noise (using realistic recordings of background noise of speech). Realistic data and problems are key to developing models viable for actual use. Standard benchmarks often investigated, such as WSJ0-2, often lack diversity in the speakers and recording conditions, have unrealistic mixing process (e.g., with too high overlaps), have no consideration of reverberation, use a fixed number speakers, or are based on non-ecological speech material (i.e., speech recorded, for example, while reading written material aloud as opposed to during a natural conversation) [7]. Later datasets have since addressed some of these issues, e.g. by extending WSJ with realistic, reverberant noises in realistic conditions [23, 24], integrating multiple corpora [25], a more diverse set of speakers [26], sparse speaker overlaps [7], and varying number of speakers and sound types [27]. In this work, we will rely on LibriMix [7], since it both includes realistic noises (from WHAM!) and a large, diverse set of speakers based on LibriSpeech [28]. While we focus on mono-channel separation, multi-channel processing that can utilize spatial information will be more relevant in reverberant environments.

Variational autoencoders (VAEs) A VAE is a latent variable model that aims to learn a representation of high-level features of data at scale through amortized approximate inference of variational distributions using deep learning [11, 12]. For a given data point, x, a VAE optimizes a lower bound, $\mathcal{L}$, on the model evidence, or evidence lower bound (ELBO):

$$\ln p_\theta(x) \geq \mathcal{L}(x; \varphi, \theta) = \mathbb{E}_{q_\varphi(z|x)} \left[\ln p_\theta \left(x|z\right)\right] - D_{\mathrm{KL}}(q_\varphi(z|x)||p(z)), \tag{2}$$

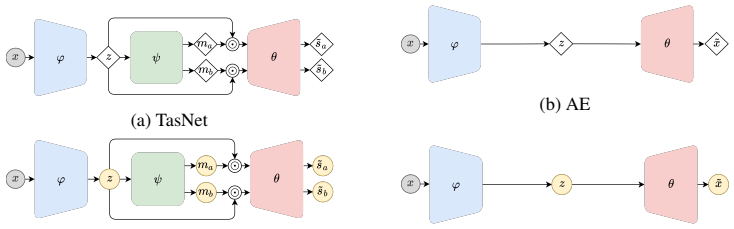
where θ, φ are the encoder and decoder parameters, respectively, $q_\varphi(z|x)$ is a variational approximation to a true, but intractable, posterior $p_\theta(z|x)$, $\mathbb{E}_{q_\varphi(z|x)}$ denotes an expectation w.r.t. the variational distribution, and D_{KL} is the Kullback-Leibler (KL) divergence. The distributions $p_\theta(x|z)$ and $q_\varphi(z|x)$ are parameterized by the decoder/generative network and by the encoder/inference network, respectively. The encoder maps from a given data point, x, to a latent representation, z, and the decoder learns to map from the latents back to the data. Often the parameterized distributions are Gaussians, and an isotropic Gaussian is used for the prior, $p(z)$. By optimizing $\mathcal{L}$, we ensure that the model in z, learns—in some sense—a well-behaved, compressed representation of x (ensured by the "KL divergence term") while still being able to reconstruct the data (ensured by $p_\theta(x|z)$, the "reconstruction term").

Priors, rate-distortion and compression Commonly, an isotropic Gaussian prior imposing less co-varying, smaller magnitude latents is used. This prior can be prohibitively restrictive, and using more expressive priors can improve learning [29–31]. Other characteristics can be imposed by using non-Gaussian priors, such as a directional or non-negative distribution [32, 33]. Balancing reconstruction (the negative log-likelihood of the data, or distortion, D) and deviations from the prior (the KL between approximate posterior and the prior, or rate, R), for instance with a re-weighting (as in β-VAEs [34]) can provide different behaviours of the representation. Targeting lower rates under isotropic Gaussian priors can provide better disentanglement by some measures [35] and, notably, affects generalization [36]. We can visualize these trade-offs as RD curves, and Alemi et al. [37] discuss the RD-trade-off, feasible and realizable models, their relation to the entropy of the data and the capacity of the model. Notably, the information bottleneck principle shows how there is an optimum rate when seeking to improve generalization [38, 39]. We discuss these aspects in greater detail in Appendix E. In particular, we also discuss the representation learning aspects of RD curves, how these trade-offs relate to compression [40], and how these trade-offs are also related to further quantities regarding matching the overall data distribution (similar to a standard generative adversarial learning objective) [41], and how, e.g., over-completeness and sparsity fits into VAE-based representation learning [42].

3 Variational Inference Encoder-Masker-Decoder Separation

We propose a probabilistic modelling variant of EMDs. We introduce VI-TasNets that closely resemble a well-studied [7] (deterministic, Conv-)TasNet [3]. Note that we refer to the Conv-TasNet as the deterministic TasNet, or just TasNet, for brevity. We stress that the VI-EMD framework can readily be applied to later extensions that are e.g. more performant or more explicitly consider reverberations. Like TasNets, VI-TasNets use simple, single-layer, convolutional encoders and decoders. For TasNet, the outputs of the encoder directly provide encodings, and in VI-TasNet, the same outputs instead parameterize encoding distributions, e.g. the means and variances of Gaussians. Similarly, while the TasNet masker and decoder directly output masks and the estimated sources, the VI-TasNet parameterizes distributions over the masks and the estimated sources. An overview of the model is shown on Figure 1c. Example model outputs are given as visualizations in Appendix C, and as audio in the supplementary material.

Encoder and masker distributions For the encoder distribution, we consider a large, over-complete K-dimensional latent space ($K = 512$), matching the TasNets. For an input time-series x of length T to the encoder, we obtain a distribution of a latent time-series of length T'. The sampling frequency of the two is related through the strides of the encoder-decoder structure; we choose a lower latent space tick, such that $T/T' = 8$. Note that the input time-series is one-dimensional, $x \in \mathbb{R}^{1\times T}$, while the latents have K dimensions per latent time step. We opt for factorized Gaussian approximate posterior for the encodings parametrized by f_φ; similarly, we opt for log-normal masks parametrized



(a) TasNet (b) AE

(c) VI-TasNet (d) VAE

Figure 1: EMD models, such as TasNets depicted in (a), share an encoder-decoder structure with AEs in (b) and their variational inference (VI) counterparts VAEs in (d). VI-EMDs, like VI-TasNets shown in (c), are robust, variational extensions. VI-EMDs learn distributions—over encodings z of audio and multiplicative masks m—by reconstructing single sources s from the input mixture x. Deterministic parametrizing networks blue, green, red: encoder (φ), masker (ψ), decoder (θ). Variables: explanatory (gray), stochastic (yellow), deterministic (diamond). $\odot$: Hadamard product.

by h_ψ (mimicking non-negative properties of a ReLU activated TasNet mask):

$$q_\varphi(z|x) = \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathcal{N}(z_{k,t'}; \mu^{\varphi}_{k,t'}, (\sigma^{\varphi}_{k,t'})^2), \tag{3}$$

$$q_\psi(\mathcal{M}|z) = \prod_{n=0}^{N} \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathcal{LN}(m_{n,k,t'}; \mu^{\psi}_{n,k,t'}, (\sigma^{\psi}_{n,k,t'})^2), \tag{4}$$

where e.g. $\mu^{\varphi}_{k,.}$, $\sigma^{\varphi}_{k,.}$ are time-series of length T' with distribution parameters (location and scale) for the k'th latent dimension output. For further specifics and prior specifications, we refer to Appendix F and Appendix L for discussion of other choices of prior.

Decoder distribution In standard VAEs, a typical choice would be to parameterize a per time-step Gaussian for the generative network: $p_\theta(\tilde{s}|z) = \prod_{t=0}^{T} \mathcal{N}(\tilde{s}_t; \mu_t, \sigma_t^2)$, where $\mu_{.}, \sigma^2$ are time-series output of the decoder, g_θ. Given a fixed unit variance, this would correspond to an MSE loss, which, as discussed, does not train performant TasNets. In Appendix I, we discuss a multivariate Cauchy objective (MVC) since the log-likelihood conceptually enables us to minimize a log-error similar to the log-MSE [21]. The SI-SDR is not a likelihood, but we are interested in a likelihood that is similarly invariant to a re-scaling of the time-series. Where the SI-SDR uses α (Eq. 1), we take the view that a similar factor γ is a regression coefficient. Using Bayesian linear regression (BLR), we wish to model both γ and a noise-scale parameter, σ^2. For the target time-series s with steps s_t, we consider a linear regression model in which the approximation $\tilde{s}$ takes the role of predictor: $s_t = \tilde{s}_t\gamma + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$. Conjugate priors for σ^2 and γ take the form $p\left(\gamma, \sigma^2\right) = p\left(\sigma^2\right) p\left(\gamma|\sigma^2\right)$, where $p\left(\sigma^2\right)$ is an inverse-gamma distribution with parameters a_0 and b_0, Inv-Gamma(a_0, b_0), and the conditional prior distribution for the regression coefficient is a normal distribution with a mean μ_0, and variance $\sigma^2\lambda_0^{-1}$, where λ is a scalar prior precision for the regression coefficient. Update rules for these parameters enable us to determine a likelihood for given time-series given the input $\tilde{s}$ when integrating over γ and σ^2, $p_\theta\left(s|\tilde{s}\right) = \int p\left(s|\tilde{s}, \gamma, \sigma^2\right) p\left(\gamma, \sigma^2\right) d\gamma d\sigma^2$. This provides an analytical expression for the likelihood which we use in the optimization of VI-TasNets as the distortion measure. For further details, see Appendix H, where we show a comparison to the SI-SDR and show how they optimize related quantities.

Evidence lower bound The ELBO optimzed with VI-TasNets takes the form (derivation in Appendix B): $\log p_\theta(\mathcal{S}) \geq \mathcal{L}(\theta, \varphi, \psi; \mathcal{S}) = -D_\mathcal{S} - R_z - R_\mathcal{M}$, where, $D_\mathcal{S}$ is the distortion of the single sources (how well they are reconstructed), R_z is the divergence of the encodings from their prior (the encoding rate), and R_m is the divergence of the masks from their prior (the mask rate). Different from an actually auto-encoding VAE, the VI-TasNet reconstructs single sources instead of the original input to the encoder. To explore the RD trade-off, we use various modified losses based on the ELBO during training (by means of free bits, adaptive reweighing of the rate, or a β-coefficient; see details in Appendix E).

Flexible priors We can consider priors that are more flexible than the standard, static ones. In Appendix J, we discuss a learnt auto-regressive flow prior. While such a prior constitute a powerful, domain-agnostic approach to a more flexible prior, we also introduce a domain-inspired adaptive prior which promotes that mixture encodings resemble an addition of single source encodings. Such a prior is in part motivated by the separation task, seeing as the separation-by-masking is assuming a similar underlying generation mechanism. In itself, the use of a masker in the encoded space with element-wise masks on the mixture encoding can be thought of as an inductive bias, or architectural prior, under which the model is learning to separate. The addition of random variables corresponds to a convolution of their distributions, and, for the encoding distributions we use, we have closed-form convolutions (see Appendix F). Consider now encodings of single sources for the known true single sources, s_i, in Figure 1; using a notational shorthand for the encodings of the single sources, we define an adaptive prior for the mixture encoding as:

$$q_{\varphi,a} = q_\varphi\left(z_a|s_a\right), \quad q_{\varphi,b} = q_\varphi\left(z_b|s_b\right), \quad p_\varphi\left(z|\mathcal{S}\right) = \left(q_{\varphi,a} \circledast q_{\varphi,b}\right)(z), \tag{5}$$

where $\circledast$ denotes a convolution operation, which amounts to enforcing that the mixture encoding is a sum of the single source encodings $z = z_a + z_b$. We add corresponding rate terms to reflect the single source encodings divergence from a prior. This is needed to use the adaptive prior to ensure that we actively promote that $q_{\varphi,a}$ and $q_{\varphi,b}$ resemble the distributions that we convolve in making the adaptive prior.

Multitasking VI-TasNet with autoencoding objectives A model that utilizes the described adaptive prior already obtain encodings of the single sources. We can utilize the single sources as a target signal, too, combining the VI-TasNet separation task with a single source AE task, sharing the encoder and decoder parameters. This amounts to adding a distortion term stemming from decoding the single source encodings instead of masked mixture encodings. We can further augment such a multitasking model with a mixture AE task, especially relevant since we do not have the true single sources in a real-world scenario. We reconstruct the original mixture from the obtained mixture encodings and add a mixture reconstruction distortion term. An overview of this setup is shown in Appendix D. Adding mixture AE also provides a quantification of how well the model fits the mixture, and we investigate whether the mixture AE performance is indicative of separation performance, which would provide a principled uncertainty quantification for the separation system.

Comparison with later TasNet variant Finally, we can investigate how the variational inference procedure affects later developments of EMD models. We choose to investigate the "successive downsampling and resampling of multi-resolution features" (SuDoRMRF) model [43], as this model—like the original (Conv-)TasNets—seeks to have a minimal footprint. The SuDoRMRF model uses a more efficient, convolutional separation module relying on an architecture reminiscent of U-net [44]. We isolate the effect of the VI framework and enable a comparison to the (VI-)TasNet results by only replacing the masking network, and otherwise keeping everything the same. For further details on this, we refer to Appendix J.

4 Related work

Later EMDs similar to TasNets remain competitive models for speaker separation. Improvements include improved dilated temporal convolution blocks and multi-scale modelling [45, 43], specialized RNN architectures [4], and utilizing transformers/attention mechanisms in the masker [46, 5]. Other separation networks rely on e.g. explicitly modeling speakers and perform separation using on a learned speaker stack [47], or—in line with Luo et al. [4]—use specialized RNNs [19]. Early work separated using a clustering approach [18], and recently attractor-based systems have been proposed [48]. Beyond the currently considered anechoic problems, recent works have more explicitly considered reverberant environments and spatialized problems e.g. incorporating neural beamformers [49, 50]. While we use VI to improve generalization, pre-training tasks have been considered; some show improvements with speech enhancement pre-training [51], and a range of self-supervised learning approaches [52]. Other improvements to the training procedures include MixIT [53] and ReMixIT [54].

Spectral speaker separation VAEs Speaker separation in the frequency domain using VAEs has seen many studies in recent years [55–64] including also multi-channel models [65, 66], and models

integrating visual information [67]. Girin et al. [68] present more general considerations on modelling audio spectrograms with VAEs. While these methods employ VAEs towards speaker separation, they operate in a spectral representation (often using the short-term Fourier transform), where the presently considered work is concerned with directly modelling the time-domain audio waveform. The difference in TasNets modelling spectral or temporal representations is considered in Bahmaninezhad et al. [69]. It can be beneficial to use spectral losses to optimize AEs even if operating in the time domain [70]. Hybrid TasNets are proposed that use both spectral and temporal representations [71]. Such hybrids provide an avenue for leveraging the extensive study of spectrogram VAEs with the present work.

Time-domain modelling Deep learning speech separation in the time-domain has been done using WaveNets, e.g. in a generative modelling framework [72], or in discriminative, non-autoregressive variants [73]. As audio modelling with generative adversarial networks [74] and flows [75] improves, similar concepts are applied in speech separation and enhancement including both adversarial [76–78] or flow-based methods [79]. In all cases, the models produce speech separation with high-quality outputs but they rely on a high model complexity as compared to TasNets to achieve these results. Structured State Space sequence (S4) models [80] have improved audio modelling [81], but have yet to be applied to speaker separation. VAE applied to time-domain audio notably include e.g. VAEs with WaveNet decoders [15, 82], and more recently NaturalSpeech, used for text to speech [83]. Time-domain VAEs are less studied for speech separation (speaker separation and/or speech enhancement). One example is the variance constrained (VC) AE for speech enhancement [14]; the VC AE focuses on a different task and does not optimizing a VAE objective, and the VC AE model differs from the present work in that the VI-TasNet relies on the same architecture as the well-studied EMDs/TasNets (using masking of the encodings for separation).

5 Discussion

Synthethetic data We introduce a simple synthetic dataset for source separation; the dataset constitutes a controlled, simplified speaker separation problem, see Appendix K for further details on the data. The datasets consist of "2-speaker" mixtures, where the sources are generated as a number of overlapping randomly-generated sinusoidal Gauss pulses in a speaker-specific frequency range with overtones. On this dataset, we fit a series of VI-TasNets using an adaptive re-weighting of the total rate (encoding rate and the masking rate) towards a desired target rate (see Appendix E) for specifics on the procedure. For the synthetic data, we reduce the capacity of both the encoder-decoder and the masker to fit the complexity of the problem by reducing the number of filters/channels, but otherwise we use the same architecture overall as are used in later experiments.

LibriMix/VCTK data We consider $N = 2$ talker mixtures with a sampling frequency of 8 kHz from the LibriMix dataset [7]. We use the same processing and splits and consider mixtures of length matching the shortest single source (the "min" mode). We train models either on the clean or the noisy 100 hour variant (Libri2Mix train-100). Single replicates are reported for these models considered, and randomness in the initialization is not characterized; for further specifics on training/compute-requirements and model/results limitations, see Appendix J. We evaluate the performance of these models in their ability to generalize to both a "familiar" Libri2Mix test set and to the "unfamiliar" VCTK-2mix test.

RD and generalization In Figure 2 (top) we show *test-set* RD curves for models trained on a given level of Gaussian additive noise (dashed, black, medium noise) on the *synthetic* data (lines

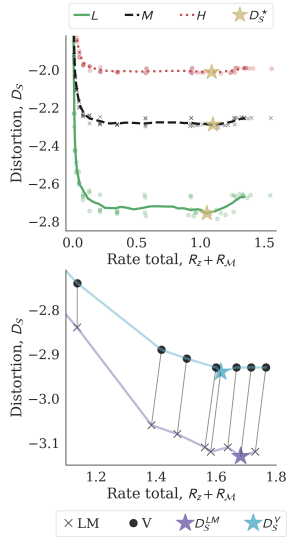


Figure 2: RD curves for speaker separation on synthetic (top) and real data (bottom). Stars denote optimum distortion within a dataset. LM: LibriMix, V: VCTK, L/M/H: low/mid/high noise setting.

are running means of distortion as a function of sorted rates). The curves show the expected trade-off of poor distortion at low rates. Note that rates are normalized by the number of latent dimensions. The models with the highest rates (above 1.25 nats/dim) display poorer separation performance than models with slightly lower rates, indicating that high-rate models learn sub-optimal representations. We evaluate the same models in lower (full, green) and higher noise settings (dotted, red), and we see that an optimal rate lower than the highest rates exists in all conditions. Furthermore, the distortion difference between optimum and high-rate models grows in a low noise setting compared to the training domain. This is evidence of how expressive models that optimize single source distortion over rate produce poorer generalization. In particular, these results show that the variational framework for separation displays trade-offs for rate, distortion and generalization that match the information bottleneck perspective. This, in turn, provides an avenue for improving the generalization of EMD models, since we can learn models that jointly optimize rate and distortion in a principled manner. Further details and conditions are discussed in Appendix M.

We investigate the same concept in the real speaker separation data. We evaluate test-RD curves for models trained on *clean* LibriMix. These models were trained using various levels of free-bits, details in Appendix E. Models with very low rates and higher distortions were trained but omitted in the visualization for clarity. In Figure 2 (bottom), we see that models with rates that are too low perform poorly, and we see indications that the model with the lowest distortion in LibriMix does not achieve the best distortion in VCTK, where instead a lower-rate model performs the best. The deterministic models will solely optimize a distortion and do not quantify the informational rate, and so the results for generalization (from a LibriMix-trained model to VCTK) are consistent with the need for considering models that can quantify and optimize the distortion and rate jointly to achieve improved generalization.

Improved generalization to unseen data and conditions We train a VI-TasNet with Gaussian encodings and log-normal masks using the BLR likelihood optimizing an adaptively reweighted ELBO towards a target total rate. As the deterministic baseline, we train a TasNet using the standard SI-SDR objective. Having trained the models on the *noisy* LibriMix, we contrast performance seen/unseen conditions (noisy, clean) and seen/unseen datasets (LibriMix/VCTK). The performance of these models is shown in Table 1. We see that the VI-TasNet is a strict improvement over the deterministic TasNet both in seen and unseen datasets and conditions. Specifically, noisy LibriMix/VCTK test-performance is improved by 0.37 and 0.55 dB respectively, with a 2 %-points lower relative drop. Similar generalization improvements to the unseen clean condition are observed with the VI-TasNet. While this improvement comes at negligible increases in parameter count and inference time computations, it does increase the training time. In Appendix L we show results for TasNets trained with both SI-SDR and BLR on the *clean* condition, which we compare to VI-TasNet with different priors. Similarly, when we evaluate the framework on a different masking network, the SuDoRMRF, we see consistent, but small improvements in generalization.

Table 1: SI-SDRi (and performance drop from LibriMix to VCTK) for models trained on the 100 hr noisy 2-speaker LibriMix data. Models are *not* trained on the clean condition, it is an unseen test condition, similar to how VCTK is an unseen dataset/domain. See Appendix N for further metrics. †using only the SuDoRMRF separation module. N: noisy, C: clean.

	Model	*LibriMix*	*VCTK*	*Drop*
N	TasNet	11.6	9.9	1.8 (0.15)
	VI-TasNet	**12.0**	**10.4**	**1.6 (0.13)**
C	TasNet	13.0	10.4	2.6 (0.20)
	VI-TasNet	**13.6**	**11.4**	**2.1 (0.16)**
N	SuDoRMRF†	11.1	9.2	2.0 (0.18)
	VI-SuDoRMRF†	**11.5**	**9.5**	**1.9 (0.17)**
C	SuDoRMRF†	12.5	9.8	2.7 (0.22)
	VI-SuDoRMRF†	**12.9**	**10.4**	**2.5 (0.19)**

Uncertainty quantification We train a multitasking VI-TasNet on the *clean* condition. We present further results and discussions of the multitasking VI-TasNets, including a discussion of the adaptive prior in Appendix D. Importantly, besides separation, this model also learns to do mixture AE. We leverage the mixture AE task to quantify how well the shared encoder-decoder structure models the input mixture (input density). We investigate whether the mixture AE ELBO is informative of the separation performance, as measured by the single source dis-

tortions in the separation task. Figure 3 shows the separation performance measured as BLR estimated single source distortion as a function of the negative mixture ELBO for each mixture in the LibriMix and VCTK test sets—each dot is a mixture, and contours are from a kernel density estimator used solely for visualization. Lower values for both values indicate improved performance (either better separation for the distortion or higher evidence for the mixture AE). Mixture AE is informative of the separation task performance, and the lowest distortions also model the mixture the best—conversely, when the mixture is poorly modelled, the separation distortion increases. We note that there is a slightly increased number of mixtures (dots) at lower mixture AE and higher distortion for VCTK, indicating mixtures for which the input density is a poorer predictor of separation performance. The positive correlation effectively provides a model that can quantify its uncertainty in performing the separation task. The mixture is available in a real scenario, and so if the AE task is performing poorly, the results suggest that this would be indicative of poor separation performance, too. While the two quantities are not normally distributed or linearly related, we can quantify the correlation as a Pearson correlation coefficient, or we can quantify it with a non-parametric Spearman rank correlation coefficient, instead. Here, we report both and see a weak to moderate correlation that decreases slightly in the new domain (Pearson correlation r_p, Spearman, r_s, and $p < 1\mathrm{e}-50$, $n = 3000$ per data set and condition): $r_p = .37, r_s = .45$ (LibriMix) and $r_p = .28, r_s = .35$ (VCTK).

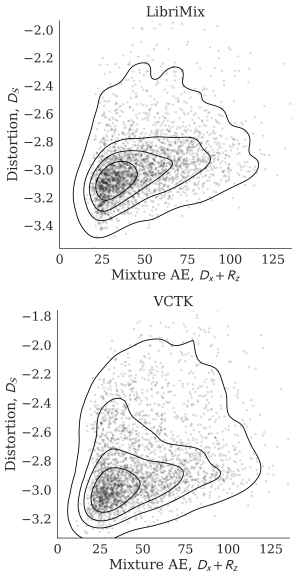


Figure 3: Separation performance versus mixture AE for LibriMix (top) and VCTK (bottom).

Broader impact Improved speaker separation systems will facilitate the improvement of hearing aids and thus improve the treatment of hearing loss. DL systems in hearing aids, however, will likely impose increased demands on e.g. internet connectivity or hardware capabilities, meaning these improvements will likely reach listeners with access to more resources first. Improved speaker separation improves automatic speech recognition systems and, for example, transcription automation. Probabilistic models that more explicitly impose priors enable interpretability, while also potentially improving learning in the face of more scarce data (e.g. smaller non-English corpora). Generative modelling enables the production of deep fakes and systems that can, for example, mimic a given speaker's voice can be used to conduct fraud and produce fake media.

6 Conclusions

We have presented variational inference encoder-masker-decoder models, and particular instantiations in the variational inference time-domain audio separation network, VI-TasNet, and VI-SuDoRMRF. The VI-EMDs effectively learns to separate audio while learning distributions of latent encodings and latent masks in a manner that improves tests performances on seen conditions and test data from seen datasets, but also improves generalization to unseen conditions and new datasets. The VI-TasNet uses a Bayesian linear regression likelihood, which enables likelihood-based training with scale-invariance similar to scale-invariance signal-distortion-ratio. The probabilistic formulation of the model provides the means of imposing priors on the learning, we discuss an adaptive prior and provide results in the supplementary on how various priors impact the separation performance. We show how the generalization of VI-TasNet can be characterized using rate distortion trade-offs; we show indications that, while trading off increased rates sometimes improves performance within the condition and within the same dataset, this does not necessarily generalize to a new dataset. Lastly, we show that a multitasking VI-TasNet performing mixture autoencoding can quantify its uncertainty in a manner informative of the separation performance without requiring access to the single sources.

Acknowledgments and Disclosure of Funding

RMTH is partially funded by the Innovation Fund Denmark. RMTH and JBBN are employed by WS Audiology and they declare no conflicting interests for the work presented as a consequence of this or any other affiliation.

References

[1] DeLiang Wang and Jitong Chen. Supervised Speech Separation based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10): 1702–1726, 2018.

[2] Yi Luo and Nima Mesgarani. TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2018.

[3] Yi Luo and Nima Mesgarani. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019.

[4] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE, 2020.

[5] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in Speech Separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE, 2021.

[6] Berkan Kadioglu, Michael Horgan, Xiaoyu Liu, Jordi Pons, Dan Darcy, and Vivek Kumar. An Empirical Study of Conv-Tasnet. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7264–7268. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9054721. URL `https://doi.org/10.1109/ICASSP40776.2020.9054721`.

[7] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. LibriMix: An Open-Source Dataset for Generalizable Speech Separation. *arXiv preprint arXiv:2005.11262*, 2020.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. *preprint, available at openai.com*, 2018.

[9] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA, 2016. URL `http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_DS-4_van_den_Oord.html`.

[10] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional Image Generation with PixelCNN Decoders. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4790–4798, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html`.

[11] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6114`.

[12] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference On Machine Learning*, pages 1278–1286. PMLR, 2014.

[13] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019. doi: 10.1561/2200000056. URL https://doi.org/10.1561/2200000056.

[14] Daniel T Braithwaite and W Bastiaan Kleijn. Speech Enhancement with Variance Constrained Autoencoders. In *INTERSPEECH*, pages 1831–1835, 2019.

[15] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

[16] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[17] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.

[18] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[19] Eliya Nachmani, Yossi Adi, and Lior Wolf. Voice Separation with an Unknown Number of Multiple Speakers. In *International Conference on Machine Learning*, pages 7164–7175. PMLR, 2020.

[20] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR–Half-Baked or Well Done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE, 2019.

[21] Jens Heitkaemper, Darius Jakobeit, Christoph Boeddeker, Lukas Drude, and Reinhold Haeb-Umbach. Demystifying TasNet: A dissecting approach. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6359–6363. IEEE, 2020.

[22] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). 2019.

[23] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending Speech Separation to Noisy Environments. In Gernot Kubin and Zdravko Kacic, editors, *INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1368–1372. ISCA, 2019. doi: 10.21437/Interspeech.2019-2821. URL https://doi.org/10.21437/Interspeech.2019-2821.

[24] Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux. WHAMR!: Noisy and Reverberant Single-Channel Speech Separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2020.

[25] Matthew Maciejewski, Gregory Sell, Yusuke Fujita, Leibny Paola García-Perera, Shinji Watanabe, and Sanjeev Khudanpur. Analysis of Robustness of Deep Single-Channel Speech Separation using Corpora Constructed from Multiple Domains. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2019, New Paltz, NY, USA, October 20-23, 2019*, pages 165–169. IEEE, 2019. doi: 10.1109/WASPAA.2019.8937153. URL https://doi.org/10.1109/WASPAA.2019.8937153.

[26] Zhuo Chen, Takuya Yoshioka, Liang Lu, Tianyan Zhou, Zhong Meng, Yi Luo, Jian Wu, Xiong Xiao, and Jinyu Li. Continuous Speech Separation: Dataset and Analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7284–7288. IEEE, 2020.

[27] Scott Wisdom, Hakan Erdogan, Daniel PW Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, and John R Hershey. What's all the Fuss about Free Universal Sound Separation Data? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 186–190. IEEE, 2021.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: an ASR Corpus based on Public Domain Audio Books. In *2015 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[29] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved Variational Inference with Inverse Autoregressive Flow. *Advances in neural information processing systems*, 29:4743–4751, 2016.

[30] Xi Chen, Diederik P. Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=BysvGP5ee`.

[31] Jakub Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.

[32] Jiacheng Xu and Greg Durrett. Spherical Latent Spaces for Stable Variational Autoencoders. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4503–4513. Association for Computational Linguistics, 2018. URL `https://aclanthology.org/D18-1480/`.

[33] Steven Squires, Adam Prügel Bennett, and Mahesan Niranjan. A Variational Autoencoder for Probabilistic Non-Negative Matrix Factorisation. *arXiv preprint arXiv:1906.05912*, 2019.

[34] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

[35] Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *CoRR*, abs/1804.03599, 2018. URL `http://arxiv.org/abs/1804.03599`.

[36] Alican Bozkurt, Babak Esmaeili, Jean-Baptiste Tristan, Dana Brooks, Jennifer Dy, and Jan-Willem Meent. Rate-Regularization and Generalization in Variational Autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 3880–3888. PMLR, 2021.

[37] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a Broken ELBO. In *International Conference on Machine Learning*, pages 159–168. PMLR, 2018.

[38] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015.

[39] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[40] James Townsend, Thomas Bird, and David Barber. Practical Lossless Compression with Latent Variables using Bits Back Coding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=ryE98iR5tm`.

[41] Yochai Blau and Tomer Michaeli. Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR, 2019.

[42] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

[43] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis. Sudo RM -RF: Efficient Networks for Universal Audio Source Separation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2020. doi: 10.1109/MLSP49062.2020.9231900.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[45] Liwen Zhang, Ziqiang Shi, Jiqing Han, Anyan Shi, and Ding Ma. FurcaNeXt: End-to-End Monaural Speech Separation with Dynamic Gated Dilated Temporal Convolutional Networks. In *MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part I*, volume 11961 of *Lecture Notes in Computer Science*, pages 653–665. Springer, 2020. doi: 10.1007/978-3-030-37731-1_53. URL `https://doi.org/10.1007/978-3-030-37731-1_53`.

[46] Jingjing Chen, Qirong Mao, and Dong Liu. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. pages 2642–2646, 10 2020. doi: 10.21437/Interspeech.2020-2205.

[47] Neil Zeghidour and David Grangier. Wavesplit: End-to-End Speech Separation by Speaker Clustering. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2840–2849, 2021. doi: 10.1109/TASLP.2021.3099291. URL `https://doi.org/10.1109/TASLP.2021.3099291`.

[48] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Paola Garcia. Encoder-Decoder Based Attractors for End-to-End Neural Diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022.

[49] Chenda Li, Jing Shi, Wangyou Zhang, Aswin Shanmugam Subramanian, Xuankai Chang, Naoyuki Kamo, Moto Hira, Tomoki Hayashi, Christoph Boeddeker, Zhuo Chen, and Shinji Watanabe. ESPnet-SE: End-To-End Speech Enhancement and Separation Toolkit Designed for ASR Integration. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 785–792, 2021. doi: 10.1109/SLT48900.2021.9383615.

[50] Teerapat Jenrungrot, Vivek Jayaram, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. The Cone of Silence: Speech Separation by Localization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[51] Sung-Feng Huang, Shun-Po Chuang, Da-Rong Liu, Yi-Chen Chen, Gene-Ping Yang, and Hung-yi Lee. Self-supervised Pre-training Reduces Label Permutation Instability of Speech Separation. *arXiv preprint arXiv:2010.15366*, 2020.

[52] Zili Huang, Shinji Watanabe, Shu-wen Yang, Paola García, and Sanjeev Khudanpur. Investigating Self-Supervised Learning for Speech Enhancement and Separation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6837–6841. IEEE, 2022.

[53] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. Unsupervised Sound Separation using Mixture Invariant Training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3846–3857. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/28538c394c36e4d5ea8ff5ad60562a93-Paper.pdf`.

[54] Efthymios Tzinis, Yossi Adi, Vamsi Krishna Ithapu, Buye Xu, Paris Smaragdis, and Anurag Kumar. RemixIT: Continual Self-Training of Speech Enhancement Models via Bootstrapped Remixing. *arXiv preprint arXiv:2202.08862*, 2022.

[55] Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, and Tatsuya Kawahara. Statistical Speech Enhancement based on Probabilistic Integration of Variational Autoencoder and Non-Negative Matrix Factorization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 716–720. IEEE, 2018.

[56] Yoshiaki Bando, Kouhei Sekiguchi, and Kazuyoshi Yoshii. Adaptive Neural Speech Enhancement with a Denoising Variational Autoencoder. In *Proceedings of Interspeech*, pages 2437–2441, 2020.

[57] Jen-Tzung Chien, Kuan-Ting Kuo, et al. Variational Recurrent Neural Networks for Speech Separation. In *18TH Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Vols 1-6: Situated Interaction*, pages 1193–1197, 2017.

[58] Simon Leglaive, Laurent Girin, and Radu Horaud. A Variance Modeling Framework Based on Variational Autoencoders for Speech Enhancement. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

[59] Laxmi Pandey, Anurendra Kumar, and Vinay Namboodiri. Monoaural Audio Source Separation Using Variational Autoencoders. In *INTERSPEECH*, pages 3489–3493, 2018.

[60] Ertuğ Karamatlı, Ali Taylan Cemgil, and Serap Kırbız. Audio Source Separation using Variational Autoencoders and Weak Class Supervision. *IEEE Signal Processing Letters*, 26(9): 1349–1353, 2019.

[61] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Audio-Visual Speech Enhancement Using Conditional Variational Auto-Encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1788–1800, 2020.

[62] Hao Duc Do, Son Thai Tran, and Duc Thanh Chau. Speech Source Separation using Variational Autoencoder and Bandpass Filter. *IEEE Access*, 8:156219–156231, 2020.

[63] Katerina Zmolíková, Marc Delcroix, Lukás Burget, Tomohiro Nakatani, and Jan Honza Cernocký. Integration of Variational Autoencoder and Spatial Clustering for Adaptive Multi-Channel Neural Speech Separation. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 889–896. IEEE, 2021. doi: 10.1109/SLT48900.2021.9383612. URL `https://doi.org/10.1109/SLT48900.2021.9383612`.

[64] Guillaume Carbajal, Julius Richter, and Timo Gerkmann. Guided Variational Autoencoder for Speech Enhancement with a Supervised Classifier. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 681–685. IEEE, 2021.

[65] Li Li, Hirokazu Kameoka, and Shoji Makino. Fast MVAE: Joint Separation and Classification of Mixed Sources based on Multichannel Variational Autoencoder with Auxiliary Classifier. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 546–550. IEEE, 2019.

[66] Shogo Seki, Hirokazu Kameoka, Li Li, Tomoki Toda, and Kazuya Takeda. Underdetermined Source Separation based on Generalized Multichannel Variational Autoencoder. *IEEE Access*, 7:168104–168115, 2019.

[67] Viet-Nhat Nguyen, Mostafa Sadeghi, Elisa Ricci, and Xavier Alameda-Pineda. Deep Variational Generative Models for Audio-Visual Speech Separation. In *31st IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2021, Gold Coast, Australia, October 25-28, 2021*, pages 1–6. IEEE, 2021. doi: 10.1109/MLSP52302.2021.9596406. URL `https://doi.org/10.1109/MLSP52302.2021.9596406`.

[68] Laurent Girin, Fanny Roche, Thomas Hueber, and Simon Leglaive. Notes on the use of Variational Autoencoders for Speech and Audio Spectrogram Modeling. In *DAFx 2019-22nd International Conference on Digital Audio Effects*, pages 1–8, 2019.

[69] Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Meng Yu, and Dong Yu. A Comprehensive Study of Speech Separation: Spectrogram vs Waveform Separation. In *Proc. Interspeech 2019*, pages 4574–4578, 2019. doi: 10.21437/Interspeech.2019-3181. URL `http://dx.doi.org/10.21437/Interspeech.2019-3181`.

[70] Jesse H. Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable Digital Signal Processing. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1x1ma4tDr.

[71] Gene-Ping Yang, Chao-I Tuan, Hung-yi Lee, and Lin-Shan Lee. Improved Speech Separation with Time-and-Frequency Cross-Domain Joint Embedding and Clustering. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1363–1367. ISCA, 2019. doi: 10.21437/Interspeech.2019-2181. URL https://doi.org/10.21437/Interspeech.2019-2181.

[72] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson. Speech Enhancement Using Bayesian WaveNet. In *Interspeech*, pages 2013–2017, 2017.

[73] Dario Rethage, Jordi Pons, and Xavier Serra. A WaveNet for Speech Denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[74] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial Audio Synthesis. In *International Conference on Learning Representations*, 2018.

[75] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: A Flow-based Generative Network for Speech Synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3617–3621. IEEE, 2019. doi: 10.1109/ICASSP.2019.8683143. URL https://doi.org/10.1109/ICASSP.2019.8683143.

[76] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. SEGAN: Speech Enhancement Generative Adversarial Network. In Francisco Lacerda, editor, *INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 3642–3646. ISCA, 2017. URL http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1428.html.

[77] Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 5024–5028. IEEE, 2018. doi: 10.1109/ICASSP.2018.8462581. URL https://doi.org/10.1109/ICASSP.2018.8462581.

[78] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. MetricGAN: Generative Adversarial Networks based Black-box Metric Scores Optimization for Speech Enhancement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2031–2041. PMLR, 2019. URL http://proceedings.mlr.press/v97/fu19b.html.

[79] Martin Strauss and Bernd Edler. A Flow-Based Neural Network for Time Domain Speech Enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 5754–5758. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413999. URL https://doi.org/10.1109/ICASSP39728.2021.9413999.

[80] Albert Gu, Karan Goel, and Christopher Re. Efficiently Modeling Long Sequences with Structured State Spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

[81] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's Raw! Audio Generation with State-Space Models. *arXiv preprint arXiv:2202.09729*, 2022.

[82] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

[83] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality. *arXiv preprint arXiv:2205.04421*, 2022.

[84] Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5885–5892, 2019.

[85] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The Autoencoding Variational Autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 15077–15087. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/ac10ff1941c540cd87c107330996f4f6-Paper.pdf`.

[86] Sander Dieleman, Charlie Nash, Jesse H. Engel, and Karen Simonyan. Variable-Rate Discrete Representation Learning. *CoRR*, abs/2103.06089, 2021. URL `https://arxiv.org/abs/2103.06089`.

[87] Kevin P Murphy. Conjugate Bayesian Analysis of the Gaussian distribution. Technical report, The University of British Columbia, 2007.

[88] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. VoxCeleb: Large-scale Speaker Verification in the Wild. *Computer Science and Language*, 2019.

[89] Manuel Pariente, Samuele Cornell, Joris Cosentino, Sunit Sivasankaran, Efthymios Tzinis, Jens Heitkaemper, Michel Olvera, Fabian-Robert Stöter, Mathieu Hu, Juan M Martín-Doñas, et al. Asteroid: the Pytorch-based Audio Source Separation Toolkit for Researchers. *arXiv preprint arXiv:2005.04132*, 2020.

[90] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624.

[91] Chris Donahue, Ian Simon, and Sander Dieleman. Piano Genie. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 160–164, 2019.

[92] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked Autoencoder for Distribution Estimation. In *International Conference on Machine Learning*, pages 881–889. PMLR, 2015.

Supplementary Materials: Improving Speaker Separation Generalization with Variational Inference

A Overview

- Appendix A: this overview.
- Appendix B: a derivation of the ELBO for the VI-TasNet, building on standard result of the ELBO for standard VAEs.
- Appendix C: visualization of the example VI-TasNet, and a comparison to a spectrogram
- Appendix D: overview of the multitasking VI-TasNet with discussion of the adaptive prior
- Appendix E: introduction of modified VI-TasNet ELBO and results on clean speaker separation condition.
- Appendix F: definition of used notation for distributions, the used priors, and expressions for convolutions.
- Appendix G: derivation of different expression for the SI-SDR based on unit-vectors.
- Appendix H: derivation of a scale-invariant likelihood based on Bayesian linear regression, and comparison to the SI-SDR.
- Appendix I: introduction and discussion of the multivariate Cauchy objective.
- Appendix J: various details on data, model and training setup.
- Appendix K: details on the synthetic data experiment.
- Appendix L: results for models trained on clean LibriMix for VI-TasNets with various types of priors.
- Appendix M: rate-distortion trade-off results for further test conditions in the synthetic problem.
- Appendix N: an expanded table with further metrics for the noisy LibriMix/VCTK evaluation.

B VI-TasNet evidence lower bound

We start by considering the standard VAE evidence lower bound (ELBO), and use this as a starting point for showing the slightly more notationally involved bound for the VI-TasNet. Conceptually, the extra latents simply add an extra KL divergence term in the VI-TasNet formulation.

We will write the Kullback-Leibler divergence between distributions $a(x)$ and $b(x)$ as (note that we are using lower x here to denote a stochastic variable):

$$D_{KL}(a(x)||b(x)) = \int_{-\infty}^{\infty} a(x) \log \frac{a(x)}{b(x)} dx = \mathop{\mathbb{E}}_{x\sim a(x)} \left[\log \frac{a(x)}{b(x)} \right] \geq 0 \tag{6}$$

B.1 VAE evidence lower bound

In practice, we consider data from a particular dataset, $\mathcal{D}$, and consider an empirical data distribution $p_{\mathcal{D}}(x) = \frac{1}{||\mathcal{D}||} \sum_{x' \in \mathcal{D}} (\delta(x - x'))$ (where $||\mathcal{D}||$ denotes the cardinality of the dataset), which we hope reflects some true data distribution. The derivations below follow for a single sample from $p_{\mathcal{D}}(x)$, and we derive a lower bound on the evidence conditioned on that sample, $\mathcal{L}(\theta, \varphi; x)$, dependent on the generative and inference network parameters. We optimize the bound for such samples, but this can be also be extended to considering an expectation over the dataset and in a batch setting, too, such that we can extend this to consider a total bound over the dataset, $\mathcal{L}(\theta, \varphi) = \mathbb{E}_{x\sim p_{\mathcal{D}}(x)} \left[\mathcal{L}(\theta, \varphi; x) \right]$ (as discussed in e.g. Zhao et al. [84] and Cemgil et al. [85]).

We consider random variables x and z (adopting lower case notation for this similar to the one used in Zhao et al. [84]). Note that $p_\theta(x, z) = p_\theta(z|x)p_\theta(x) = p_\theta(x|z)p_\theta(z)$, and e.g. $p_\theta(x) = \frac{p_\theta(x,z)}{p_\theta(z|x)}$ (assuming here that the denominator is non-zero everywhere), where the subscript θ denotes the dependency on generative network parameters. An expression for the ELBO can be arrived at by introducing an expectation over the variational distribution $q_\varphi(z|x)$ (dependent on the inference network parameters φ) and re-arranging[2]:

$$\begin{aligned}
\log p_\theta(x) &= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(p_\theta(x) \right) \right] && (7)\\
&= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(p_\theta(x) \frac{q_\varphi(z|x)}{q_\varphi(z|x)} \right) \right] && (8)\\
&= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(\frac{p_\theta(x,z)}{p_\theta(z|x)} \frac{q_\varphi(z|x)}{q_\varphi(z|x)} \right) \right] && (9)\\
&= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(\frac{p_\theta(x,z)}{q_\varphi(z|x)} \right) + \log \left(\frac{q_\varphi(z|x)}{p_\theta(z|x)} \right) \right] && (10)\\
&= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(\frac{p_\theta(x,z)}{q_\varphi(z|x)} \right) \right] + \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(\frac{q_\varphi(z|x)}{p_\theta(z|x)} \right) \right] && (11)\\
&= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(\frac{p_\theta(x,z)}{q_\varphi(z|x)} \right) \right] + D_{KL} \left(q_\varphi(z|x) || p_\theta(z|x) \right) && (12)\\
&\geq \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)} \left[\log \left(\frac{p_\theta(x,z)}{q_\varphi(z|x)} \right) \right] = \mathcal{L}_{\text{VAE}}(\theta, \varphi; x) && (13)\\
& && (14)
\end{aligned}$$

Where we used that the KL divergence between the approximate posterior over the latents z and the true (but intractable) posterior is a non-negative quantity, such that the model evidence is lower bounded by the expression $\mathcal{L}_{\text{VAE}}$. This expression can be rewritten as:

[2] While this derivation follows e.g. Sec. 2.2 in Kingma and Welling [13], we note that the bound can be derived using Jensen's inequality by introducing the variational distributions in a very similar manner.

$$\mathcal{L}_{\text{VAE}}(\theta, \varphi; x) = \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\log\left(\frac{p_\theta(x|z)p(z)}{q_\varphi(z|x)}\right)\right] \tag{15}$$

$$= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\log\left(p_\theta(x|z)\right) + \log\left(\frac{p(z)}{q_\varphi(z|x)}\right)\right] \tag{16}$$

$$= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\log\left(p_\theta(x|z)\right)\right] - D_{KL}\left(q_\varphi(z|x)||p(z)\right), \tag{17}$$

where the first term corresponds to the negative distortion (the distortion is the negative log-likelihood), and the second term corresponds to a rate (the KL-divergence between approximate posterior and prior).

B.2 VI-TasNet evidence lower bound

We denote a set of single sources of time-series $\mathcal{S} = \{s_0, \ldots, s_N\}$. Figure 1 shows a two-speaker scenario, matching the data used for the various experiments (i.e., on the figure we have $\mathcal{S} = \{s_a, s_b\}$). We assume a joint distribution over single sources ($\mathcal{S}$), their mixture (x), corresponding masks ($\mathcal{M} = \{m_0, \ldots, m_N\}$), and the mixture encoding (z) like follows:

$$p_\theta(\mathcal{S}, x, \mathcal{M}, z) = p_\theta(x|\mathcal{S})p_\theta(\mathcal{S}|z, \mathcal{M})p_\theta(\mathcal{M}|z)p_\theta(z), \tag{18}$$

where we assume that $p_\theta(\mathcal{S}|z, \mathcal{M}) = \prod_{i=0}^{N} p_\theta(s_i|z, m_i)$ and $p_\theta(\mathcal{M}|z) = \prod_i^N p_\theta(m_i|z)$. We also assume that these distributions factorize over the temporal dimension (latent or original data space), as e.g. shown for the mask prior later (cf. Eq. 36). With repeated application of the product rule, we have that:

$$p_\theta(\mathcal{S}, x, \mathcal{M}, z) = p_\theta(x, \mathcal{M}, z|\mathcal{S})p_\theta(\mathcal{S}) \tag{19}$$

$$p_\theta(\mathcal{S}) = \frac{p_\theta(\mathcal{S}, x, \mathcal{M}, z)}{p_\theta(x, \mathcal{M}, z|\mathcal{S})} \tag{20}$$

$$p_\theta(x, \mathcal{M}, z|\mathcal{S}) = p_\theta(\mathcal{M}|\mathcal{S}, x, z)p_\theta(z|\mathcal{S}, x)p_\theta(x|\mathcal{S}). \tag{21}$$

The assumption for the set of masks might be worth exploring, seeing as e.g. self-consistency in the deterministic TasNet under the sometimes used softmax/sum-to-one-like constraints enforce a dependency between masks. The formulation with $p_\theta(x|\mathcal{S})$ could accommodate some stochastic mixing process, which is not the case for the data and model we are considering[3], and instead $x = \text{Mix}(\mathcal{S})$ is a deterministic mapping from sources to a mixture, or $p_\theta(x|\mathcal{S}) = \delta(x - \text{Mix}\left(\mathcal{S}\right))$. Realistic mixture generation functions are not noise-free or additive, but should e.g. take into account a reverberant environment with different spatial locations of speakers and multiple noise sources. In this setting, multi-channel recordings are of value in enabling resolving different spatial locations, and deep learning systems in general can benefit from utilizing systems that have traditional been used to improve performance (see e.g. Jenrungrot et al. [50], which uses a deep learning version of beam forming).

Following a similar derivation of the ELBO for the VAE, we start by introducing an expectation over the variational approximation to the latent encodings arising from the inference network with parameters φ, and an expectation over the variational approximation to the latent masks arising from the masker network with parameters ψ. We also, for the purpose of illustration, include an expectation over the mixing process, initially:

[3]For the LibriMix data considered, the dataset is static, and while the mixing process randomly samples a SNR in a particular range during the creation of the dataset, this mixing SNR does not change after the dataset is made.

$$\log p_\theta(\mathcal{S}) = \mathop{\mathbb{E}}_{x\sim\delta(x-\text{Mix}(\mathcal{S}))}\left[\mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\mathop{\mathbb{E}}_{\mathcal{M}\sim q_\psi(\mathcal{M}|z)}\left[\log\left(p_\theta(\mathcal{S})\right)\right]\right]\right] \tag{22}$$

$$= \mathop{\mathbb{E}}_{\substack{x\sim\delta(x-\text{Mix}(\mathcal{S}))\\ z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\log\left(p_\theta(\mathcal{S})\frac{q_\varphi(z|x)}{q_\varphi(z|x)}\frac{q_\psi(\mathcal{M}|z)}{q_\psi(\mathcal{M}|z)}\right)\right] \tag{23}$$

$$= \mathop{\mathbb{E}}_{\substack{x\sim\delta(x-M(\mathcal{S}))\\ z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\log\left(\frac{p_\theta(\mathcal{S},x,\mathcal{M},z)}{p_\theta(\mathcal{M}|\mathcal{S},x,z)p_\theta(z|\mathcal{S},x)p_\theta(x|\mathcal{S})}\frac{q_\varphi(z|x)}{q_\varphi(z|x)}\frac{q_\psi(\mathcal{M}|z)}{q_\psi(\mathcal{M}|z)}\right)\right] \tag{24}$$

Dropping the mixing process expectation and using x instead of $\text{Mix}(\mathcal{S})$ to highlight it as the input to the inference network, the expression can be written as:

$$\log p_\theta(\mathcal{S}) = \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\mathop{\mathbb{E}}_{\mathcal{M}\sim q_\psi(\mathcal{M}|z)}\left[\log\left(\frac{p_\theta(\mathcal{S},x,\mathcal{M},z)}{p_\theta(\mathcal{M}|\mathcal{S},x,z)p_\theta(z|\mathcal{S},x)}\frac{q_\varphi(z|x)}{q_\varphi(z|x)}\frac{q_\psi(\mathcal{M}|z)}{q_\psi(\mathcal{M}|z)}\right)\right]\right] \tag{25}$$

$$= \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\mathop{\mathbb{E}}_{\mathcal{M}\sim q_\psi(\mathcal{M}|z)}\left[\log\left(\frac{p_\theta(\mathcal{S},x,\mathcal{M},z)}{q_\varphi(z|x)q_\psi(\mathcal{M}|z)}\frac{q_\varphi(z|x)}{p_\theta(z|\mathcal{S},x)}\frac{q_\psi(\mathcal{M}|z)}{p_\theta(\mathcal{M}|\mathcal{S},x,z)}\right)\right]\right] \tag{26}$$

Splitting the factors within the logarithm out as addends, and noting that the expression with encodings does not depend on the mask latent, we get:

$$\mathop{\mathbb{E}}_{\substack{z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\log\left(\frac{p_\theta(\mathcal{S},x,\mathcal{M},z)}{q_\varphi(z|x)q_\psi(\mathcal{M}|z)}\right)\right] + \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[\frac{q_\varphi(z|x)}{p_\theta(z|\mathcal{S},x)}\right] + \mathop{\mathbb{E}}_{\substack{z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\frac{q_\psi(\mathcal{M}|z)}{p_\theta(\mathcal{M}|\mathcal{S},x,z)}\right] \tag{27}$$

$$= \mathcal{L}(\theta,\varphi,\psi;\mathcal{S}) + D_{KL}\left(q_\varphi(z|x)||p_\theta(z|\mathcal{S},x)\right) + \mathop{\mathbb{E}}_{z\sim q_\varphi(z|x)}\left[D_{KL}\left(q_\psi(\mathcal{M}|z)||p_\theta(\mathcal{M}|\mathcal{S},x,z)\right)\right] \tag{28}$$

Since the divergences are non-negative quantities, measuring how close the variational approximation for the latent encodings and latent masks are to the true posteriors, the last term is a lower bound on the evidence over the set of speakers, $\mathcal{L}(\theta,\varphi,\psi;\mathcal{S})$. We can write this as:

$$\log p_\theta(\mathcal{S}) \geq \mathcal{L}(\theta,\varphi,\psi;\mathcal{S}) \tag{29}$$

$$= \mathop{\mathbb{E}}_{\substack{z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\log\left(\frac{p_\theta(\mathcal{S},x,\mathcal{M},z)}{q_\varphi(z|x)q_\psi(\mathcal{M}|z)}\right)\right] \tag{30}$$

$$= \mathop{\mathbb{E}}_{\substack{z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\log\left(\frac{p_\theta(\mathcal{S}|z,\mathcal{M})p_\theta(\mathcal{M}|z)p_\theta(z)}{q_\varphi(z|x)q_\psi(\mathcal{M}|z)}\right)\right] \tag{31}$$

We can also split the bound in three contributing terms:

$$\log p_\theta(\mathcal{S}) \geq \mathcal{L}(\theta,\varphi,\psi;\mathcal{S}) \tag{32}$$

$$= \mathop{\mathbb{E}}_{\substack{z\sim q_\varphi(z|x)\\ \mathcal{M}\sim q_\psi(\mathcal{M}|z)}}\left[\log\left(p_\theta(\mathcal{S}|z,\mathcal{M})\right)\right] - D_{KL}\left(q_\varphi(z|x)||p_\theta(z)\right) - D_{KL}\left(q_\psi(\mathcal{M}|z)||p_\theta(\mathcal{M}|z)\right) \tag{33}$$

$$= -D_\mathcal{S} - R_z - R_\mathcal{M} \tag{34}$$

Here, $D_\mathcal{S}$ is the distortion of the single sources (how well they are reconstructed), R_z is the divergence of the mixture encoding from their prior (the encoding rate), and $R_\mathcal{M}$ is the divergence of the masks

from their prior (the mask rate). Note that since we assume the single sources conditioned on their mask and the mixture encoding factorize, the distortion is the sum of the log-likelihoods for each individual source. As is common when training VAEs, we estimate the expectation with single samples from the approximate posteriors for z and $\mathcal{M}$, and have defined priors for these latents, which enable us to evaluate the objective. For the models we consider, we are also using a prior on the masks that is independent of z, both for the log-normal masker (used in main paper) and the beta masker discussed in this supplementary:

$$p_\theta^{\mathcal{LN}}(\mathcal{M}|z) = p_\theta^{\mathcal{LN}}(\mathcal{M}) = \prod_{i=0}^{N} \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathcal{LN}(m_{i,k,t'}; 0.9, 1), \tag{35}$$

$$p_\theta^{\mathcal{B}}(\mathcal{M}|z) = p_\theta^{\mathcal{B}}(\mathcal{M}) = \prod_{i=0}^{N} \prod_{k=0}^{K} \prod_{t'=0}^{T'} \text{Beta}(m_{i,k,t'}; 1, 1), \tag{36}$$

where k indicates a particular latent dimension, and t' indicates the latent time-step, and i is speaker index. We hypothesize that e.g. a dependency on the strength of the prior mask based on the "energy" in the encodings (the distance from zero in the Gaussian case, or the concentration parameter in the gamma case) might improve learning. In the case of the multi-tasking VI-TasNet, we hypothesize that using a mask prior that explicitly depend on the relative energies in single source encodings at particular times and latent dimensions could be useful, too.

C Visualize model outputs

We visualize the outputs the model with a gamma approximate posterior in Table 3. Figure 4 shows the example output of a network for approximately 50 ms of input audio. Similarly, Figure 5 shows a comparison over approximately four seconds of input mixture of the learnt representation (encodings) compared to a spectrogram.

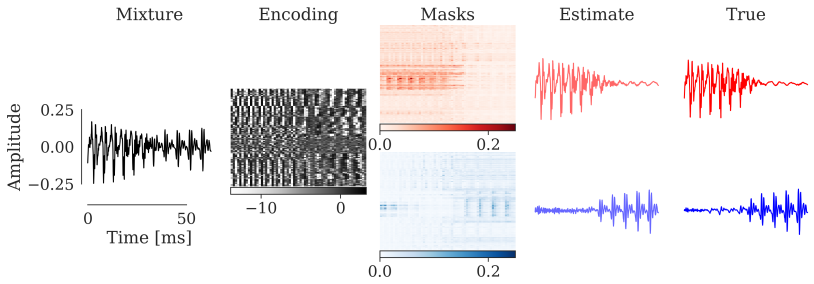


Figure 4: Example output of a VI-TasNet. From left to right: input mixture, mixture encoding sample, masks samples, estimates, and ground truth single sources. Based on an audio mixture input, the inference network provides an encoding of the mixture. These encodings are here visualized as the log-value of a sample from a gamma approximate posterior. The latent dimensions have been sorted using an agglomerative clustering over the full sentence input (for visualization purposes solely). The encodings are processed by the masker, and the masker provides distributions for a fixed set of speakers, here two. The red and blue masks shown are samples from beta mask approximate posteriors for two different sources. The generative network sees the multiplication of the encodings and the masks to parametrize distributions of the estimated separated signals, and we visualize a sample from this distribution alongside the known ground truth.

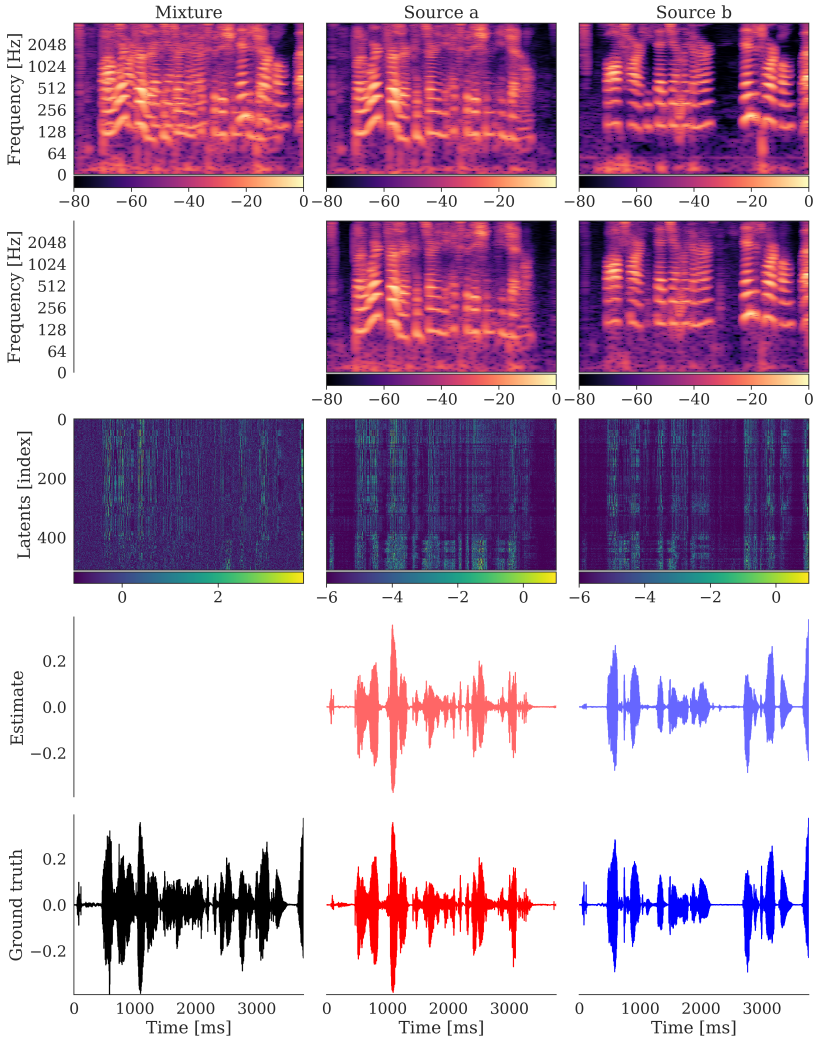


Figure 5: Comparison of encodings and spectrogram representation. Like Figure 4, the latents were sorted with a clustering. From top to bottom: ground truth spectrograms, estimated signal spectrograms, latents, estimated time-series, ground truth time-series. From left to right: input mixture, source A, and source B. The latents shown for the mixture are the output of the encoder network, while the latents shown for the single sources are the masked mixture encodings. These results or not for a multitasking VI-TasNet, so not visualizations are present for the mixture in estimated spectrograms or time-series.

D Multitasking VI-TasNet and prior adaptivity

Figure 6 shows the multitasking VI-TasNet. The yellow arrows enable the auto-encoding task on the single sources, whereas the purple arrows enable the mixture autoencoding task, which can be done without knowledge of the single sources. For the adaptive prior, we note that the visualized z_a and z_b are added (their distributions convolved) in making the adaptive prior for z.

This, we hypothesized, would improve learning since the model would have a direct path (circumventing the masker) for how encodings that would reconstruct a single source should look like. That is, with the masked encodings $\hat{z}_a = z \odot m_a$ we aim to reconstruct the single source, s_a, and with the distortion from the single sources, we add to the objective that it should try to reconstruct it directly from z_a, as well.

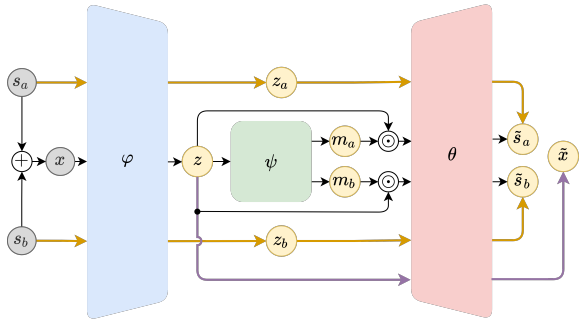


Figure 6: Multitasking VI-TasNet. The figure uses the same notation as Figure 1, and introduces the autoencoding tasks. The added single source autoencoding tasks are highlighted with orange arrows, and the added mixture autoencoding task is highlighted with purple arrows. The use of $+$ operator in mixing s_a and s_b is a simplification to the $\text{Mix}(\circ)$-mixing process. The encoder and decoder are not, in parameter counts, larger—the larger size is to visualize that the parameters are shared across tasks.

E Priors, rate-distortion and modified VI-TasNet ELBO

Priors in VAEs A central part of the VI is the prior distributions used for the latent variables. The most common prior for a simple VAE latent encoding is an isotropic Gaussian distribution (i.e., a multivariate normal distribution with identity matrix covariance). Conceptually, it can be argued that this prior forces the VAE to learn latent dimensions with activations that generally tend to zero and that do not co-vary. The disentanglement is obtained by implicitly penalizing off-diagonal elements in the approximate posterior covariance. This prior is, in some scenarios, prohibitively restrictive. The benefits of more flexible priors are underlined by the improvements seen e.g. using a VampPrior [31] or by learning the prior distributions using normalizing flows [29, 30]. Other choices than the isotropic Gaussian prior and Gaussian approximate posterior provide tools for enforcing other characteristics on the learned encoding. For instance, by using von Mises-Fisher distributions [32], a learned representations reside on a unit hyper-sphere, forcing latents describing directions without considerations of magnitudes. Similarly, a non-negative encoding (parts based, akin to non-negative matrix factorization) can be achieved using log-normal, gamma or Weibull distributions [33].

Rate-distortion analysis Optimizing a modified loss different from the ELBO facilitates adjustment of the trade-off between accurate generation (i.e., reconstruction) and deviation from the prior (i.e., more tightly constrained). For instance, for an isotropic Gaussian prior, a lower rate is related to more disentangled latents [35]). Specifically, we can adjust the prioritization of the KL-term with a coefficient β (β-VAEs [34]). For $\beta < 1$, the model is less restricted by the prior, freeing the model to produce better reconstructions, and for $\beta > 1$ the models are forced to learn representations more aligned with the prior. This trade-off can be thought of as a trade-off between a distortion D, the decoded negative log-likelihood, and a rate R, the KL divergence between encoding approximate posterior and prior, since $-\mathcal{L} = D + R$.

While higher capacity models can generally achieve better model evidences (i.e., higher ELBOs), it is only up to a limit of the complexity of the data. An unconstrained ($\beta \ll 1$) model with sufficient capacity could reconstruct the inputs perfectly ($D = 0$) to the limit of the entropy of data by having a high R (i.e., the "auto-decoding limit") [37]. Similarly, a tightly constrained model with sufficient capacity for encoding can map to something with $R = 0$ (i.e., the auto-encoding limit) but high D. There is a gap between models that are feasible (i.e., within auto-encoding and -decoding limits) and models that are realizable. For realizable models, there exists an optimal trade-off (in terms of lowest ELBO) between rate and distortion, but the relative capacity of the encoder and decoder alter the optimal trade-off. The trade-offs can be visualized using RD curves (i.e., phase diagrams in the RD-plane) by optimizing different trade-offs (e.g., using β). Work in exploring rate-regularization and the role of the prior (e.g., concerning generalization) shows how an isotropic Gaussian prior might not be the best inductive bias in general [36].

The information bottleneck principle, as discussed in the context of deep learning by Tishby and Zaslavsky [38], provides a framework for understanding representation learning and generalization. In particular, their qualitative representation visualized in their Figure 2, shows how—under a finite data sample—an optimal rate exists to minimize a generalization gap. This is further investigated in the work by Alemi et al. [39], where they show how the information bottleneck principle applies to deep learning-based variational inference, and that an optimal rate exists to improve model test performance (generalization).

Representation learning and compression The mutual information between a learned latent representation and the observed data is lower bounded by the difference between the entropy of the data and distortion and upper bounded by the rate [37]. The RD trade-offs made are comparable to trade-offs in data compression. From the perspective of lossy compression, the RD trade-off can be thought of as reducing the complexity of the latents (i.e., more compression) at the expense of poorer reconstructions (i.e., increased distortion)—or the other way around. Note that, for VAEs, this analogy is less directly related to dimensionality of the latents, and more so a matter of prior divergence. In fact, latent variable models can be turned into (lossless) compression models [40], and for these models, the rate term is indeed related to the achievable compression rate. Lastly, recent studies of RD analysis have shown how an inherent trade-off exists between not just rate and distortion, but also a quantity measuring divergence between the encoder-decoder induced distribution over the data and the true data distribution [41]. This can be linked to e.g. the generative adversarial network objectives and naturalness—or potentially overall audio quality in the speaker separation

setting. Mostly, VAEs learn representations that are a compression in terms of e.g. dimensionality of the learned representation. However, in some settings learning over-complete representations (i.e., higher dimensionality of representation than input signal), under constraints or regularized, can be a sensible approach to representation learning, and Bengio et al. [42] discuss various general approaches to learning representations through over-complete AEs in relation to robustness of the representations, such as sparse, contractive, or denoising AEs.

E.1 Free bits and reweighting

Attempting to optimize the ELBO without any modifications often yields models effectively stuck in a local minimum of low rate, without a driving force of distortion sufficient to overcome the loss incurred of moving away from the prior. This is often true for simple, standard VAEs, but is, in particular, the case with high-dimensional over-complete representations like the ones for the VI-TasNet.

One solution to this is to e.g. introduce the rate term gradually, using KL annealing. Instead of annealing, we can opt for a "free information" approach to enable learning, as described in e.g. Kingma et al. [29, C.9]. When we measure the rate in bits/shannon, introducing free bits means that the model is always penalized some value of bits, providing a set "budget" (we measure the information in nats, here). This causes the model to effectively have the freedom to operate without being penalized within this budget. This should allow the model to trade off distortion and rate early on in training—and even go beyond the free budget. Since we have rate terms both for the encodings and the mask, we can investigate the effect of free bits, λ, for both individually, denoted by a subscript, and we will denote by $r_{z,k,t'}$ the rate contribution from encoding dimension k at latent time step t', and $r_{m_i,k,t'}$ the rate contribution from the k'th dimension in the mask for speaker i at latent time step t'.

We also make use of the re-weighting introduced with β-VAEs, and use the same subscript notation to denote encoding or mask specific re-weightings. We can write the modified ELBO as, using both β (a multiplicative factor on the rate term) and λ (free bits), with slight notation abuse (letting the output of maximum$(\circ, \circ)$ be the largest of the arguments):

$$\mathcal{L}_{\beta,\lambda}(\theta, \varphi, \psi; \mathcal{S}) = -D_{\mathcal{S}} - \beta_z \sum_{k}^{K} \sum_{t'}^{T'} \text{maximum}(\lambda_z, r_{z,k,t'}) - \beta_m \sum_{i}^{N} \sum_{k}^{K} \sum_{t'}^{T'} \text{maximum}(\lambda_m, r_{m_i,k,t'}) \tag{37}$$

E.2 Dynamic adaption towards target

We have also used a form of ELBO modification which uses adaption of the β_z and β_m terms based on two fixed, target rates for the sum over all K and T' rates of encodings and sum over all N, K and T' rates of the masks. We can combine the two, as well, and consider a total rate (sum of mask and encoding rates, with one, shared adapted β value). This is inspired by the target rate in Alemi et al. [37] and automatic penalty weighting used in Dieleman et al. [86]. Alemi et al. [37] showed how an objective which directly optimizes towards a target rate for a VAE learns a better model in a synthetic experiment with a known ground truth generative process. We investigated using the same approach as Alemi et al. [37] which can actively promote increased rates with the gradients (not just penalize), but for reasons of stability, we opted for using a version that incorporated an adaption similar to the one presented in Dieleman et al. [86]. We note that, with ways of ensuring stability in losses that promote increased rates, we might see better performances.

Dieleman et al. [86] showed that adaptively increasing or decreasing a re-weighting of a regularizing term enables optimizing towards a desired target value for a quantity of interest in an auto-encoder setting (in their work this is not variational auto-encoders, and not a rate/KL-term). When using adaptive re-weighting, we can use the same formulation as in their Sec. 3.1.3, Eq. 7, for both adapting the β_z and β_m, towards target rates for R_z and $R_{\mathcal{M}}$, or adapting a target total rate (the sum of the two rates) and sharing the adapted β weight. Compared to the free-bits and the fixed β approach, this approach has the benefit that we can control the rates towards a very specific point on the RD-curve. Starting with a low initial value for the adapted weight also provides something similar to annealing in the beginning of training. We note that this dynamic must be adjusted to the training; if too quick or too slow adaption happens, the learning can be needlessly slowed down. If e.g. early stopping or

learning rate annealing is used, a poorly specified adaption rate can cause the model to prematurely lower the learning rate or stop too early—even if the monitored objective is not the re-weighted ELBO but e.g. SI-SDR.

E.3 Numerical experimental results for RD-curve in clean condition

We now provide further details on the RD-curve analysis in the main paper for the real (i.e. no synthetic) results on the clean condition of LibriMix and VCTK. In the main paper, Table 1, we opted to primarily discuss a model with Gaussian encodings and log-normal masks to align the most with how a "standard" TasNet operates. In the results presented here, and on Figure 2 (bottom), we used a q_ϕ^Γ/p^Γ-model with beta-distribution maskers, also shown in Table 3, with the BLR objective. This was chosen because this formulation of the model, initially, more readily put information in the encodings (over a Gaussian encoding formulation) when using the free-bits and fixed β setup which we used for these earlier experiments; that is, with fewer optimization steps, the q_ϕ^Γ/p^Γ-model saw higher rates than the Gaussians. The VI-TasNet in Table 1 was ultimately trained using the adaptive re-weighting scheme (instead of a fixed), which lessened this advantage of the gamma distribution formulation over the Gaussian, causing us to opt for a formulation more closely resembling the linear encoder outputs from a deterministic TasNet in the later experiments.

In this section, we present results where we modified the VI-TasNet ELBO with varying levels of free bits and with varying (but static for a given model) weights on both the masker rate and encoding rate. Figure 7 shows a visualization where the different models fall in the RD-plane, and how they generalize from the LibriMix test set to the VCTK dataset. The RD-plane plots show how the VI-TasNet display an expected trade-off between rate and distortion; generally, by increasing the rate, the distortion is reduced. The generalization from LibriMix to VCTK push the RD-curve trade-off up (poorer rate) and to the right (poorer reconstruction), for a generally overall poorer performance. Increases in rate alone is primarily driven by the encoding rates (the masking rates are largely unchanged between test sets). The increased rates (contributing to increased ELBOs) indicate that the differences in the datasets mostly affect the encoder and decoder, and less so the masker. The model with the lowest distortion on the LibriMix also has a higher rate, but this trade-off did not result in an improved distortion on VCTK, indicating a relation between the overall rate and the dataset generalization gap.

Table 2: Rate distortions for VI-TasNets with gamma encoding and BLR-likelihood for various levels of free-bits, λ, and β-values (as introduced in Eq. 37, but since we only use $\beta < 1$, we give the reciprocal in the table).

λ_z	λ_m	$1/\beta_z$	$1/\beta_m$	Libri2Mix test				VCTK-2mix test			
				SI-SDRi	D	R_z	R_m	SI-SDRi	D	R_z	R_m
0	0	1	1	-32.6	-1.95	4.08	0.00	-29.4	-1.90	4.09	0.00
8	8	1	1	8.09	-2.84	6.28	288	7.78	-2.74	6.49	288
64	64	1	1	12.0	-3.08	133	310	10.7	-2.91	149	310
128	128	1	1	11.6	-3.06	119	295	10.4	-2.89	136	295
128	128	8	8	12.5	-3.11	163	338	11.1	-2.93	182	336
128	128	64	64	12.9	-3.13	195	333	11.3	-2.93	215	332
128	128	128	128	12.7	-3.12	198	344	11.2	-2.93	218	343
128	128	8	128	12.7	-3.12	178	316	11.4	-2.94	197	315
128	128	128	8	12.5	-3.11	170	315	11.3	-2.93	189	315

We show numerical results in Table 2 that are the basis for the visualizations in Figure 7 (which in turn is the full version of the bottom figure in Figure 2 in main paper). We note that the un-modified ELBO results in a model stuck at a low masking and encoding rate, with poor distortion/SI-SDRi. This model corresponds to a point in the far right-hand side and bottom of the RD plane, and for visualization purposes, it was left out on the RD curves. Note that all values are normalized by T in the plots and that the rates are normalized by the number of latent dimensions (K), too. The table gives the average mask rate (over $N = 2$ speakers as in all problems considered here), and total rate is $R_z + R_{\mathcal{M}} = R_z + N \cdot R_m$.

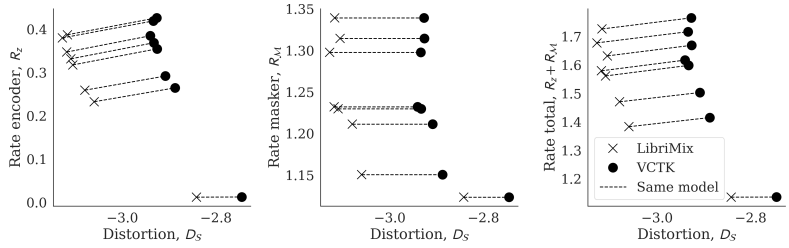


Figure 7: Rates versus distortion. Rate and distortions values are normalized by T, and the rates are normalized additionally by K.

The last three rows in Table 2 provide a comparison between a model that has overall low encoding and masking β-values with one that penalizes each more (≈ 0.1 versus ≈ 0.01). Penalizing both relatively little (with $\beta_m = \beta_z = 1/128$) resulted in the highest encoding and mask rates shown in the table, whereas increasing the penalty on *either* lowered the rates for *both*. We show one run for each model configuration, as also discussed in Appendix J, and so to further resolve the uncertainty in these estimates multiple runs would be needed.

The improved performance (in terms of SI-SDR) from lower β-values is presumably largely attributable to the learning of an over-complete representation, but might additionally partially be due to a need to compensate for the differences in overall scale in the distortions and rates considered. The distortion for a very poor model is at approximately -2 nats, a poor model at approximately -2.8, and the best models at about -3.1, whereas the same models e.g. have encoding rates ranging from about 4 to 200 nats. We use a continuous output distribution; there is a difference in differential entropy/continuous entropy and the (discrete) entropy, and we note that e.g. a discrete output distribution would enable the use of the theoretical results provided in Alemi et al. [37] concerning the relationships between the entropy of the modelled data to the rate and distortion. We hypothesize that a discrete output distribution, e.g. a discretized (mixture of) logistic(s), might provide a suitable alternative to counter this rate-distortion scale difference, but the standard variants lack the scale-invariance and logarithmically scaled error measurement; for this, the Cauchy distribution (discussed in the supplementary, too) provides an alternative solution worth considering, although it does not have scale-invariance.

We provide a view of the RD curves where the negative SI-SDR replaces the BLR distortion on Figure 8. In Appendix H, we discuss how the BLR objective, while similar to the SI-SDR, reweights terms of the objective depending on the length of the signals considered. This, in part, causes the shift to the right on the distortion axis on Figure 7 in going from LibriMix to VCTK, since the overall average length of sentences in the two datasets is different. The overall conclusions regarding the shape and trade-offs are, however, still valid, e.g. supported by the Figure 8 with SI-SDR, which does not have this T-dependency.

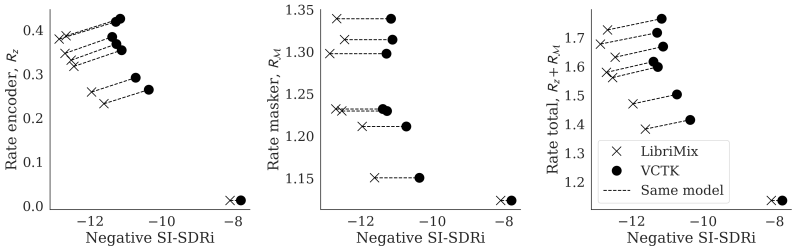


Figure 8: Rate versus negative SI-SDRi. Left: normalized encoding rate. Center: normalized masker rate. Right: normalized total rate.

F Distributions

With $\mathcal{N}(x; \mu, \sigma^2)$ we denote a (univariate) Gaussian with scalar mean μ and scalar variance σ^2. The gamma distribution density function we write as $\mathrm{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$, where Γ denotes the gamma function. Note that the α and β here has no connection to the β-VAE, nor the scaling in the SI-SDR. We refer to the α for the gamma distribution as the concentration, and β as the gamma rate[4].

The encoder distributions are specified as:

$$q_\varphi^{\mathcal{N}}(z|x) = \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathcal{N}(z_{k,t'}; \mu_{k,t'}^\varphi, (\sigma_{k,t'}^\varphi)^2), \tag{38}$$

$$q_\varphi^{\mathcal{LN}}(z|x) = \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathcal{LN}(z_{k,t'}; \mu_{k,t'}^\varphi, (\sigma_{k,t'}^\varphi)^2), \tag{39}$$

$$q_\varphi^{\Gamma}(z|x) = \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathrm{Gamma}(z_{k,t'}; c_{k,t'}^\varphi, 1), \tag{40}$$

where $\mu_{k,.}^\varphi$, $\sigma_{k,.}^\varphi$ are time-series of length T' with distribution parameters for the k'th latent dimension output. Similarly, $c_{k,.}^\varphi$ is a concentration parameter time-series output. We also define a Gaussian prior, $p^{\mathcal{N}}(z) = \prod_{k=0}^{K} \prod_{t'}^{T'} \mathcal{N}(z_{k,t'}; 0, 1^2)$ (and the equivalent log-normal version), and a gamma prior $p^{\Gamma}(z) = \prod_{k=0}^{K} \prod_{t'}^{T'} \mathrm{Gamma}(z_{k,t'}; \frac{1}{2}, 1)$.

Masker distribution The output of the masker parameterizes stochastic masks for all N speakers, where we opted for a variational approximation using either a beta distribution or a log-normal distribution:

$$q_\psi^{\mathcal{B}}(\mathcal{M}|z) = \prod_{n=0}^{N} \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathrm{Beta}(m_{n,k,t'}; \kappa_{n,k,t'}^{\psi,0}, \kappa_{n,k,t'}^{\psi,1}), \tag{41}$$

$$q_\psi^{\mathcal{LN}}(\mathcal{M}|z) = \prod_{n=0}^{N} \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathcal{LN}(m_{n,k,t'}; \mu_{k,t'}^{\psi}, (\sigma_{k,t'}^{\psi})^2), \tag{42}$$

where $\kappa_{n,k,.}^{\psi,0}$, $\kappa_{n,k,.}^{\psi,1}$ are time-series of parameters output from the masking network, h_ψ, that correspond to the k'th latent dimension for the n'th source. The $\kappa^{\psi,0}$ models increasing the likelihood of the mask being closer to 0, and $\kappa^{\psi,1}$ the same but for a value of 1. The log-normal parameters are transformed versions of corresponding Gaussian parameters. We opt for a flat, uniform prior for the Beta-distributed masks, such that $p_\theta(\mathcal{M}) = \prod_{n=0}^{N} \prod_{k=0}^{K} \prod_{t'=0}^{T'} \mathrm{Beta}(m_{n,k,t'}; 1, 1)$, but note that the inductive bias of an e.g. Jeffrey's prior towards either exclusion or inclusion of encoding elements is worthwhile investigating. For the log-normal masks, we use a standard lognormal distribution: $\mathcal{LN}(0, 1)$ prior.

Convolutions of standard distributions For Gaussians, we have that: $\sum_i \mathcal{N}(\mu_i, \sigma_i^2) \sim \mathcal{N}\left(\sum_i \mu_i, \sum_i \sigma_i^2\right)$, and for gamma distributions with one fixed gamma rate parameter, we have that: $\sum_i \mathrm{Gamma}\left(c_i, r\right) \sim \mathrm{Gamma}\left(\sum_i c_i, r\right)$.

[4]Following the naming convention of `torch.distributions.gamma.Gamma`.

G Scale-invariance

We consider time series of length T, $s \in \mathbb{R}^{1\times T}$, and we approximate true single source speakers, s, with the the approximation $\tilde{s}$. In training TasNets, a standard approach is to optimize the SI-SDR (or minimize the negative SI-SDR). First, we provide a view of this as related to projection of estimates onto the true sources. Following this, we introduce a likelihood invariant to a scaling.

G.1 Negative scale-invariant signal-distortion-ratio

The *negative* SI-SDR (NSISDR, for convenience) is[5]:

$$\text{NSISDR}(s,\tilde{s}) = -10\log_{10}\frac{\|\alpha s\|^2}{\|\alpha s-\tilde{s}\|^2} = -10\log_{10}\frac{\|\frac{\langle s,\tilde{s}\rangle}{\langle s,s\rangle}s\|^2}{\|\frac{\langle s,\tilde{s}\rangle}{\langle s,s\rangle}s-\tilde{s}\|^2}, \tag{43}$$

Since the numerator is:

$$\|\alpha s\|^2 = \sum_i (\alpha s_i)^2 = \alpha^2\sum_i s_i^2 = \alpha^2\langle s,s\rangle = \frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}, \tag{44}$$

and the denominator is:

$$\|\alpha s-\tilde{s}\|^2 = \sum_i\left((\alpha s_i)^2+\tilde{s}_i^2-2\alpha s_i\tilde{s}_i\right) = \frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}+\sum_i\tilde{s}^2-2\alpha\sum_i s_i\tilde{s}_i \tag{45}$$

$$= \frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}+\langle\tilde{s},\tilde{s}\rangle-2\frac{\langle s,\tilde{s}\rangle}{\langle s,s\rangle}\langle s,\tilde{s}\rangle = \langle\tilde{s},\tilde{s}\rangle-\frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle} \tag{46}$$

We can rewrite the NSISDR in Eq. 43, as:

$$-10\log_{10}\left(\frac{\frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}}{\langle\tilde{s},\tilde{s}\rangle-\frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}}\right) = 10\log_{10}\left(\frac{\langle\tilde{s},\tilde{s}\rangle-\frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}}{\frac{\langle s,\tilde{s}\rangle^2}{\langle s,s\rangle}}\right) = 10\log_{10}\left(\frac{\langle s,s\rangle\langle\tilde{s},\tilde{s}\rangle}{\langle\tilde{s},s\rangle^2}-1\right) \tag{47}$$

Furthermore, we can consider a version where we rescale s and $\tilde{s}$ to be unit vectors:

$$\text{NSISDR}(s,\tilde{s}) = 10\log_{10}\left(\langle\hat{e}_{\tilde{s}},\hat{e}_s\rangle^{-2}-1\right), \tag{48}$$

where $\hat{e}_a = a/\sqrt{\langle a,a\rangle}$. That is, for a given s and its estimate, the NSISDR is related to the squared projection of $\hat{e}_{\tilde{s}}$ onto $\hat{e}_s$

[5]Here we let $\|\circ\|$ be the 2-norm and $\langle\circ,\circ\rangle$ is the inner product operator, such that $a^\top a = \langle a,a\rangle = \|a\|^2 = \sum_i a_i^2 = \sum_i a_i a_i$, where a_i is the i'th element of the column-vector a and $\sum_i$ implies a sum over all indices.

H Bayesian linear regression likelihood

Alternatively, we can consider a loss using a Bayesian linear regression likelihood. We are interested in a likelihood which is invariant to a re-scaling of the whole time-series. The SI-SDR handles this using the α, and in the following we take the view that a similar factor γ is a regression coefficient, which we will marginalize out. In classic Bayesian linear regression, we are interested in modelling an unknown regression coefficient, γ, and an unknown noise-scale parameter, σ^2. For the target time-series s with steps s_t, we consider a linear regression model, where the approximation $\tilde{s}$ takes the role of predictor[6]:

$$s_t = \tilde{s}_t \gamma + \epsilon_t, \tag{49}$$

where $\epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$. Here $\tilde{s}_t, \beta \in \mathbb{R}^{k \times 1 = 1 \times 1}$ are a scalar predictor and a regression coefficient, not vectors. The entire predictor time-series $\tilde{s}$ of length T corresponds to a design matrix of size $T \times 1$. The corresponding likelihood is proportional to:

$$p\left(s|\tilde{s}, \gamma, \sigma^2\right) \propto \left(\sigma^2\right)^{-T/2} \exp\left(-\frac{1}{2\sigma^2}\left(s - \tilde{s}\gamma\right)^\top \left(s - \tilde{s}\gamma\right)\right) \tag{50}$$

The least-squares estimate of the γ coefficient is given by:

$$\hat{\gamma} = \left(\tilde{s}^\top \tilde{s}\right)^{-1} \tilde{s}^\top s \tag{51}$$

Conjugate priors for σ^2 and γ take the form $p\left(\gamma, \sigma^2\right) = p\left(\sigma^2\right) p\left(\gamma|\sigma^2\right)$, where $p\left(\sigma^2\right)$ is an inverse-gamma distribution with parameters a_0 and b_0, Inv-Gamma(a_0, b_0), and the conditional prior distribution for the regression coefficient is a normal distribution with a mean μ_0, and variance $\sigma^2 \lambda_0^{-1}$, where λ is a scalar prior precision for the regression coefficient. Updates rules based on this take the form:

$$\mu_T = \left(\tilde{s}^\top \tilde{s} + \lambda_0\right)^{-1} \left(\lambda_0 \mu_0 + \tilde{s}^\top \tilde{s} \hat{\gamma}\right), \quad \lambda_T = \tilde{s}^\top \tilde{s} + \lambda_0 \tag{52}$$

$$a_T = a_0 + T/2, \quad b_T = b_0 + \frac{1}{2}\left(s^\top s + \mu_0^2 \lambda_0 - \mu_T^2 \lambda_T\right) \tag{53}$$

For this, the model, m, evidence is given by:

$$p\left(s|m\right) = \int p\left(s|\tilde{s}, \gamma, \sigma^2\right) p\left(\gamma, \sigma^2\right) d\gamma d\sigma^2 = \frac{1}{(2\pi)^{T/2}} \sqrt{\frac{\lambda_0}{\lambda_T}} \frac{b_0^{a_0}}{b_T^{a_T}} \frac{\Gamma\left(a_T\right)}{\Gamma\left(a_0\right)}, \tag{54}$$

and the log-likelihood, $\log p\left(s|m\right)$, becomes:

$$-\frac{T}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\lambda_0\right) - \frac{1}{2}\log\left(\lambda_T\right) + a_0 \log\left(b_0\right) - a_T \log\left(b_T\right) + \log\left(\Gamma\left(a_T\right)\right) - \log\left(\Gamma\left(a_0\right)\right) \tag{55}$$

Using this log-likelihood as the objective for the generative network (the decoder), we will refer to as using the Bayesian linear regression (BLR) likelihood. Up to a constant (depending on prior parameters and for a fixed length T), the negative of the above is equal to:

$$\mathcal{L}_{\text{BLR}} = \frac{1}{2}\log\left(\lambda_T\right) + a_T \log\left(b_T\right) \tag{56}$$

$$= \frac{1}{2}\log\left(\tilde{s}^\top \tilde{s} + \lambda_0\right) + \left(a_0 + T/2\right) \log\left(b_0 + \frac{1}{2}\left(s^\top s + \mu_0^2 \lambda_0 - \mu_T^2 \lambda_T\right)\right) \tag{57}$$

[6]The results utilized here are presented in e.g. Sec. 6 of Murphy [87]. Here we will follow the notation in `en.wikipedia.org/wiki/Bayesian_linear_regression` (as of January 2022) for convenience, but to avoid confusion with β-VAEs and the SI-SDR α we will denote the regression coefficient γ. The design matrix we consider (analogous to $\mathbf{X} \in \mathbb{R}^{n \times k}$) is $\tilde{s}$ with $k = 1$ predictor variables. Note that for the scalars considered, we remove various transpositions, and e.g. replace $\det\left(\mathbf{\Lambda}_0\right)$ with simply the scalar value λ_0.

H.1 Relationship between the BLR objective and SI-SDR

If we make some simplifications by assuming e.g. a weak prior, we can relate the BLR to SI-SDR. For $\lambda_0 \ll \tilde{s}^\top \tilde{s}$ and $\lambda_0 \ll s^\top s$, we have that Eq. 57 is approximately equal to (note that the λ_0 plays much the same role as a constant added to ensure numerical stability of the log operation as e.g. used in Asteroid):

$$\mathcal{L}_{\text{BLR}} \approx \frac{1}{2} \log\left(\tilde{s}^\top \tilde{s}\right) + (a_0 + T/2) \log\left(b_0 + \frac{1}{2}\left(s^\top s - \mu_T^2 \lambda_T\right)\right) \tag{58}$$

Under the same assumption on λ_0, we have that μ_T and λ_T are:

$$\mu_T = \left(\tilde{s}^\top \tilde{s} + \lambda_0\right)^{-1} \left(\lambda_0 \mu_0 + \tilde{s}^\top \tilde{s} \hat{\beta}\right) = \left(\tilde{s}^\top \tilde{s} + \lambda_0\right)^{-1} \left(\lambda_0 \mu_0 + \tilde{s}^\top \tilde{s} \left[\left(\tilde{s}^\top \tilde{s}\right)^{-1} \tilde{s}^\top s\right]\right) \tag{59}$$

$$= \left(\tilde{s}^\top \tilde{s} + \lambda_0\right)^{-1} \left(\lambda_0 \mu_0 + \tilde{s}^\top s\right) \approx \left(\tilde{s}^\top \tilde{s}\right)^{-1} \left(\tilde{s}^\top s\right) \tag{60}$$

$$\lambda_T = \tilde{s}^\top \tilde{s} + \lambda_0 \approx \tilde{s}^\top \tilde{s} \tag{61}$$

$$\mu_T^2 \lambda_T \approx \left(\frac{\tilde{s}^\top s}{\tilde{s}^\top \tilde{s}}\right)^2 \tilde{s}^\top \tilde{s} = \frac{\left(\tilde{s}^\top s\right)^2}{\tilde{s}^\top \tilde{s}} \tag{62}$$

And so, inserting this $\mu_T^2 \lambda_T$ into the expression for the loss in Eq. 58, we have:

$$\mathcal{L}_{\text{BLR}} \approx \frac{1}{2} \log\left(\tilde{s}^\top \tilde{s}\right) + (a_0 + T/2) \log\left(b_0 + \frac{1}{2}\left(s^\top s - \frac{\left(\tilde{s}^\top s\right)^2}{\tilde{s}^\top \tilde{s}}\right)\right) \tag{63}$$

For small a_0 and b_0, we have:

$$\mathcal{L}_{\text{BLR}} \approx \frac{1}{2} \log\left(\tilde{s}^\top \tilde{s}\right) + \frac{T}{2} \log\left(\frac{1}{2}\left(s^\top s - \frac{\left(\tilde{s}^\top s\right)^2}{\tilde{s}^\top \tilde{s}}\right)\right) \tag{64}$$

Up to a constant arising from the $\frac{1}{2}$ factor within the $\log$ in the second term, this is equal to:

$$\frac{1}{2} \log\left(\tilde{s}^\top \tilde{s}\right) + \frac{T}{2} \log\left(s^\top s - \frac{\left(\tilde{s}^\top s\right)^2}{\tilde{s}^\top \tilde{s}}\right) \tag{65}$$

$$= \frac{1}{2}\left(\log\left(\tilde{s}^\top \tilde{s}\right) + T \log\left(s^\top s - \frac{\left(\tilde{s}^\top s\right)^2}{\tilde{s}^\top \tilde{s}}\right)\right) \tag{66}$$

When T is large (e.g. as during training $T = 3\,\text{s} \cdot 8\,\text{kHz} = 24000$), the second term dominates, which, in isolation, looks like:

$$\log\left(s^\top s - \frac{\left(\tilde{s}^\top s\right)^2}{\tilde{s}^\top \tilde{s}}\right) = \log\left(\langle s, s\rangle\right) + \log\left(1 - \frac{\langle \tilde{s}, s\rangle^2}{\langle \tilde{s}, \tilde{s}\rangle \langle s, s\rangle}\right) \tag{67}$$

W.r.t. the model parameters the first term is constant, ignoring this we have:

$$\log\left(1 - \frac{\langle \tilde{s}, s\rangle^2}{\langle \tilde{s}, \tilde{s}\rangle \langle s, s\rangle}\right) = \log\left(1 - \langle \hat{e}_{\tilde{s}}, \hat{e}_s\rangle^2\right) \tag{68}$$

We had that the NSI-SDR was:

$$\text{NSISDR}(s, \tilde{s}) = 10 \log_{10}\left(\frac{\langle s, s\rangle \langle \tilde{s}, \tilde{s}\rangle}{\langle \tilde{s}, s\rangle^2} - 1\right) = 10 \log_{10}\left(\langle \hat{e}_{\tilde{s}}, \hat{e}_s\rangle^{-2} - 1\right), \tag{69}$$

and we see that the objectives are based on measuring the square of the inner product between the directions of the target and the estimate, albeit in slightly different manners.

This difference stems partly from the difference in the view of re-scaling the target to fit the estimate, or the other way around (compare α with $\hat{\gamma}$). We note, especially, that the BLR objective varies with T in how the power of the estimated signal is taken into account. This dependency on T stems from the update rule for the a_t for the inverse gamma distribution, and thus, in part, from an i.i.d. assumption on ϵ_t, and it is worthwhile considering a model that addresses this differently.

I Multivariate Cauchy objective

Besides using the BLR objective, we also consider a model which can model a per time-step scale (similar to the per time-step Gaussian with a modelled scale/variance), but which measures the error in a log-manner. We consider a Cauchy distribution, since the log-likelihood conceptually enables us to minimize a log-error similar to the log-MSE [21]. A T dimensional Student-t with one degree of freedom ($\nu = 1$) corresponds to a multivariate (T-dimensional) Cauchy (MVC) distribution, with a density function:

$$\text{Student-}t_{\nu=1}(\tilde{s}; \mu, \Sigma) = \text{MVC}(\tilde{s}; \mu, \Sigma, T) = \frac{\Gamma\left(\frac{1+T}{2}\right)}{\Gamma\left(\frac{1}{2}\right) \pi^{\frac{T}{2}} |\Sigma|^{\frac{1}{2}} \left[1 + (\tilde{s} - \mu)^\top \Sigma^{-1} (\tilde{s} - \mu)\right]^{\frac{1+T}{2}}}, \tag{70}$$

where we model $\tilde{s}$ using a mean time series, μ, and scale matrix, Σ. The MVC model we consider will only parametrize a diagonal scale matrix. For an identity scale matrix, the log-likelihood of the MVC is proportional to the log of one plus the squared difference between $\tilde{s}$ and μ (i.e. similar to a log-MSE objective).

We present and discuss some early results for using this likelihood in Appendix L.

J Data, model and training specifics

General specifics on data foundation We use the open-source LibriMix, as introduced in Cosentino et al. [7], which combines speech from LibriSpeech [28] (CC BY 4.0) and noise samples from WHAM! [23] (CC BY-NC 4.0). Cosentino et al. [7] also introduces a dataset with VCTK speech [22] (CC BY 4.0) mixed with WHAM! noise, which we also use.

Impacts of improved generative models With increasingly powerful generative models on audio, as we discuss in Section 5, the problems of misuse of models, and misuse of available data to clone a voice without consent should be taken into consideration. While the presented VI-TasNet does not enable e.g. voice conversion, it is a generative model. The use of the VI-TasNet is solely focused on recreating, as closely as possible, the original speech of the single sources in the mixture. A key aspect is the level of temporal abstraction on which the generative model works; higher-capacity models, such as the one considered in e.g. van den Oord et al. [15], operate on a considerably higher temporal abstraction in the latent variables than a VI-TasNet. Simplistically, a VI-TasNet learns something akin to an efficient version of the average speech spectrum with the addition of some knowledge of phase, whereas models with higher temporal abstraction will learn more high-level components of speech, like phonemes, words, sentences and speaker identity, allowing them to also reproduce or coherently alter such aspects of audio. We further refer to Dieleman et al. [86, Sec. 6.1] for a discussion on such considerations concerning e.g. imitating voice identities in the datasets.

Consent and identity It should be noted that only the VCTK dataset was explicitly made with an aim related to voice synthesis. Yamagishi et al. [22] presents "the Voice Bank corpus, specifically designed for the creation of personalised synthetic voices for individuals with speech disorders". Yamagishi et al. [22] outlines that participants were given "a consent form detailing the conditions of use of their recordings", but the details of this consent—to the best of our knowledge—are not given in the paper, nor at the dataset current web page[7]. The LibriSpeech data is a curation of speech from the LibriVox project, which collects free public domain audiobooks. Volunteer audiobook recorders are instructed that audio enters the public domain, and examples are currently provided to inform the volunteers of various potential consequences[8]. The VCTK speakers are given an anonymous numeric ID, which is available alongside their age, gender, accents, and region of England that they came from. The LibriVox speakers are identified by the name under which the reader is registered in LibriVox alongside the sex specified. The LibriSpeech dataset and the VCTK dataset, however, both contain audio recordings of speech, which can be used to identify a person. For the WHAM! noise dataset, the data web page[9] states that the noise datasets "have been processed to remove any segments containing intelligible speech".

Written material foundation The content for the VCTK dataset was curated from relatively recent newspaper articles ("3000 articles of the Scottish Herald newspapers", and additionally "The Rainbow passage", and "Accent elicitation passage from the Speech Accent Archive") [22]; while the creation was based on coverage optimisation, we expect a very limited amount of explicitly offensive material, although the study does not mention filtering on a such a parameter. LibriMix, which uses LibriSpeech, is built using LibriVox. For LibriVox, the written content is old books in the public domain. The content of the LibriVox books, being older books, contain sentiments prevalent at the time of writing of the original works; this is an especially important consideration if learning e.g. a generative language model based on the data, but the VI-TasNet under consideration does not enable that higher level of representation learning of underlying language. It is important to note the limitations of training on a dataset of predominantly older material mainly from (non-conversational) English book reciting. The representation learnt of such audio does not reflect well the diversity of spoken English, and it is very unlikely that a model trained on English performs well on different languages altogether, and poorer representation fit for "non-standard" English and non-English causes reduced separation performance. Testing on VCTK enables testing of the algorithm on a dataset explicitly attempting to include diversity in British dialects, and further efforts in this direction could include evaluating on larger datasets with e.g. more nationalities (such as the VoxCeleb [88]).

[7] `datashare.ed.ac.uk/handle/10283/2950` (accessed June 2021)

[8] cf. `librivox.org/pages/volunteer-for-librivox` (accessed June 2021) under the heading "What Can Other People Do with LibriVox Recordings".

[9] `wham.whisper.ai` (accessed June 2021)

Realistic mixtures As discussed in Cosentino et al. [7], speech separation algorithms like TasNet perform poorly on more sparsely overlapping data. The mixtures in LibriMix are both densely overlapping and are unrelated sentences from various audiobooks. Towards a better understanding of the speech separation performance, evaluation on more realistic, conversation-like dataset would be valuable; the sparsely overlapping version of LibriMix, SparseLibriMix test set [7], represents one such dataset. This present study focus on simple mixtures ("anechoically mixed"). Future investigations should address how e.g. reverberant environments affect the learning (using e.g. WHAMR! as considered for LibriMix in [53], using simulated room impulse responses, or even actual, reverberant/realistic mixtures).

Various training specifics We aligned with the Asteroid training recipes for the ConvTasNet[10]. We used a smaller batch size of 4 (limited by the memory of available hardware) for the LibriMix experiments in the clean condition (that is, for results in Table 2 and Table 3). Another deviation from the Asteroid recipes is a reduced learning rate to $3 \cdot 10^{-4}$. Training the variational network parameters with higher learning rates for that particular batch size resulted in training instabilities. For the results on the noisy condition (in Table 1), we used four GPUs in parallel with the same batch size of 4 with an effective batch size of 16. We could alleviate this instability for higher learning rates by increasing the batch size, but this, however, reduces the experimental throughput significantly. For the noisy condition VI-TasNet model, the total rate per time step (i.e. $\left(R_z + R_{\mathcal{M}}\right)/T$) was adaptively optimized towards a value of 256 nats. This is a heuristically chosen hyper-parameter, and it is worth tuning and exploring, e.g. by resolving the RD-curves for the problem. Here, it was chosen through initial exploratory runs in trying to balance on the one hand being too restrictive while on the other hand ensuring that the value is imposing the needed regularization. Too low of a target rate would result in never reaching a performance comparable to the TasNet. Similarly, too high target rates would result in distributions collapsing onto deterministic distribution, using the freedom to remove all variance

For the SuDoRMRF results, we use only the separation modules of the SuDoRMRF [43] implementation in Asteroid [89] (or more specifically, their SuDORMRFImproved). We compare a deterministic version and a variational inference version (SuDoRMRF and VI-SuDoRMRF). The only changes for these models (compared to the TasNet and VI-TasNet models) were the networks called in the masker module (the separation module) that parametrizes the distribution over the masks (or directly outputs the masks in the deterministic setting). The SuDoRMRF was started with the package's standard parameters matching the improved version configuration, and otherwise, the same setup was used as the VI-TasNet. To not include too many new factors, we did not, for instance, use the sum-to-one masking activation used in [43] (we have also discussed the implications of this in the original Appx. L).

We also did a similar experiment with the Sepformer [5] using the SpeechBrain [90] implementation of the Sepformer for the separation models. With the size of the Sepformer, we needed to reduce the batch size. We used the reduced learning rate and magnitude of gradient clipping reported in the original paper and used the same configuration as in the original paper as available on the SpeechBrain repository for the masking network, and otherwise the same parameters for all other VI-EMD-related parameters (i.e., the same as the TasNet/VI-TasNet configurations). The VI-Sepformer, notably, used the same target total rate of 256 nats, which caused the VI-Sepformer to more quickly be strictly more regularized than the VI-versions of TasNet or SuDoRMRF. Seeing the Sepformer is a considerably more expressive model (on parameter count it is more than 5–10x larger than the TasNet and SuDoRMRF) and operating on a different mechanism (convolutional versus attention), it is unsurprising that the model has significantly different rate-distortion trade-offs than the TasNet and SuDoRMRF. The VI-Sepformer with the same target total rate was heavily over-regularized (i.e., closer to the first rows in Table 2). We concluded that it would be beyond the scope of this paper to also contrast these additional trade-offs, especially considering the increasing training times of the larger models, but we consider it a promising future investigation to illuminate how generalization and RD-trade-offs are related to model architectures.

Learning rate annealing We trained with increased "patience" in the learning rate scheduler (50 epochs/passes over the full training dataset before reducing the learning rate at validation SI-SDR plateaus). This is, we believe, the cause of improved results on the deterministic TasNet reported

[10] `github.com/asteroid-team/asteroid`, MIT License.

compared to the Asteroid repository reported results[11] (13.0 dB SI-SDRi versus the performance shown in Table 3 of 14.4 dB using Libri2Mix train-100 on the clean separation task). Similarly, Asteroid reports an SI-SDR improvement of 10.8 dB on the same noisy LibriMix (2 speakers, 8 Khz, min-mode, 100 hr), whereas the increased patience improved the deterministic baseline to a performance of 11.6 dB) in the noisy condition as reported in the main paper Table 1 (and the corresponding expanded table here in the supplementary, Table 4). While we otherwise use the same architecture as provided in Asteroid (time-dilated convolutions architecture) masker, we saw some stability improvements by normalizing the summed skip connections by the depth of the masking network.

Computational resources A single experiment with the VI-TasNet doing 200 epochs over the 100 hour Libri2Mix training set with a batch size of 4 on a single GPU took approximately 5-6 days on an NVIDIA GeForce GTX 1080 Ti (or approximately 120-144 GPU hours). The adaptive model and flow models took slightly longer with extra encoding/decoding of single sources and mixture for the adaptive model and more model parameters in the learnt flow. The seven models in Table 3 alongside the eight models in Figure 7 (numerical results in Table 2) thus required approximately 2000 GPU hours total, disregarding an equivalent, at least, amount of GPU hours in developing the model before running the experiments. As we stress in Section 5, the reported results are all performances of one run of the model (one random seed). Accordingly, we have no basis for comparing the model performances rigorously, e.g. evaluating whether the flow prior is statistically performing better than the other variants of the VI-TasNet. While such a characterization is valuable information, we balanced the available time and compute against the value of resolving various priors. For the noisy condition results (in Table 1), the TasNet and VI-TasNet were trained on four GPUs for 1000 epochs (280 and 300 hours, respectively), or about 1.2k GPU hr per model. The TasNet, even with the increased patience, converged faster and could likely be stopped as early as 75–100 hours, whereas the VI-TasNet could likely have been stopped at about 200 hours. The best validations loss checkpoints (used in testing) were in both models, however, from the last 50 epochs of the 1000 epochs. We stress that the results are, under the compute available, single repetitions of a model training, and the results are limited in not addressing the uncertainty in final model performance given the stochastic initialization and optimization. Given the significant compute involved in training a single of these models, we chose to focus on the more challenging noisy condition for larger models (presented in Table 1), as we expected differences between deterministic and VI-based models to be bigger in this condition. Some experiments were done in the clean condition (presented in e.g. Table 2, Table 3, Figure 2, and Figure 3), and we chose to retain these findings and present them as initially done instead of repeating the experiments in the noisy condition.

Parameter counts and model size The VI-TasNet has a parameter count (and prediction time complexity) similar to the base TasNet counterpart we consider which has 5.1 million parameters. When using the gamma encoding, no extra parameters are added to the encoder. In the results for Table 3, using a Gaussian encoding or the MVC objective adds an extra dimension to the encoder or decoder output, respectively, which adds 8192 parameters to model the variances (i.e. a very negligible about 1–2 ‰ relative increase in parameters). The extra parameters in the masking network to enable two parameters per time and latent dimension increases the total model parameter count to 5.2 million trainable parameters, for the Gaussian encoding. The presented flow prior model, due to the large size of the latent space and relatively large size of chosen MADE configurations, has a (potentially needlessly large) total of 17.8 million trainable parameters. We chose to focus on a model closely aligned with a (deterministic, Conv-)TasNet to more readily compare to a well-studied model. Notably, the standard TasNet has a single-layer encoder and decoder—and can benefit from deeper structures [6]. We saw in initial exploratory investigations that a VI-TasNet will similarly benefit from deeper encoders and decoders, possibly to a greater extent than a TasNet depending on how restrictive the utilized prior is. For a Gaussian prior and posterior, we are essentially forcing a linear mapping from windows of 16 samples of raw audio to closely resemble a Gaussian. Even slightly deeper, non-linear mappings might be beneficial.

Approximate KL-divergences In evaluating the rate terms, we have analytical expressions for the KL-divergence between two Gaussians, two betas or two gammas, but this is not the case

[11] Asteroid results available at `https://github.com/asteroid-team/asteroid/tree/master/egs/librimix/ConvTasNet` (accessed May 2022).

for e.g. the flow prior. In training the models, we initially used an approximation of the KL-divergence, as is common practice in training VAEs, and this was used in the clean condition results (Table 2 and Table 3). With the KL-divergence as defined in Eq. 6, a single sample Monto Carlo estimate corresponds to estimating the divergence using a single sample drawn from the approximate posteriors of the latent variables to determine the expectation of the log density ratios. While these approximations are generally close to the analytical expressions, we saw that the model learned better using the approximation than (when available) when the analytical expression. Investigating this, we saw that in some cases the estimated KL saturates when the parameters of the variational distribution are much lower than the prior value (e.g. trying to have very little activation in a gamma latent dimension), whereas the analytical expression does not. While an unintended consequence of using the estimated KL-divergence, this behaviour enabled the model to perform better and could indicate that a prior more flexible in allowing to "turn off" latents is useful for the VI-TasNet specification considered.

In the synthetic and noisy condition experiments, we used the analytical expressions, facilitated by the use of using adaptive re-weighting, instead of the free bits and static re-weighting used in the clean condition results. While the noisy condition results use an adaptive re-weighting, the training also included a (potentially inconsequential) free encoding nat (1.0 nat) across all encoding dimensions as well as one free mask nat across all mask dimensions and speaker masks.

Numerical stability: dequantization, initialization, clamping, and margin loss Since the audio signal is a 16-bit audio discrete-time signal, we used a (potentially inconsequential) uniform dequantization. Traditionally, float representations of audio are scaled to be in the $[-1, 1]$ range. Using a standard PyTorch initialization for the encoder and decoder resulted in initial estimates that were much outside this range, so we used a uniform initialization on $[-10^{-2}; 10^{-2}]$ for the encoder and decoder weights. While values outside the $[-1, 1]$ range is not a problem for the scale-invariance objective (it simply re-scales the signals), it is inconvenient to have the model operate in this range needlessly. We employ a margin loss entirely similar to the one used in Donahue et al. [91] and Dieleman et al. [86]. We saw little to no effect from dequantization, and its use was dropped for the noisy condition and synthetic experiments (these experiments are later than the clean condition results). The synthetic experiments also do not use a margin loss. The parameterizations of the various distributions rely on transforming the direct output of the networks in some manner; for the gamma concentration parameter and Gaussian scale, we e.g. pass the network output through a softplus function, and clamp it to a minimum value of 10^{-6} to have the required strictly positive parameter. The adaptive KL re-weighting was in the noisy condition started at a factor of 10^{-9} and adapted by a factor of 10^{-4} at each step (otherwise using the same formulation and e.g. minimum threshold for change as Dieleman et al. [86]). For the synthetic experiments, a higher initial value and adaption rate of 10^{-6} and 10^{-2}, respectively, were used, to accommodate the faster training of simpler models. The data was standardized based on the mean and variance of the waveforms across time and across all mixtures in the dataset, and the same standardization was used in testing on both LibriMix and VCTK (i.e. LibriMix train set values of mean and variance were used for standardization).

Autoregressive flow prior The learnable autoregressive flow (AF) prior introduced with the variational lossy autoencoder (VLAE, in Chen et al. [30]) is equivalent to the inverse-autoregressive flow (IAF) introduced in [29]. We investigate how this type of more flexible AF prior can be used in the VI-TasNet, by learning a mapping, or flow, from a base distribution ("noise source"), $u(\epsilon)$, to the latent encodings, $p_\xi(z)$. We use a Gaussian base distribution for ϵ, and learn a series of flows. These flows together make up a mapping $z = \omega_\xi(\epsilon)$, and they use a series of invertible mappings parametrized by autoregressive networks with parameters ξ. While one such possible mapping is a series of affine transformations, where a scaling and a translation are learnt, we adopt the approach from the VLAE to use a mean-only flow. Similarly, we also make use of a series of masked autoencoder density estimations (MADE) networks [92] as the autoregressive networks parameterizing the flows[12].

Sampling frequency We chose to work with the 8 kHz version of the datasets to reduce the computational requirements. Investigations on whether the findings presented hold for increased

[12] In particular, we make use of the implementation available at `github.com/karpathy/pytorch-made`, MIT license.

sampling frequencies would be important for real-world use since many use cases require a higher audio quality than achievable with 8 kHz.

Dataset size Similar to the choice of the 8 kHz variant of the data, we worked with the 100 hr (smaller) version of LibriMix to reduce computational requirements. Using the larger versions of the datasets, or more datasets, tends to increase performance; for instance, Asteroid reports 10.8 dB SI-SDRi in the noisy, 2-speaker LibriMix condition when trained on 100 hrs, but a performance of 12.0 dB SI-SDRi when trained on the larger 360 hrs variant of LibriMix. Within the dataset (i.e. on LibriMix), this means that the TasNet trained on 360 hrs of data matches the performance of a VI-TasNet trained on only 100 hrs of data (this, of course, does not say anything about the generalization to new domains or conditions of VI-TasNet versus TasNets on larger datasets). For scenarios where examples are scarce, we would propose the investigation of VI-TasNets—and especially the multitasking version for learning from more abundant audio without available single sources. Characterizing the performance as a function of dataset sizes/number of examples with single sources (exploring learning curves) was not the focus of the present study; in such scenarios, we stress that the baseline would not solely be a deterministic TasNet, but rather a model trained with methods such as MixIt and methods with similar aims [53].

K Synthetic experiment

K.1 Gaussian pulse mixtures

We construct a simple source separation problem which makes mixtures of single sources that themselves are overlapping sinusoidal Gaussian pulses from a specific frequency region with overtones. We specify a frequency range for the "fundamental frequency of given speaker". In the experiments shown, this was set to 300-400 Hz for "Speaker/Target 0" and 100-200 Hz for "Speaker/Target 1". For an example set of targets and input, see Figure 9. To create one of the synthetic single sources, we create three sinusoidal pulses with a Gaussian envelope and add them to create one single source in the mixture. For each pulse, we sample a fundamental frequency in the given "speaker's" range. We also sample an overall amplitude in a given range (this range is shared between targets), a phase, and pulse delay within a specified time axis of 3 seconds so at least half of the pulse is within the segment. Additionally, we add four overtones to the sampled fundamental sine wave and add Gaussian noise.

We create distinct examples/datasets by keying the randomness/sampling procedures to integer base seed/ranges; for the replicates in the synthetic experiments, the first replicate had indices ranging from $[100000, 102048[$ for 2048 training examples, the next 128 indices were validation examples, and after that came 1024 indices for each test data configuration (e.g. different noise levels). Similarly, the next replicate had training examples from $[200000, 202048[$, and so on.

While the frequency region of the fundamental frequencies is not overlapping, the inclusion of the overtones results in a problem where regions of the spectrogram will share energy between the two targets.

K.2 Model and optimization

We reduce the number of encoding dimensions with respect to the models we consider on LibriMix from 512 to 32. With reference to the naming from the temporal convolutional network used in Conv-TasNet [2] and as available in the Asteroid framework [89], we similarly reduce the masker bottleneck channels down from 128 to 16, the skip channels from 128 to 8, the hidden channels from 512 to 16. We retain the same number of blocks per repeat but reduce the number of repeats/cycles to 1. We parametrize Gaussian encodings and lognormal masks and we use standard priors for both (mean/location and standard deviation/scale of 0.0 and 1.0, respectively). We use a BLR likelihood, and we re-weigh the rate terms to match varying levels of rates to resolve the RD-curve using the adaptive re-weighting described in Appendix E. We use an initial value for the adaptive factor for the total rate (sum of both rate and encodings) of 10^{-6} and adapt with 1 % at each step if needed ($\delta = 0.01$).

We use a learning rate of $3 \cdot 10^{-4}$ (with the same higher patience scheduling as described for larger models) and optimize the model using a permutation invariance evidence lower bound loss (i.e. while using PIT, we minimize the negative ELBO). The models are trained for a maximum of 800 epochs (parses over the 2048 data examples), with potential early stopping if no improvement in validation SI-SDR is seen for 60 consecutive epochs. We do not use the model with the highest validation SI-SDR for the evaluation but instead use the last model to make the RD curves, since the adaptation can produce early models that had high rates with better performance than later, more tightly regularized versions. Monitoring the actual loss (modified ELBO) instead of the SI-SDR is non-trivial, because the adaptive reweighing continuously changes the values, e.g. potentially increasing the loss in periods where a rate is being re-weighted towards lower rate values without it necessarily indicating a plateau and a needed "early stop".

We expected to find the models could over-fit, which would be evident as a (significant, especially for higher rates) gap between training and validation/test performances. However, with the specifications detailed above, the models did not display significant over-fitting to the training data, even for the highest rates. This, we hypothesize, is a consequence of simultaneously (i) having reduced the complexity of the model (fewer latents, fewer filters, etc.) and (ii) having employed regularizing elements (such as early stopping and learning rate annealing). Provided that e.g. a more over-complete model was trained without learning rate annealing, we hypothesize that the generalization behaviour and rate-generalization trade-offs would be even more pronounced.

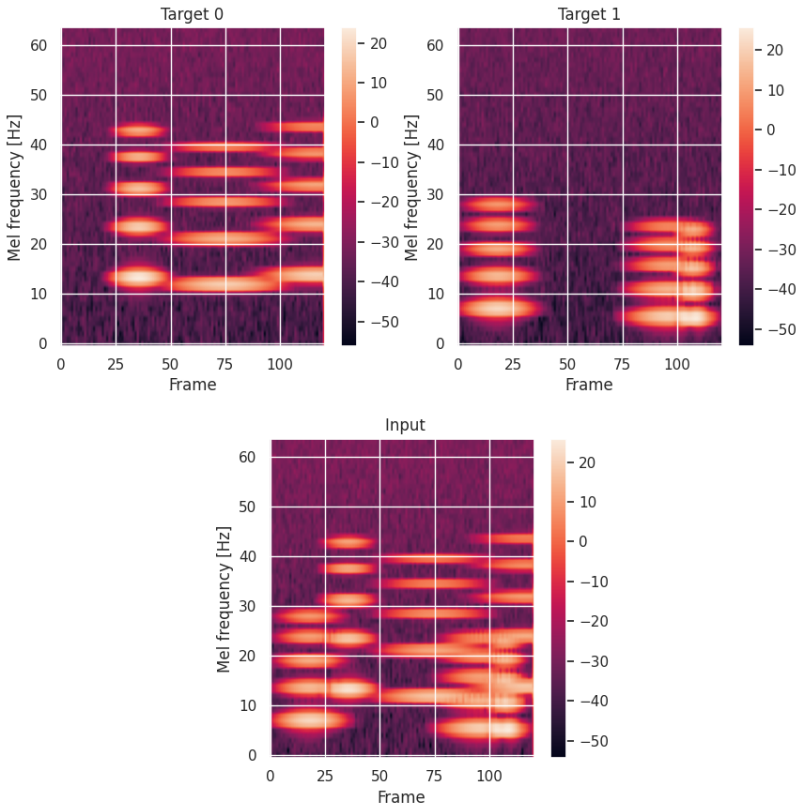


Figure 9: An example of the synthetic Gaussian pulses dataset. Top row: spectrogram of single sources targets in isolation. Bottom row: input mixture to the model.

L Different priors on clean LibriMix

In this section, we present earlier results for TasNets and VI-TasNets trained on the *clean* version of LibriMix (min-mode, 2 speakers, 8 kHz, clean, 100 hrs). As the deterministic baseline, we trained a TasNet using both the standard SI-SDR objective and the BLR. We compare these TasNets to four variations of the VI-TasNet to investigate the effect of encoding posterior and priors, all using the BLR objective, and all using a beta masker. Firstly, we train a VI-TasNet with a Gaussian encoder distribution and prior, and a similar gamma version. In addition to these, we train a model that uses a Gaussian approximate posterior with a flow prior. We train a VI-TasNet that uses a gamma posterior in conjunction with adaptive prior, which also uses the multitasking objective. Finally, we train a VI-TasNet with a more flexible decoder distribution, the MVC, using a gamma encoder and prior. The performance of these models is shown in Table 3.

Training a deterministic TasNet with the BLR objective (unsurprisingly) reduced the SI-SDR performance compared to directly optimizing SI-SDR, but the BLR does produce reasonable SI-SDRi scores while additionally providing actual probabilities/normalized densities. The VI-TasNets learn to perform the separation task well, albeit not—in this earlier version, without e.g. adaptive reweighting—to the same performance as the deterministic counterparts. These models are different from the models outperforming the TasNets in Table 1 in the amount of (target total) rate they achieve and the masker distribution used. Here, the best performing VI-TasNet uses the flow-based prior, but both the simple Gauss and gamma models attain performances near the 13 dB SI-SDRi mark.

The multitasking, adaptive model and the MVC model perform the poorest of the VI-TasNet. These models were not converged within a 200 epoch limit (about 150 GPU hours of training), and they would likely see improved performances with longer training. While the TasNets here display the highest difference between the LibriMix and VCTK test sets (generalization gap), this result does not support this being attributed to differences in variational versus deterministic models, seeing as the TasNets also display higher overall SI-SDRi. For the models in Table 3, the differences in model performance are smaller in the VCTK-2mix test, where e.g. the difference between the BLR TasNet and the flow-based VI-TasNet is 0.38 dB, as opposed to their difference of 0.73 on LibriMix. We note, in contrast to the findings in this section, that other configurations of the VI-TasNet models (such as the ones presented in the main paper in Table 1) produce both better overall performance of the VI-TasNet and notably also better generalization to VCTK.

Table 3: Model SI-SDR improvements in dB for LibriMix and VCTK test sets in noise-free/clean condition. $q_\varphi^{\mathcal{N}}/q_\varphi^{\Gamma}$: Gaussian/gamma approximate posterior, respectively; similar notation for Gaussian and gamma prior; p_ξ: flow prior; $p_{\varphi,\mathcal{S}}$: adaptive prior. SI-SDR is parenthesized to denote it as an objective rather than a likelihood.

Model	Encoding q/prior	Likelihood	Libri2Mix test	VCTK-2mix test	Difference
TasNet	-	(SI-SDR)	14.36	12.91	1.45
TasNet	-	BLR	13.86	12.13	1.73
VI-TasNet	$q_\varphi^{\mathcal{N}}/p^{\mathcal{N}}$	BLR	12.75	11.52	1.23
VI-TasNet	$q_\varphi^{\Gamma}/p^{\Gamma}$	BLR	12.68	11.40	1.28
VI-TasNet	$q_\varphi^{\mathcal{N}}/p_\xi$	BLR	13.13	11.74	1.38
VI-TasNet	$q_\varphi^{\Gamma}/p_{\varphi,\mathcal{S}}$	BLR	10.21	9.11	1.10
VI-TasNet	$q_\varphi^{\Gamma}/p^{\Gamma}$	MVC	11.31	10.19	1.11

In the main paper, we report results for the model with Gaussian encodings and log-normal masks (to align with the most standard TasNet formulation), but we here show the viability of considering other distributions. With these results, we show that the VI-TasNets support the incorporation of different types of structure in the latent encodings and different observation models (decoder distributions, likelihoods) and that the choice of these affects performance.

The gamma formulation can produce a non-negative encoding, which could potentially draw strengths from a parts-based representation, and similarly, we show that the BLR and MVC are viable avenues of exploration for imposing certain characteristics (like scale-invariance) in a manner compatible with variational inference.

The flow model is expensive during training, but we do not need to evaluate the prior during a call to the model if we are using it to do separation after training is completed. In this case, training time complexity might potentially be traded off for increased performance We would, however, need it if we wanted to do uncertainty estimation.

The variational inference formulation enables future work to incorporate various other well-known approaches from probabilistic modelling. Some examples are: we could address problems with an unknown number of speakers using a stick-breaking/Dirichlet process for the masks distribution; we could address the permutation problem by modelling target and estimated speakers with a mixture; or, we could provide a stronger learning signal to the masks by incorporating knowledge of the single source encodings as masking targets through an adaptive mask prior.

The adaptive model in Table 3 (second to last row) is, importantly, an example of a functioning multitasking model. We showed how such a model, and its input density estimates, can be useful from an uncertainty quantification perspective in the main paper with Figure 3. It is worthwhile stressing, however, that such a multitasking model can also learn directly from mixtures without reference single target sources in isolation.

The results shown here do not investigate the effect of the masker distribution choice, but we note that these results consider a beta masker (more similar to a sigmoidally gated TasNet mask), whereas e.g. Table 1 consider log-normal masks (more similar to a ReLU gated/rectified TasNet mask), showing how both are viable possibilities even if they have very different ways of masking.

M Further synthetic rate-distortion results

In this section, we provide further results on the RD-curve analysis of the synthetic problem and models considered in Figure 2 and Appendix K. We train VI-TasNets with the adaptive rate-regularization towards a range of rates on the 2-speaker synthetic problem. The target rates log-spaced from 10^{-3} to 10^{3}.

In Figure 10, we show how a model trained on a domain characterized by a particular level of noise amplitude, "NA" of 0.5—this corresponds to a standard deviation of Gaussian noise added on top of the single sources after additive mixing. The test performance on this seen/familiar-condition data is shared across all plots in black. A line is drawn as the running average distortion as a function of (sorted) rates, and a dot is drawn for each model. Alongside the black line in each plot, we show a corresponding red line which shows the performance of the model evaluated on different test conditions.

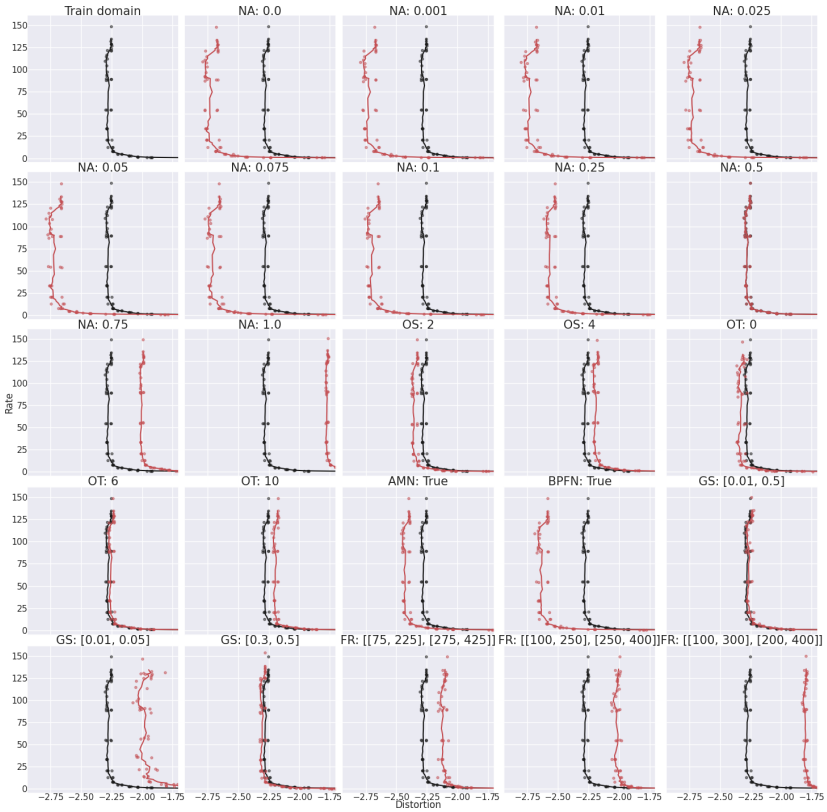


Figure 10: RD curves for various test conditions performance of models with varying target rates trained in a particular version of the synthetic Gaussian pulses separation test. Note that the x-axis and y-axis are shared across all plots. Details in the text. NA: noise amplitude, OS: (number of) overlapping sines, OT: (number of) over-tones, AMN: amplitude-modulated noise, BPFN: band-pass filtered noise, GS: Gauss-pulse scale, FR: frequency range.

Firstly, in going from top to bottom and from left to right, we start with the training domain and then ranging from models with no noise to increasing amounts of noise, all the way to a noise amplitude

of 1.0 (a factor 2 above the training domain of 0.5). We see that models with too high rates generalize less well to low noise settings, but the same is not immediately the case for higher noise settings. Following this, we see how changing the number of over-lapping sines ("OS")—by either adding or removing one from the training domain amount of 3 sines—produces either slightly better separation or slightly poorer separation, but no clear differences for high rates versus other models in their generalization abilities. A similar conclusion holds for the number of overtones ("OT"); removing them altogether makes it easier, whereas adding 1 (from five to six) or doubling (five to ten) makes the problem harder. We see that a slow (2 Hz) amplitude modulation of the noise signal ("AMN") or band-pass filtering of the noise ("BPFN", to have the noise only be in the region of the speaker Gaussian pulse frequencies) both produce slightly easier problems. We can control the width of the Gaussian pulses with a scale ("GS"). Evaluating with pulses that can be slightly longer and shorter ([0.01, 0.5] versus [0.05, 0.3] in the training domain), does not significantly change the performance in expectation. Testing on shorter pulses ([0.01, 0.05]), however, is a harder task, and only longer ([0.3, 0.5]) is an easier one. Lastly, we can change the frequency ranges ("FR") that define the speakers, to be either slightly expanded (from 100–200 Hz and 300–400 Hz to 75–225 Hz/275–425 Hz), nearly overlapping, or actually overlapping. In each case, this produces models with poorer distortion, but no clear difference in optimal versus higher rate models in the generalization abilities.

While the target total rates in some cases were as high as 1000, no model achieve rates over 150 nats. We hypothesize that the limited capacity coupled with e.g. learning rate annealing, early stopping and stochastic optimization might produce models that do not over-fit to the same extent and thus do not produce very high rates, even if the model has the freedom to do it. Generally, we have found that, when the models can achieve the target rate they do so with a higher consistency across replicates, whereas the models that cannot achieve the set rate tend to display a larger variance in the final expected rate over the test set and a similarly large variance in achieved distortion (some might do well, others do very poorly).

N All metrics LibriMix/VCTK evaluation

In Table 4 we provide extra numerical results for the Table 1 from the main paper. The full table shows (in addition to the already reported SI-SDRi), the BLR objectives and the SI-SDR. The table also highlights what is a familiar ("intra-dataset" or "intra-condition") evaluation versus an unfamiliar one ("inter-"). These extra metrics show how the VI-TasNet and VI-SuDoRMRF is an improvement both in SI-SDR, SI-SDRi and BLR over the deterministic counterpart on both familiar and unfamiliar datasets and conditions. The only exception is that the BLR better for the deterministic model in the LibriMix conditions.

Table 4: TasNet and VI-TasNet on noisy separation task. Full version of Table 1 with SI-SDR and BLR metrics, and indication of whether the performance is an inter- and intra- dataset or condition evaluation.

Condition	*Model*	*SI-SDRi*			*SI-SDR*			*BLR*	
		LibriMix/intra	*VCTK/inter*	*Drop*	*LibriMix/intra*	*VCTK/inter*	*Drop*	*LibriMix/intra*	*VCTK/inter*
Noisy/intra	TasNet	11.61	9.86	1.75 (0.15)	9.61	7.96	1.66 (0.17)	0.33	0.15
	VI-TasNet	**11.98**	**10.41**	**1.58 (0.13)**	**9.99**	**8.50**	**1.49 (0.15)**	**0.36**	**0.20**
Clean/inter	TasNet	12.96	10.38	2.58 (0.20)	12.96	10.37	2.59 (0.20)	0.67	0.39
	VI-TasNet	**13.56**	**11.42**	**2.14 (0.16)**	**13.56**	**11.42**	**2.14 (0.16)**	**0.73**	**0.49**
Noisy/intra	SuDoRMRF†	11.12	9.15	1.97 (0.18)	9.12	7.24	1.88 (0.21)	0.28	0.08
	VI-SuDoRMRF†	**11.46**	**9.53**	**1.93 (0.17)**	**9.47**	**7.62**	**1.84 (0.19)**	**0.30**	**0.11**
Clean/inter	SuDoRMRF†	12.46	9.75	2.71 (0.22)	12.46	9.75	2.71 (0.22)	0.62	0.33
	VI-SuDoRMRF†	**12.86**	**10.37**	**2.49 (0.19)**	**12.86**	**10.36**	**2.50 (0.19)**	**0.64**	**0.37**

APPENDIX C

CD-VAE: Hierarchical Variational Auto-Encoders using Latent Neural Stochastic Differential Equations

Hierarchical Variational Auto-Encoders using Latent Neural Stochastic Differential Equations

Rasmus M. Th. Høegh[*,1,2] *rmth@dtu.dk*

Aditi S. Krishnapriyan[3] *aditik1@berkeley.edu*

Liam Hodgkinson[4] *lhodgkinson@unimelb.edu.au*

Michael W. Mahoney[3,5,6] *mmahoney@berkeley.edu*

[*]*Corresponding author,* [1]*Technical University of Denmark,* [2]*WS Audiology,* [3]*University of California, Berkeley,* [4]*The University of Melbourne,* [5]*International Computer Science Institute,* [6]*Lawrence Berkeley National Laboratory*

Abstract

Variational auto-encoders are competitive likelihood models when constructed with a deep hierarchy of latent variables. We introduce a variational auto-encoder using a hierarchy of neural stochastic differential equations as latent objects. Building on the Very Deep Variational Auto-Encoder (Child, 2021) architecture, the introduced model replaces discrete top-down blocks with stochastic differential equation blocks. Using the stochastic process formulation and flexible numerical integration procedures, we present experiments on simple computer vision tasks and explore continuity in the learned model representation. The experiments show how the depth of the hierarchy can be continuously varied to produce different trade-offs between computational complexity and the model performance (number of numerical integration steps versus achieved evidence lower bound). Depth-continuity of this type allows a single trained model to be used across a range of computational restrictions in downstream uses of generative models. Finally, we discuss possible extensions of the proposed model as well as the potential implications on efficiency and generalization of using a depth-shared parametrization.

1 Introduction

Deep generative models seek to learn representations of data from large, readily available unlabelled data sets, often to use the learned representation in other tasks (Bond-Taylor et al., 2021). The representations can, for instance, be a starting point for training models where only a smaller data set of labeled data is available, or a generative task can also be used in a setting where a labeled and unlabelled data set is used concurrently during learning (Kingma et al., 2014). Generative models also enable tasks such as increasing the resolution of an input (super-resolution), image and audio synthesis, out-of-distribution detection, and compression (Bond-Taylor et al., 2021; Townsend et al., 2018).

Variational auto-encoders (VAEs) VAEs are a particular type of deep generative model (Rezende et al., 2014; Kingma & Welling, 2013), falling under a broader category of likelihood-based models, to which also belongs, e.g., flow-based models (Dinh et al., 2015; 2017), autoregressive models (Germain et al., 2015; van den Oord et al., 2016a;b) and diffusion models (Sohl-Dickstein et al., 2015; Rombach et al., 2022). VAEs are latent variable models that learn to model data distribution using variational inference. A VAE jointly optimizes an inference and a generative network to solve a bottle-necked auto-encoding task—i.e.,

reconstructing an input after parsing it through a restricted information channel. The inference network takes a datum and produces distributions over a latent space. These encodings are then decoded using the generative network, producing distributions over the original data space. The models are optimized towards ensuring (i) that the distortions introduced by the composition of the encoding and the decoding are as small as possible (that the input has a high likelihood under the output distribution) and (ii) that the encoding distributions produced by the inference network diverge as little as possible from a set prior. Many variations and improvements exist to the standard VAE framework, such as using more expressive priors (Kingma et al., 2016; Tomczak & Welling, 2018) and employing a hierarchical structure (Ranganath et al., 2016; Burda et al., 2016; Sønderby et al., 2016; Salimans, 2016; Maaløe et al., 2019; Vahdat & Kautz, 2020). In particular, increasing the stochastic depth greatly (number of levels in the hierarchy)—producing very deep variational auto-encoders (VD-VAEs) (Child, 2021)—makes for competitive likelihood models. Child (2021) showed how model performance increased by increasing the stochastic depth even with an otherwise fixed parameter count.

Dynamical systems in deep learning Incorporating differential equations in machine learning models provides expressive modeling components with the added benefit of being able to rely on a long history of research on dynamical systems, numerical integration, etc. The intersection of differential equations and deep learning modeling can, e.g., provide avenues of understanding and improving existing models through the lens of dynamical systems and make more expressive versions of existing models (Haber & Ruthotto, 2017; Chang et al., 2018; Raissi et al., 2019; Erichson et al., 2021; Hodgkinson et al., 2021). In particular, a dynamical systems view enables modeling with inductive biases towards continuity in modeling temporal data (Krishnapriyan et al., 2022) or as continuity of model feature representation (Xu et al., 2022; Queiruga et al., 2020; 2021).

Continuity Haber & Ruthotto (2017) explored the connection between residual neural networks (ResNets) (He et al., 2016) and an Euler discretization of an ordinary differential equation (ODE). Building on this, Queiruga et al. (2020) introduce continuous-in-depth networks, relying on the ODE perspective of ResNets to construct networks that incorporate approximate numerical solvers in deep supervised learning. They show how higher-order solvers induce representations that generalize to step sizes different from the training step size (the step size used in the numerical integration). This depth "continuity" allows the model to train with adaptive step sizes—for instance, coarsely during initial training and finer and finer as training progresses. It enables using the model at different levels of computational complexity after training, trading off accuracy for reducing computational requirements (using an increased number of step sizes to produce better performance or vice versa). Similarly, Krishnapriyan et al. (2022) consider neural physics models and explore continuity properties. They introduce a convergence test that can verify whether a learned model reflects the continuity of a modeled systems dynamics, and they show how higher-order solvers induce continuous representations.

Differential equations in generative modeling The incorporation of differential equations extends to deep generative modeling. For example, Chen et al. (2018) introduced the neural ODE and showed how to use a neural ODE to produce a continuous normalizing flow (CNF). This was later extended with FFJORD (Grathwohl et al., 2018) allowing for more efficient training and fewer restrictions in architecture. We consider hierarchical variational auto-encoders in a dynamical systems framework. In this context, commonalities and unifying frameworks for the seemingly distinct likelihood-based deep generative models like flows, VAEs, autoregressive, and diffusion models are worth noting. To name some: VAEs and flows have been unified in frameworks like SurVAE flows (Nielsen et al., 2020) and AEF (Silvestri et al., 2022)); variational diffusion models are an infinitely deep limit of a standard VAE (Kingma et al., 2021); Child (2021) discuss how very deep VAEs generalize autoregressive models; and score-based diffusion models can be transformed into continuous normalizing flows (Song et al., 2021). Regarding stochastic differential equations (SDEs) specifically, they have been used to extend CNFs producing stochastic continuous normalizing flows (Hodgkinson et al., 2021), and a specific SDE formulation is the basis for score-based diffusion models (Song et al., 2021). Tzen & Raginsky (2019), also using SDEs, consider the diffusion limit of deep latent variable models that use Gaussian approximate posteriors in a variational inference framework. Continuously, or "infinitely", deep models can be constructed using the dynamical systems perspective; Xu et al. (2022)

present continuous-depth Bayesian neural networks using latent SDE, and they also discuss the diffusion limit of discrete-time models.

Latent SDEs A latent SDE (Liu et al., 2019; Jia & Benson, 2019; Li et al., 2020) resembles a VAE in which the usual Gaussian latent object is replaced by a latent process defined as solutions to an SDE under the stochasticity of a Brownian motion. In this work, we will build, in particular, on the latent SDE formulation presented by Li et al. (2020). General improvements to how well neural SDEs can be incorporated in deep learning model systems include, e.g., the stochastic adjoint method allowing for accurate gradients and constant space complexity in determining gradients (Li et al., 2020), efficient algorithms for simulating the Brownian motion (Li et al., 2020; Kidger et al., 2021b), improving the calibration of uncertainty estimation (Look et al., 2022), and improvements to SDE solvers, such as the introduction of the reversible Heun solver (Kidger et al., 2021b). While Li et al. (2020) consider the incorporation of neural SDEs in a VAE framework, they can also be part of a generative adversarial network (GAN)-like model (Kidger et al., 2021a;b).

Contributions In the following, we introduce a continuously deep variational auto-encoder (CD-VAE). The model can be seen as an extension of the VD-VAE (Child, 2021), using a hierarchy of latent SDEs as latent processes thus replacing the discrete latent variables within levels of VD-VAEs. Utilizing the dynamical systems perspective, we explore whether the model displays continuity properties by investigating how performance is affected by changes to the step size used in numerical integration. We start by introducing VAEs and provide details on the VD-VAE in Section 2. Following this, we introduce the proposed CD-VAE in Section 3 alongside specifics of the latent SDE as introduced by Li et al. (2020). Finally, we present and discuss experiments applying the model to two simple computer vision tasks in Section 4.

2 Hierarchical Variational Auto-Encoders

A standard variational autoencoder (Kingma & Welling, 2013; Rezende et al., 2014) has an inference network with parameters φ that takes an input, $\boldsymbol{x}$, and parameterizes and approximate posterior distribution, $q_\varphi(\mathbf{z}|\boldsymbol{x})$, over the latent random variable, $\mathbf{z}$. The generative network, with parameters θ, parameterizes distributions, $p_\theta(\mathbf{x}|\boldsymbol{z})$, of the reconstructed input, $\mathbf{x}$, and we specify some prior over the latent distribution, $p_\theta(\mathbf{z})$—typically an isotropic Gaussian. VAEs are then trained to optimize a bound on the marginal likelihood, the evidence lower bound (ELBO) $\mathcal{L}$:

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\theta, \varphi; \boldsymbol{x}) = \mathbb{E}_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\mathbf{z})\right] - D_{\mathrm{KL}}\left(q_\varphi(\mathbf{z}|\boldsymbol{x}) \,||p_\theta(\mathbf{z})\right) = -(D+R), \tag{1}$$

where we introduce the negative log-likelihood of the datum, the distortion, D, and the Kullback–Leibler (KL) divergence between the approximate posterior and the prior, the rate, R.

Very Deep Variational Auto-Encoders In hierarchical VAEs, the basic VAE framework is extended to consider a hierarchy of latent variables (Sønderby et al., 2016; Child, 2021). In VD-VAEs, a hierarchy of N latents, $\mathbf{z}_0, \ldots, \mathbf{z}_N$, is used (collectively referred to as $\mathbf{z}$). Letting the top-most latent variable be $\mathbf{z}_0$, we let each variable be dependent on the latent variables higher in the hierarchy, defining the prior as $p_\theta(\mathbf{z}_0)\, p_\theta(\mathbf{z}_1|\,\mathbf{z}_0), \ldots, p_\theta(\mathbf{z}_N|\mathbf{z}_{<N})$, where $\mathbf{z}_{<N}$ is a shorthand for all latents higher in the hierarchy than N. Similarly, we define the approximate posterior as $q_\varphi(\mathbf{z}_0|\boldsymbol{x})\, q_\varphi(\mathbf{z}_1|\mathbf{z}_0, \boldsymbol{x}) \ldots q_\varphi(\mathbf{z}_N|\mathbf{z}_{<N}, \boldsymbol{x})$. The architecture for parameterizing the prior and approximate posteriors in a VD-VAE is visualized in Figure 1 and Figure 2. Using the hierarchical structure for the VD-VAE, we, more explicitly, get the following ELBO for the VD-VAE, $\mathcal{L}_{\mathrm{VD}}$:

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &\geq \mathcal{L}_{\mathrm{VD}}(\theta, \varphi; \boldsymbol{x}) \\ &= \mathbb{E}_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\mathbf{z})\right] - \sum_{n=1}^{N} D_{\mathrm{KL}}\left(q_\varphi(\mathbf{z}_n|\mathbf{z}_{<n}, \boldsymbol{x})||p_\theta(\mathbf{z}_n|\mathbf{z}_{<n})\right) - D_{\mathrm{KL}}\left(q_\varphi(\mathbf{z}_0|\boldsymbol{x})||p_\theta(\mathbf{z}_0)\right). \end{aligned} \tag{2}$$

A VD-VAE has a deterministic bottom-up (BU) path (see left side of Figure 1, green box) and a stochastic top-down (TD) path (middle red box). The hierarchy of latent variables is the yellow blocks in the top-down

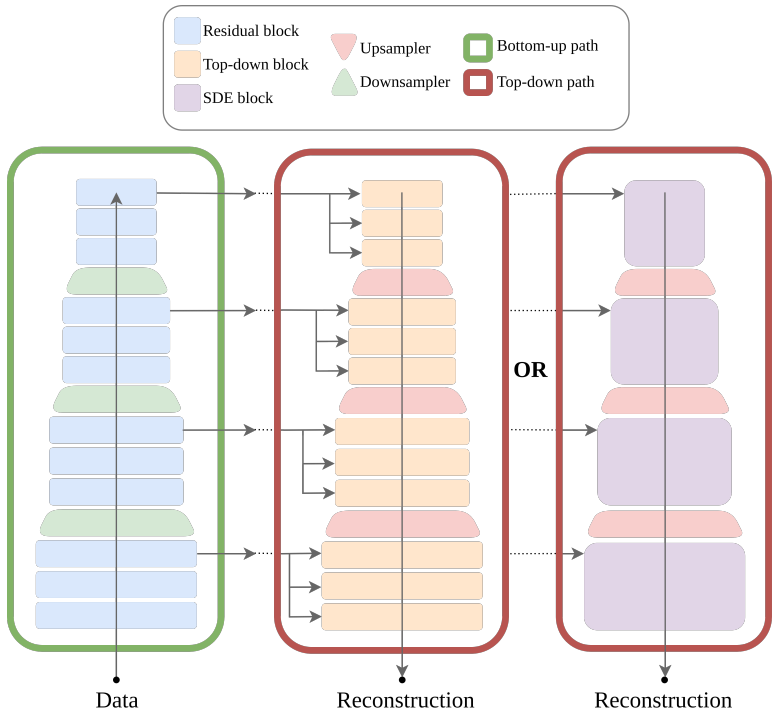


Figure 1: Overview of considered type of hierarchical variational auto-encoders. Data is encoded using a bottom-up path (left) and reconstructed using a top-down path, which is either the version used in a VD-VAE (middle) or the one used in this work, the CD-VAE (right). The two models share the same structure for the bottom-up path (green), and the difference between them is the replacement of top-down blocks (yellow) in the VD-VAEs with SDE blocks (purple) in the CD-VAE for their respective top-down paths (red boxes).

path. The paths are organized into levels, where each level operates on a particular spatial resolution. In the BU path, going from the bottom and up, the levels are separated by averaging pooling layers to reduce the resolution. In the TD path, going from top to bottom, the levels are separated by nearest neighbor up-sampling layers to increase the resolution. The BU levels consist of a series of residual blocks, and the TD levels consist of (discrete) TD blocks, which are residual blocks adapted to parameterize a prior and approximate posterior.

The TD blocks are visualized in Figure 2 (left). Each TD block takes as input two signals. The first is the BU path signal from the corresponding level at the same resolution ("from the left"), which can be thought of as a data conditioning signal used in conditioning the approximate posterior on the input datum. The second signal is the TD signal from the block above in the hierarchy ("from the top"), which can be considered conditioning from more spatially abstract parts of the representation. The TD signal is used to parameterize the prior for the current level with a convolution block, thus conditioning the prior distributions on latents above in the hierarchy. When we concatenate the TD and BU contexts, we get the input to a convolutional block that parameterizes the approximate posterior for that specific latent variable. The pathway on the right side of the block propagates the TD signal all through the TD path. It adds two components for

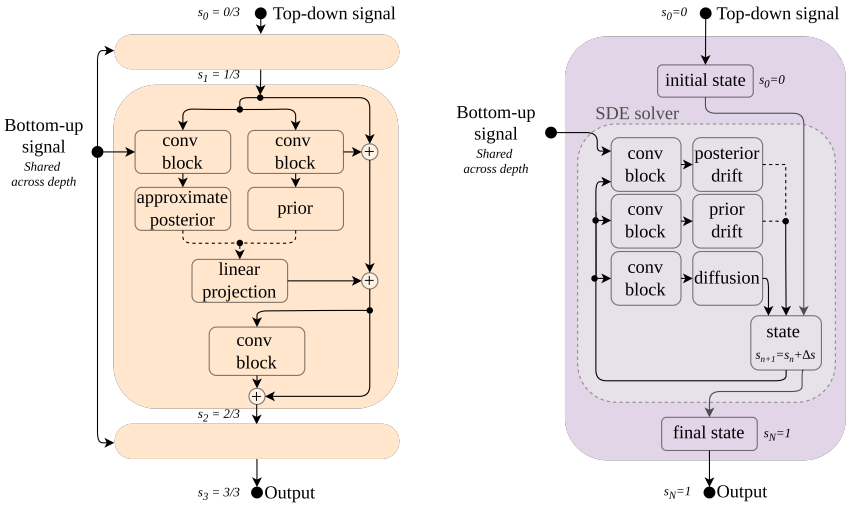


Figure 2: Stochastic blocks used in considered hierarchical VAEs; either a set of VD-VAE TD blocks (left, yellow) or an SDE block (right, purple). Convolutional blocks ("conv blocks") are four successive two-dimensional convolutions with square kernels of sizes 1, 3, 3, and 1 with final input and output dimensions matching the latent variable dimensionality. Dashed lines indicate that only one of the signals is used in updating the state; during training, the approximate posterior (for the VD-VAE) or the posterior drift (for the CD-VAE) updates the state, but the prior and prior drift are evaluated such that we can evaluate the rates. In the CD-VAE, we replace the three discrete levels (indicated by the $s = 0/0, \ldots, 1/3$) with a single SDE block with a flexible number of steps, n, or equivalently, step size, Δs. For further details, see the text.

each block: one is a signal depending on the prior block, and another is a linear projection of a sample from *either* the prior or the posterior (not both). The sample to be added to the residual TD pathway is either a sample from the approximate posterior or the prior, depending on whether we want to condition on data. During training, we do not need to sample the prior. However, we need to measure the KL divergence between the approximate posteriors and the priors—when training, the sample used is always a sample from the approximate posterior. Note that the VD-VAE takes an extra signal depending on the same features that parameterize the prior and adds that to the TD pathway—the signal is not a sample from the prior, but a deterministic, additive component (extra filters are added to the final convolution in the convolutional block that also provides the parameterization of the prior). In addition to this particular architecture, Child (2021) uses various training specifics for stabilization in training, such as gradient skipping. Within a level, several blocks share the same bottom-up signal. For example, in Figure 1, there are four levels with three blocks corresponding to twelve latent variables (each a vector of some latent dimensionality).

3 Continuously Deep Variational Auto-Encoder

Latent SDEs The following aligns with the presentation in Li et al. (2020), and we refer to Li et al. (2020, Sec. 5) for a detailed description of the latent SDEs. We can, loosely, interpret the TD blocks within a level as a discretization of a (latent, neural) SDE (Li et al., 2020). Figure 2 (right) shows an SDE block. The central element in the SDE block is an SDE solver which finds a numerical solution given the latent process defined by a prior and posterior drift and a shared diffusion. The TD signal provides the initial value for the SDE solve in the block. The solver determines a solution given the initial value and three functions: the

posterior drift, the prior drift, and the shared diffusion, each defined by neural networks mapping from the current state and depth to the terms in the differential equation. Just like the prior in the VD-VAE does not depend on data, the prior drift and diffusion do not depend on the data through BU path—only the posterior drift depends on data.

Prior and approximate posterior processes We interpret the dimension over which we integrate as a depth in a (continuously deep) hierarchy, such that we have depth, s, and a "depth horizon", $\mathbb{S} = [0, S]$, corresponding to the depth of the hierarchy. We have levels, $l = 0, 1, \ldots, L$, each defined by one set of SDEs (one SDE block). Each level spans a unit length depth horizon $\mathbb{S}_l = [l, l + 1]$, such that the total depth of the hierarchy is $S = L + 1$. We let each level have independent parameterizations of the drifts and the diffusion. Thus, the model can viewed as a hierarchy of SDEs in which the terminal state of each level defines the initial value for the next. Within each level, defined by an SDE block, we have two processes. These two processes share a diffusion term controlled by a k_l-dimensional Brownian motion, $\{W_s^l\}_{s\in\mathbb{S}_l}$. The two processes defined by the SDEs are:

$$\mathrm{d}\tilde{Z}_s^l = h_\xi^l(\tilde{Z}_s^l, s)\mathrm{d}s + \sigma_\psi^l(\tilde{Z}_s^l, s)\mathrm{d}W_s^l, \tag{3}$$

$$\mathrm{d}Z_s^l = h_\zeta(Z_s^l, s)\mathrm{d}s + \sigma_\psi^l(Z_s^l, s)\mathrm{d}W_s^l, \tag{4}$$

where $\{\tilde{Z}_s^l\}_{s\in\mathbb{S}_l}$ and $\{Z_s^l\}_{s\in\mathbb{S}_l}$ are the prior and approximate posterior stochastic process, respectively, each having separate drift terms, h, but sharing a diffusion, σ, for the level, l. Here $h_\xi^l : \mathbb{R}^{k_l} \times \mathbb{R} \to \mathbb{R}^{k_l}$ denotes a neural network that parameterizes the prior drift based on the current depth, s, and the current latent prior state, $\tilde{Z}_s^l \in \mathbb{R}^{k_l}$, where k_l denotes the dimensionality of the state. Similarly, $h_\zeta^l : \mathbb{R}^{k_l} \times \mathbb{R} \to \mathbb{R}^{k_l}$ is a neural network that takes the depth and state and provides the posterior drift; the parameters ζ depend on a deterministic, data-dependent context from the BU path. The shared diffusion is parameterized by a network $\sigma_\psi^l : \mathbb{R}^{k^l} \times \mathbb{R} \to \mathbb{R}^{k^l}$ in a manner ensuring that the diffusion is weakly diagonal. The above general description allows for a depth- and state-dependency of all the functions; however, we opted for a constant, scalar, learned diffusion independent of both depth and state (this simplification also ensures that there is no difference between Itô and Stratonovich solutions to the SDEs (Kidger, 2021, Sec. 4.1)).

CD-VAE compared to the VD-VAE The stochasticity from sampling (either in the prior or the approximate posterior) in a VD-VAE corresponds to the stochasticity of the controlling Brownian motion in the SDEs' diffusion term. In Figure 2, we compare the two types of stochastic blocks—the right-hand side shows an SDE block. The SDE block takes as input a BU (data dependent signal) and a TD signal, entirely like the VD-VAE stochastic block. For the top-most block alone, the top-down signal stems from a (prior, posterior)-pair like that in a regular VAEs—here, an isotropic Gaussian prior and a Gaussian approximate posterior dependent on the top-most BU signal. The CD-VAE uses a shared parameterization across the depth of a given level; the VD-VAE has independent weights for each TD block, but the CD-VAE uses the same parameterization for all depths within a level.

Comparison to earlier latent SDEs A few central aspects of our proposed latent SDE blocks differ from that of Li et al. (2020). Li et al. (2020) model temporal sequences and encode them into temporally evolving latent states. In contrast, we consider images with latent spatial states; using the spatial structure in the latent state, the CD-VAE vector fields are convolutional neural networks parameterizing the drifts and diffusions. The model that Li et al. (2020) present has a time series of contexts for the approximate posterior process, where we use a single, shared bottom-up signal for each level ("shared across depth" within a level). The CD-VAE has a hierarchy of latent SDEs, multiple SDEs operating at different spatial resolutions resulting in different dimensionalities of the latent state—and of the Brownian motion—for each level.

Mock spatial latent process visualization We keep the number of latent dimensions constant *per spatial dimension* across levels, but we increase the number of latent spatial channels. Figure 3 shows a mock example of three levels of a CD-VAE using a simplified set of processes for visualization purposes. The evolution is governed by prior, posterior, and diffusion functions. At the top level, the spatial resolution of the latents results in a single latent spatial dimension, i.e., the latents are a 1-by-1 "latent image", where

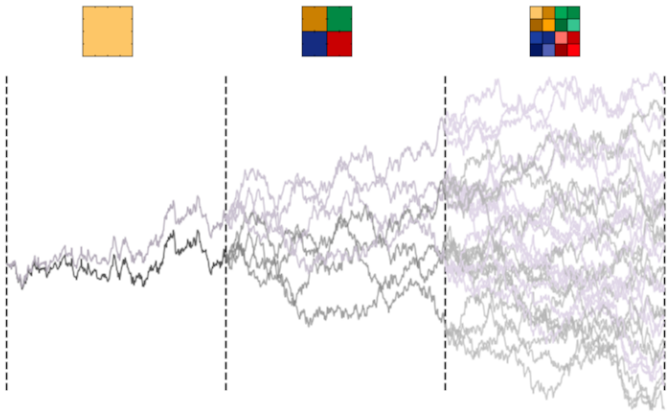


Figure 3: Mock example of a CD-VAE with three levels of different resolution, starting with a 1x1 spatial configuration and ultimately producing a 4x4 latent state. Grey shows a prior process and purple shows a similar posterior process affected by a different drift. From left to right, we go through three levels of different spatial resolution separated by dashed lines; the first dashed line corresponds to a depth $s = 0$, and the final dashed line corresponds to a depth $s = 3$.

each pixel has a process of some dimensionality. Here, this dimensionality is 1, for visualization purposes, corresponding to one set of lines "per latent pixel": a grey prior process and a purple posterior process. The two processes are similar since they share the same diffusion (the same stochastic evolution controlled by a sampled Brownian motion and σ-functions) but differ since they have different drifts. When training and evaluating the model, we obtain the values for the latent states by using a numerical integrator within a level, i.e., to obtain the solution between each pair of dashed lines, we call an SDE solver. Specifically, we make use of Diffrax (Kidger, 2021) to handle the SDEs and Equinox (Kidger & Garcia, 2021) to construct the neural networks. At the top level, the processes share the same initial value (at the left-most dashed line). In lower levels, the initial values are dictated by the higher level's evolution and the current level's spatial configuration. The final state of the previous approximate posterior process is upsampled, and the resulting latent image is used as the initial value for the next level. For example, when the second level (denoted by the second dashed line) starts, the spatial 1x1 latents are up-sampled to be a 2x2 configuration, resulting in a prior and an approximate posterior process with four dimensions that share the same initial value. Each of the dimensions "within a latent pixel" evolves differently because of the different Brownian motions and—in the case of the posterior—because of different spatial contexts from the data. The upsampling is, for this figure, repeated another time, and the lower-most level shows the evolution of a prior and a posterior for a 4x4 spatial configuration of latents. Finally, the final state of the lowermost level is sent to a convolutional decoder. This decoder uses a set of convolutional blocks to parameterize an observation model/distribution, which in our case is either a per-pixel Bernoulli or Gaussian distribution.

CD-VAE ELBO The CD-VAE shares the same BU path as the VD-VAE. In both, the top BU level output is used to parameterize a Gaussian approximate posterior. This distribution's divergence to a standard Gaussian prior is the first component in the ELBO, which in the CD-VAE we refer to as the "initial rate". In the VD-VAE, all remaining TD blocks function in this manner. However, only the topmost level uses this structure in the CD-VAE, where this block parameterizes the initial value for the entirety of the hierarchy. Instead of a KL divergences between discrete Gaussian distributions, we determine a "path rate" when training CD-VAEs (during the solver call in the SDE blocks). The path rate is the KL between the prior and posterior processes and takes on a form distinct from the discrete case. For the CD-VAE, the evidence

lower bound, $\mathcal{L}_{\text{CD}}$, is:

$$\log p_{\xi,\psi,\theta}\left(\boldsymbol{x}\right) \geq \mathcal{L}_{\text{CD}}\left(\varphi,\theta,\xi,\zeta,\psi;\boldsymbol{x}\right)$$
$$= \mathbb{E}\left[\log p_{\xi,\psi,\theta}\left(\boldsymbol{x}|\boldsymbol{z}_S^L\right) - \sum_{l=0}^{L}\int_l^{l+1}\frac{1}{2}|u^l\left(\boldsymbol{z}_s^l,s\right)|^2\mathrm{d}s - \mathrm{KL}\left(q_\varphi(\mathbf{z}_0^0|\boldsymbol{x})||p_\theta(\mathbf{z}_0^0)\right)\right], \quad (5)$$

where the expectation is under the approximate posterior distribution for the initial value and the approximate posterior processes distributions. The first term corresponds to the distortion or the reconstruction error. The second term is a a sum over each level tallying the integration of a function u, which measures the difference in the posterior and prior drift scaled by the diffusion (defined by $\sigma_\psi^l\left(\boldsymbol{z}_s^l,s\right)u^l\left(\boldsymbol{z}_s^l,s\right) = h_\zeta^l\left(\boldsymbol{z}_s^l,s\right) - h_\xi^l\left(\boldsymbol{z}_s^l,s\right)$). During training, we augment the SDE solves with a dimension that has a diffusion term corresponding to the integrand and with a diffusion term of zero. The SDE solve thus provide the path rate (i.e., the integral in the second term of the ELBO) alongside the evolution of the latent processes. The last term corresponds to a more standard VAE KL between the two Gaussians defining the initial values, the "initial rate", where the φ,θ are the parameters of the networks parameterizing the initial values of the approximate posterior and prior, respectively. Compared to the ELBO for the VD-VAE in Equation (2), the main change is the replacement of the sum over N discrete KL-divergences in the second term in the VD-VAE ELBO with a sum over path rates in the second term of Equation (5). The derivation of the ELBO as presented in Li et al. (2020) relies on relating the change of (probability) measure for the posterior process to the prior process using Girsanov's Theorem II (Øksendal, 2013, Theorem 8.6.6). For details on the derivation of the variational bound and prerequisite assumptions of regularity, we refer to Li et al. (2020, Sec. 9).

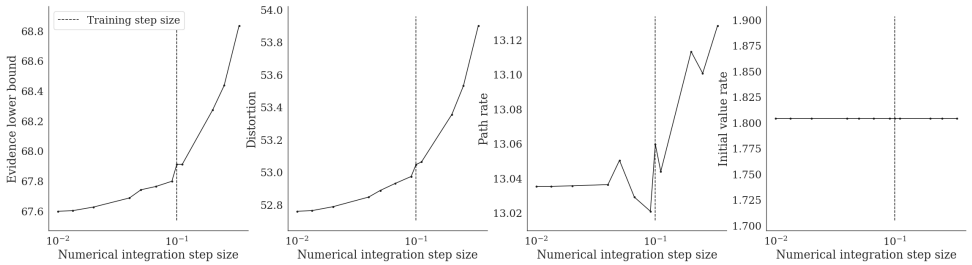
ODE-like CD-VAE comparison We compare the CD-VAE to a model constructed like the CD-VAE but with two changes: (1) the bottom-up signals are all removed (multiplied by zero) except for the signal that parameterizes the initial value approximate posterior, and (2) the path rate term in the loss is removed (multiplied by zero during training). This model is free to increase the path rate without penalization. Consequently, the magnitude differences between the drift term and the diffusion term are much higher, yielding a high (drift-)signal-to-(diffusion-)noise ratio. This effectively produces an ODE by ignoring the diffusion, and we will refer to this model as an ODE-like CD-VAE. For this model, the approximate posterior evolution is only affected by data in its initial value. We introduce this comparison, not as a fair baseline or a competitive model, but to compare the effect on representation continuity of minimizing the KL divergence between the approximate posterior process and the prior process in the CD-VAE.

4 Results and discussion

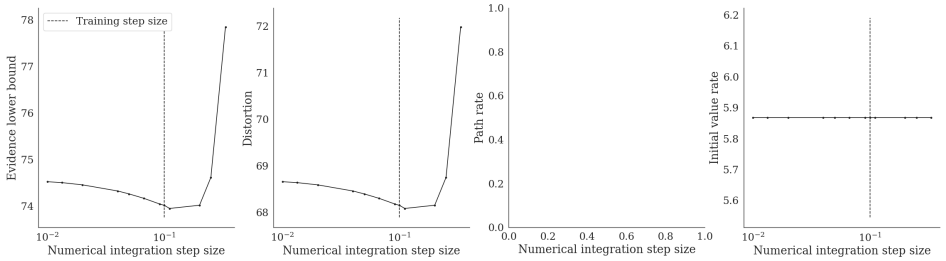
In the following, we present results for a simple Poisson equation data set and a binarized version of MNIST. We provide details on hyper-parameters and model construction in Appendix A.1.

Synthetic Poisson equation data set We construct a data set consisting of 16-by-16 pixels images corresponding to noisy observations of two different solutions to simple Poisson equations. We use exact solutions to $-\Delta h\left(x,y\right) = f\left(x,y\right)$, where $f\left(x,y\right) = \sin\left(a\pi x\right)\sin\left(b\pi y\right)$ for $(x,y) \in [0,1]^2$ with Dirichlet boundary conditions $h\left(x,0\right) = h\left(x,1\right) = h\left(0,y\right) = h\left(1,y\right) = 0$. We use either $a = b = 1$ or $a = b = 3$ to construct the ground truths. The data set consists of 1024 observations—each observation is either of the two ground truths (in equal proportions) with an additive Gaussian noise of scale $5 \cdot 10^{-3}$. The data set is split into partitions of proportions 80 %, 10 %, 10 % used for training, validation, and testing, respectively. We show the ground truth solutions, their noisy observations, example reconstructions, and samples from the learned prior in Appendix A.2 in Figure 6. There, we also show a detailed visualization of the processes and numerical integration in Figure 7 (i.e., corresponding to a detailed non-mocked version of Figure 3).

In Figure 4a, we show results for varying the number of steps used during the call to the numerical solver for the CD-VAE and the ODE-like CD-VAE. The figure shows how the step size affects the loss components when evaluating the model on a held-out test set. The CD-VAE improves modestly (see left-most plot) when increasing the number of steps taken (i.e., decreasing the step size) compared to the step size used during



(a) CD-VAE



(b) Comparison ODE-like CD-VAE model (the path rate is ignored in the loss, so no curve is shown).

Figure 4: Loss components as a function of integration step size, Δs, for both the CD-VAE and a simple example of a model not displaying continuity. From left to right, we show the evidence lower bound, the distortion, the path rate, and the initial value rate. All are given in nats.

training. The improvement amounts to about 0.3 nats when decreasing the step size by a magnitude. When increasing the step size, we see modest but slightly more significant decreases in performance; there is a drop of 1 nat at a step size of 1/3, which is the largest step size considered. The initial value rate is constant over the range of steps, and the path rate changes slightly, but the changes in achieved ELBO are primarily driven by changes in distortion.

In Figure 4b, we show the results of running the same experiment on the ODE-like CD-VAE model. The bound for this model does not consider the path rate, so no curve is shown. The ODE-like CDVAE puts the information needed for reconstruction into the initial value rate. The model does not achieve a lower bound as good as the CD-VAE; while the initial value rate of about 5.9 nats is much lower than the CD-VAE's combined path and initial values rates of about 14 nats, the distortions are much higher, thus producing an ELBO of about 74 nats (about 6.4 nats higher than the CD-VAE using the smallest step size). In the absence of a term in the loss that penalized divergence from a well-behaved prior process, the ODE-like CD-VAE displays its lowest ELBO around the training step size and, in contrast to the CD-VAE, does not improve by decreasing the step size.

Binarized MNIST We train a CD-VAE on a binarized MNIST using the code provided with the Efficient VD-VAE paper (Hazami et al., 2022) to construct the data set. We provide example outputs in the appendix, Figure 8. We show example reconstructions, samples from the learned prior, and samples conditioned on only the top level in the hierarchy (showing the variation in produced samples when only the most spatially abstract level is informed about the datum).

We conduct the same experiment of changing the numerical integration step size after training and show the results for the ELBO on the test set in Figure 5. While the binarized MNIST is still a simple problem, it is twice the spatial resolution and more diverse than the simple Poisson data set. We see that the CD-VAE

improves its ELBO when decreasing the step size. As points of comparison, we show the performance of the consistency-regularized nouveau VAE (CR-NVAE) (Sinha & Dieng, 2021), the efficient VD-VAE (Hazami et al., 2022), the PixelCNN (van den Oord et al., 2016b), and MADE (Germain et al., 2015). The CD-VAE attains a performance between the level of MADE and a PixelCNN. The improvements attained by decreasing the step size by one or two magnitudes amount to about one nat, or about half the improvement seen in going from, e.g., a PixelCNN to an Efficient VD-VAE.

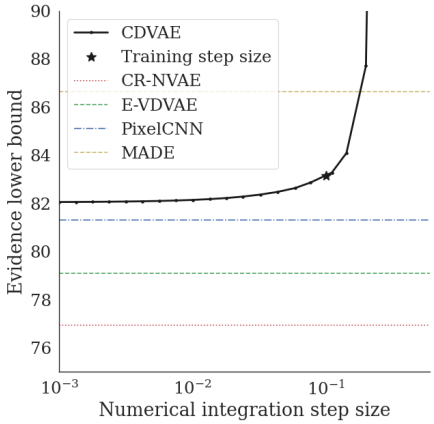


Figure 5: Evidence lower bound as a function of numerical integration step size for a CD-VAE trained on binarized MNIST and comparisons to other deep generative models.

Extensions Hierarchical VAEs have various ties to flows, auto-regressive models and diffusion models, especially with an SDE perspective. Models like the CD-VAE are possible avenues of drawing on the strengths of multiple frameworks, such as combining lower dimensionality latent spaces of VAEs and the expressive power of recent diffusion models (Kingma et al., 2021; Rombach et al., 2022). The CD-VAE share learned features across the depth of the hierarchy, much like convolutional neural networks share features over the spatial dimensions of their input. The presented CD-VAE uses a single set of convolutional blocks for a given level, sharing the parameters across all depths within a level—a level that would otherwise correspond to multiple distinct blocks in a VD-VAE, each block with each their own set of parameters. Sharing the parameters in this manner might be an effective parametrization, up to some depth horizon, and extensions to view the weights as continuous-in-depth functions using basis functions, as presented by Queiruga et al. (2021), could further the benefits of a continuously deep hierarchy. As seen in Figure 7, the learned dynamics are simple; allowing more expressive neural networks in the neural SDEs could improve performance. We use a simple SDE solver with a constant step size during training. The CD-VAE might be improved with a stronger solver, such as the Reversible Heun solver (Kidger et al., 2021a) or by using an adaptive step size controller. The current CD-VAE formulation could be viewed as a particular discretization of a neural stochastic partial differential equation (Salvi & Lemercier, 2021); this view could enable dynamically adapting the current static resolution changes in the CD-VAE's bottom-up and top-down paths, thus learning the inherent relevant scales of the data. In this work, we make use of a discretize-then-optimize approach. However, the stochastic adjoint method (Li et al., 2020) could allow for deep hierarchies at constant memory requirements, which—in conjunction with the observation by Child (2021) that greater stochastic depth is beneficial—might allow the CD-VAEs to be highly expressive models. We use single-sample estimates of the initial values and processes. Yet, approaches like the importance-weighted auto-encoder (Burda et al., 2016) show that moving beyond a single-sample estimate of the expectation in the ELBO can be beneficial. We retain a (discrete) ResNet for the bottom-up path, but the model could be adapted to use a ContinuousNet-style formulation with an ODE (Queiruga et al., 2020). Similarly, we improved inference in hierarchical VAEs

using BIVA (Maaløe et al., 2019) could be adapted to the CD-VAE structure, in which case the CD-VAE bottom-up path would be constructed as another set of neural SDEs.

5 Conclusion

We present a hierarchical variational autoencoder using latent stochastic differential equations. The model extends the architecture of the very deep variational auto-encoder; using a hierarchy of stochastic processes as latent objects, we introduce a continuously deep variational auto-encoder. We show results for training the model on two simple computer vision tasks, and we show how the learned representation displays properties of continuity in the sense that the model generalizes to step sizes beyond its training step size. Depth-continuity of this type allows a single trained model to be used across a range of computational restrictions in downstream uses of generative models. The CD-VAE has a number of possible extensions that have proven helpful in similar previous models, such as using more expressive parametrizations of the functions defining the differential equation, adaptive step size controllers, etc. The presented results add to ongoing explorations of depth-continuity of model representations, extending conclusions on improved performance with decreased step sizes to generative models.

Acknowledgements

The authors would like to thank Morten Mørup for feedback on the manuscript. Through RH, this work was partly funded by the Innovation Fund Denmark (IFD, grant number: 9065-00077B).

References

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris George Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. Importance Weighted Autoencoders. In *ICLR (Poster)*, 2016.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H Chi. AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks. In *International Conference on Learning Representations*, 2018.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Rewon Child. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=RLRXCV6DbEJ`.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL `http://arxiv.org/abs/1410.8516`.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=HkpbnH9lx`.

N. Benjamin Erichson, Omri Azencot, Alejandro Queiruga, Liam Hodgkinson, and Michael W. Mahoney. Lipschitz Recurrent Neural Networks. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=-N7PBXqOUJZ`.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pp. 881–889. PMLR, 2015.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *International Conference on Learning Representations*, 2018.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1): 014004, 2017.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-VDVAE: Less is more. *arXiv preprint arXiv:2203.13751*, 2022.

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

Liam Hodgkinson, Chris van der Heide, Fred Roosta, and Michael W. Mahoney. Stochastic continuous normalizing flows: training SDEs as ODEs. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1130–1140. PMLR, 27–30 Jul 2021. URL `https://proceedings.mlr.press/v161/hodgkinson21a.html`.

Junteng Jia and Austin R Benson. Neural Jump Stochastic Differential Equations. *Advances in Neural Information Processing Systems*, 32, 2019.

Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2021.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

Patrick Kidger, James Foster, Xuechen Li, and Terry J Lyons. Neural SDEs as Infinite-Dimensional GANs. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5453–5463. PMLR, 18–24 Jul 2021a. URL `https://proceedings.mlr.press/v139/kidger21b.html`.

Patrick Kidger, James Foster, Xuechen (Chen) Li, and Terry Lyons. Efficient and Accurate Gradients for Neural SDEs. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18747–18761. Curran Associates, Inc., 2021b. URL `https://proceedings.neurips.cc/paper/2021/file/9ba196c7a6e89eafd0954de80fc1b224-Paper.pdf`.

Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations*, 2013.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational Diffusion Models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=2LdBqxc1Yv`.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Aditi S Krishnapriyan, Alejandro F Queiruga, N Benjamin Erichson, and Michael W Mahoney. Learning continuous models for continuous physics. *arXiv preprint arXiv:2202.08494*, 2022.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable Gradients for Stochastic Differential Equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.

Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh. Neural SDE: Stabilizing neural ode networks with stochastic noise. *arXiv preprint arXiv:1906.02355*, 2019.

Andreas Look, Melih Kandemir, Barbara Rakitsch, and Jan Peters. A Deterministic Approximation to Neural SDEs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2022. doi: 10.1109/TPAMI.2022.3202237.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling. *Advances in neural information processing systems*, 32, 2019.

Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. SurVAE flows: Surjections to bridge the gap between VAEs and flows. *Advances in Neural Information Processing Systems*, 33:12685–12696, 2020.

Alejandro Queiruga, N Benjamin Erichson, Liam Hodgkinson, and Michael W Mahoney. Stateful ODE-Nets using basis function expansions. *Advances in Neural Information Processing Systems*, 34:21770–21781, 2021.

Alejandro F Queiruga, N Benjamin Erichson, Dane Taylor, and Michael W Mahoney. Continuous-in-depth Neural Networks. *arXiv preprint arXiv:2008.02389*, 2020.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International conference on machine learning*, pp. 324–333. PMLR, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Tim Salimans. A structured variational auto-encoder for learning deep hierarchies of sparse features. *arXiv preprint arXiv:1602.08734*, 2016.

Cristopher Salvi and Maud Lemercier. Neural stochastic partial differential equations. *CoRR*, abs/2110.10249, 2021. URL `https://arxiv.org/abs/2110.10249`.

Gianluigi Silvestri, Daan Roos, and Luca Ambrogioni. Closing the Gap: Exact maximum likelihood training of generative autoencoders using invertible layers. *arXiv preprint arXiv:2205.09546*, 2022.

Samarth Sinha and Adji Bousso Dieng. Consistency Regularization for Variational Auto-Encoders. *Advances in Neural Information Processing Systems*, 34, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder Variational Autoencoders. *Advances in neural information processing systems*, 29, 2016.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=AkltWFnxS9`.

Jakub Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pp. 1214–1223. PMLR, 2018.

James Townsend, Thomas Bird, and David Barber. Practical lossless compression with latent variables using bits back coding. In *International Conference on Learning Representations*, 2018.

Belinda Tzen and Maxim Raginsky. Neural Stochastic Differential Equations: Deep Latent Gaussian Models in the Diffusion Limit. *arXiv preprint arXiv:1905.09883*, 2019.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*, 2016a.

Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel Recurrent Neural Networks. In *International conference on machine learning*, pp. 1747–1756. PMLR, 2016b.

Winnie Xu, Ricky T. Q. Chen, Xuechen Li, and David Duvenaud. Infinitely Deep Bayesian Neural Networks with Stochastic Differential Equations. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 721–738. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/xu22a.html`.

Bernt Øksendal. *Stochastic Differential Equations: an Introduction with Applications*. Springer Science & Business Media, 2013.

A Appendix

A.1 Model details

For the Poisson model, we train the model with a batch size of 128 for a fixed 4000 updates using an Adam optimizer and a learning rate of $3 \cdot 10^{-4}$.

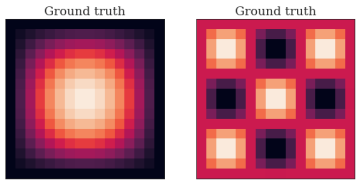
We have four channels per latent dimension in the latent processes and use a total of three levels with a change in resolution of a factor of two between the levels. The bottom-up encoder uses 128 hidden channels in the ResBlocks. The final part of the top-down path parametrizing the observation distribution uses three convolution blocks with 128 hidden channels. For the Poisson data, the final number of parameters matches the scale and location (mean and standard deviation) of a Gaussian per pixel.

The neural SDEs are trained with a constant step size of 0.1. The drift parametrizing convolutional blocks use 128 hidden channels each (both prior and posterior), and the drifts are a single, learned scalar per level. The final activations on the drift terms are linear. The drift terms are both state and depth-dependent. We use a Euler-Heun solver.

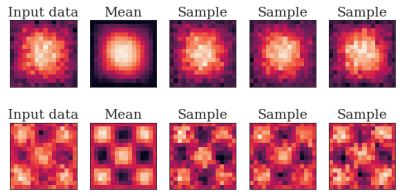
The ELBO uses "free nats", 1.0 nat across all dimensions for both the initial values and for each level in the hierarchy. We also use a sigmoidal warm-up over 1000 steps of the initial rate and path rates (slowly annealing in the KL terms).

For the MNIST models, we reduce the batch size to 32 to allow for a larger model. The MNIST model trains for 50000 steps and uses 16 latent per spatial dimensions but retains three levels separated by resolution changes of a factor of two. The bottom-up path channels are increased to 256 filters, and the same was done for the final part of the top-down path parametrizing the observation distribution. For the binarized MNIST data, we use a per-pixel Bernoulli distribution. The neural SDEs use the same construction of the processes as the Poisson data set model, and the loss uses the same free nats and KL-term warmups. The model starts with a learning rate of 10^{-3} decayed with a cosine decay throughout training to 10^{-4}.

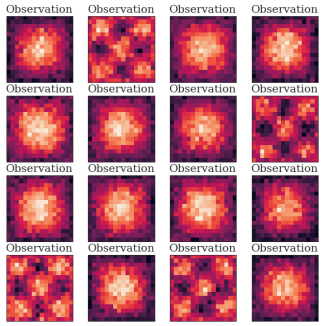
A.2 Poisson data set



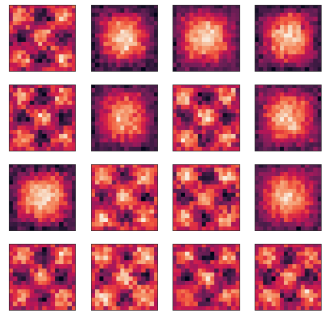
(a) Ground truth of two types of observations in the data set.



(b) Example reconstructions of the two types of data. 1st column is the input data, 2nd column is the observation distribution mean, and the last three columns are three samples.



(c) 16 examples from the training data set. Each observation is a noisy observation of one of the ground truth images.



(d) Samples from the prior of the learned model.

Figure 6: Overview of Poisson equation data set.

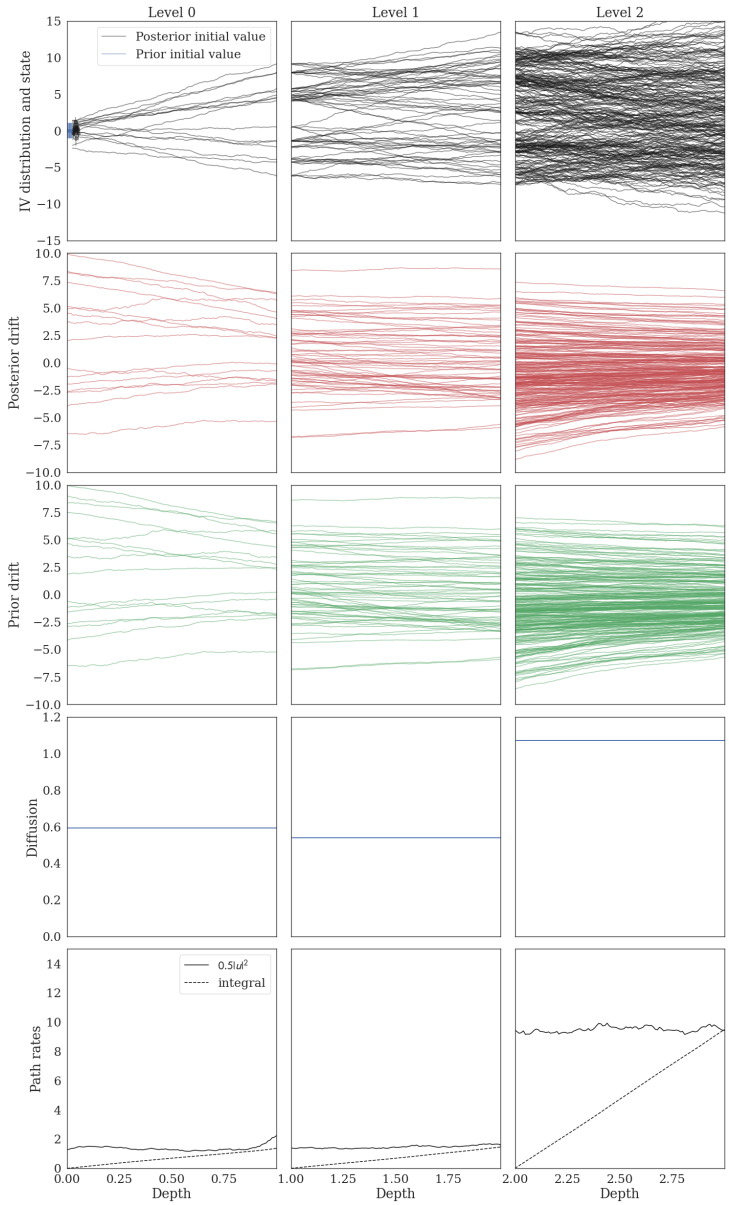
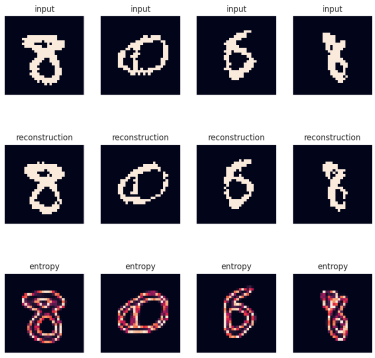
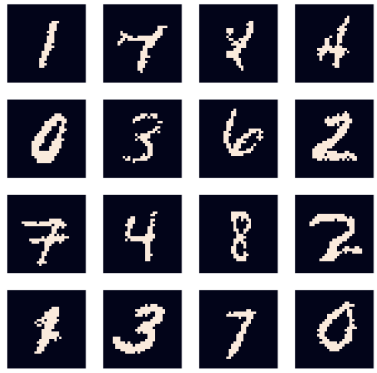


Figure 7: Detailed visualization of the latent processes in a Poisson data example. Each row shows one of three levels in the hierarchy. The first row shows the initial value distribution (black and blue) and the state (black) evolving over the depth. The second, third, and fourth rows show the posterior drift (red), prior drift (green), and diffusion (blue) at the approximate posterior state/depth inputs corresponding to the first row. The last row shows the path rates (both the instantaneous value, fully drawn, and the integral of the path rate, dashed).

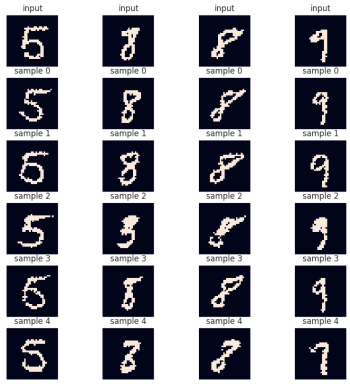
A.3 Binarized MNIST



(a) Example reconstructions. 1st row shows input data, 2nd row shows a sample from the recosntruction, and the 3rd row shows the entropy of the observation distribution.



(b) Samples from the learned prior.



(c) Example reconstructions where only the top level in the hierarchy sees the data conditioning signal. 1st row shows the input samples, a different one in each column, and the remaining rows show different samples.

Figure 8: Overview of the binarized MNIST data set and example model outputs.

Bibliography

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. (2018). Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning*, pages 159–168, Stockholmsmässan, Stockholm Sweden.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2017). Deep variational information bottleneck. In *International Conference on Learning Representations*.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*, volume 4. Springer.

Blau, Y. and Michaeli, T. (2019). Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff. In *International Conference on Machine Learning*, pages 675–685. PMLR.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bond-Taylor, S., Leach, A., Long, Y., and Willcocks, C. G. (2021). Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Burda, Y., Grosse, R. B., and Salakhutdinov, R. (2015). Importance Weighted Autoencoders. In *International Conference on Learning Representations*.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in (β)-VAE. *2017 NIPS Workshop on Learning Disentangled Representations*, abs/1804.03599.

Centers for Disease Control Prevention (CDC) and National Center for Health Statistics (NCHS) (1999). National health and nutrition examination survey data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. cdc.gov/nchs/nhanes.

Chaitin, G. J. (2004). Meta Math! The Quest for Omega. *https://arxiv.org/abs/math/0404335*.

Chang, B., Chen, M., Haber, E., and Chi, E. H. (2018). AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*.

Che, T., Zhang, R., Sohl-Dickstein, J., Larochelle, H., Paull, L., Cao, Y., and Bengio, Y. (2020). Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. *Advances in Neural Information Processing Systems*, 33:12275–12287.

Chen, J., Mao, Q., and Liu, D. (2020). Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. In *Interspeech 2020*, pages 2642–2646.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. (2018). Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. (2017). Variational Lossy Autoencoder. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5):975–979.

Child, R. (2021). Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*.

Chorowski, J., Weiss, R. J., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using WaveNet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2041–2053.

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., and Vincent, E. (2020). LibriMix: An Open-Source Dataset for Generalizable Speech Separation. *arXiv:2005.11262 [eess]*.

Cutler, A. and Breiman, L. (1994). Archetypal Analysis. *Technometrics*, 36(4):338–347.

Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: Non-linear independent components estimation. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017). Density estimation using real NVP. In *International Conference on Learning Representations*.

Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430.

Erichson, N. B., Azencot, O., Queiruga, A., Hodgkinson, L., and Mahoney, M. W. (2021). Lipschitz recurrent neural networks. In *International Conference on Learning Representations*.

Gardner, J. R., Song, X., Weinberger, K. Q., Barbour, D. L., and Cunningham, J. P. (2015). Psychophysical detection testing with bayesian active learning. In *UAI*, pages 286–295.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). MADE: Masked Autoencoder for Distribution Estimation. In *International Conference on Machine Learning*, pages 881–889. PMLR.

Ghiasi, A., Kazemi, H., Huang, R., Liu, E., Goldblum, M., and Goldstein, T. (2021). Feature sonification: An investigation on the features learned for automatic speech recognition.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Granger, R. (2017). How brains are built: Principles of computational neuroscience.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. (2018). FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*.

Haber, E. and Ruthotto, L. (2017). Stable architectures for deep neural networks. *Inverse Problems. An International Journal on the Theory and Practice of Inverse Problems, Inverse Methods and Computerized Inversion of Data*, 34(1):014004.

Hawking, S. W. (2001). *The Universe in a Nutshell*. Bantam, London.

Hazami, L., Mama, R., and Thurairatnam, R. (2022). Efficient-VDVAE: Less is more. *arXiv preprint arXiv:2203.13751*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Heitkaemper, J., Jakobeit, D., Boeddeker, C., Drude, L., and Haeb-Umbach, R. (2020). Demystifying TasNet: A dissecting approach. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6359–6363. IEEE.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Hodgkinson, L., van der Heide, C., Roosta, F., and Mahoney, M. W. (2021). Stochastic continuous normalizing flows: Training SDEs as ODEs. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1130–1140. PMLR.

Hoogeboom, E., Gritsenko, A. A., Bastings, J., Poole, B., van den Berg, R., and Salimans, T. (2022). Autoregressive diffusion models. In *International Conference on Learning Representations*.

Ipsen, N. B., Mattei, P.-A., and Frellsen, J. (2021). Not-MIWAE: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations*.

Jia, J. and Benson, A. R. (2019). Neural jump stochastic differential equations. *Advances in Neural Information Processing Systems*, 32.

Kadioglu, B., Horgan, M., Liu, X., Pons, J., Darcy, D., and Kumar, V. (2020). An Empirical Study of Conv-Tasnet. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 7264–7268. IEEE.

Kates, J. M. (2008). *Digital Hearing Aids*. Plural publishing.

Kidger, P. (2021). *On Neural Differential Equations*. PhD thesis, University of Oxford.

Kidger, P., Foster, J., Li, X., and Lyons, T. J. (2021a). Neural SDEs as infinite-dimensional GANs. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5453–5463. PMLR.

Kidger, P., Foster, J., Li, X. C., and Lyons, T. (2021b). Efficient and accurate gradients for neural SDEs. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 18747–18761. Curran Associates, Inc.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M. (2014). Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved Variational Inference with Inverse Autoregressive Flow. *Advances in neural information processing systems*, 29:4743–4751.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. (2021). Variational diffusion models. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2019). An Introduction to Variational Autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392.

Krishnapriyan, A. S., Queiruga, A. F., Erichson, N. B., and Mahoney, M. W. (2022). Learning continuous models for continuous physics. *arXiv preprint arXiv:2202.08494*.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al. (2021). On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). SDR–Half-Baked or Well Done? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Li, X., Wong, T.-K. L., Chen, R. T. Q., and Duvenaud, D. K. (2020). Scalable gradients and variational inference for stochastic differential equations. In Zhang, C., Ruiz, F., Bui, T., Dieng, A. B., and Liang, D., editors, *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–28. PMLR.

Look, A., Kandemir, M., Rakitsch, B., and Peters, J. (2022). A deterministic approximation to neural SDEs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15.

Luo, Y., Chen, Z., and Yoshioka, T. (2020). Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. *arXiv:1910.06379*.

Luo, Y. and Mesgarani, N. (2018). TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE.

Luo, Y. and Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266.

Ma, C., Gong, W., Hernández-Lobato, J. M., Koenigstein, N., Nowozin, S., and Zhang, C. (2018). Partial VAE for hybrid recommender system. In *NIPS Workshop on Bayesian Deep Learning*, volume 2018.

Ma, C., Tschiatschek, S., Palla, K., Hernandez-Lobato, J. M., Nowozin, S., and Zhang, C. (2019). EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *International Conference on Machine Learning*, pages 4234–4243. PMLR.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. (2019). BIVA: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32.

Mahomed, F., Swanepoel, D. W., Eikelboom, R. H., and Soer, M. (2013). Validity of automated threshold audiometry: A systematic review and meta-analysis. *Ear and hearing*, 34(6):745–752.

Margolis, R. H. and Morgan, D. E. (2008). Automated Pure-Tone Audiometry: An Analysis of Capacity, Need, and Benefit. *American Journal of Audiology*, 17(2):109–113.

Mattei, P.-A. and Frellsen, J. (2019). MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4413–4423. PMLR.

Moore, B. C. (2012). *An Introduction to the Psychology of Hearing*. Brill.

Morgan, N., Bourlard, H., and Hermansky, H. (2004). Automatic Speech Recognition: An Auditory Perspective. In *Speech Processing in the Auditory System*, volume 18, pages 309–338. Springer-Verlag, New York.

Nachmani, E., Adi, Y., and Wolf, L. (2020). Voice Separation with an Unknown Number of Multiple Speakers. In *International Conference on Machine Learning*, pages 7164–7175. PMLR.

Nazábal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, 107:107501.

Nielsen, D., Jaini, P., Hoogeboom, E., Winther, O., and Welling, M. (2020). SurVAE flows: Surjections to bridge the gap between VAEs and flows. *Advances in Neural Information Processing Systems*, 33:12685–12696.

Øksendal, B. (2013). *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3):10.23915/distill.00010.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. (2019). On variational bounds of mutual information. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR.

Queiruga, A., Erichson, N. B., Hodgkinson, L., and Mahoney, M. W. (2021). Stateful ODE-Nets using basis function expansions. *Advances in Neural Information Processing Systems*, 34:21770–21781.

Queiruga, A. F., Erichson, N. B., Taylor, D., and Mahoney, M. W. (2020). Continuous-in-Depth Neural Networks. *arXiv preprint arXiv:2008.02389*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Salimans, T. (2016). A structured variational auto-encoder for learning deep hierarchies of sparse features. *arXiv preprint arXiv:1602.08734*.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. (2017). PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications. *arXiv:1701.05517 [cs, stat]*.

Salvi, C. and Lemercier, M. (2021). Neural stochastic partial differential equations. *CoRR*, abs/2110.10249.

Schlittenlacher, J., Turner, R. E., and Moore, B. C. (2018). Audiogram estimation using Bayesian active learning. *The Journal of the Acoustical Society of America*, 144(1):421–430.

Settles, B. (2009). Active learning literature survey.

Shannon, C. E. et al. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1.

Sinha, S. and Dieng, A. B. (2021). Consistency regularization for variational auto-encoders. *Advances in Neural Information Processing Systems*, 34.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). Ladder variational autoencoders. *Advances in neural information processing systems*, 29.

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., and Barbour, D. L. (2015). Fast, Continuous Audiogram Estimation Using Machine Learning. *Ear & Hearing*, 36(6):e326–e335.

Song, Y., Durkan, C., Murray, I., and Ermon, S. (2021). Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). Attention is all you need in Speech Separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE.

Tomczak, J. and Welling, M. (2018). VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR.

Townsend, J., Bird, T., and Barber, D. (2019). Practical Lossless Compression with Latent Variables using Bits Back Coding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tzen, B. and Raginsky, M. (2019). Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*.

Tzinis, E., Adi, Y., Ithapu, V. K., Xu, B., Smaragdis, P., and Kumar, A. (2022). RemixIT: Continual Self-Training of Speech Enhancement Models via Bootstrapped Remixing. *arXiv preprint arXiv:2202.08862*.

Tzinis, E., Wang, Z., and Smaragdis, P. (2020). Sudo RM -RF: Efficient Networks for Universal Audio Source Separation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Espoo, Finland. IEEE.

Vahdat, A. and Kautz, J. (2020). NVAE: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016a). WaveNet: A Generative Model for Raw Audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA.

van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016b). Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR.

van den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

von Gablenz, P. and Holube, I. (2015). Prevalence of hearing impairment in northwestern Germany. Results of an epidemiological study on hearing status (HÖRSTAT). *HNO*, 63(3):195–214.

Wang, D. and Chen, J. (2018). Supervised Speech Separation based on Deep Learning: An Overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., and Hershey, J. (2020). Unsupervised Sound Separation using Mixture Invariant Training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3846–3857. Curran Associates, Inc.

Xu, W., Chen, R. T. Q., Li, X., and Duvenaud, D. (2022). Infinitely deep bayesian neural networks with stochastic differential equations. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 721–738. PMLR.

Yamagishi, J., Veaux, Christophe, and MacDonald, Kirsten (2019). CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. (2017). Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE.

Zeghidour, N. and Grangier, D. (2021). Wavesplit: End-to-End Speech Separation by Speaker Clustering. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2840–2849.

Zenil, H. (2019). Compression is Comprehension, and the Unreasonable Effectiveness of Digital Computation in the Natural World. *arXiv preprint arxiv:1904.10258*.

Zhang, G., Qian, J., Chen, J., and Khisti, A. J. (2021). Universal Rate-Distortion-Perception Representations for Lossy Compression. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

{2.7182818284}